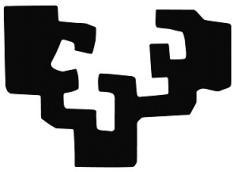
eman ta zabal zazu



Universidad del País Vasco
Euskal Herriko Unibertsitatea

Department of Computer Architecture and Technology

Speech Technologies for the Audiovisual and Multimedia Interaction Environments

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy by:

Aitor Álvarez Muniain

1. Supervisor Prof. Antonio Arruti Illarramendi
Department of Computer Architecture and Technology
University of the Basque Country (UPV/EHU)

2. Supervisor Ph.D. María Arantzazu del Pozo Echezarreta
Human Speech and Language Technologies department
Vicomtech-IK4

Aitor Álvarez Muniain
Speech Technologies for the Audiovisual
and Multimedia Interaction Environments
Supervisors: Prof. Antonio Arruti Illarramendi and Ph.D. María Arantzazu del Pozo Echezarreta
University of the Basque Country
Department of Computer Architecture and Technology
Vicomtech-IK4
Human Speech and Language Technologies Department
Donostia - San Sebastián

Abstract

The progress of technology, the easy access to powerful machines and electronic devices, social networks, the unlimited storing space on the Internet and ultimately, all that encompasses the new Digital Era, have driven a huge increase of the amount of contents that are created and publicly shared on a daily basis. These contents may include text, images, video and/or audio.

The generation of such vast amount of contents has led to the advancement of new methodologies for their optimal indexing and mining and for the automatic extraction of semantic information in different applications and domains, such as the security, surveillance, information access and retrieval, audiovisual or forensics sectors, among others.

Concerning audio analysis, it can be used in a wide range of applications considering the large amount of information that can be extracted from each audio content. Depending on the type of application, audio analysis can encompass information extraction at different levels, such as the linguistic level (speech transcription), language identification, the paralinguistic level (e.g. emotions), the speaker level (number, genre, segmentation, identification), the acoustic level (background or isolated noises, etc.), classification of audio segments (e.g. music, noise, speech) or music analysis. Audio analysis has to continually deal with the variability created by the particularities of each speaker, the acoustic environment, volume changes, accents, types of speech, overlappings, etc. Most of these aspects still pose a great challenge for the speech community. Besides, given their statistical nature, most of the solutions implemented for audio analysis are still highly domain-dependent and require adaptation when the application domain notably differs from the training data conditions.

This dissertation work involves several advanced audio and speech processing technologies that can be applied to the audiovisual and human-computer interaction environments. It includes an analysis of their applicability, their current state and details of the main contributions made to the fields. Finally, various of the developed

technological solutions are described, as well as their transfer to several companies for use in Industry.

Resumen

El progreso de la tecnología, la disponibilidad y fácil acceso a potentes máquinas y dispositivos electrónicos, las redes sociales, el espacio ilimitado que ofrece Internet para almacenar contenidos y, en definitiva, todo lo que engloba la nueva era digital, ha provocado un incremento ingente de los contenidos multimedia que se generan cada día y que son compartidos públicamente para su consumo. Estos contenidos multimedia pueden contener texto, imágenes, videos y/o audio.

La creación de estas enormes cantidades de contenidos ha provocado la necesidad de avanzar en nuevos mecanismos de almacenaje, en la gestión e indexación de estos contenidos y en la extracción de información semántica a través de análisis automáticos para diferentes aplicaciones y dominios, como el de seguridad, vigilancia, consulta y recuperación, el sector audiovisual o el ámbito forense, entre otros.

En lo referente al análisis del audio, su campo de aplicación es muy amplio, así como la cantidad y tipo de información que puede extraerse de un sólo contenido. Dependiendo del tipo de aplicación, el análisis del audio puede englobar la extracción de información a nivel lingüístico (transcripción del habla), idiomático, paralingüistico (emociones, estados de ánimo), de hablantes (número, género, segmentación, identificación), acústico (entorno, ruidos de fondo, ruidos específicos o aislados), clasificación de audio (por tipo de segmentos) o de análisis musical. El análisis del audio debe lidiar continuamente con la amplia variabilidad originada por la diversidad en las fuentes del sonido a nivel de hablantes, entornos acústicos, volúmenes, acentos, tipos de habla, solapamientos, etc. Muchos de estos aspectos representan todavía un reto científico en la comunidad. La naturaleza estadística de las tecnologías desarrolladas para el análisis del audio hace además que la necesidad de adaptación de la tecnología sea todavía una necesidad cuando ha de aplicarse en dominios dispares al de los datos de entrenamiento.

En esta memoria de tesis, se analiza el estado actual de algunas tecnologías de análisis del audio y procesamiento del habla aplicadas a sectores como el audiovisual y el de interacción persona-máquina, y se describen tanto su aportación a las mismas como las nuevas contribuciones realizadas al estado del arte. Finalmente, se describen soluciones tecnológicas desarrolladas y su transferencia a diferentes entidades para su uso en la Industria.

Agradecimientos

Considero éste un espacio de total libertad en el que poder expresar mis impresiones tras haber terminado de redactar este documento de tesis, y agradecer a todos los que de una u otra forma han colaborado para que haya podido culminar este proceso doctoral.

Muchas personas a mi alrededor, sabedoras ya de lo complicado de este camino, me advertían de la pesadez del mismo en este último tramo de escritura, dándome continuos ánimos para que no cayera en agujeros negros de insatisfacciones e inseguridades. Lo cierto es que, afortunadamente, apenas los sufrí. El proceso de escritura fue un camino mayoritariamente agradable que me sirvió para rememorar todo el trabajo realizado en los últimos años. Este momento de pausa, de mirar hacia atrás y contemplar todo la labor bien hecha, me ayudó a entender lo importante que a veces resulta parar, respirar, y valorar todo lo realizado hasta el momento sin que el día a día nos siga devorando a su ritmo frenético.

Después de algunos años de experiencia laboral iniciada en la Universidad y continuada en Vicomtech-IK4, sois muchos a los que querría agradecer que me hayáis querido acompañar en este viaje. Gracias a cada uno de vosotros soy ahora mejor, y gracias a vuestro apoyo y trabajo he llegado a completar esta memoria. Así que antes de empezar a nombraros, mis agradecimientos de antemano y mis disculpas anticipadas a todos aquéllos que sin motivo he dejado fuera.

De mi primer grupo universitario, no podía dejar de nombrar a Idoia Cearreta, responsable como pocos en su trabajo y gran compañera. Espero que en algún momento nuestras vidas profesionales vuelvan a cruzarse. Puede decirse que mis inicios en Vicomtech-IK4 estuvieron marcados por el carácter indomable de Kutz Arrieta, persona de firmes principios, gran creatividad y amplios conocimientos. Cada reunión, comida y café contigo supuso un nuevo aprendizaje, además de ser una de las primeras personas en hablarme de tesis. Ya volverás ya. Más tarde llegaste tú Haritz Arzelus, y te convertiste en el mejor compañero para sacar los proyectos adelante. Gracias por tu dedicación incansable y calidad humana. Como bien sabes, parte de todo esto también es tuyo. Y cómo no nombrar al vaquero de Iparralde, a la mejor compañía para disfrutar de una buena conversación en torno a una mesa.

Gracias por tus sabios consejos Thierry, por tu amistad, y por ser también uno de los grandes impulsores de todo esto. Y a ti Montse, qué decirte, siempre tan risueña y generosa con los que te rodean. Gracias también a Santi, Andoni, Manex y Naiara por vuestra ilusión y energía.

La calidad humana es algo que siempre ha caracterizado a Vicomtech-IK4. A través de ella he podido conocer a grandes personas que han sido decisivas para que haya podido evolucionar hasta lo que soy hoy en día. Maider, Leti, Esti, María, Borja, Sara, Aritz, Labayen, Naiara, Jon Haitz, Aiala, Iñaki, Bea, Jon, Felipe, Beñat... Gracias por vuestro cariño y por los buenos momentos vividos y que aun nos quedan por vivir, tanto a nivel profesional como personal.

Gracias también al comité de dirección de Vicomtech-IK4 por darme la oportunidad de crecer como investigador y de culminar todo el trabajo realizado hasta el momento a través de esta memoria. No descansaste, Edurne Loyarte, hasta verme llegar hasta aquí. Me mostraste el camino del tan necesario pragmatismo, y su aplicación fue clave para conducir este proyecto al puerto adecuado por el camino más rápido. Tambíen fueron importantes, Jorge Posada, tus mensajes de ánimo durante este proceso. Gracias por ellos y por todos los prácticos consejos que me ofreciste desde el inicio. Julián Flórez, como bien sabemos, dos Aries están condenados a encontrarse dialécticamente si los puntos de vista iniciales no coinciden. Aun así, gracias a nuestro compromiso con el centro y a nuestro carácter detallista y perfeccionista, espero que podamos seguir encontrándonos en pro de un crecimiento imparable.

Y, por supuesto, gracias a mis supervisores Andoni Arruti y Arantza del Pozo por ayudarme a que por fin este documento de tesis haya visto la luz. Mención aparte en este aspecto mereces tú, Basi, uno de los mayores artífices en impulsar este tramo final de mi proceso doctoral. Tu apoyo y confianza fueron definitivos para alcanzar la meta. Mila esker benetan.

A nivel personal, respetaré el deseo de la mayoría de vosotr@s de no querer ser nombrad@s ni load@s. Vuestra humildad y discreción os honra y por eso me siento tan orgulloso de estar a vuestro lado. Aun así, haré una excepción contigo, Ama, ya que nada de esto habría sido posible sin tu labor incansable desde que dos enanos quedaron a tu disposición. Fuiste fuerte, superaste mil y una barreras, nos enseñaste el valor de la responsabilidad, del esfuerzo, de la honestidad y la humildad, y gracias a eso y mucho más conseguiste que cada uno de esos dos enanos adquiriera unos principios inquebrantables para poder caminar con orgullo y firmeza por la vida. Desde aquí, mi más sincero e inagotable agradecimiento por haberme ayudado a convertirme en lo que fundamentalmente soy.

Mila esker bihotzez.

Contents

1 Introduction

"*The possession of speech is the grand distinctive character of man*" (T.H. Huxley, 1871). Speech involves great cognitive skills, which include not only the ability to invent words and construct grammars for a language, but also the capacity of pronouncing them, processing them, and thinking and understanding the world with them. All the issues above make human beings unique in the animal kingdom at communication level.

Since speech fossils are not available, there is no real consensus regarding the origin of speech and language, which remains a mystery without clues. Most researches agree that language may have gradually emerged about a couple of hundred thousand years ago. And that, considering found evidence of creation arts and burials of dead [Hol+04], fluent language should have been somehow present at least 50,000 years ago.

As more complex languages were generated by human beings, speech became the most natural communication mode between people, since it provides the fastest rate of information transfer. At scientific level, the mechanisms for the realization and perception of speech have intrigued engineers and researches for centuries. In this sense, the first known attempt dates from the 2nd half of the 18th century, when the Russian scientist Christian Kratzenstein, a professor of physiology in Copenhagen, succeeded in producing vowel sounds using resonance tubes connected to organ pipes [Kra82]. After a few more advanced solutions, in the 1930's the electric speech synthesizer VODER (Voice Operating Demonstrator) [DRW39] was developed, considered an important milestone in the evolution of speaking machines. Following these advances and those achieved in the theory of acoustic-phonetics, it was not until the mid-twentieth century when the research activity on speech recognition started. In 1952, Bell Laboratories pioneered a system that could recognize single digits from a single speaker. Since then and to date, great efforts have been done in the scientific community to improve speech technologies, thanks to important contributions from IBM and AT&T laboratories, Carnegie Mellon University (CMU), SRI International, Massachusetts Institute of Technology (MIT), and the Defense Advanced Research Projects Agency (DARPA) through different programs, among many others.

Nowadays, speech technologies have evolved enough to be usual in our common tasks. Applications like Google Voice Search or Siri are used daily by millions of people in different languages. Nevertheless, speech technologies still decrease performance when dealing with more challenging situations, as it will be addressed in this work.

1.1 Context of this research work

This PhD dissertation is a compendium of both basic and applied research activities, which have been carried out at Vicomtech-IK4 and the University of the Basque Country over the last 12 years. These activities are focused on speech processing and encompass two main application environments and four fields of contribution: Automatic Subtitling and Rich Transcription within the Audiovisual Environment (Chapter 3) and Speech Emotion Recognition and Speech-driven Facial Animation within the Multimedia Interaction Environment (Chapter 4).

Whereas work on Speech Emotion Recognition was performed as basic research in the Computer Engineering Faculty in collaboration with the departments of Computer Science and Artificial Intelligence and Computer Architecture and Technology, the Automatic Subtitling, Rich Transcription and Speech-driven Facial Animation work was carried out at Vicomtech-IK4 as a combination of basic research and R&D projects. This way, some basic research activities have been published in journals and conference proceedings, whilst some other technologies developed in R&D projects have been transferred to Industry as new services and/or through the creation of new companies that are exploiting the implemented solutions.

1.2 Vicomtech-IK4

Most of the research activities described in this work have been realized at Vicomtech-IK4, an applied Research Center located in San Sebastian (Basque Country, Spain) that combines basic and applied research with the aim of transferring technology to Industry. More specifically, one of the main missions of Vicomtech-IK4 involves meeting the applied research, technology development and innovation requirements of local companies and institutions in Computer Graphics, Visual Computing and Multimedia fields to enhance their competitiveness and improve society's economic development and quality of life. To this end, Vicomtech-IK4 promotes the development of innovative visual interaction and communications technologies, creating product prototypes and applications in collaboration with Industry. Besides,

Vicomtech-IK4 aims at contributing to universal knowledge through the publication of its scientific results.

Nowadays, Vicomtech-IK4 is composed of seven technological departments, each seeking to develop and apply technology in different fields and industries, as listed below:

- Industry and Advanced Manufacturing
- Digital Media
- Human Speech and Language Technologies
- eTourism and Cultural Heritage
- Intelligent Transport Systems and Engineering
- Interactive Computer Graphics
- eHealth and Biomedical Applications

This dissertation work corresponds to the first PhD project originated within the Human Speech and Language Technologies (HSLT) group and greatly reflects the evolution and growth of the speech technologies area, which has become one of the main four research lines of the department. This evolution has been given mainly by (1) the expertise acquired throughout the execution of several local and European projects, (2) the strong collaboration with other technological fields, departments and the University, (3) the knowledge of the market and Industry needs of such technologies, and (4) all the linked research and development activities.

1.3 Human Speech and Language Technologies department

In response to the growing demand to integrate language engineering in other applications, Vicomtech-IK4 opened in 2008 a new emerging department focused on the research, development and integration of speech and language technologies in other areas. This new department aimed at offering added value and at contributing to the intelligence of the applications developed in the fields of Digital TV, Multimedia

Services, Tourism and Cultural Heritage, Biomedical Sciences, Industrial Applications and Human-Computer Interaction.

At the beginning, the emerging department was focused on integrating solutions mainly related to speech processing technologies. It did not aim to develop proprietary synthesizers and/or recognition engines, but to integrate existing commercial solutions and adapt them to the context. In this sense, the initial main technological solutions that the department offered were as follows:

- A subtitling prototype: which included a commercial Automatic Speech Recognition engine connected to a professional subtitling software to generate subtitles automatically from Spanish TV contents.

- A Voice Transformation toolkit: which provided the possibility of transforming a source voice into a target voice.

- A Speech-driven animation system: which was capable of producing suitable data for the animation of a virtual character from natural voice in Basque using open source technologies.

As new markets and research interests continued to grow, the emerging department kept gaining more relevance, until it became the current Human Speech and Language (HSLT) department.

Nowadays, the objectives of the HSLT group are more ambitious and involve research and development of proprietary technology in the fields of speech processing, machine translation, natural language processing and dialogue systems, with the aim of transferring such technology to Industry through innovative applications and solutions.

Regarding the Speech Processing research line, its more prominent technologies correspond to Automatic Subtitling, Rich Transcription and Voice Biometrics. In addition, other technologies such as Speech Emotion Recognition and personalised Text-to-Speech systems are part of the technological roadmap as well.

During its professional path, the Speech Processing research line has transferred technology to different customers with the aim of improving their internal production workflows. Some of such clients are Mixer Servicios Audiovisuales S.L., Ubertitles S.L., Irekia and Serikat Consultoría e Informática S.A., among others. In addition, Mira lo Que te Digo S.L.U. (MQD), Natural Vox, Synthema and Voice Interaction are some of its local, national and European technological partners.

1.4 Main publications

The current PhD project is mostly based on the contributions published in the journals and conference proceedings compiled in Section 7. Such publications include scientific contributions to the Automatic Subtitling, Rich Transcription, Speech Emotion Recognition and Speech-driven Facial Animation fields. The list of the main publications per field is presented below.

1.4.1 Automatic Subtitling

All the activities related to the research and development of technology for Automatic Subtitling were performed at Vicomtech-IK4 within R&D and Industrial projects.

Automatic live and batch subtitling systems

- **[ÁPA10]** Álvarez, A., del Pozo, A., and Arruti, A. (2010). *APyCA: Towards the automatic subtitling of television content in Spanish.* In Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT) (pp. 567-574). IEEE. (see Section 7.1)

  This paper describes the subtitling prototype developed within the APyCA project (Section 2.1.1). This protoype was developed in an attempt to automate the process of subtitling television content in Spanish through the application of state-of-the-art speech and language technologies.

- **[Álv+15]** Álvarez, A., Mendes, C., Raffaelli, ..., and del Pozo, A. (2015). *Automating live and batch subtitling of multimedia contents for several European languages.* Multimedia Tools and Applications, 1-31. (see Section 7.2)

  This article describes the contributions carried out along the FP7 EU-funded SAVAS project (Section 2.1.4). It contains a detailed description of the live and batch automatic subtitling applications developed by the SAVAS consortium for several European languages based on Large Vocabulary Continuous Speech Recognition (LVCSR) technology specifically tailored to the subtitling needs, together with results of their quality evaluation.

SAVAS corpus

- **[Poz+14]** del Pozo, A., Aliprandi, C., Álvarez, A., ... , and Raffaelli, M. (2014). *SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling.* In LREC Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 432-436). (see Section 7.3)

  This paper describes the data collection, annotation and sharing activities carried out within the European SAVAS project. The project aimed to collect, share and reuse audiovisual language resources from broadcasters and subtitling companies to develop LVCSR engines in specific domains and new languages, with the purpose of solving the automated subtitling needs of the media industry.

Long Audio Alignment

- **[Bor+16]** Bordel, G., Penagarikano, M., Rodriguez-Fuentes, L., Alvarez, A., and Varona, A. (2016). *Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks. In IEEE Signal Processing Letters, vol. 23, no. 1, pp. 126-129.* (see Section 7.4)

  In this article, an alternative approach to deal with text-to-speech alignment of long audio tracks is presented. To this end, probabilistic criteria based on the confusion matrix provided by a phone decoder are used as informative kernels to align the phone transcription of the reference text and the phone recognition of the input audio. These probabilistic kernels outperform our baseline kernels and other alternatives, including a reference ASR-based approach and a knowledge-based kernel, in experiments on the Hub4-97 dataset.

- **[ÁRA14]** Álvarez, A., Ruiz, P., and Arzelus, H. (2014). *Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque.* In Text, Speech and Dialogue (pp. 473-480). Springer International Publishing. (see Section 7.5)

  In this work, the system presented in [ÁAR14] for long audio alignment in an automatic subtitling scenario is improved and extended. Phone-decoder accuracy is enhanced using context-dependent acoustic models, besides implementing an adaptation of the generic language models to the script of the contents to subtitle. The system is also extended to Basque, its original

languages being English and Spanish, and additional linguistic resources are created for the Spanish aligner.

- **[ÁAR14]** Álvarez, A., Arzelus, H., and Ruiz, P. (2014). *Long audio alignment for automatic subtitling using different phone-relatedness measures.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014 (pp. 6280-6284). IEEE. (see Section 7.6)

  In this paper, long audio alignment systems for Spanish and English are presented in an automatic subtitling scenario. Pre-recorded contents are automatically recognized at phoneme level by language-dependent phone decoders. A dynamic-programming alignment algorithm finds matches between the automatically decoded phones and the ones in the phonetic transcription from the script. The accuracy of the alignment algorithm is evaluated when applying three non-binary scoring matrices based on phone confusion-pairs from each phone decoder, on phonological similarity and on human perception errors.

- **[RÁA14]** Ruiz, P., Álvarez, A., and Arzelus, H. (2014). *Phoneme similarity matrices to improve long audio alignment for automatic subtitling.* In LREC, Ninth International Conference on Language Resources and Evaluation. (see Section 7.7)

  This work was our first attempt to enhance the work presented in [Bor+12] for long audio alignment. In this paper, we showed that, as compared with results for a binary matrix, scoring alignment operations with a matrix based on phoneme-similarity improved alignment results at phoneme level, word level and subtitle level.

Automatic Segmentation of Subtitles

- **[Álv+16c]** Álvarez, A., Martínez-Hinarejos Carlos-D., Arzelus H., Balenciaga M., and del Pozo A. (2016) *Improving the Automatic Segmentation of Subtitles through Conditional Random Field*. Speech Communication, Elsevier. ***Status: In 2nd revision.*** (see Section 7.8)

  In this article, a method based on Conditional Random Field (CRF) is presented to deal with automatic subtitling segmentation. This is a continuation of a previous work [ÁAE14] in the field, which proposed a method based on Support Vector Machines and Logistic Regression classifiers to generate possible candidates for breaks. For this study, two corpora in Basque and Spanish are

used for experiments, and the performance of the CRF-based method is tested and compared with the previous solution through several evaluation metrics.

- **[Álv+16b]** Álvarez, A., Balenciaga, M., del Pozo, A., Arzelus, H., Matamala, A., Martínez-Hinarejos, Carlos-D. (2016). *Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles.* In Proceedings of the 10th international conference on Language Resources and Evaluation (LREC2016), pp. 3049-3053. (see Section 7.9)

  This paper describes the evaluation methodology followed to measure the impact of using a CRF-based algorithm to automatically segment intralingual subtitles. The segmentation quality, productivity and self-reported post-editing effort achieved with such approach are shown to improve those obtained by the technique based in counting characters, mainly employed for automatic subtitle segmentation currently.

- **[ÁAE14]** Álvarez, A., Arzelus, H., and Etchegoyhen, T. (2014). *Towards customized automatic segmentation of subtitles.* In Advances in Speech and Language Technologies for Iberian Languages (pp. 229-238). Springer International Publishing. (see Section 7.10)

  In this study, we presented a novel approach to automate the segmentation of subtitles through machine learning techniques, allowing the creation of customized models adapted to the specific segmentation rules of subtitling companies. Support Vector Machines and Logistic Regression classifiers were trained over a reference corpus of subtitles manually created by professionals and used to segment the output of speech recognition engines.

1.4.2 Rich Transcription

All the activities linked to the research and development of technology for Rich Transcription were performed at Vicomtech-IK4 within R&D and Industrial projects.

Proprietary Rich Transcription systems for Basque and Spanish

- **[Álv+16a]** Álvarez, A., Arzelus, H., Prieto, S., and del Pozo, A. *Rich Transcription and Automatic Subtitling for Basque and Spanish.* In: Advances in Speech

and Language Technologies for Iberian Languages. 2016. **Status: Submitted**. (see Section 7.11)

In this paper, complete rich transcription and automatic subtitling systems for Basque and Spanish are described. They enable the automatic transcription and/or subtitling of bilingual contents, through the integration of a language tracker that discriminates between segments spoken in Basque and Spanish. The technology is accessible through a web platform hosted on the Internet. The paper details the architecture of the systems and focuses on the description and evaluation of each technological component. Performance results are reported for the parliamentary domain.

1.4.3 Speech Emotion Recognition

The basic research done within this field was carried out in collaboration with the University of the Basque Country, as a result of the work begun in the RekEmozio project (see Section 2.3.1).

Feature Subset Selection in Basque and Spanish

- **[Arr+14]** Arruti, A., Cearreta, I., Álvarez, A., Lazkano, E., and Sierra, B. (2014). *Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction.* PloS one, 9(10), e108975. (see Section 7.12)

  This article shows an attempt to select the most significant features for emotion recognition in spoken Basque and Spanish using different methods for feature selection. Experiments were executed in three phases, using different sets of features as classification variables in each phase. Besides, Feature Subset Selection technique through the Estimation of Distribution Algorithm (EDA) was applied at each phase in order to seek for the most relevant feature subset.

- **[Álv+07a]** Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007, September). *A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms.* In Text, Speech and Dialogue (pp. 423-430). Springer Berlin Heidelberg. (see Section 7.13)

  This paper presents a study where, using a wide range of speech parameters, improvement in emotion recognition rates is analyzed. Using an emotional

multimodal bilingual database for Spanish and Basque, emotion recognition rates in speech have significantly improved for both languages comparing with previous studies.

- **[Álv+07b]** Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007). *Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech.* In Advances in Nonlinear Speech Processing (pp. 273-281). Springer Berlin Heidelberg. (see Section 7.14)

  In this paper, we present a study performed to analyze different machine learning techniques validity in automatic speech emotion recognition area. In this particular case, techniques based on evolutive algorithms (EDA) have been used to select speech feature subsets that optimize automatic emotion recognition success rate.

- **[Álv+06]** Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2006, September). *Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in spoken Spanish and Standard Basque Language.* In Text, Speech and Dialogue (pp. 565-572). Springer Berlin Heidelberg. (see Section 7.15)

  In this work, different speech parameters have been computed for each audio recording of a bilingual data set in Basque and Spanish. Then, several Machine Learning techniques were applied to evaluate their usefulness in speech emotion recognition. We improved the initial results through the use of feature subset selection techniques, and a study of the most relevant features was performed for each language in the data set.

Classifier Subset Selection for the Stacked Generalization

- **[Álv+16d]** Álvarez, A., Sierra, B., Arruti, A., López-Gil, J. M., and Garay-Vitoria, N. (2015). *Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech.* Sensors, 16(1), 21. (see Section 7.16)

  In this article, a new supervised classification paradigm, called classifier subset selection for stacked generalization (CSS stacking), is presented to deal with speech emotion recognition. The new approach consists of an improvement of a bi-level multi-classifier system known as stacking generalization by means of an integration of an estimation of distribution algorithm (EDA) in the first

layer to select the optimal subset from the standard base classifiers. The good performance of the proposed new paradigm was demonstrated over different configurations, speech features and datasets.

1.4.4 Speech-driven Facial Animation

All the activities related to the research and development of technology for this field were performed at Vicomtech-IK4 within R&D and Industrial projects.

- **[Oya+10]** Oyarzun, D., Mujika, A., Álvarez, A., Legarretaetxeberria, A., Arrieta, A., and del Puy Carretero, M. (2010). *High-realistic and flexible virtual presenters.* In Articulated Motion and Deformable Objects (pp. 108-117). Springer Berlin Heidelberg. (see Section 7.17)

  This paper presents the mixed reality prototype called PUPPET, which provides a 3D virtual presenter that is embedded in a real TV scenario and is driven by an actor in real time. The key modules of this prototype improve the state-of-the-art in such systems in four different aspects: real time management of high-realistic 3D characters, less equipment needs, and flexibility in the real/virtual integration, and animations generated from actor's speech.

- **[Muj+13]** Mujika, A., Diez, H., Alvarez, A., Urteaga, M., and Oyarzun, D. (2013). *Realistic visual speech synthesis in WebGL.* In Proceedings of the 18th International Conference on 3D Web Technology (pp. 207-207). ACM. (see Section 7.18)

  This paper presents the work carried out to develop a web application that shows the face of a virtual character pronouncing the sentences the user sets. The level of realism was high and the performance was fast enough. The application makes use of WebGL, speech processing, text to speech and co-articulation technologies to obtain the virtual pronunciation.

1.5 Thesis Structure

This PhD memory is divided into seven main chapters, considering this introductory chapter and the remaining six that are introduced below.

Chapter 2

This chapter is focused on describing the R&D projects which have mainly led to this

dissertation work. An initial description and the results obtained are included per project, in addition to the publications that have resulted from each project. The description contains the main objectives and the context in which the project was created. The results describe the work and the technological solution developed.

Chapter 3

The aim of this chapter is to describe the work that has been done within the audiovisual environment and the technological contributions made in terms of speech processing technologies to this domain. This environment encompasses the technological fields linked to Automatic Subtitling and Rich Transcription. The chapter is divided into five sections, including (1) an Introduction; (2) a State-of-the-Art section; (3) the main Challenges faced by the speech technologies proposed to enhance solutions within the audiovisual environment; (4) the main Contributions of this work; and (5) final Conclusions from the technological point of view.

Chapter 4

In this chapter, two main technologies are combined to describe their contribution to the Multimedia Interaction environment. More specifically, it explains how Speech Emotion Recognition and Speech-driven Facial Animation can contribute to improving Human Computer Interaction. This chapter is also structured in five sections. An Introduction provides context to the chapter, whilst the State-of-the-Art looks at the current state of the main technologies included. The current and future main Challenges of these technologies are also presented in the third section. Finally, the main Contributions of this dissertation work and the final technological Conclusions are detailed in the last two sections.

Chapter 5

As an applied research center, the transfer of technology to Industry is one of the main missions of Vicomtech-IK4. In this chapter, some examples of how the speech solutions based on the technology described in this memory have been integrated within several companies and entities are given. In some cases, technology has served to enhance internal processes. In others, the developed technological solutions have become the main technological core of the companies.

Chapter 6

This chapter describes the main conclusions of this PhD project and provides a look to the future of the addressed fields, both from the market and technological points of view.

Chapter 7

The main publications which support the current dissertation work are presented

in this last chapter. The articles and conference proceedings are listed in the same order as they have been mentioned in Section 1.4.

2 R&D Projects

This dissertation work is the consequence of many basic research activities and several R&D projects, which started in 2004 at the University of the Basque Country. Many of these research activities are still running and the technology keeps evolving, as there is still room for improvement.

In the following Sections, the R&D projects linked to this work are presented for each of the four main contribution fields. Each project includes a brief description, looks at its objectives and describes its main contributions through a summarized conclusion and a list of the resulting scientific publications.

It should be noted that not all the projects have reported scientific publications. The technology implemented in some of the projects has served to develop state-of-the-art components or complete solutions to be transferred to companies. These developments have also helped to grow, enhance and evolve the speech processing line of the HSLT department in Vicomtech-IK4. In addition, basic research activities apart from the R&D projects at Vicomtech-IK4 have also reported scientific publications that are referenced in the Section 2.1.7.

2.1 Automatic Subtitling

2.1.1 APyCA

- Title: Hacia la subtitulación automática: Asignación de Puntuación y Color Automática
- Tipology: Industrial Project supported by the Innotek program of the Basque Government
- Period: 2007-2009
- Consortium: MIXER Servicios Audiovisuales, S.L. Irusoin, S.A., Euskal Irrati Telebista (EiTB)

Description and Objectives

The main objective of the APyCA project was the development of a platform for the automatic subtitling of audiovisual contents. The platform had to be constructed over an open and modular architecture in order to allow the integration of new modules in the future.

From the technologically point of view, the platform had to include three main modules:

- A commercial state-of-the art speech recognition engine.
- An Autopunctuation module, which included methods based on acoustic, prosodic, and linguistic characteristics.
- A Speaker Diarization module, including speaker segmentation and clustering algorithms.

From the technical and application side, the platform had to accomplish the following issues:

- It had to include the most adequate commercial software for subtitle generation.
- The final platform had to be ready to work correctly in real subtitling scenarios.
- The technology had to work in batch and semi-live subtitling modes. In addition, the performance of the platform when dealing with live contents had to be evaluated.

Results

Four main components were developed and integrated into the APyCA prototype for the automatic subtitling of Spanish audiovisual contents:

1. Voice Activity Detection (VAD): This module segmented and classified the input audio into four acoustic categories (speech, speech and noise, noise and silence). It was based on the speech detection functionality of the open source LIUM_SpkDiarization tool [MM10].

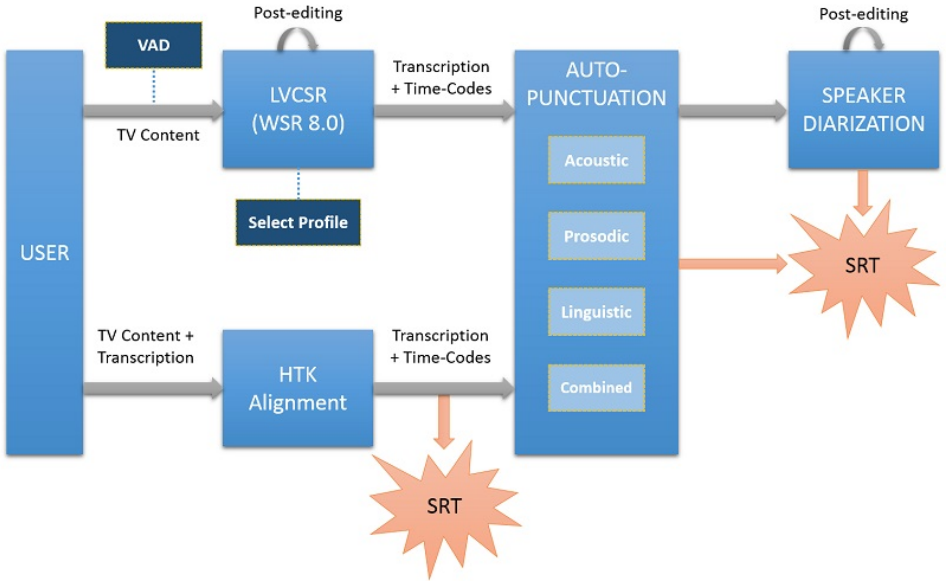


Fig. 2.1: APYCA prototype architecture

2. Automatic Speech Recognition (ASR) and Alignment: the Windows Speech Recogniser (WSR version 8.0) was employed as the principal ASR engine. In addition, in those cases for which transcriptions already existed, an alignment module built with the HTK toolkit [You+97] was constructed to automatically synchronize audio and text.

3. Discourse Segment Detection (DSD): The output of the ASR engine could be segmented into discursive segments using four different techniques based on (1) acoustic information, (2) prosodic information, (3) linguistic information and (4) combined information from the previously described techniques.

4. Speaker Diarization (SD): the LIUM_SpkDiarization tool was used to perform the speaker diarization task.

Objective evaluation of the different modules showed that the developed prototype was capable of generating automatically time-coded and colour-assigned Spanish subtitles for post-editing in bounded domains.

More information about this project can be found in the related publication presented in Section 7.1.

Related publication

- [ÁPA10] Álvarez, A., del Pozo, A., and Arruti, A. (2010). *APyCA: Towards the automatic subtitling of television content in Spanish.* In Proceedings of the

2010 International Multiconference on Computer Science and Information Technology (IMCSIT), (pp. 567-574). IEEE.

2.1.2 BerbaTek, Ber2Tek, Elkarola

- Title: BerbaTek, Ber2Tek, Elkarola
- Tipology: Strategic research projects supported by the Elkartek program of the Basque Government
- Period: 2009-2016
- Consortium: Elhuyar Foundation, IXA Group (UPV/EHU), Aholab (UPV/EHU), Vicomtech-IK4, Tecnalia Research & Innovation.

Objectives

The aim of the BerbaTek project (2009-2011) was the research and development of language, voice and multimedia technologies so that they could provide the technological basis to support the economic sector of the Basque language industries, composed by the translation, teaching and content sectors.

Ber2Tek (2012-2014) aimed at advancing the research and development of the technologies of analysis of cross-media contents, high quality machine translation and natural spoken multimodal interaction. Ber2Tek also focused on the generation of knowledge and training of a qualified critical mass which would allow to face the future of R&D in this strategic line.

Finally, Elkarola (2015-2016) is focused in basic research for the application of speech and language technologies to the strategic priorities for smart specialization (RIS3) of the Basque Country, such as advanced Manufacturing, Energy and Biosciences, and additionally some niches related to Territory.

Results

Throughout these strategic projects, the HSLT department of Vicomtech-IK4 has investigated, developed and evolved speech processing technologies mainly in the fields of automatic speech recognition, rich transcription, automatic subtitling and spoken term detection.

These activities have driven the implementation of the speech processing demonstrators described below.

- Video search engine demo
  http://bideobilatzailea.ber2tek.eus:8086/Ber2Tek_Web/eu/demo.html
  This web application included many speech and language technologies with many applications which are detailed in the following points:

  - Automatic transcription: the integration of proprietary ASR technology based on the KALDI toolkit [Pov+11], allowed the automatic generation of transcriptions from the original videos in Basque, Spanish and English, and thereby the creation of subtitles. In addition, it enabled keyword search to retrieve specific videos or to automatically point to an exact time position within the videos.

  - Machine Translation: it allowed the translation of the original automatic subtitles of the videos to other languages (Basque, Spanish and English).

  - Speech Synthesis: this technology enabled vocalizing the translated subtitles in Basque, Spanish and English.

  - Voice Transformation: allowed the synthetic voices in the translated languages to imitate the voice of the original speaker.

  - Image Processing: this module allowed searching information related to lightness, colour, change of scenes or faces within the videos.

- Cross-lingual subtitling demo
  http://212.81.220.68:8086/berbatek/eu/

  This web application integrated speech technology to (1) automatically align Basque video contents and transcriptions in order to generate word level time-codes, (2) generate subtitles using such time level information, (3) translate the original Basque subtitles into Spanish, and (4) synthesize the Spanish subtitles using text-to-speech technology with the aim of listening to videos in the translated language. This demo mainly promoted accessibility generating intra- and inter-lingual subtitles of Basque videos.

  More information about the Berbatek project can be found in the related publication included in Section 7.19.1.

Related Publication

- [Azk+13] Azkarate, I. L., Cordón, E. N., Moncalvillo, I. S., Baranda, D., Iturraspe, U., Gabiola, K. M. S., Arregi, X., de Ilarraza, A., del Pozo, A. and Álvarez, A. (2013). *BerbaTek: euskararako hizkuntza teknologien garapena itzulpengintza, edukien kudeaketa eta irakaskuntza arloetan.* Euskalingua, (23), 66-76.

2.1.3 SUSA

- Title: Subtitulación Semi-Automática
- Tipology: Industrial project
- Period: 2010-2013
- Consortium: MIXER Servicios Audiovisuales S.L.

Objectives

The main goal of SUSA was the development of a system to speed up the manual creation of subtitles. MIXER is an audiovisual company which provides dubbing and subtitling services to broadcasters. Its professional subtitlers generated subtitles manually from scratch, taking them between 8 and 10 hours to generate subtitles per hour of content. With the SUSA solution, MIXER was seeking to improve productivity of its subtitle creation process.

Results

Most of the contents MIXER had to subtitle also had to be dubbed. This implied the creation of the target language transcription, which could be exploited for automatic subtitle generation. Hence, the developed solution involved the integration of technology which allowed the automatic alignment between the dubbed audio and its existing transcription.

Following the approach used in APyCA, an HTK-based alignment system was built for Basque and Spanish. This enabled to obtain time-codes at word level automatically for each content and transcription. Then, subtitles were generated following several specific rules defined by MIXER. These rules were mainly related to spacing features

(number of lines and maximum number of characters per line), timing features (minimum and maximum time of each subtitle on screen), segmentation features (mainly related to punctuation marks), and conventions regarding the colors to be assigned to each speaker. The main advantage of developing a system based on automatic alignment was that there were no text errors to be corrected by post-editors. The post-edition work was mainly focused on the correction of the possible desynchronization errors due to imperfect transcriptions or highly noisy segments. This system helped MIXER professionals subtitle each hour of content in 3-4 hours, being 50-60% faster and therefore more productive.

The SUSA solution inspired the creation of UBERTITLES S.L., a company which was born in 2012 to provide an automated solution to the subtitling community. UBERTITLES launched a product based on the SUSA technological solution, which allowed subtitling English, Spanish and Basque videos in a fast, simple and automatic way based on their existing transcriptions.

More information about this company and its technological solution can be found in Section 5.1.

2.1.4 SAVAS

- Title: Sharing AudioVisual language resources for Automatic Subtitling
- Tipology: European Project supported within the Seventh Framework Programme for Research and Technological Development (FP7-ICT-2011-SME-DCL)
- Period: 2012-2014
- Consortium: Vicomtech-IK4(ES), Voice Interaction (PT), Synthema (IT), EiTB - Euskal Irrati Telebista (ES), MIXER Servicios Audiovisuales S.L. (ES), RTP - Radio e Televisao de Portugal S.A. (PT), RAI - Radiotelevisione Italiana (IT), SWISSTXT - Schweizerishe Teletext A.G. (CH).

Objectives

The SAVAS project was focused on providing an automatic subtitling solution to broadcasters and subtitling companies, which were seeking for more productive alternatives than the traditional method of subtitling due to the high quantity of

the demand and the high cost of the manual process. In this context, the SAVAS partners aimed to acquire and share data, and develop automatic subtitling systems based on LVCSR (Large Vocabulary Continuous Speech Recognition) technology for the following nine European languages: Basque, Spanish, Portuguese, Italian, French, German and the swiss variants of the latter three. The SAVAS consortium built a META-SHARE repository with the audiovisual language resources collected and annotated within the project.

From the technological point of view, the SAVAS project aimed to advance the state-of-the-art in the following fields: (a) automatic data collection, transcription and annotation; (b) data sharing; (c) LVCSR technology for automated subtitling; (d) punctuation and capitalization technology to enrich transcriptions; (e) real-time subtitling systems based on LVCSR technology.

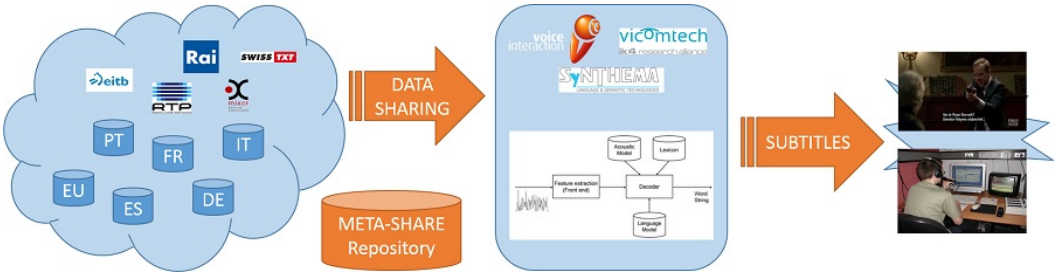


Fig. 2.2: SAVAS project

Results

Data resources suitable to train acoustic and language models of LVCSR systems in addition to technology for automatic subtitling were compiled and developed for nine European languages within the SAVAS project [Poz+14; Álv+15].

Concerning data resources, 200 hours of annotated audio and 1000 million words of text were collected for each language, with the exception of Basque, due to the more limited availability of Basque digital news texts, and Portuguese. All the audio and text data collected from broadcasters and subtitling companies was shared through the META-SHARE repository under research and commercial licenses.

Regarding the technology developed, three types of applications were implemented for each language; (1) S.Scribe!, a batch Speaker Independent Transcription system for offline subtitling; (2) S.Live!, a Speaker Independent Transcription System, with real-time performance for live subtitling; and (3) S.Respeak!, a dictation engine for live and batch production of subtitles. The systems were developed following subtitling requirements, and considering the needs of the subtitling companies and

market. Besides, they were further evaluated using several metrics related to LVCSR technology and subtitling quality specific features.

More information about this project, the development of the systems and the compiled SAVAS corpora can be found in the related publications included in Sections 7.2 and 7.3.

Related Publications

- [Álv+15] Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., anddel Pozo, A. (2015). *Automating live and batch subtitling of multimedia contents for several European languages.* Multimedia Tools and Applications, 1-31.

- [Poz+14] del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J. P., Paulo, S., Piccinini, N., and Raffaelli, M. (2014). *SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling.* In LREC Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 432-436).

- [Ali+14a] Aliprandi, C., Gallucci, I., Piccinini, N., Raffaelli, M., del Pozo, A., Álvarez, A., Cassaca, R., Neto, J., Mendes, C. and Viveiros, M. (2014). *Assisted subtitling: a new opportunity for access services.*The International Broadcasting Conference 2014 (IBC2014), September 10-15 2014, Amsterdam, The Netherlands

- [Ali+14b] Aliprandi, C., Scudellari, C., Gallucci, I., Piccinini, N., Raffaelli,M., del Pozo, A., Álvarez, A., Arzelus, H., Cassaca, R., Luis, T., Neto, J., Mendes, C., Paulo, S., Viveiros, M. (2014). "*Automatic Live Subtitling: state of the art, expectations and current trends.*",NAB Broadcast Engineering Conference, April 5-10 2014, Las Vegas.

2.1.5 HBB4ALL

- Title: Hybrid Broadcast Broadband for All

- Tipology: European Project supported within the Seventh Framework Programme for Research and Technological Development (CIP-ICT-PSP-2013-7)

- Period: 2013-2016

- Consortium: UAB - Universitat Autonoma de Barcelona (ES), RBB - Rundfunk Berlin-Brandenburg (DE), IRT - Institut fuer RundfunkTechnik (DE), RTP - Radio e Televisao de Portugal (PT), TVC - Televisio de Catalunya (ES), Schweizerische Teletext (CH), Vsonix (DE), Vicomtech-IK4 (ES), Screen Subtitling Systems (UK), Holken Consultants and Partners (FR), People's Playground (NL), UPM - Universidad Politécnica de Madrid (ES).

Objectives

The project HBB4ALL addressed media accessibility possibilities in different platforms, including the new Hybrid broadcast-broadband TV (HbbTV) environment and Internet-based video services for PC and mobile devices. The project has tested access services thorough the following four interlinked pilot implementations (from the definition to the operational phase):

- Pilot-A: Multi-platform subtitle services.
- Pilot-B: Alternative audio production and distribution.
- Pilot-C: Automatic user Interface (UI) adaptation.
- Pilot-D: Sign-language translation services.

Besides, implicit and explicit user feedback has been gathered in order to assess the acceptance and the achievable quality of service across the several scenarios.

Regarding the Pilot-A, a new trend known as automatic multilingual generation of subtitles has been explored, considering the maturity of technologies such as Automatic Speech Recognition (ASR) and Machine Translation (MT) to support professional multilingual subtitling and increase productivity of broadcasters and subtitling companies. This pilot has integrated technology developed in the European SME-DCL SAVAS[1] and CIP-PSP SUMAT[2] projects, both led by Vicomtech-IK4, which addressed the development of technology for the automatic generation and automatic translation of subtitles respectively for several European languages. The HBB4ALL project has served as a new use-case scenario to test the integration of both technologies into a unique platform with the aim of generating multilingual subtitles in real-time on the news domain.

[1] http://www.fp7-savas.eu/
[2] http://www.sumat-project.eu/

Results

Three main components have been integrated into the developed pilot for the automatic generation of HbbTV compatible multilingual subtitles in real-time, as they are described below:

- Automatic Subtitling Component. It generates subtitles from the input audio signal. It applies Large Vocabulary Continuous Speech Recognition (LVCSR) technology from an English audio and creates EBU-TT-D format intralingual subtitles.

- Machine Translation Component. It works on top of Statistical Machine Translation technology. This component takes the English transcription created by the Automatic Subtitling Component and translates it into Spanish, creating EBU-TT-D Spanish subtitles.

- Live Broadcast-Internet Subtitle Viewer and Synchroniser Component. It creates an HTTP media stream from a broadcast video and audio source, and injects the automatically generated EBU-TT-D subtitles. The output of this component is an MPEG-DASH stream with the synchronized video, audio and the automatically generated intra- and inter-lingual subtitles in English and Spanish respectively.

In Figure 2.3 an overview of the integration of the main components into the developed pilot is presented.

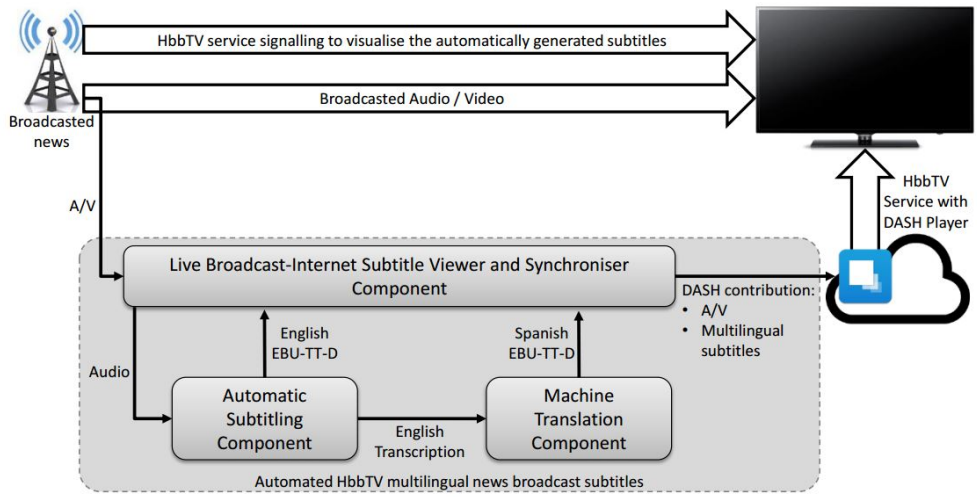


Fig. 2.3: High-level integration overview of the Automatic HbbTV multilingual subtitles pilot

The Automatic Subtitling Component has been developed to work in real-time and as it is shown in Figure 2.4, it is composed of a pipeline of processing modules, each providing a set of operational capabilities needed to automatically subtitle live audio-visual contents. The main LVCSR component is based on a DNN (Deep Neural Network) acoustic model trained using KALDI [Pov+11] and following the C++/CUDA DNN implementation written by Karel Vesely [Ves+13]. The language model (LM) is a trigram language model trained with the KenLM toolkit [Hea11].

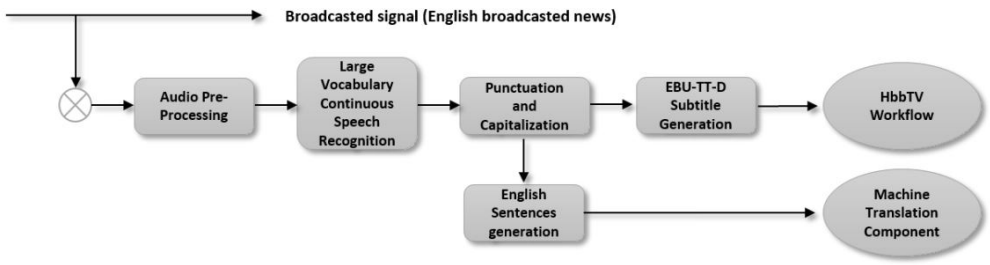


Fig. 2.4: Automatic Subtitling Component system

More information about this pilot can be found in the publication included in Section 7.19.4. This paper also presents the results of a preliminary experiment aimed to determine whether the automatically generated interlingual and intralingual subtitles help to better understand news content.

Related Publication

- [Mat+15] Matamala, A., Álvarez Muniain, A., Azpeitia Zaldua, A., and Oliver Moreno, A. (2015). *The reception of intralingual and interlingual subtitling.* In Translating and the Computer Conference (No. 37).

2.1.6 SSAB

- Title: Servicio de Subtitulación Automática Bilingüe
- Tipology: Industrial project
- Period: 2016
- Consortium: Irekia (Basque Government)

Objectives

The goal of the SSAB project has been the adaptation and transfer of the automatic subtitling technology to the Irekia use case. Irekia, as the embryo of the Open Government policy development in the Basque Country, serves as the direct communication channel between the general public and the Administration. Among other activities, Irekia records video contets with Government information and makes them available on its website. Although Irekia promotes web and content accessibility, it does not have enough resources to generate subtitles manually and sought for technology that could help them automate the process.

Results

Technology for automatic subtitling of video contents has been transferred from Vicomtech-IK4 to Irekia. The technology has been adapted with in-domain data, gathered from the contents produced by Irekia over previous years. The solution has followed the architecture of the Web Platform for automatic transcription and subtitling developed by Vicomtech-IK4, described in more detail in Section 3.4.6, and it is composed by the following five main technological components: Language Tracker, Large Vocabulary Continuous Speech Recognition, Capitalization and Punctuation, Normalization and Segmentation. The components were developed using the Transkit toolkit (Section 3.4.6), but exploiting texts from Irekia in order to adapt the technology to its domain.

2.1.7 Internal Research Activities

- Tipology: Internal Basic Research projects
- Period: 2012-Present

Objectives

These activities are not linked to any particular project, but are part of the work carried out at Vicomtech-IK4 to address a new research field and/or to improve proprietary technology with the aim of being more competitive. These research activities are commonly driven when a market need is detected and the proprietary technology is not mature enough to meet it.

Results

In the last years, we performed promising advances in the fields of automatic forced-alignment of long audios (see Section 3.4.4) and automatic segmentation of subtitles (see Section 3.4.5). Even if the automatic forced-alignment corresponds to a technology that has been extensively studied in the speech processing community over decades, the automatic segmentation of subtitles can be considered as a novel research field without known references in the literature. Research activities in both technologies have enabled us to improve proprietary baseline systems for automatic subtitling (see Section 3.4.6), and also to obtain promising results that have been published in scientific papers.

Related Publications - Long Audio Alignment

- [Bor+16] Bordel, G., Penagarikano, M., Rodriguez-Fuentes, L., Alvarez, A., and Varona, A. (2016). *Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks.* In IEEE Signal Processing Letters, vol. 23, no. 1, pp. 126-129.

- [ÁRA14] Álvarez, A., Ruiz, P., and Arzelus, H. (2014). *Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque.* In Text, Speech and Dialogue (pp. 473-480). Springer International Publishing.

- [ÁAR14] Álvarez, A., Arzelus, H., and Ruiz, P. (2014). *Long audio alignment for automatic subtitling using different phone-relatedness measures.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014 (pp. 6280-6284). IEEE.

- [RÁA14] Ruiz, P., Álvarez, A., and Arzelus, H. (2014). *Phoneme similarity matrices to improve long audio alignment for automatic subtitling.* In LREC, Ninth International Conference on Language Resources and Evaluation.

Related Publications - Automatic Segmentation of Subtitles

- [Álv+16c] Álvarez, A., Martínez-Hinarejos Carlos-D., Arzelus H., Balenciaga M., and del Pozo A. (2016) *Improving the Automatic Segmentation of Subtitles through Conditional Random Field*. Speech Communication, Elsevier. **Status: In 2nd revision.**

- [Álv+16b] Álvarez, A., Balenciaga, M., del Pozo, A., Arzelus, H., Matamala, A., Martínez-Hinarejos, Carlos-D. (2016). *Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles.* In Proceedings of the 10th international conference on Language Resources and Evaluation (LREC2016), pp. 3049-3053.

- [ÁAE14] Álvarez, A., Arzelus, H., and Etchegoyhen, T. (2014). *Towards customized automatic segmentation of subtitles.* In Advances in Speech and Language Technologies for Iberian Languages (pp. 229-238). Springer International Publishing.

2.2 Rich Transcription

2.2.1 CAPER

- Title: Collaborative information Acquisition Processing Exploitation and Reporting for the prevention of organized crime

- Tipology: European Project supported within the Seventh Framework Programme for Research and Technological Development (FP7-SEC-2010-1)

- Period: 2011-2014

- Consortium: S21sec (ES), Vicomtech-IK4 (ES), Synthema (IT), Fraunhofer-IGD (GR), Voice Interaction (PT), ALTIC (FR), Technion (GR), IKUSI (ES), ALMA (FR), Cnr-IIT (IT), UAB (ES), Studio Professionale Associaton a Baker & McKenzie (IT), Ministero dell'Interno (IT), Ministério da Justicia (PT), Ministerio del Interior (ES), Departament d'Interior - Generalitat de Catalunya (ES).

Objectives

CAPER's objective was to build a common collaborative and information sharing platform for the detection and prevention of organised crime. The techniques and technologies developed in CAPER had to be applied to Open Source Intelligence (OSI), including tools associated with the Social or Semantic Internet, and Close Source Intelligence, i.e., existing information systems in use by the Low Enforcement Agencies (LEAs). The technical components responsible of analyzing contents had to support multilingual and multimedia content processing, including speech, audio, text, image, video and biometrics. CAPER had to provide LEAs with a

common operational platform for OSI complemented by standards-based interface sets. This platform also had to allow easy integration with legacy systems and future applications.

Results

Within the whole architecture of CAPER, the Audio Analysis component was responsible of obtaining information from audio data sources, with the aim of being useful for inferring intelligence and to prevent organized crime. The Audio Analysis component supported the following 12 languages: English, Spanish, Basque, Catalan, Italian, Portuguese, French, German, Hebrew, Russian, Arabic and Romanian. Both proprietary (Audimus [Net+08]) and commercial Large Vocabulary Continuous Speech Recognition (LVCSR) systems were included in the CAPER platform, in order to obtain text transcriptions from the audio sources in all the supported languages. With the aim of selecting the appropriate LCVSR system for each audio file, a Language Identification module was included in the Audio Analysis component. This module identified the language in which a particular audio file was spoken.

The output of the Audio Analysis component was an XML file including information concerning audio segmentation and classification, speaker clustering and automatic speech transcription.

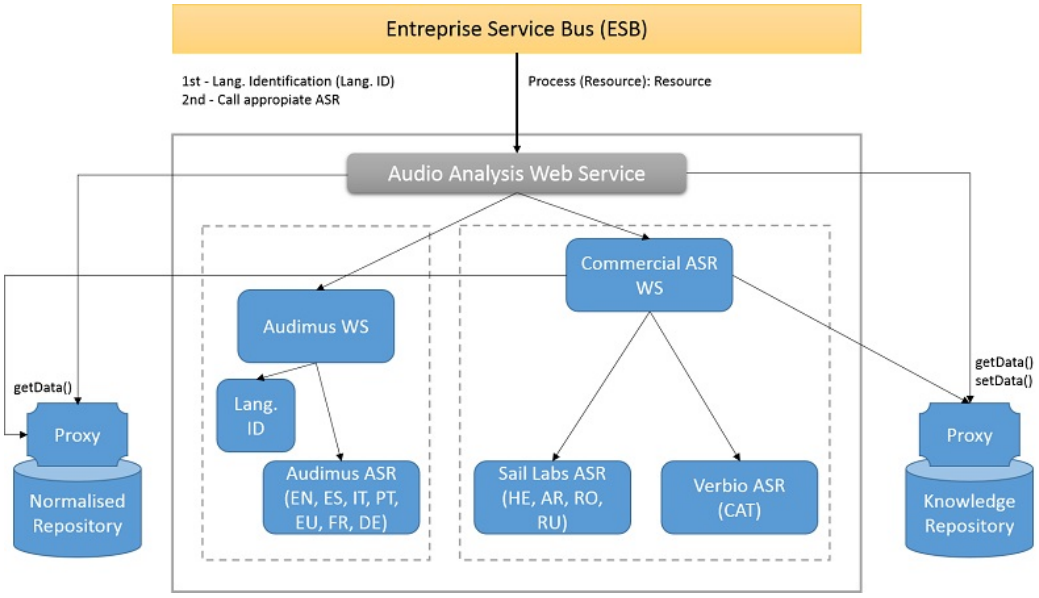


Fig. 2.5: CAPER Audio Analysis' workflow

The workflow of the Audio Analysis component of the CAPER prototype is presented in Figure 2.5. Considering the Service-Oriented Architecture (SOA) implemented within the CAPER project, the Enterprise Service Bus (ESB) orchestrator model was responsible of calling the Audio Analysis Web Service (WS) component to launch any

process in which an audio file had to be transcribed. This WS retrieved the content to be processed from the Normalized Repository. First, a Language Identification process was done using proprietary technology. Seven languages (English, Spanish, Italian, Portuguese, Basque, French and German) were transcribed through the Audimus system. For Hebrew, Russian, Arabic and Romanian the SailLabs ASR engine was integrated, whilst Verbio ASR technology was employed to transcribe Catalan contents. The XML result file was finally saved in the Knowledge Repository.

2.2.2 TE-PARLA

- Title: Transcripción Enriquecida de contenidos PARLAmentarios
- Tipology: Industrial project supported by the Hazitek programme
- Period: 2014-2016
- Consortium: SERIKAT Consultoría e Informática S.A.

Objectives

TE-PARLA aims at developing proprietary technology for the automatic rich transcription of contents generated during the Basque Parliament sessions. Within this project, SERIKAT is seeking for new methods to make its manual transcription processes more productive. In order to be useful, the TE-PARLA solution has to deal with the following constrains imposed by the Basque Parliament:

- Speed. For every 15 minutes of audio content, the transcription has to be ready and sent back in 75 minutes.
- Accuracy. The final transcriptions delivered to the Basque Parliament cannot contain mistakes. Hence, a post-editing process by SERIKAT is necessary to verify and/or correct possible automatic errors.
- Format. The transcriptions have to be delivered following specific format guidelines.

In addition to transcription, the technology has to provide information about the speaker turns and if possible identify specific speakers.

Results

The TE-PARLA solution has to provide technology for automatic rich transcription in the Basque Parliament domain including information about speakers (identification), and it has to work over bilingual contents, in which Spanish and Basque can be mixed interchangeably. To this end, a data set containing audio tracks and text minutes from the Basque Parliament sessions has been first collected. The compiled audio contains 10 hours and 19 minutes of Basque and 13 hours and 44 minutes of Spanish content annotated at transcription level. In addition, in-domain texts totaling 7.2 and 12.4 million words for Basque and Spanish were gathered.

Most of the integrated components have been taken from the proprietary rich transcription systems for Basque and Spanish (see Section 3.4.6) developed at Vicomtech-IK4; such as, the speech-non-speech detection, language tracker, capitalization and punctuation, and Normalization modules. All these components were trained and evaluated using the SAVAS corpus [Poz+14] and the data set collected within this project.

The technological components for the TE-PARLA solution have been adapted or generated from scratch as follows:

- Large Vocabulary Continuous Speech Recognition. Specific LVCSR engines adapted to the final domain have been developed for the TE-PARLA solution. While the acoustic models are the same models integrated in the proprietary rich transcription systems (see Section 3.4.6), the language model and, therefore, the vocabulary have been adapted using the in-domain texts collected from the Basque Parliament sessions. A WER (Word Error Rate) of 16.73% has been reached for Basque on a test set of 91 minutes, while a WER of 9.47% has been obtained for Spanish in 164 minutes of test set.

- Speaker Identification. Given that the Transkit SDK provides technology just for speaker segmentation and clustering, a speaker identification module has been developed for the TE-PARLA solution with the aim of identifying specific speakers in the audio tracks. This component has been constructed following the i-vectors (identity vectors) paradigm, for which a Universal Background Model (UBM) and a Total Variability (TV) matrix estimated with an in-domain corpus of 16 hours and 50 minutes and 105 speakers, have been used to collect statistics for i-vector extraction, and a probabilistic linear discriminant analysis (PLDA) back-end computes the similarity between i-vectors. With this approach, an accuracy of 85.34% has been achieved during the identification of speakers in an in-domain test set of 110 minutes.

A pipeline of the TE-PARLA solution is shown in Figure 2.6.

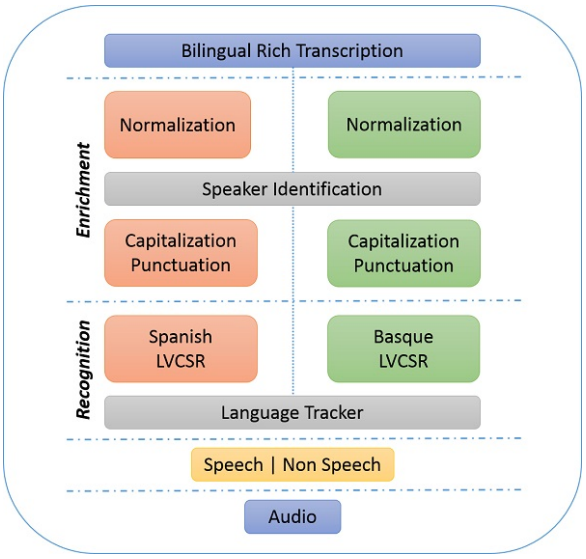


Fig. 2.6: TE-PARLA rich transcription solution pipeline

More information about this project can be found in the recently submitted paper included in Section 7.11.

Related Publication

- [Álv+16a] Álvarez, A., Arzelus, H., Prieto, S., and del Pozo, A. Rich Transcription and Automatic Subtitling for Basque and Spanish. In: Advances in Speech and Language Technologies for Iberian Languages. 2016. **Status: Submitted**.

2.3 Speech Emotion Recognition

2.3.1 RekEmozio

- Title: Reconocimiento automático de emociones humanas basado en el análisis de la voz y la expresión facial

- Tipology: University-Industry call funded by the University of the Basque Country (UPV/EHU)

- Period: 2004-2005

- Consortium: University of the Basque Country (UPV/EHU) and Innovae Vision S.L.

Description and Objectives

The RekEmozio project was focused on the emotional computation field. With the idea of developing more naturalistic interfaces, it aimed at recognizing human emotions automatically in real-time. The emotion recognition system had to be multimodal, including face expressions, speech and a combination of both sources. The technology was intended to be integrated into a modular prototype, which had to be evaluated and validated by end users and subsequently employed in commercial applications developed by Innovae Vision S.L.

The main objectives of the RekEmozio project were the following:

- To develop an emotion recognition system for Basque and Spanish based on the analysis of features extracted from speech.
- To develop an emotion recognition system based on the analysis of visual facial expressions of human beings.
- To implement a decision-making process to combine the data obtained from both of the above recognition systems. Following the studies presented in [PP97], it was estimated that this combination was useful to improve the emotion perception and recognition.
- To validate the emotion recognition systems with end-users, as well as to integrate them in a final prototype to be delivered to Innovae Vision S.L.

Results

Due to the scarcity of public data resources to train and evaluate emotion recognition systems, an affective dataset called RekEmozio [Lóp+09] was recorded in Spanish and Basque using actors. The recognition systems were trained and evaluated over this multimodal dataset, which was composed by audios, images and videos.

Several Machine Learning (ML) algorithms were applied to evaluate their performance on the Speech Emotion Recognition (SER) field. More specifically, Feature Subset Selection (FSS) techniques based on Estimation of Distribution Algorithms (EDA) were employed with the main objective of selecting the most significant

features for SER in Basque and Spanish. More recently, EDA was applied for the selection of the most relevant classifiers within a bi-level multi-classifier system known as Stacking Generalization.

Regarding facial expression analysis, some advances were carried out on the combination of frontal and profile facial image landmark tracking, facial feature fusion and combined facial expression recognition.

Related publications

- [Álv+16d] Álvarez, A., Sierra, B., Arruti, A., López-Gil, J. M., and Garay-Vitoria, N. (2015). *Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech.* Sensors, 16(1), 21.

- [Arr+14] Arruti, A., Cearreta, I., Álvarez, A., Lazkano, E., and Sierra, B. (2014). *Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction.* PloS one, 9(10), e108975.

- [Álv+07a] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007, September). *A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms.* In Text, Speech and Dialogue (pp. 423-430). Springer Berlin Heidelberg.

- [Álv+07b] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007). *Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech.* In Advances in Nonlinear Speech Processing (pp. 273-281). Springer Berlin Heidelberg.

- [Álv+06] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2006, September). *Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in spoken Spanish and Standard Basque Language.* In Text, Speech and Dialogue (pp. 565-572). Springer Berlin Heidelberg.

2.4 Speech-driven Facial Animation

2.4.1 PUPPET

- Title: Presentadores virtuales para aplicaciones de realidad mixta
- Tipology: Industrial project supported by the Gaitek program of the Basque Government
- Period: 2008-2010
- Consortium: Pausoka, Delirium Studios

Objectives

PUPPET aimed at embedding a virtual character in the real world to be easily handled by an actor/actress. This project mixed virtual reality and real image with real-time voice analysis. A simplified architecture of PUPPET is presented in Figure 2.7.

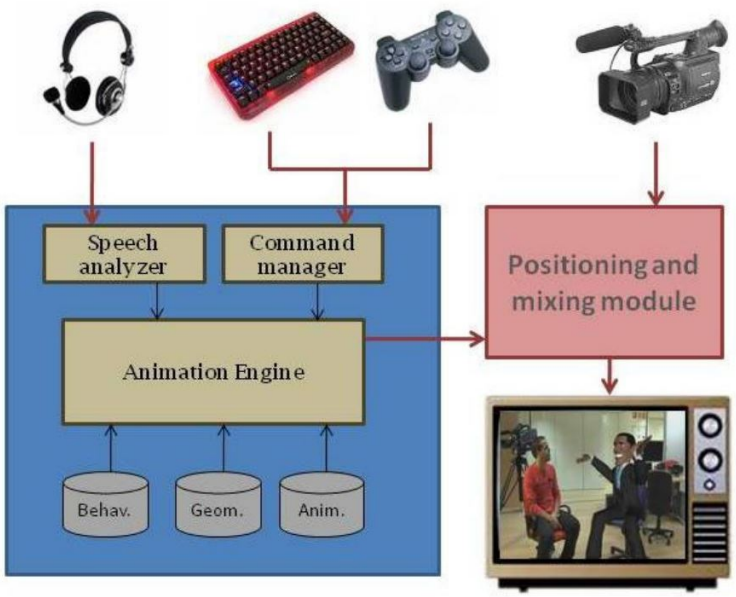


Fig. 2.7: Simplified PUPPET architecture

In order to reach the aim of the project, its main objectives were defined as follows:

- Easy-to-use: given that the platform could be employed by non-expert users, the system had to be non-intrusive and easy to use, including a simple and fast configuration process.

- Real-time avatar animation: besides having to be animated in real-time, the virtual character had to be of high quality in order to achieve a smooth mix between the virtual and real images.

- Speech-driven facial animation: the avatar had to be animated by the speech of the user in real-time.

Results

The PUPPET prototype provided a 3D virtual presenter that was integrated in a real TV scenario and could be handled by an actor in real-time, so that it could interact with real presenters and/or the public. It included a Speech Analyzer component, which provided the synchronization between the actor's speech and the avatar's lips as well as some facial expressions and animations.

The analyzer captured the speech signal from the standard audio input using a microphone and recognized the phonemes in real-time through a developed application based on the ATK API toolkit [You07]. As phonemes were recognized, they were mapped to their corresponding visemes, which correspond to the mouth shapes that represent phonemes. In a parallel process, the speech signal was processed by a pitch and energy tracking algorithm, in order to analyze its behavior and decide nonverbal facial movements.

More information about this component and the whole project can be found in the related publication included in Section 7.17.

Related Publication

- [Oya+10] Oyarzun, D., Mujika, A., Álvarez, A., Legarretaetxeberria, A., Arrieta, A., and del Puy Carretero, M. (2010). *High-realistic and flexible virtual presenters.* In Articulated Motion and Deformable Objects (pp. 108-117). Springer Berlin Heidelberg.

2.4.2 SPEEP

- Title: Sistema Para el Estudio y Entrenamiento de la Pronunciación

- Tipology: Industrial project supported by the Gaitek program of the Basque Government

- Period: 2011-2013

- Consortium: Mondragon Lingua, The Movie, Conexia

Objectives

The main objective of the SPEEP project was the development of a Computer Assisted Language Learning (CALL) platform to help students improve their pronunciation of a foreign language through the use of speech, NLP and virtual reality technologies. The SPEEP platform is presented in Figure 2.8.

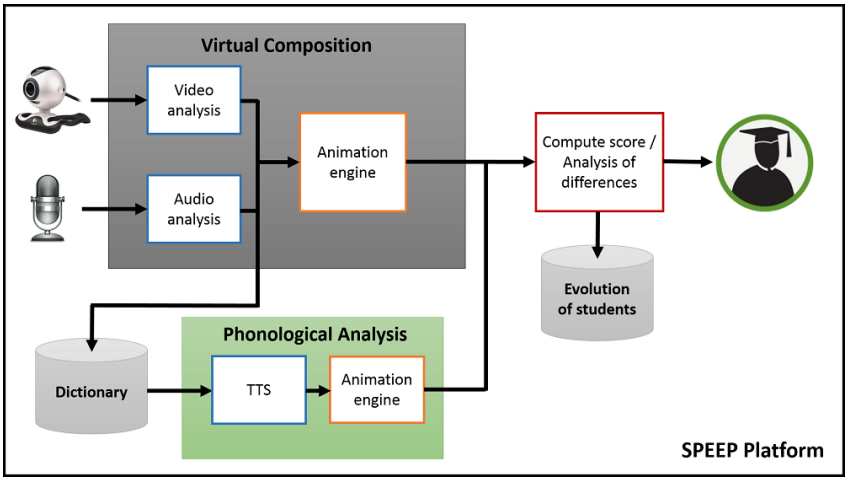


Fig. 2.8: The main platform of the SPEEP solution

The platform had to deal with the following requirements:

- **Input**. The solution had to allow voice and/or text as standard inputs. In the case of voice, it had to be processed by a speech recognition engine to obtain the transcription automatically.

- **Languages**. The platform had to operate for Basque and English.

- **3D reconstruction**. The animation had to be done at two levels; the first one employing the natural videos and audio recorded from the webcam and the microphone, and the second one using a Text-to-Speech engine and an animated avatar. This way, through the use of advanced techniques of co-articulation and 3D graphics, the platform had to generate 3D virtual reconstructions considering both the correct way to pronounce the input and the way in which the user had pronounced it.

- **Error analysis**. The solution had to compare the canonical pronunciation with that generated by the user, and it had to report pronunciation errors automatically at graphical and phoneme levels.

- **Progress and history**. The user could practice indefinitely. The platform had to evaluate the progress of the student and propose new sentences and words to help improve his/her most common pronunciation errors.

Results

With regard to speech processing, pronunciation evaluation was performed at alignment level following the work presented in [SS12b]. This work was based on computing the distance of the acoustic alignment scores and durations of phonemes between the input pronunciation and its related canonical representation. To this end, 10 native speakers recorded 21 sentences per language that were stored in the platform. These canonical recordings were force-aligned to obtain the two main parameters corresponding to the alignment scores and durations at phoneme level. Afterwards, for each sentence and main parameter, statistics such as the mean, standard deviation and the maximum and minimum $z - score$ values (see Formula 2.1) were computed at phoneme level including all the native speakers in the database. The resulting values were finally stored in a reference matrix.

During evaluation, the alignment score, duration and $z - score$ values for each phone in the test were computed. The first two values were calculated through a force-alignment process, whilst the $z - score$ was estimated using the following formula:

$$z - score = \frac{x - \mu_i}{\sigma_i} \tag{2.1}$$

where x describes the value of the main parameter (alignment score or duration) of the current phone, and μ_i and σ_i correspond to the mean and standard deviation of this particular phone and parameter in the reference matrix. This way, one $z - score$ value was computed for each parameter and phone in the test file. This value was then normalized from 1 to 5 for each parameter considering the maximum and minimum $z - scores$ values of each phone in the reference. The final evaluation score corresponded to the sum of the two normalized $z - score$ values of each main parameter. Forced-alignment was carried out using the CMU Sphinx recognition system [Lam+03] and triphone-based HMM acoustic models trained

with the TIMIT corpus [Gar+93] for English and an internal database of 21 hours of Basque recordings.

On the other hand, in order to deal with the co-articulation effect, a syllabification process was performed over the phones, and the animation was launched using 3-grams of syllables and following the facial synthesis method described in [CM93].

Related Publication

- [Muj+13] Mujika, A., Diez, H., Alvarez, A., Urteaga, M., and Oyarzun, D. (2013). *Realistic visual speech synthesis in WebGL*. In Proceedings of the 18th International Conference on 3D Web Technology (pp. 207-207). ACM.

2.5 Other Related Projects

2.5.1 SABioV

- Title: SABioV
- Tipology: Industrial project
- Period: 2015-2016
- Consortium: Bantec Group

Objectives

The aim of the SABioV project was to delevop a Speaker Verification system for secure authentication applications. This system had to be implemented for mobile environments and devices, allowing users to use it from iOS and Android operating systems. The main constrains of the solution to be developed were as follows:

- Enrollment and Authentication. The SABioV solution had to integrate technology for both Enrollment and Authentication processes. During Enrollment, the speaker's voice is recorded, analyzed and a number of features are extracted to create a voice print or model. This model will then be employed in the Authentication phase, in which a new voice will be recorded, processed and compared with the previously generated model.

- Language and Domain. The SABioV solution had to work for English utterances, which had to be composed of five digits from 0 to 9.

- Performance. The system had to work fast (less than five seconds for each process), and obtain accuracy rates higher than 90%.

Results

The SABioV solution was developed following a distributed architecture, composed by a mobile device (front-end) and a server (back-end). The Mobile App included the components needed to manage users and to provide them with the functionalities to perform the enrollment and authentication processes. In addition to the user interface, the front-end also included the functions needed to record the user's voice and to extract and normalize the Mel-Frequency Cepstral Coefficients (MFCC), as they were the features selected for the parametrization of the audio signals.

The MFCC coefficients were processed on the server (back-end), where the biometric technology was installed, and i-vectors were generated for both Enrollment and Authentication processes.

The Figure 2.9 describes the main distributed architecture used for SABioV.

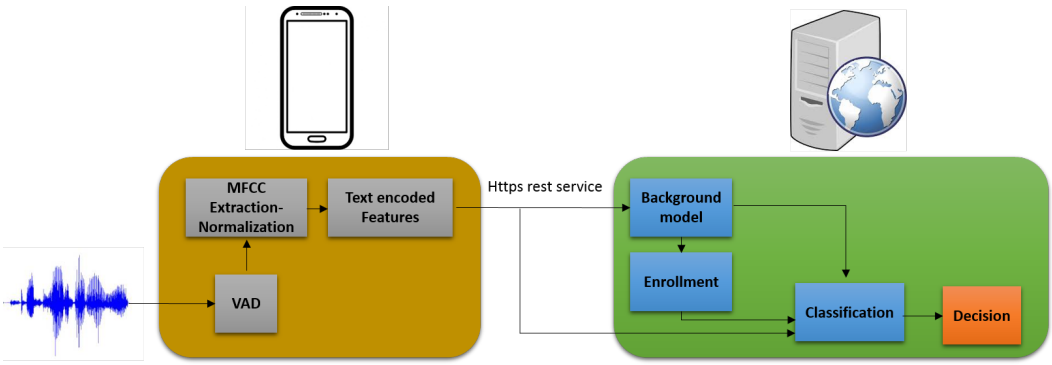


Fig. 2.9: SABioV distributed architecture for Voice Biometrics

The technology was developed using Spro [Gra03] for parametrization and the ALIZE toolkit [BWM05] to train models and implement the Enrollent and Authentication processes. The models were trained and evaluated using the RSR2015 [Lar+12] corpus, in addition to an internally recorded dataset, which was created following the project specific needs. One hundred speakers volunteered to participate in the generation of this dataset, each speaker recording 10 sentences in two acoustic conditions (clean and noise). The biometric technology was constructed following the i-vector paradigm and several configurations and normalization techniques were tested within the project. The model with best performance was estimated using the

Eigen Factor Radial (EFR) normalization function and the Probabilistic Discriminant Analysis (PLDA) scoring technique. This model achieved accuracies of 97.51% and 97.25% for the True Acceptance Rate and True Rejection Rate metrics on a test set composed by 601,400 one-to-one comparisons of speaker models.
Regarding the VAD (Voice Activity Detection) module, it was based on a time-frame analysis and the computation of the Fundamental Frequency (autocorrelation method) and the Energy (Root Mean Square) values, which were taken as the basis features to estimate the presence of speech.

2.6 Summary

The projects described above encompass the two main technological environments and four contribution fields presented in the current dissertation work. Following a chronological order, RekEmozio on the one hand, and PUPPET and SPEEP on the other, supplied the Multimedia Interaction Environment with thorough Speech Emotion Recognition and Speech-driven facial animation technologies respectively. RekEmozio was a first attempt to the Speech Emotion Recognition field, which triggered many posterior research activities related to different feature extraction and classification methods. PUPPET allowed the development of a real-time phone decoder based on the ATK/HTK speech recognition toolkit with the aim of synchronizing the natural voice of the user with the lips of a virtual character. In addition, and employing the knowledge acquired in RekEmozio, an online prosody analysis module was integrated for facial animation. The co-articulation and facial animation techniques were further improved in the SPEEP project, in which a module for the evaluation of the pronunciation for Basque and English was also implemented.

The rest of the projects belong to the Audiovisual Environment and are mainly related to the Automatic Subtitling and Rich Transcription fields. APyCA can be considered the first project in which a whole platform for automatic subtitling was developed in Spanish, employing state-of-the-art and commercial tools. The developed HTK-based alignment module was then used and extended to Basque within the SUSA solution, in which automatic subtitles were created from aligning audio and its transcript following predefined specific rules. The European SAVAS project allowed the development of pioneer systems for live and batch automatic subtitling for several European languages, including Basque and Spanish, in addition to the collection of large Basque and Spanish corpora, prepared and annotated for the development of several speech processing components. Exploiting such data resources, proprietary automatic subtitling and rich LVCSR transcription technology was developed within the TE-PARLA, HBB4ALL and SSAB projects, in which the corresponding adaptation tasks were performed in order to adjust the technology

to each specific use case and domain. SABioV served as a framework to carry out research in Speaker Verification technology based on the state-of-the-art i-vector paradigm. The knowledge acquired in this project was employed for the Speaker Identification component developed in TE-PARLA.

The complementarity of all the projects above has driven the continuous improvement of the technological components integrated within the various automatic subtitling, rich transcription and speaker verification systems developed. As a result, two Software and Development Kits (SDK) have been implemented within the Speech Technologies research line at Vicomtech-IK4: Transkit[3] and BioVoice[4], which include the functions needed to train, build and evaluate all the technological components involved in the development of Automatic Subtitling, Rich Transcription and Voice Biometrics applications. Both toolkits are licensed by Vicomtech-IK4 for their transfer to Industry.

A high-level schema of the relationship between the aforementioned projects and the developed technological components can be seen in Figure 2.10.

[3]http://www.vicomtech.org/resources/archivosbd/sdks_documentos/Transkit.pdf
[4]http://www.vicomtech.org/resources/archivosbd/sdks_documentos/BioVoice.pdf

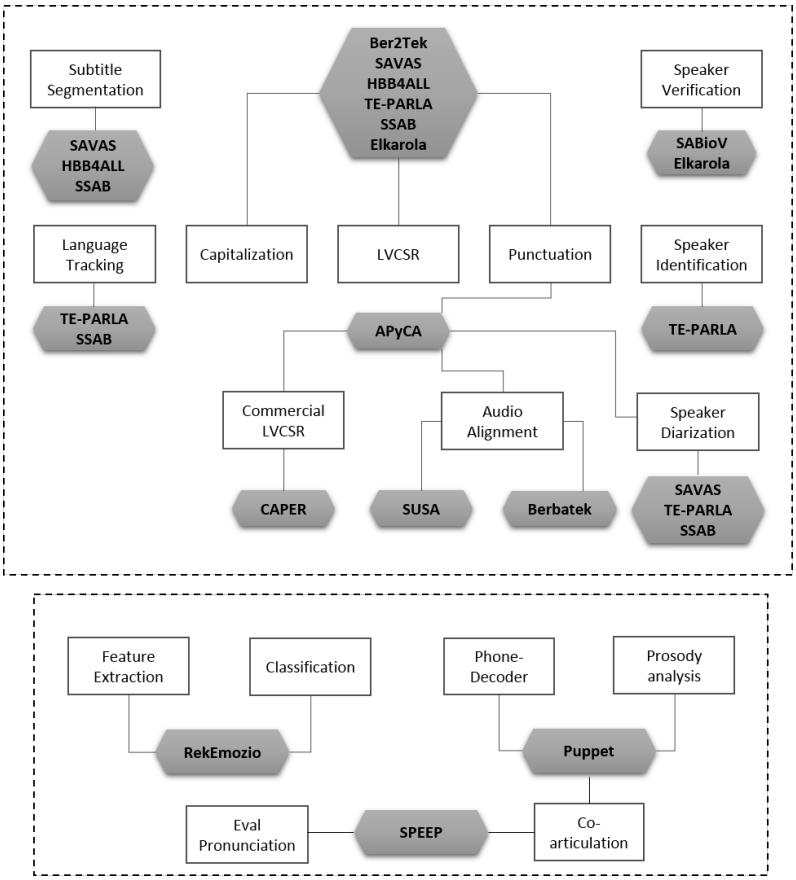


Fig. 2.10: Schema of the relationship between the R&D projects and the technological components

3 Audiovisual Environment

3.1 Introduction

As technology progresses, increasingly competitive open source tools are available and novel modeling paradigms overcome previous techniques, more sophisticated speech and audio applications are being developed and gently introduced to extract information from audio contents. In this sense, Audio Content Analysis (ACA) can be defined as the automatic extraction of semantic information from audio clips [Bur+08], which may include segments of speech, music and/or noise. One of the aims of ACA is therefore to implement a set of technologies such as audio classification, speech recognition, speaker segmentation or music analysis, among others, which solve specific needs of semantic access. Rich Transcription and Automatic Subtitling can be considered two applications of ACA, since they focus on annotating spoken content and optionally recovering information about speakers and some required particular noises.

Whereas classical speech recognition focuses on converting a sequence of spoken words into a sequence of textual words, Rich Transcription aims to produce more highly annotated and informative output. Rich Transcription can be defined as the technology that allows the automatic generation of a transcript from an audio clip along with metadata to enrich the word stream with useful information such as punctuation, capitalization, speaker identification, sentence units and proper names, among others. Due to the increasing amount of digital media on the Internet, Rich Transcription has become a widely used technology for applications such as spoken document retrieval, spoken term detection, summarization, semantic navigation, speech data mining and annotated automatic transcription.

Automatic Subtitling refers to the technique of producing subtitles without human intervention. It was born in response to a high subtitling demand, as a more productive alternative to the manual process. The aim of Automatic Subtitling is, on the one hand, to assist professionals produce subtitles faster, either generating final subtitles or post-editing automatic draft subtitles, and on the other, to enable subtitling in challenging situations, such as live broadcasts, where traditional subtitling was not directly applicable. Automatic Subtitling includes many technological components, mostly related to speech processing technologies, in order to automatically generate

subtitle text from audio. Its core component corresponds to a Large Vocabulary Continuous Speech Recognition (LVCSR) engine, which is in charge of transcribing the spoken content into text.

Subtitles have acquired great relevance during the last few years, mainly after the adoption of the new audiovisual directive (Article 7 of the Audiovisual Media Services Directive[1]) of the European Parliament and Council in March 2010. This legislation regulates the rights of people with visual or hearing disabilities, and is pushing member states to take the necessary measures to guarantee that the services of audiovisual providers under their jurisdiction are gradually more accessible by means of sign-language, audio-description, easily menu navigation and subtitling.

As a result of this new legal framework, broadcasters and subtitling companies have been moved to subtitle a high percentage of their broadcast content, and this has forced them to seek for automatic solutions to speed up the subtitle generation process. The large effort in research and development of LVCSR over the last decade has resulted in significant improvements on multimedia data transcription, making it the most powerful technology available to increase productivity in several automated intralingual subtitling tasks.

In order to comply with the new audiovisual legislation, broadcasters started focusing their increased subtitling effort on quantity. However, an increasing demand driven by society and disability organizations to improve the quality of subtitles has arised recently as well. The quality of subtitles involves several features linked to subtitle layout, duration and text editing [Álv+15], subtitle segmentation being one of the most relevant ones. As proved in [Raj+13], a correct segmentation by phrase or by sentence significantly reduces the time needed to read subtitles. Furthermore, the strong need for proper segmentation is supported by the psycholinguistic literature on reading [DR89], where the consensual view is that subtitle lines should end at natural linguistic breaks to improve readability and reduce the cognitive effort produced by poorly segmented text lines [Per+10].

This chapter presents the main contributions made to the fields of automatic subtitling and rich transcription. Regarding automatic subtitling, the work has been mainly focused on (1) the implementation of pioneer subtitling systems for several European languages, (2) the collection of a vast amount of audiovisual resources to train speech processing based systems, (3) the definition of a new metric to evaluate the overall quality of subtitles, (4) new methods to improve long audio alignment, and (5) novel methodologies for the automatic segmentation of subtitles.

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0013&from=EN

Concerning rich transcription, proprietary systems for Basque and Spanish adapted to the Basque Parliament domain are presented.

3.2 State of the Art

There is extensive research focused on automatic subtitling, mainly through the use of LVCSR technology for the recognition and alignment tasks [Net+08; Bor+12; Álv+15]. Most of the work in the area has been focused on improving recognition accuracy and producing well-synchronized subtitles. In this sense, Audimus [Net+08] can be considered as a pioneer and reference system in the field, as it provides a complete framework for the automatic subtitling of broadcast news in both batch and live modes in several languages. This framework is composed by a pipeline of technological modules, including (1) an Audio Pre-Processing block, which discriminates between Speech and Non-speech segments and performs speaker clustering, speaker gender and identification of relevant speakers, (2) a LVCSR engine, (3) a Normalization module that converts sequences of words representing digits, connected digits and numerals into numbers and additionally capitalizes proper names and introduces punctuation marks, and finally (4) a Subtitling Generation component that generates subtitles according to the subtitling rules and specifications of each broadcaster. The current system is described in more detail in the work presented in [Álv+15].

Recently, Google has also started to support the automatic generation of time-aligned draft transcriptions and subtitles of the videos uploaded to Youtube [Goo09; LMS13]. Nevertheless, for the moment Youtube's automatic transcriptions do not include punctuation and capitalization marks nor follow standard professional subtitling practices [Álv+15]. Other companies such as Koemei[2], SailLabs[3], Vecsys[4] and Verbio[5] commercialise automated transcription solutions for varying pools of languages and application scenarios, but do not produce subtitles.

LVCSR technology has been exploited commercially mainly for dictation and command-based interaction applications through well-known solutions, such as Microsoft Windows Speech Recognition[6] and Nuance Dragon Naturally Speaking[7]. However, the unsuitability of these dictation engines for automatic audio transcription has limited their integration into automatic subtitling systems so far.

[2]https://koemei.com/
[3]https://www.sail-labs.com/
[4]https://www.bertin-it.com/vecsys/
[5]http://www.verbio.com
[6]http://www.windows.microsoft.comen-uswindows7dictate-text-using-speech-recognition
[7]http://www.nuance.com

Over the last few years, advances in both machine learning algorithms and computer hardware have led to significant improvements in LVCSR technology, mainly through the use of Deep Learning. Thereby, different research groups have shown that DNNs (Deep Neural Networks) can outperform traditional GMMs (Gaussian Mixture Models) at acoustic modeling for speech recognition on a variety of data sets [Hin+12].

A wide range of approaches have made important contributions in the past decade to the four main components of LVCSR systems, which correspond to (1) front-end processing, (2) acoustic modeling, (3) language modeling, and (4) back-end search and system combination.

1. Regarding front-end processing, much work has been done on several areas such as feature transformation using LDA (Linear Discriminant Analysis) [Sao+00] and STC (Semitied Covariance Transform) [Gal98], noise robust features using SPLICE [LG08] or QE (Quantile-based histogram Equalization) [HN06], speaker-adaptative features using VTLN (Vocal Tract Length Normalization) [LR98] and fMLLR (feature-space Maximum Likelihood Linear Regression) [Gal98], and discriminative features using fMPE (feature-space Minimum Phone Error) [Pov+05], Neural Networks [HES00] and bottleneck features [Gré+07]. All these techniques have led to improvements on various LVCSR tasks.

2. In the field of acoustic modeling, feature- and model-space discriminative training based on BMMI (Boosted Maximum Mutual information) [Pov+08] on the one hand, and MPE-based speaker adaptation [WW04] on the other, achieved competitive recognition results. More recently, DNN acoustic models using generalized maxout networks [Zha+14] have brought significant improvements.

3. Regarding language modeling, backoff smoothing using HPYLM (Hierarchical Pitman-Yor Language Model) [Teh06], large-span modeling using Maximum Entropy [Ros96], and syntactic and Recurrent Neural Network language models [Mik+11] obtained competitive performances.

4. Finally, in the area of hypothesis search, dynamic and WFST Viterbi decoding [SS09] and system combination using boosting [SS12a] have recently demonstrated their suitability for LVCSR systems.

All these technological advances and the availability of toolkits such as Kaldi [Pov+11], which includes recipes for front-end processing, acoustic model training and for the construction of WFST decoders following the last paradigms of the

research community, plus other tools such as RNNLM [Mik+11] that allow the estimation of Recurrent Neural network based language models, have fostered the development of sophisticated LVCSR engines, enabling their integration into automatic subtitling frameworks with high performances and low error rates in bounded domains.

Another efficient approach when the script or transcription of the content is available, is the application of a speech-text alignment algorithm, which relies on aligning audio with its script to automatically recover time-stamps. The synchronization of text transcripts and audio tracks is typically solved by a forced-alignment algorithm at phonetic level. However, forced-alignment becomes challenging with long audio signals, because of the widely-used Viterbi algorithm, which forms very large lattices during decoding, requiring a lot of memory. Besides, acoustically imperfect transcripts demand more complex methods.

Much work has been done in the field of automatic alignment of audio tracks and their transcripts. Many of the related studies have taken the work presented in [Mor+98] as reference, where the forced-alignment task was turned into a recursive and iteratively adapted speech recognition process. They used dynamic programming to align the hypothesis text and the reference transcript at word level. Subsequent works proposed improvements of this system to deal with imperfect transcripts [Lec+06; HK07]. More recently, other approaches have been focused on developing simple alignment procedures for long audio recordings [HP13; Ahm+13; ALG14]. In this context, an efficient and simple long audio alignment approach was presented in [Bor+12]. They developed a system based on Hirschberg's dynamic programming algorithm [Hir75] to align the phone decoder output with the transcription at phoneme level, using a binary matrix to score alignment operations.

Concerning the quality of subtitles, several studies collect good subtitling practices [Kar98; IC98; DR07; For09; Aen12; Ofc15]. Although according to our experience each broadcaster and subtitling company uses its own subtitling conventions and rules, the above studies agree that the main characteristics of good subtitles can be classified into the following features:

- Spacing features: distributing the text in one or two lines between 4 and 43 characters.

- Timing features: showing subtitles at a speed of 130-170 words per minute and keeping them on screen between 1 and 6 seconds, while synchronizing with the audio and inserting short pauses between consecutive subtitles.

- Linguistic features: keeping the original terms and avoiding more than two sentences per subtitles, one per line.

- Orthotypographic features: following the general guidelines of printed text.

As technology is increasingly integrated into the subtitling field, objective metrics need to be defined to evaluate the quality of the automatically generated subtitles. In this context, the NER model [RM11] has been used in recent years to measure live subtitle errors. Based on the NERD model presented in [Rom11], it was adapted to suit the needs of different live subtitling techniques, and re-speaking in particular. The NER model employs the following formula to determine the quality of live re-spoken subtitles:

$$NER = \frac{N - E - R}{N} \times 100 \tag{3.1}$$

where N is the number of words in the re-spoken text, E corresponds to the edition errors caused by the respeaker's strategies, and R are the errors committed by the recognizer. Computing the formula, a NER value of 100 indicates that the content was subtitled entirely correctly. Good quality live subtitles are expected to go beyond 98% accuracy according to this formula.

Since it was devised mainly for quality assessment of re-spoken subtitles, the NER model only considers recognition and editing errors. However, speaker color, timing and splitting information is also relevant in automatic subtitling and helps to establish subtitle quality.

Finally, concerning methods to perform automatic subtitle segmentation, a survey of the literature provides no references in the field. A few studies were carried out on related topics, as for instance [Per08], which explores the way line-breaking is commonly performed or [Per+10], which studies the impact of arbitrary segmented subtitles on readers. The importance of segmentation has been noted by [Raj+13], a study whose aim was to verify whether text chunking over live re-spoken subtitles had an impact on both comprehension and reading speed. They concluded that even though significant differences were not found in terms of comprehension, a correct segmentation by phrase or by sentence significantly reduced the time spent reading subtitles. None of the related works includes technology to automatically segment subtitles properly.

3.3 Challenges in the fields

The current challenges linked to automatic subtitling and rich transcription involve improving the various technological components which compose the common pipeline of these systems. These components correspond to (1) the module responsible for generating raw text transcriptions with time-stamps from the source audio using a LVCSR engine or audio-to-text forced-alignment technology, (2) the enrichment of the raw text with capitalization and punctuation marks, (3) the segmentation and identification of specific speakers, (4) text normalization, (5) subtitle segmentation, in the case of automatic subtitling, and final presentation. Besides, these systems can work in batch and live modes, providing subtitles and transcriptions for pre-recorded contents or in real-time respectively. The latter mode holds more difficulties, since all the information has to be extracted in the shortest time with the smallest possible delay.

In the following subsections, the current main challenges for each of the technological modules described above are presented.

LVCSR and Audio-To-Text Alignment

Even if automatic recognition technology has evolved enough to provide competitive results for speech dictation in clean environments, its performance still degrades for spontaneous speech and in noisy environments. Its low performance in such situations can result in automatic subtitles with serious errors that can change the meaning of a sentence or turn it unintelligible.

Several techniques and approaches have been proposed in the last decades to deal with robust speech recognition in noise environments, both at feature- [SC12; Yu+08] and model-levels [Tac+13; SYW13; Wen+14], but there is still room for improvement on the construction of acoustically noise robust LVCSR systems. In a similar way, recognition error rates for spontaneous speech are still unacceptably high [Sin+13].

Since forced-alignment is based on speech recognition technology, this technique also degrades performance under acoustically noisy environments and when facing spontaneous speech. Furthermore, forced audio-to-text alignment becomes problematic with long audios due to the associated higher memory demand, increased processing time and lower reliability of the commonly employed Viterbi search algorithm when aligning long sequences.

In this memory, significant contributions are proposed in the field of long-audio forced-alignment for several languages and acoustic conditions (see Section 3.4.4), with the aim of overcoming the computational and technical problems described above.

Capitalization and Punctuation

Recovering capitalization and punctuation marks of spoken transcripts is still a challenging task for the speech and natural language processing community. Automatic capitalization is commonly context-dependent and has been studied in several works through different approaches based on language models [GJB09], rule-based taggers [Bri94], maximum entropy Markov models [CA06] and condition random fields [WKM06]. However, the ambiguity of the context and unseen words during training produce more errors than desired, especially in open domains.

Autopunctuation is even more domain-dependent than Capitalization, especially with different types of speech (expressive, planned, dictate, etc.) and if acoustics and prosody are employed as features to train models. Therefore, in order to guarantee a correct performance, autopunctuation models are usually built per domain and/or speech type. Although different punctuation marks can be used, most studies have focused on recovering the most frequent full stop and comma, although comma is quite problematic due to its multi-functionality [Bat+12].

The Capitalization and Punctuation results achieved by the automatic subtitling systems presented in [Álv+15] can be considered a reference for languages such as Basque and Spanish. In these systems, both Capitalization and Punctuation were treated as a classification problem, sharing the same technical approach based on logistic regression classification models, which corresponded to the maximum entropy classification for independent events. F1-score values of 83.5% and 85.7% were achieved for Basque and Spanish respectively, concerning automatic capitalization. With regard to the automatic recovery of punctuation marks (full stops and commas), F1-score values of 50,90% were reached for Basque and 39,90% for Spanish. New processing techniques with more competitive results for automatic Capitalization and Punctuation have been developed and integrated into the proprietary automatic subtitling and rich transcription systems presented in Section 3.4.6.

Speaker Segmentation and Identification

As it was pointed out by [HW07] some years ago, results of a speaker diarization system were (and still are) dependent on both evaluation data and the performance of other integrated components, such as the Voice Activity Detection module. Five

years later, [Mir+12] reported that Speaker Diarization was not yet mature enough for models to be portable across domains without suffering a performance drop. However, the main challenge nowadays is probably dealing with overlapping speech, not only during the clustering process, but also when identifying speakers within such type of complex segments. In this sense, systems which do not consider overlapping speech will always produce significant segmentation and speaker identification errors [SK13]. Besides, issues such as background noise, reverberation, short speaker turns and interruptions may also degrade the performance of a speaker diarization system. Recently, work in the field has focused on two other open challenges: speeding up the diarization process and performing cross-show speaker diarization [Del+15]. The first line of work derives from the need of processing the increasingly large amount of audiovisual content being produced, whilst cross-show speaker diarization expands the task to a broader context, in which recurrent speakers participate in different audio tracks of a specific database and have to be identified with the same label.

In most of the state-of-the-art systems, Gaussian Mixture Models (GMMs) and metrics like Bayesian Information Criterion (BIC) or Generalized Likelihood Ratio (GLR) are commonly used for the detection of speaker change points, and for speaker cluster modeling, cluster merging and as the stopping criterion. Recently, factor-analysis based techniques, such as i-vectors, which are popular in the speaker verification domain, have been adapted to the speaker diarization task with the aim of discriminating the variability posed by channel characteristics, ambient noises and spoken phonemes. This way, a state-of-the-art diarization module based on the i-vector paradigm was developed and integrated in the proprietary automatic subtitling and rich transcription system presented in Section 3.4.6.

Text Normalization

The Text Normalization module aims at converting numbers, numerals, dates and amounts (e.g. money and percentage) to their digit representation. This operation is commonly performed through rule-based functions per each language, and it does not mean any particular challenge for the community.

Subtitle Segmentation

Automatic segmentation of subtitles is a novel research field which aims at providing syntactically coherent breaks so that viewers can read subtitles as quick as possible. Although it can be seen as a text segmentation problem, there are several features related to subtitling that have to be considered at the same time, and that makes this task challenging. Apart from the text analysis, a correct segmentation of subtitles

depends on (1) the amount of characters allowed per line, (2) timing issues related to long pauses and speech rhythm, (3) speaker changes, (4) the preceding and posterior words in order to select the most appropriate type and point of break and (5) the persistence of the subtitle on screen, which has a real impact in its readability.

To date, most of the developed automatic subtitling solutions have not been able to discriminate the natural pauses, syntactic and semantic information relevant for quality segmentation and, thus, automatic segmentation in stenotyping, respeaking or audio transcription applications is mostly performed only considering the maximum number of characters allowed per line or through manual intervention [Álv+16b]. With the aim of automating this task and providing the first automatic subtitle segmentation solutions with syntactically coherent breaks to the community, two machine learning based approaches have been proposed as main contributions to the field (see Section 3.4.5).

3.4 Main contributions

The main contributions of this dissertation work to the fields of automatic subtitling and rich transcription are described in the following subsections.

3.4.1 Automatic Live and Batch Subtitling systems

As technological part of the European SAVAS project, LVCSR based full systems for automatic subtitling of audiovisual contents were developed for several European languages, such as Portuguese, Spanish, Basque, Italian, French, German and the Swiss variants of the latter three. All the systems were initially trained and adapted to the Broadcast News domain, whilst the Portuguese system was further extended to interviews and debates, which encompass a more difficult domain due to artifacts such as repetitions, hesitations, disfluences, unfinished sentences, overlapping speech etc.

Three types of applications were built for several subtitling purposes. The first application was a batch Speaker Independent Transcription and Subtitling application (S.Scribe!), capable of automatically transcribing pre-recorded audio and video files into time-aligned enriched subtitles. The second application corresponded to an Online Subtitling System (S.Live!) and was able to automatically transcribe live audio into configurable and well-formatted subtitles. The final application involved a Respeaking engine (S.Respeak!) for dictation, which could be easily integrated into

Tab. 3.1: LVCSR based SAVAS systems and applications per language

Language	S.Scribe!	S.Live!	S.Respeak!
Portuguese	Broadcast News Interview/debate	Broadcast News Interview/debate	News –
Spanish	Broadcast News	Broadcast News	News
Basque	Broadcast News	Broadcast News	Sports and News
Italian	Broadcast News	Broadcast News	Sports and News
Swiss Italian	Broadcast News	Broadcast News	News
French	Broadcast News	Broadcast News	News
Swiss French	Broadcast News	Broadcast News	News
German	Broadcast News	Broadcast News	News
Swiss German	Broadcast News	Broadcast News	News

any commercial subtitling solution and was capable of producing subtitles with an acceptable delay. Table 3.1 summarizes the LVCSR based systems and applications that were developed per language and domain within the SAVAS project.

All the systems were trained using the SAVAS corpus [Poz+14] compiled within the project (see Section 3.4.2). As it is shown in Figure 3.1, the SAVAS systems can be represented as a pipeline of processing blocks, which represent the different technological components involved. Each of the technological components is described in detail in [Álv+15].

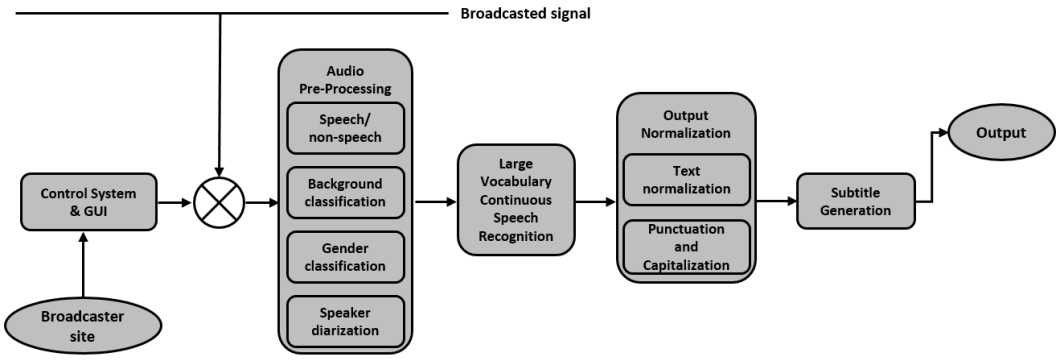


Fig. 3.1: Pipeline of the SAVAS subtitling systems

On the other hand, a number of metrics were employed, some of which were designed specifically for this work, with the aim of measuring the performance of the SAVAS applications with regard to the subtitling quality features accepted by the Industry, including the delay of showing subtitles and their persistence on the screen. Besides, each technological component was individually evaluated, and measures of the global quality of the systems were computed following the metrics described

in [Álv+15]. In addition, a productivity gain evaluation was performed with post-editors. In this sense, subtitling professionals in all languages were asked to post-edit automatic subtitles and to create them from scratch, using their usual subtitle editing software and quality standards. Finally, the S.Scribe! systems developed for Spanish, Italian, French and German were compared to the Google speech transcription technology integrated in the Youtube platform. Since the Google transcriptions do not include punctuation and capitalizaton, results were compared at WER level, improving the error rates from more than 4 points in German to 12 in French, as it can be seen in Table 3.2.

Tab. 3.2: WER metrics for S.Scribe! and Youtube application

Language	Duration	S.Scribe!	Youtube
Spanish	2 H	16.50%	24.91%
Italian	2.5 H	17.81%	27.80%
French	2 H	25.81%	37.61%
German	2 H	26.92%	31.10%

The SAVAS systems were considered pioneer offering a solution for full transcription and subtitle generation for different types of applications and languages, and covering live and batch working modes.

The main publication describing the development of the SAVAS systems is given below:

- [Álv+15] Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., anddel Pozo, A. (2015). *Automating live and batch subtitling of multimedia contents for several European languages.* Multimedia Tools and Applications, 1-31.

3.4.2 SAVAS corpus

Besides the development of technology for automatic subtitling, one of the main goals of the SAVAS European project was to collect and annotate a huge amount of audio and text corpora in several European languages to develop LVCSR systems for automatic transcription.

The development of robust LVCSR systems for automatic transcription and subtitling in the audiovisual domain requires considerably large audio and text corpora for acoustic and language modeling. Based on previous experience [Net+08], it was estimated that the development of good performance transcription systems would ideally require at least 200 hours of audio and 1000 million words of text. Besides,

Tab. 3.3: Collected audio and text corpora per language and domain within the SAVAS project

Language	Variant	Domain	Audio	Text
Portuguese	European	Broadcast News	113H	1012M
		Interview/debate (adaptation)	20H	200K
Spanish	European	Broadcast News	200H	1009M
Basque	Standard Basque	Broadcast News	200H	329M
		Sports (adaptation)	20H	500K
Italian	Italian	Broadcast News	162H	950M
		Sports (adaptation)	–	500K
	Swiss Italian	Broadcast News	50H	100M
French	European	Broadcast News	150H	932M
	Swiss French	Broadcast News	50H	100M
German	European	Broadcast News	151H	808M
	Swiss German	Broadcast News	51H	100M

the adaptation of an already existing transcription systems to a new domain was estimated to be achievable with 20 hours of audio and at least 500k words.

As it can be seen from the Table 3.3, most of the targeted amounts were almost reached, except for Basque text corpora in the broadcast news domain and Portuguese in the interview/debate domain for adaptation purposes. As a minority language, the availability of Basque text corpora in the news domain was limited.

For each audio content, transcriptions, speaker turns and identification, background conditions, isolated noises and linguistic information (name entities, foreign words, syntactic errors and mispronunciations, among others) were annotated. Transcriber 1.5.1 [Bar+01] was used as annotation tool, and the methodology employed aimed at making the annotation process as productive as possible with the help of automatic tools to produce annotation drafts that annotators post-edited to correct mistakes [Poz+14]. The corpora collected has been shared with the community through the SAVAS META-SHARE[8] repository, which has became one of the biggest available audio and data sources exploitable for LVCSR development.

Collecting the Spanish and Basque corpora has allowed the research and development of the speech processing technology integrated in Transkit SDK and in most of the developed proprietary subtitling and transcription solutions (see Sections 3.4.6 and 5).

The main publication related to the collection of the SAVAS corpus is shown bellow:

[8] `http://www.meta-net.eu/meta-share`

- [Poz+14] del Pozo, A., Aliprandi, C., Álvarez, ..., and Raffaelli, M. (2014). *SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling.* In LREC Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 432-436).

3.4.3 Metric for Subtitling Quality

As explained in Section 3.2, the NER model can be considered the unique metric in the community to measure the quality of live subtitles generated through respeaking. This metric takes into account the errors made by the respeaker and the recognition engine. Nevertheless, other parameters such as punctuation, speaker turns, timing and splitting also need to be considered when measuring the quality of automatic subtitles.

With the aim of taking these features into account, the NER model was extended as follows within the SAVAS project:

$$eNER = \frac{(N \times P) - \sum_{i=1}^{N}(R + SD + T + S)}{(N \times P)} \times 100 \tag{3.2}$$

where N is the number of test subtitles, P is the number of parameters to be evaluated (configurable), R corresponds to the recognition errors, SD represents the speaker change errors, T is the timing persistence errors that can score 0 (no error) or 1 (error) values, and S represents the splitting errors, which can be 0 (no error), 0.5 (inter-[9] or cross-[10] subtitle error) or 1 (inter- and cross-subtitle errors). All the parameters have a maximum value of 1.

This metric was employed to evaluate the overall quality of the developed SAVAS systems [Álv+15]. During such evaluation, recognition (R) and speaker change (SD) errors were calculated using Word Error Rate and F1-measure metrics respectively. Timing (T) and splitting (S) parameters were evaluated manually by subtitling professionals.

The extended eNER metric was integrated in the commercial NERStar tool[11] in order to assess the quality of automatic subtitles, and it is described in detail in [Álv+15].

[9] It denotes a splitting error between the first and the second line of the same subtitle
[10] It corresponds to a splitting error between the last line of the current subtitle and the first line of the next subtitle
[11] http://www.nerstar.com

3.4.4 Long-audio alignment

The need of an efficient method for the alignment of long audio tracks in the automatic subtitling context is established by the fact that there are many cases in which the transcript exists or otherwise it is created manually with the aim of avoiding recognition errors. This transcripts have to be further synchronized with the original audio to recover time-stamps at word level. This methodology is still currently employed by many companies which provide methods and technology to automatically align audios and transcripts to generate well-synchronized subtitles, such as SyncWords[12], eCaption[13], 3Play Media[14], WebCaptioning[15] and Ubertitles[16], which was created as a spin-off of Vicomtech-IK4 in the automatic subtitling field.

The technological starting point was the work in [Bor+12], in which alignment results were reported for 3-hour long audios. They turned the audio-to-text alignment task into a text-to-text alignment problem, by applying a phonetic decoder to the long speech signal and aligning the recognized sequence of phones to the phonetic transcript derived from the reference text. Their alignment method was based on Hirschberg's algorithm [Hir75], originally used for genetic sequence alignment, and they managed to develop a system which worked fast and was light in terms of the required computational cost and resources. They used a binary scoring matrix for alignment operations, where all the insertions, deletions and substitutions bore a cost of 1, and matches bore a cost of 0.

Taking this system as a reference, an attempt to improve it was carried out employing the information provided by several phoneme-relatedness scoring matrices, instead of using binary scoring matrices. Systems and matrices were developed for Spanish, English and Basque. It was an incremental work that was published over three publications [RÁA14; ÁAR14; ÁRA14].

In the final system described in [ÁRA14], context-dependent phone decoders, based on cross-word triphone acoustic models trained with the HTK [You+97] tool were included. In addition, with the aim of reducing linguistic variability, the language model consisted of an interpolation of a generic language model and a specific model created for each transcript to be aligned.

Regarding the phoneme-relatedness matrices, they provided information to the aligner about how likely it was for an alignment between two phonemes to be

[12] https://www.syncwords.com/
[13] https://www.ecaption.eu/
[14] http://www.3playmedia.com/
[15] http://webcaptioning.com/
[16] http://www.ubertitles.com/

correct. The matrices favored aligning similar phonemes, by giving such alignments higher scores than to alignments between less similar phonemes. The matrices gave the lowest scores to alignments between highly dissimilar phones, which were unlikely to be correct. We created different scoring matrices for each language, applying different phoneme-relatedness criteria. The first scoring matrix was decoder-dependent, based on errors made by the phone decoder. The second matrix was decoder-independent, and based on phonological similarity, assessed by comparing largely articulatory features. The final matrix was also decoder-independent, and relied on phoneme confusion in human perception.

Matrices based on Phone-decoding Errors

These matrices were based on the phone decoders' confusion matrices. They were created based on HTK's HResults logs, when aligning the phone-decoding output and the phone transcription for sequences of a huge quantity of phonemes per language. For each phone in the phone-set, the matrices contained the percentages of misrecognitions and correct recognitions by the decoder.

Matrices based on Phonological Similarity

The phonological similarity scores were based on the metric devised by Kondrak in [Kon02]. Phonemes were described with Ladefoged's [LJ14] multivalued features, and a salience factor weighted each feature according to its impact for phoneme similarity.

The scoring function used to define the similarity between two phonemes is presented in Figure 3.2.

$$\sigma_{\text{sub}}(p,q) = (C_{\text{sub}} - \delta(p,q) - V(p) - V(q))/100$$
$$\text{where } V(p) = \begin{cases} 0 & \text{if } p \text{ is a consonant or } p = q \\ C_{\text{vwl}} & \text{otherwise} \end{cases}$$
$$\delta(p,q) = \sum_{f \in R} diff(p,q,f) \times salience(f)$$
$$\sigma_{skip}(p) = ceiling(|C_{\text{sub}}/400|)$$

Fig. 3.2: Similarity function

where C_{vwl} represents the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically and the function $diff(p,q,f)$ yields the similarity score between phonemes p and q for feature f, and the feature-set R is configurable. Last, $\sigma_{skip}(p)$ returns the penalty for insertions and deletions used in the aligner. We

defined heuristically a C_{sub} value of 3,500 (i.e. a maximum similarity score of 35), and a gap penalty of 9 for alignment, which corresponds to $ceiling(|C_{sub}/400|)$.

Matrices Based on Perceptual Errors

The English matrices were based on human perceptual error data from [Cut+04]. They performed a phoneme identification study with native speakers of American English, asking them to identify the initial or final phoneme of 645 syllables of types CV (ConsonantVowel) and VC, at signal-to-noise ratios (SNR) of 0, 8 and 16. The Spanish matrix was based on an extended version, provided by the authors directly, of the corpus of human misperceptions in noise developed in [Lec+13]. The methodology involved presenting 69 native speakers of Spanish with over 20,000 single-word stimuli, under different masking-noise conditions, and asking the speakers to write the word they had heard. Perceptual-relatedness matrices were not created for Basque, since we were not aware of appropriate data that could be exploited for their creation.

The phonesets used for the three languages, in addition to samples for all types of matrices and a discussion about the features, values and saliences employed are given in more detail in the projectś website[17].

The results show the effectiveness of our long audio alignment system and the non-binary similarity matrices, even with contents containing noisy-speech and imperfect transcriptions [ÁRA14]. The improvements using non-binary matrices are clearly proved comparing the accuracies obtained with the binary matrix at word and subtitle-level. Within an acceptable maximum deviation of 1 second, near-perfectly aligned subtitles were obtained for the three languages. In fact, alignment accuracies of 96.72%, 91.30% and 95.18% at subtitle level were reached for Spanish, English and Basque respectively at this maximum deviation time [ÁRA14].

All the work described above was even further extended through matrices based on probabilistic kernels, as it is explained in the following subsection.

Matrices based on probabilistic kernels

In the study [Bor+16] we also took advantage of the information provided by a confusion matrix from a phone decoder, but in a probabilistic way. In this sense, several probabilistic kernels were defined and evaluated, understanding a kernel as a similarity function used by a sequence alignment algorithm to evaluate different

[17] https://sites.google.com/site/similaritymatrices/

possible subsequence alignments. Each kernel was composed by three values, which were the piece of information needed by the aligner to perform one editing operation: K_{rh} for pairing, K_r for deleting r, and K_h for inserting h, where $r \epsilon R$ and $h \epsilon H$ are phonetic symbols from the reference R and hypothesis H respectively, and the deletion and insertion values K_{del} and K_{ins} will not depend on their corresponding context symbols, h and r, respectively. Thus, each of the kernels evaluated in this work was defined as:

$$K = \{K_{pair}, K_{del}, K_{ins}\} \tag{3.3}$$

The computation of each kernel value was based on the probabilities of the three editing operations; being P_{pair} the probability of pairing the symbols r and h, P_{del} the probability of deleting the symbol r and P_{ins} the probability of inserting h. These three values can be estimated as the frequency of occurrence of each event in the master alignment:

$$P_{pair} = P(pair|r,h) = \frac{c_{rh}}{c_{rh} + \frac{c_{r\epsilon}}{N} + \frac{c_{\epsilon h}}{N}} \tag{3.4}$$

$$P_{del} = P(del|r) = \frac{c_{r\epsilon}}{c_{r\epsilon} + \sum_{\forall h}(c_{rh} + \frac{c_{\epsilon h}}{N})} \tag{3.5}$$

$$P_{ins} = P(ins|h) = \frac{c_{\epsilon h}}{c_{\epsilon h} + \sum_{\forall r}(c_{rh} + \frac{c_{r\epsilon}}{N})} \tag{3.6}$$

where N is the total amount of phonetic symbols, c_{rh} corresponds to the number of times the pair of symbols r and h have been paired together, $c_{r\epsilon}$ denotes how many times a symbol r has been deleted and $c_{\epsilon h}$ is the number of times a symbol h has been inserted.

The sequence alignment algorithm aims to find the path that minimizes or maximizes the accumulated kernel value in terms of *costs* and *benefits* respectively. Both representations could be valid and we can switch from one to the other by inverting the sign of the values. In this sense, the two baseline kernels considered in this work were as follows:

$$K_{MaxMatch} = \{\delta_{rh}, 0, 0\} \tag{3.7}$$

$$K_{MinDist\_Cost} = \{1 - \delta_{rh}, 1, 1\} \tag{3.8}$$

where $K_{MaxMatch}$ was a benefit kernel, $K_{MinDist\_Cost}$ was a cost kernel, and δ_{rh} took the value 1 if $r = q$, and 0 otherwise. In addition, we inverted the values of the $K_{MinDist\_Cost}$ kernel to adopt its benefit maximization representation as other kernel for this work:

$$K_{MinDist} = \{\delta_{rh} - 1, -1, -1\} \tag{3.9}$$

Besides, we generalized the $K_{MaxMatch}$ kernel by using the previously defined probability P_{pair} instead of δ_{rh}, creating the $K_{ExpectedMatch}$ kernel:

$$K_{ExpectedMatch} = \{P_{pair}, 0, 0\} \tag{3.10}$$

Similarly to the $K_{MaxMatch}$ kernel generalization, the $K_{MinDist}$ kernel can be directly converted to a probabilistic kernel as follows:

$$K_{ExpectedDist} = \{P_{pair} - 1, P_{del} - 1, P_{ins} - 1\} \tag{3.11}$$

Finally, considering the symmetry of costs and benefits, we mapped the $[0, 1]$ domain to $(\infty, -\infty)$ through the logit function $logit(p) = log(\frac{p}{1-p})$, proposing the following kernel:

$$K_{logit} = \{P_{pair} - 1, P_{del} - 1, P_{ins} - 1\} \tag{3.12}$$

With the aim of comparing the results with the ones obtained in [ÁRA14], the kernel based on the Kondrak's metric was included in the experiments, which were performed over the Hub4-97 dataset [GFG02], and using a confusion matrix estimated on the Wall Street Journal [PB92] training corpus. The results in Figure 3.3 demonstrated that the probabilistic kernels $K_{ExpectedMatch}$ and $K_{ExpectedDist}$ clearly outperformed their non-probabilistic versions $K_{MaxMatch}$ and $K_{MinDist}$ respectively. Besides, the K_{logit} reached the best results because of the wide range of values $(\infty, -\infty)$ considered. Finally, given the performance of the kernel based on the $Kondrak$ metric, we concluded that the information provided by the confusion matrix is good enough to avoid the use of information related to phonological similarity.

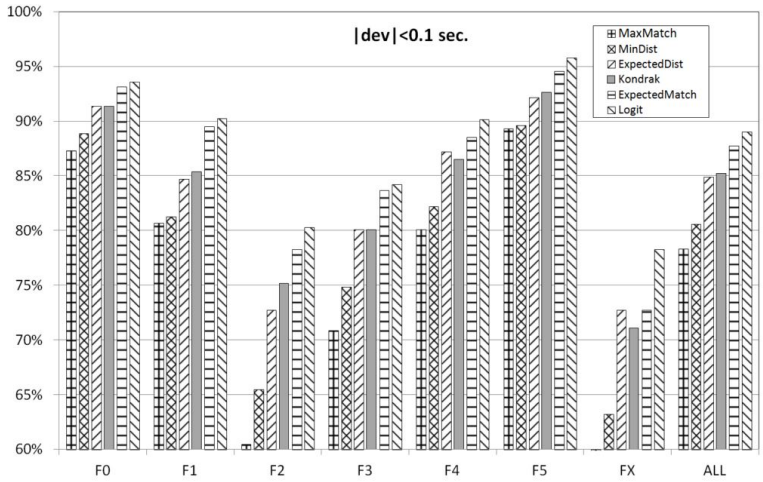


Fig. 3.3: Alignment accuracy (tolerance interval: 0.1 seconds) for all the kernels regarding the different acoustic conditions within Hub4-97 (F0:Baseline Broadcast Speech, F1: Spontaneous Broadcast Speech, F2: Telephone Speech, F3: Noisy Speech, F4: Speech under degraded acoustic conditions, F5: Speech of non-native speakers, Fx: All other speech).

The main publications related to the work performed on the long audio alignment field are listed bellow:

- [Bor+16] Bordel, G., Penagarikano, M., Rodriguez-Fuentes, L., Alvarez, A., and Varona, A. (2016). *Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks.* In IEEE Signal Processing Letters, vol. 23, no. 1, pp. 126-129.

- [ÁRA14] Álvarez, A., Ruiz, P., and Arzelus, H. (2014). *Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque.* In Text, Speech and Dialogue (pp. 473-480). Springer International Publishing.

- [ÁAR14] Álvarez, A., Arzelus, H., and Ruiz, P. (2014). *Long audio alignment for automatic subtitling using different phone-relatedness measures.* In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014 (pp. 6280-6284). IEEE.

- [RÁA14] Ruiz, P., Álvarez, A., and Arzelus, H. (2014). *Phoneme similarity matrices to improve long audio alignment for automatic subtitling.* In LREC, Ninth International Conference on Language Resources and Evaluation.

3.4.5 Automatic Segmentation of Subtitles

The contributions presented in [ÁAE14] and [Álv+16c] can be considered the two first research approaches in the field of automatic segmentation of subtitles.

In [ÁAE14], Support Vector Machine (SVM) and Logistic Regression (LR) classifiers were built over a Basque corpus consisting of TV cartoon programs and subtitles generated by professionals. To this end, subtitles with correct or incorrect segmentation were divided into two classes. Positive (correct) feature vectors were extracted from professionally-created subtitle data and contained the segmentation marks found in the corpus, whilst negative (incorrect) vectors were generated by manually inserting improper segmentation marks. Classifiers where then trained on balanced sets formed with these two types of vectors and employed for the segmentation task. The feature vectors were composed by 4 types of characteristics related to timing, number of characters, speaker change and a perplexity value given by a language model built over the training data. During decoding, an iterative algorithm was in charge of generating all the possible candidates for a break at each iteration. These candidates included sequences of consecutive words that did not exceed the maximum allowed length in characters before and after segmentation points. Feature vectors were then computed from these candidates and measured against the machine-learned classifiers and optimal candidates selected according to the obtained score. Similar performance was obtained for the two classifiers under evaluation, with an average F1-measure score of 74.71% and 76.12% for the SVM and LR classifiers, respectively.

However, in this work the possible segmentation points were only estimated, without distinguishing between different types of breaks. This implies considering line-breaks and subtitle-breaks, which is critical information to automatically generate final subtitles correctly. Finally, the computation time needed to generate all the candidates and select the optimal ones was inefficient for a real application.

With the aim of improving the results and performance of the system described in [ÁAE14], a new approach has been presented in [Álv+16c]. In this case, we have decided to turn the automatic segmentation of subtitles into a text sequence labeling problem, in which each word corresponded to a specific category defining its function and position within the subtitles. Following a statistical approximation, the objective was to obtain the optimal categorical assignment from an input sequence of words.

If we consider the sequence of words as $W = w_1^n = w_1, w_2 \cdots w_n$, and the sequence of labels as $L = l_1^n = l_1, l_2 \cdots l_n$, the problem can be statistically stated as follows:

$$\hat{L} = \underset{l_1^n}{\operatorname{argmax}} \Pr(l_1^n | w_1^n) \tag{3.13}$$

The problem can be solved by defining a model to estimate $\Pr(l_1^n | w_1^n)$ from training data (*training*) and applying a search algorithm on that model for a given sequence of words (*decoding*). In this sense, Conditional Random Fields (CRF) offered an appropriate framework to model conditional probabilities between input-output sequences, as well as search algorithms that allow obtaining the decoding results.

Following a similar representation of the linear CRF chain given in [SM11], considering that the input was a sequence of feature vectors derived from the sequence of words to be segmented, and the output corresponded to a sequence of labels that represented the position of each word in the subtitle, the final CRF model for subtitle segmentation can be stated as follows:

$$\hat{L} = \underset{l_1^n}{\operatorname{argmax}} \Pr(l_1^n | w_1^n) = \underset{l_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} \exp\left(\sum_{k=1}^{K} \theta_k f_k(l_i, l_{i-1}, \vec{w}_i) \right) \tag{3.14}$$

where $\vec{x}$ and $\vec{y}$ represent input and output sequences, respectively; f_k (with $k = 1, \ldots, K$) is the set of features functions, which established the correspondence between input and output elements, and they actually form the probability distribution; and θ_k (with $k = 1, \ldots, K$) is the set of weights associated to each feature function f_k. Feature functions f_k and weights θ_k were obtained during the training process.

In order to determine a dependence structure for automatic segmentation through a graphical CRF model, eight class labels were created to define the function of each word within the subtitle, as listed below:

- B-SU *(Begin-Subtitle):* For each first word in subtitles.
- I-LI *(In-Line):* For each word in a subtitle which is not the first or last word of a line or subtitle.
- E-LI *(End-Line):* For each word which represents the last word of a line which does not correspond to the end of a subtitle (e.g. last word of the first line in a subtitle with two lines).
- B-LI *(Begin-Line):* For each first word in a line that is not the first word in a subtitle (i.e., first word in second line for subtitles with more than one line).

- E-SU *(End-Subtitle):* Each final word of a subtitle composed by one or more lines.

- BE-SU *(BeginSubtitle-EndSubtitle):* For words in an one-word subtitle.

- BS-EL *(BeginSubtitle-EndLine):* For words in the first one-word line of subtitles with more than one line.

- BL-ES *(BeginLine-EndSubtitle):* For words in the second one-word line of subtitles with more than one line.

00:00,166 - 00:05,333
Come here,
Mum come here

Example 1: Subtitle example composed by 6 words and 2 lines

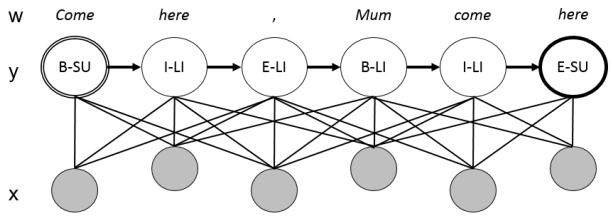


Fig. 3.4: Graphical model of the executed CRF over Example 1. Transition factors depend on the surrounding two observations.

In Figure 3.4, a practical execution of the defined Markov dependence structure is presented, given the example subtitle shown in Example 1, which is composed by 6 tokens and formatted in 2 lines. Given this input example, the target of the CRF model would be to predict an output vector $y = \{y_0,y_1,..., y_N\}$ through the observed feature vectors $\{x_0,x_1,..., x_N\}$ extracted from the sequence of words $\{w_0,w_1,..., w_N\}$. In the graphical CRF models constructed for this work, each variable y_j corresponds to one of the class labels described above for each word at position j. For its part, each x_j contains the feature vector values about the word at position j. The transition factors of the CRF models depend on the surrounding two observations on the left and the right.

The features used to describe each of the words at position j were composed of a total of 15 characteristics, including the current and the neighboring two words *(5 features)*, the current and neighboring two words' Part-Of-Speech *(5 features)*, two Boolean values to control if the amount of characters per line and subtitle has exceeded *(2 features)*, speaker change *(1 feature)*, and time difference between the current and the neighboring two words *(2 features)*.

The CRF based approach has been tested for Basque and Spanish, and compared to the SVM-based method employed in the previous work [ÁAE14]. In addition, the Counting Character (CC) method was also applied over the test sets and results were computed. The latter method can be considered as the most simple technique to perform segmentation, as it only considers the maximum number of characters allowed per line. We used two corpora composed by TV series in Basque and Spanish. Subtitles were manually created by professionals following specific rules based on keeping linguistic and syntactic coherence. The Basque corpus contained a total amount of 109,006 subtitles (80% for train and 20% for test), whilst the Spanish corpus was composed of 81,802 subtitles; 80,058 were used to train models and the remaining to test them.

Four main evaluation metrics have been used for testing purposes, as they are briefly described below:

- F1-LINE: it measures segmentation errors (false negatives and false positives) and correct segmentations (true negatives and true positives), and computes the accuracy through the F1-Score metric. It does not distinguish between line and subtitle breaks.

- NIST-SU: it computes the number of segmentation errors (missed segments and false alarm segments) divided by the number of segments in the reference. For this work, it was computed at line level (NIST-SU-LI), which included both line-breaks and subtitle-breaks, and at subtitle level (NIST-SU-SUB).

- DSER: This metric is computed dividing the number of incorrectly segmented portions in the reference by the total of segments in the reference. In this work, it was computed at line level (DSER-LI), composed by line-breaks and subtitle-breaks, and at subtitle level (DSER-SUB).

- SegER: It is computed as the edit distance between sequences of reference positions and hypothesis positions (those obtained automatically by the classifiers), using Insertion, Deletion, and Substitution as edition operations. It was also computed at line level (SegER-LI), which included both line-breaks and subtitle-breaks, and at subtitle level (SegER-SUB).

In the following Table 3.4 and Table 3.5, the results obtained for the three segmentation techniques under evaluation are presented for each language.

Comparing the results from Tables 3.4 and 3.5, the general remarkable issue is that using CRF for assigning the different subtitle labels to words is a valid alternative,

Tab. 3.4: Segmentation scores of the CRF-, SVM- and CC-based methods for the Basque corpus.

	Segmentation score						
	F1-LINE	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
CRF	83.0	26.5	28.3	47.1	47.4	22.6	21.6
SVM	74.7	81.6	56.1	110.5	79.1	59.0	33.7
CC	36.2	120.4	174.9	136.4	132.6	83.2	70.6

Tab. 3.5: Segmentation scores of CRF-, SVM- and CC-based methods for the Spanish test data, without including speaker change information

	Segmentation score - No Speaker Change						
	F1-LINE	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
CRF	80.7	39.4	38.2	58.0	61.6	38.8	28.4
SVM	41.4	146.6	119.2	159.4	140.0	82.7	66.6
CC	15.2	142.9	136.2	148.9	148.1	91.1	87.8

since in both languages average F1-Line accuracy is higher than 80%. Comparing with the previous SVM approximation employed in [ÁAE14], it supposes a large improvement on subtitling quality, specially for Spanish (where SVM results present an accuracy lower than 50%), although Basque presents a significant improvement as well. Regarding the CC-based method, as it was expected, it achieved the worst results, mainly due to the lack of a model built with training data and adapted to the characteristics of the domain.

Overall, the use of CRF has allowed improving the results obtained in the previous work in the following points: (1) differing between the type of breaks (line- and subtitle-breaks), (2) obtaining much higher scores and thus generating more and better segmented subtitles and (3) much faster processing time (less than 0.1 milliseconds per subtitle against 16 seconds per subtitle on the same machine) [Álv+16c].

The main publications related to the work performed on the automatic segmentation of subtitles are listed bellow:

- [Álv+16c] Álvarez, A., Martínez-Hinarejos Carlos-D., Arzelus H., Balenciaga M., and del Pozo A. (2016) *Improving the Automatic Segmentation of Subtitles through Conditional Random Field*. Speech Communication, Elsevier. ***Status: In 2nd revision.***

- [Álv+16b] Álvarez, A., Balenciaga, M., del Pozo, A., Arzelus, H., Matamala, A., Martínez-Hinarejos, Carlos-D. (2016). *Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles.* In Proceedings of the 10th international conference on Language Resources and Evaluation (LREC2016), pp. 3049-3053.

- [ÁAE14] Álvarez, A., Arzelus, H., and Etchegoyhen, T. (2014). *Towards customized automatic segmentation of subtitles.* In Advances in Speech and Language Technologies for Iberian Languages (pp. 229-238). Springer International Publishing.

3.4.6 Proprietary Rich Transcription and Automatic Subtitling systems for Basque and Spanish

As a result of all the developments performed in the R&D projects and research activities described in Section 2, proprietary systems for rich transcription and automatic subtitling have been implemented for Basque and Spanish. These systems comprise several technological components that can be built, improved or adapted separately depending on the final application and domain. These adaptation capabilities are achievable through the methods provided by the proprietary Transkit SDK[18]. This toolkit aims at integrating all the tools and functions required to build and evaluate the whole set of technological modules involved in these kind of subtitling and transcription solutions; such as automatic speech recognition, speech-non speech detection, language tracking, punctuation, capitalization, speaker diarization, subtitle segmentation and normalization modules.

Our baseline rich transcription and automatic subtitling systems for Basque and Spanish were trained with the SAVAS corpus (see Section 3.4.2), and they were then adapted to the Basque Parliament domain within the TE-PARLA project (see Section 2.2.2).

- Regarding automatic speech recognition, LVCSR engines for Basque and Spanish were built using the open-source Kaldi toolkit[Pov+11]. Acoustic models were trained following the implementation given in [Ves+13], which corresponds to a hybrid Deep Neural Network (DNN)-Hidden Markov Models (HMM) implementation, were DNNs are trained to provide posterior probability estimates for the HMM states. Two types of language models (LM) were integrated per language: trigram Arpa-format LM for decoding and 9-gram constant Arpa-format LM for rescoring of the final lattices. The constant Arpa LM, which has been recently implemented in Kaldi, makes LM rescoring faster and requires less memory. The decoding LM were estimated with Kneser-Ney modified smoothing using the KenLM [Hea11] toolkit.

  Baseline LVCSR engines were trained using the SAVAS corpus. Regarding the Spanish engine, it was trained using 172 acoustic hours and 1,009 million

[18] http://www.vicomtech.org/resources/archivosbd/sdks_documentos/Transkit.pdf

words from the news domain. An average WER of 15.13% was achieved on a test set of 2 hours. The Basque engine was trained using 151 acoustic hours and 329 million words from the news domain, and achieved an average WER of 18.40% on a test set of 2 hours. These results were reached for both languages over generic test sets which included a number of unseen topics, speech types and background acoustic conditions. The positive impact of domain adaptation at language and vocabulary level has been shown within the TE-PARLA project (see Section 2.2.2), in which WER values of 9.47% and 16.73% were achieved for Spanish and Basque respectively in the Basque Parliament domain.

In Figure 3.5, a schema of the tools and functions of the proprietary Transkit toolkit to train and evaluate LVCSR engines is presented.

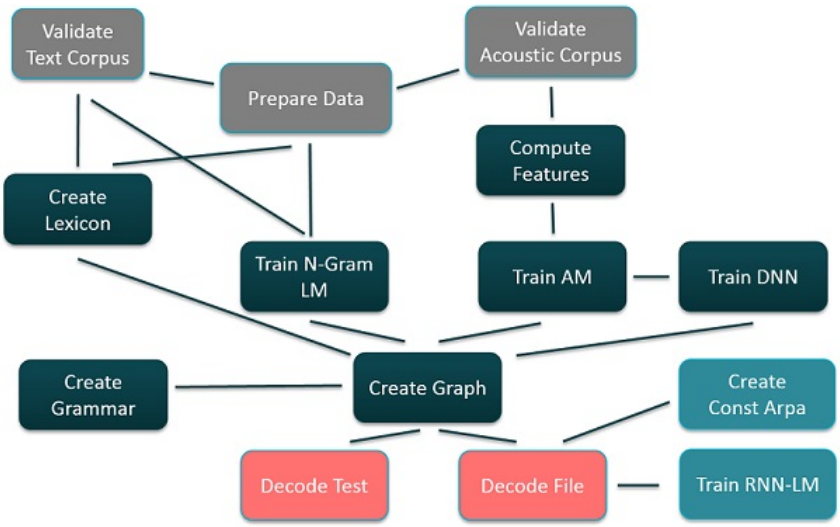


Fig. 3.5: Tools and functions to train, build and evaluate LVCSR engines based on Kaldi

Each module in the Figure 3.5 corresponds to a method that performs one specific function. The initial methods 'Validate Text Corpus' and 'Validate Acoustic Corpus' check that the data provided to train language and/or acoustic models are correct and have been provided in the required format. The function 'Prepare Data' generates the necessary auxiliary files from the input texts and audios. Once the text data is prepared, it can be used to create the lexicon and estimate LMs through the 'Create Lexicon' and 'Train N-Gram LM' methods respectively. The first method outputs the words with their corresponding phonetic transcription obtained from a Grapheme-to-Phoneme (G2P) module, whilst the second method is configurable to train N-gram LMs with any N order and the desired pruning at each order. Pruning techniques were used to produce the smallest model while keeping the performance loss as small as possible. The 'Train N-Gram LM' method works on top of the KenLM [Hea11] toolkit. Besides, constant Arpa-format and Recurrent Neural Network (RNN) LMs can also be estimated through the 'Create Const Arpa' and 'Train RNN-LM'

methods. The constant Arpa LMs are estimated using functions of Kaldi, and the RNN LMs are built with the RNNLM toolkit [Mik+11]. Both LMs are employed for rescoring purposes. Acoustic models are trained through the methods 'Train AM' and 'Train DNN'. The 'Train AM' method can be used to train GMM-HMM models, whilst the 'Train DNN' method estimates DNN-HMM acoustic models based on the implementation provided in [Ves+13]. Finally, the central 'Create Graph' method constructs the decode HCLG graph, where H contains the HMM definitions, C represents the context-dependency, L is the lexicon and G is an acceptor than encodes the grammar or language model. In addition, instead of a language model, the Transkit toolkit allows the creation of finite grammars through the 'Create Grammar' method in order to build grammar-based ASR engines. The 'Decode File' and 'Decode Test' functions enable user to process files separately or in batch mode respectively.

- Speech-Non Speech detection. This component aims at segmenting the audio input into speech and non-speech segments for posterior recognition and/or diarization processes. The Transkit toolkit includes the necessary functions based on the UBM/GMM (Universal Background Model/Gaussian Mixture Model) approximation described in [SCP15] to train and evaluate a Speech-Non Speech detection module.

  Our proprietary Speech-Non Speech detection system is capable of discriminating between three types of categories, such as speech, silence and noise. It was trained using 2 hours per audio category from the Basque Parliament sessions, and reached an accuracy of 94.1% at segment level in an in-domain test set with a total duration of 105 minutes. Its accuracy was computed using a reference ground truth which was manually split and annotated over the same evaluation segments.

- Language Tracker. This component aims at segmenting a multilingual audio track by language. In the Transkit toolkit, methods for the development of a phonotactic based language tracker have been implemented and integrated. This approximation aims at identifying phoneme boundaries which could be candidates of language turns, following the work presented in [LCL13]. The technical approach consisted in constructing an unique phone decoder combining the languages involved in the task. In our case, a language tracker to split audio tracks mixing segments in Basque and Spanish was developed. To this end, an hybrid DNN-HMM acoustic model was trained using audios in both languages, totaling an acoustic corpus of 86 hours (39 hours of Basque and 47 hours of Spanish), whilst the LM was a trigram model at phone level estimated using bilingual texts composed by 4.3 million phones (1.9 M for Basque and 2.4 M for Spanish). The phone set was composed by all the phones

in both languages, and a distinctive language tag was added to each phone to avoid mixing those that were shared along both languages.

We employed the Language Error Rate (LER) metric to evaluate the performance of the developed Language Tracker system. This metric was computed in the same way as the well-know Diarization Error Rate (DER) metric, commonly used in speaker diarization systems, but using languages instead of speakers. Thus, the LER metric aims at measuring the ratio of incorrectly detected language time to total language time. Our system was evaluated on a test set of the Basque Parliament, in which many bilingual contents are usually generated. In this case, an average LER of 10.29% was obtained on a test set of audios lasting 255 minutes, where 92 minutes were spoken in Basque and 163 minutes in Spanish.

- Punctuation. The objective of this component is to enrich the output of the LVCSR engine with punctuation marks. Within the Transkit toolkit, automatic punctuation is treated as a text sequence labelling task. As a result, the developed automatic punctuation technology has been constructed on top of a CRF (Conditional Random Field) model, in which each token is tagged with one category (NP: No punctuation; CO: Comma; FS: Full Stop) depending on whether the next token corresponds to a punctuation mark or not. In this sense, the number of labels which compose the dependency structure of the CRF graphical model depends on the amount of categories to be labelled. In our punctuation system, three categories have been defined, which correspond to Full Stop, Comma and No Punctuation.

  Regarding the features employed, each word is characterized by the following vector of values representing acoustic and linguistic features: (1) the current and the surrounding 2 words on the left and right (5 words), (2) the current and the surrounding 2 words' POS information on the left and right (5 categories), (3) time between the current and the next word (1 feature), (4) Speaker Change (1 Boolean), and (5) Language Change (1 feature).

  Two CRF punctuation models for Spanish and Basque were developed and integrated in the proprietary baseline rich transcription and subtitling systems. Models were trained on two corpora of the Basque Parliament domain containing 11 hours and 55 minutes in Spanish, and 6 hours and 30 minutes in Basque. Their evaluation was carried out using the well-known Precision, Recall and F1-measure metrics and over a test set composed of 2 hours in each language, scoring macro-average F1-measure values of 64.34% and 65.89% for Spanish and Basque respectively. CRF models were constructed using the CRFSuite tool [Oka07].

- Capitalization. The Capitalization module aims at re-casing the lower-cased text output of the LVCSR engine. Within the Transkit toolkit, automatic capitalization is done through a tool provided by the Moses open-source toolkit [Koe+07] for statistical machine translation. This tool essentially employs a word-to-word translation model and a cased language model to capitalize an input text, and requires a cased and tokenized input text and a language model to estimate the recasing model.

  Transkit's current capitalization models for Spanish and Basque have been trained on the digital newspapers collected within the SAVAS corpus, which were extended with new texts from the Politics domain composed of 12 million words in Spanish and 7 million words in Basque. These models reached a F1-score of 89.94% and 82.80% on a test set of 28K and 10K words for Spanish and Basque respectively.

- Speaker Diarization. This technological component is responsible for segmenting and clustering speakers within an audio track. Initial segmentation is done at acoustic level through the Speech-Non Speech detection module, which classifies each 10 millisecond segment as speech, silence or noise. The consecutive speech segments are then joined, and the generalized likelihood ratio (GLR) [GSR91] distance measure is applied to detect close speaker changes within the same speech segment. Each individual speech segment is modelled through an i-vector representation. To this end, an Universal Background Model (UBM) and a Total Variability (TV) matrix have been estimated using a corpus from the Basque Parliament sessions composed of 105 speakers with a total duration of 16 hours and 50 minutes. The UBM was estimated using 512 gaussians, whilst in the TV space the Linear Discriminant Analysis (LDA) compensation method was applied to reduce variabilities. LDA attempts to transform the axes to minimize the intra-class variance due to channel effects and maximize the variance between speakers. With the help of an iterative algorithm, the extracted i-vectors are compared using the cosine similarity scoring and those i-vectors with a similarity score higher than an empirically fixed threshold are clustered together as the same speaker. The developed technology has been evaluated over an in-domain test set composed of 5 contents and 110 minutes, with an average Diarization Error Rate (DER) of 4.58%.

- Subtitle Segmentation. This component is integrated only in the automatic subtitling system, and is in charge of segmenting subtitles according to syntactical coherence. The module has been built for Spanish and Basque using a Conditional Random Field (CRF) model, for which several categories to describe the function of each word within each subtitle were defined and connected through a graphical dependence model. The defined categories

represent the function of the first word of a subtitle (B-SU), the end word of a line (E-LI), the first word of the second line (B-LI), the final word of a subtitle (E-SU), inline words (I-LI), a single word in a subtitle (BE-SU), single word in the first line (BS-EL), and single word in the second line (BL-ES). The feature vectors used to describe the information extracted from each word were composed of 15 characteristics related to words, Part-of-Speech, Speaker Change, time differences between surrounding words and two parameters to control the amount of characters per line and subtitle. The Spanish corpus was composed of 98 episodes of a TV series with a total amount of 81,802 subtitles, while the Basque corpus contained subtitled TV cartoon programs totaling 109,006 subtitles. In both corpora, subtitles were manually created by professionals following specific rules to keep linguistic and syntactic coherence. The corpora were split into train and test sets, and accuracies of 80.7% and 83% were obtained for Spanish and Basque respectively.

More information regarding the construction and evaluation of the Subtitle Segmentation component can be found in the previous Section 3.4.5 and in [Álv+16c].

- Normalization. Comprises all the tasks related to (1) converting numbers into their digit representation, (2) removing filled pauses, and (3) generating abbreviations. Rule-based functions are used within Transkit and the proprietary subtitling and rich transcription systems to perform this task.

In order to make the subtitling and rich transcription systems accessible through the Internet and to allow external companies to check the performance of the technology with their own contents, a web platform has been developed as a service. This platform is further explained in the following subsection.

Web Platform for Automatic Transcription and Subtitling

The goal of this web platform is to provide companies, entities and customers an online service to test by themselves if the automatic subtitling and rich transcription technology may help improving their internal services. As it can be seen in Figure 3.6, the automatic rich transcription and subtitling technology is hosted in an internal server at Vicomtech-IK4 (internal network), whilst the web platform is installed in a server allocated on the Internet, in order to be accessed by any external user. A monitoring daemon continuously checks a shared folder where the uploaded contents are saved. In case there is a new content, it is copied to the internal server to be processed automatically. The result is then sent back to the user via email. This

type of architecture has been implemented to avoid exposing the technology in a server located in the Internet.

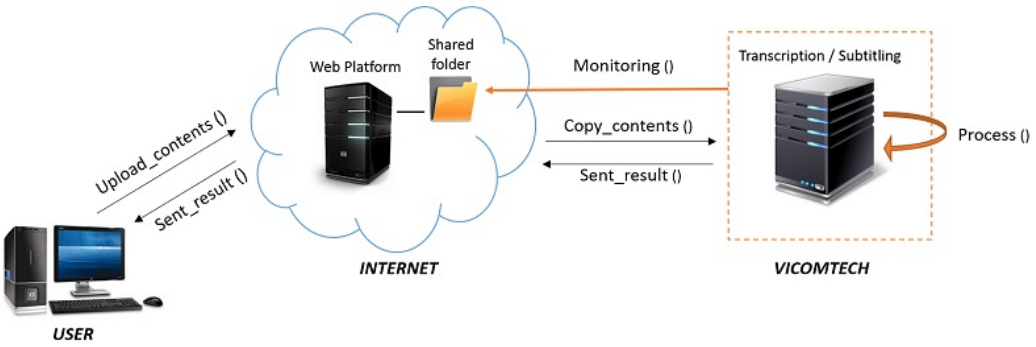


Fig. 3.6: Main architecture schema of the web platform for automatic transcription and subtitling

The user accesses the platform through a web application[19], which first presents a login interface to authenticate the user. The main characteristics of the platform are listed below:

- Multi-language: It integrates technology for English, Spanish and Basque.
- Multi-user: Depending on the user, the platform can be limited to process larger or smaller contents.
- Multi-format: the FFmpeg[20] tool has been integrated for transcoding purposes in order to support as many audio-video formats as possible. Besides, the output can be presented in four different text formats, including plain TXT, XML (eXtensible Markup Language), RTF (Rich Text Format), and SRT (SubRip).
- Configurable: The platform includes many options that are linked to each of the integrated technological components. In this sense, the user can choose whether (1) to punctuate and/or capitalize the output of the LVCSR engine automatically, (2) to apply speaker diarization, and (3) to try to automatically improve the LVCSR output through the rescoring technique. Besides, the user can also decide the format of the output file.
- Parallel Processing: The platform has been designed to allow parallel processing of many input contents.
- Transparency: The user is unaware of the technological processing involved. S/he only needs to choose the preferred options for transcription or subtitling.

[19] http://212.81.220.68:8086/SDK_web/
[20] https://ffmpeg.org/

- Accessible: The platform can be reached from any device with access to the Internet.

From the technological point of view, the platform includes the components for automatic speech recognition, language segmentation, punctuation, capitalization, speaker diarization, subtitle segmentation and normalization presented above.

The platform is always operational for any customer or user which would want to check its performance on specific audio contents and has already been transferred as-is to some entities (see Section 5).

The main publication related to the development of these proprietary rich transcription and automatic subtitling systems is listed bellow:

- [Álv+16a] Álvarez, A., Arzelus, H., Prieto, S., and del Pozo, A. *Rich Transcription and Automatic Subtitling for Basque and Spanish.* In: Advances in Speech and Language Technologies for Iberian Languages. 2016. **Status: Submitted**.

3.5 Conclusions and Future Work

Due to the increasing transcription and accessibility needs driven by the Society and the Governments during the last few years, automatic subtitling and rich transcription solutions are being requested to the scientific community. The goal of such technologies is to help professionals cover a greater amount of transcription and subtitling tasks in several sectors and domains; such as the broadcast, Internet, parliament, Healthcare or eLearning domains, among others.

In this chapter, the current state and main challenges of the principal components involved in automatic subtitling and rich transcription technologies have been presented, in addition to the main contributions of this dissertation work to the fields. In this sense, pioneer live and batch automatic subtitling systems for several European languages adapted to different type of applications have been described, in addition to a new huge corpus to train LVCSR based systems in the audiovisual domain and an extended metric to measure the quality of the automatically generated subtitles. Besides, a new method for the forced-alignment of long audio tracks has been proposed, with low computational load and high performance in several acoustical conditions. Two new methods to perform automatic subtitle segmentation have also been presented, showing that CRF-based approaches outperform SVM-based techniques in terms of precision and computational cost. Finally, complete and proprietary rich transcription and automatic subtitling systems have been described

in detail. These systems have been built using the functions and methods included in the proprietary Transkit toolkit, which serves as the basis to generate automatic subtitling and rich transcription systems adapted to other domains and sectors. A Web Platform aiming to provide a bridge between companies and the developed rich transcription and automatic subtitling technology has also been presented. Using this platform, interested companies and customers can directly subtitle and transcribe contents automatically and check if the technology can help them make their internal processes more productive.

Much work still needs to be done in the fields of automatic subtitling and rich transcription. Future work should aim to resolve current challenges, which mainly concern dealing with acoustically noisy environments, spontaneous and overlapping speech. All of these issues directly affect audio alignment, speech recognition, speech-non speech detection and speaker diarization. Following recent advances in Natural Language Processing with Neural Networks for sentence boundary detection [Col+11], new methods to improve automatic punctuation could be investigated, especially for the prediction of commas. In addition, the eNER metric could be extended to include punctuation and capitalization features. Future work on automatic subtitle segmentation could involve experimentation with Recurrent Neural Networks (RNNs) and exploiting new features such as stop words, syntactic functions or grammatical relations of the different clauses within a sentence. Finally, as an alternative to manual live subtitling, the development of future automatic subtitling implementations should be optimized to work in real-time scenarios.

4 Multimedia Interaction Environment

4.1 Introduction

Computers, electronic devices and the Internet have become an essential part of people's life since we entered the Digital Era. Technology is integrated in our most common daily activities, since we use it to work, communicate with other people, buy, seek out information, do sports and entertain ourselves. In this sense, the increasing integration of computers and technology in society has pushed the relevance and interest of the Human-Computer Interaction field.

Human–Computer Interaction (HCI) studies the way in which computer technology influences human activities [Dix09]. An important issue of the HCI discipline is the concept of *usability*, which evaluates the clarity and elegance with which a software application has been designed for its interaction with a user, providing a quality measure that assesses how easy its user interfaces are to use. Nevertheless, in systems designed for personal use, the idea of *user experience*; which refers to how people feel when they use them, is equally or even more important than the notion of *usability* [Dix09].

As introduced by Picard [PP97], Affective Computing is the discipline that develops devices to detect and respond to users' emotions. Its main objective is to capture and process affective information with the aim of enhancing the *user experience* and naturalizing the communication between the human and the computer. The area of Affective Computing covers topics such as emotion recognition, emotion understanding or emotional synthesis.

Emotion recognition by the computer is still a challenging field, especially if the recognition process is only based on the analysis of the voice. This is due to multiple issues. Sometimes not even humans are able to classify natural emotions, thus, we cannot expect machines to perform the same classification correctly. Besides, emotion expression is highly speaker, culture and language dependent. Moreover, a spoken utterance may include more than one emotion, either as a combination of different underlying emotions in the same portion or as individual expressions

of each emotion in different speech segments. Another interesting aspect is that there is no definitive consensus in the research community regarding the most useful speech features for emotion recognition. A possible reason may be the high variability introduced by different speakers in commonly-used prosodic features. Finally, selecting the set of emotions to classify is also an important decision, which can affect the performance of the speech emotion recognizer. Many works on the topic agree that any emotion is a combination of primary emotions. The primary six emotions include anger, disgust, fear, joy, sadness and surprise [Cow+01].

On the other hand, in many HCI applications, virtual characters are used as the main interface or interaction agent between the user and the computer, with the aim of making communication more pleasant and effective. In fact, as reported in [OW04], the attention of humans to the computer can increase by 30% when they communicate through avatars or talking heads.

Depending on the type of input, virtual character animations can be text- or speech-driven. Whereas text-driven animations use speech synthesizers to generate artificial voices, speech-driven systems employ natural human voices to animate the virtual characters, resulting in more natural and emotional speech. Nevertheless, the latter approach requires more advanced technology to convert the incoming audio in facial gestures and animations. In this sense, audio-to-visual conversion can be considered the core of speech-driven facial animation [LYW14].

In this section, the main contributions made to the speech emotion recognition and speech-driven facial animation fields are presented. On the one hand, research work on speech emotion recognition has been focused on feature extraction and on the selection of the best characteristics for emotion recognition in Spanish and Basque, as well as on the application of a new classification method to improve the state-of-the-art recognition accuracies achieved in two datasets. On the other hand, a real-time system for the facial animation of a virtual character is presented, which integrates technology for automatic lip synchronization and facial gesture generation and which deals with co-articulation effects.

4.2 State of the Art

The expression of emotions by humans is multimodal [Lan95]. Apart from verbal information, emotions are transmitted through speech, facial expressions, gestures and other nonverbal psycho-physiological clues. Concerning Speech Emotion Recognition (SER), it is still a very active research field [KR12; RE13; HYT14] that refers

to the analysis of spoken behavior as a marker of emotion, with a focus on the nonverbal aspects of speech [Sch+13].

Speech carries both explicit linguistic information and implicit paralinguistic information. Whereas linguistic information is linked to the words that have been spoken, the latter derives information about the way in which they were spoken. SER is commonly performed without considering linguistic information (which would require an automatic speech recognition system), but through the extraction, selection and classification of paralinguistic characteristics, mainly related to prosodic and voice quality features.

The literature regarding SER is really rich and extensive, as many studies have examined vocal expressions of emotions. The studies in [Sch03; SJK03; Sch+11; KR12] and more recently [Sch+13; RE13] and [Eyb+16] have reported wide reviews of the main related work in the field. Although a number of different types of systems haven been developed over time using a wide range of dissimilar features and machine learning algorithms, all studies have converged in a similar operative schema which includes a feature extraction front-end followed by a classification/regression module.

Concerning the speech features employed, the bibliography shows that for a long time there was no real consensus on which would be the most suitable characteristics. [GS10] demonstrated empirically that features such as intensity; mean, variability and range of the fundamental frequency; the high frequency energy, as well as the articulation rate are all linked to emotional states such as stress, anger, sadness and boredom. The study in [EWS13] also demonstrated the high impact of the MFCC (Mel-Frequency-Cepstral-Coefficients) parameters and the particular relevance of the lower order coefficients. Nevertheless, the general tendency in most studies of automatic emotion classification is the use of very large sets of features [EBS15], with the hope that the machine learning algorithms will be able to discriminate the most relevant parameters. However, large feature sets may produce an overfitting effect, reducing their generalization capabilities on unseen test data [Sch+10]. Very recently, a minimalistic set of parameters has been proposed in [Eyb+16], which is available through the OpenSmile toolkit [EWS10].

The extracted speech features are then used to train machine learning algorithms that carry out automatic classification. The literature exposes a wide range of classifiers that have been applied to speech emotion recognition. Such classifiers may be generative or discriminative. Generative classifiers model the distribution of the training data features from each emotion class individually, which means that each model is trained exclusively with data from that class, and not from any other class. The generative classifiers most commonly employed for SER include Gaussian

Mixture Models [SEA13], Hidden Markov Models [Sha13] and Probabilistic Neural Networks [Set09]. On the other hand, discriminative classifiers do not model the distribution of the entire feature space and try to maximize a function to discriminate between the different classes. Some examples of discriminative classifiers used on the SER field are Support Vector Machines [Eyb+16], Decision Trees [SR10] and Neural Networks [Wöl+09]. Even though discriminative classifiers tend to outperform the generative ones, the optimal algorithm often depends on the application scenario, the feature set and the data employed [PR11]. Another trend today is the fusion of different machine learning algorithms with the aim of combining the benefits of several classifiers [WL11; Che+12; AD13; AW14; HZG14]. Finally, some studies have also tried to model and deal with the phonetic and speaker variability that can affect SER systems designed to work on speaker-independent application domains [SEA15].

Regarding speech-driven facial animation, its importance stems from the fact that speech recognition is more accurate when auditory and visual sources are combined during the communication process. Speech understanding improves if in addition to the acoustic cues, visual cues such as lip movements or facial expressions are involved. Consequently, there has been a lot of research effort incorporating coherent facial motions linked to speech in HCI applications. In this sense, lip movements are one of the most important components in spoken facial animation [RTS06]. A convincing lip sync improves the realism of the virtual character, making it seem alive and turning the animation more credible [HYS08]. Several studies have proposed various techniques to synthesize lip sync animations, such as the use of Hidden Markov Models to manage contextual information at phoneme-level [WHS12; HYS08] or the application of Neural Networks [Tak09] and Gaussian Mixture Models [Han+12; TBT08] for direct audio-visual conversion without the use of phonetic information. Determining how to model the speech co-articulation effect is essential to generate realistic lip sync movements. Approaches based on the use of visemes conventionally employ interpolation techniques [CM93] or rule-based functions [WEF07]. Other data-driven studies focus on the optimal selection and concatenation of motion units from a pre-collected database [Tay+12; MD12], whilst some other approaches employ statistical models learned from data [EGP02]. All these approaches have shown reasonably good performance on pre-recorded audios, through the exploitation of global techniques and functions and the use of contextual information. However, in live speech-driven lip-sync scenarios the forthcoming speech is not available and global algorithms are not applicable. Several approaches have been proposed in the literature for the generation of live lip-sync animations, mainly based on pre-designed phoneme-viseme mappings [Xu+13] or statistical models like Neural Networks [HWH02] and nearest-neighbor search [Gut+05], among others.

On the other hand, empirical rules [GB06; MH07] or statistical models learnt from pre-recorded human motion data [Bus+05; CB05] have been employed for offline synthesis of facial motions such as eye, eyebrow and/or head movements. More recently, works like [LTK09; Lev+10] have implemented solutions for live speech-driven body gesture and movement generation and [LMD12] proposed a framework to generate realistic and synchronized eye and head movements on-the-fly.

4.3 Challenges in the fields

For any research field to evolve properly within the scientific community, an initial agreement of the formulation of the problem to be resolved needs to be defined. Despite considerable previous work has been done in emotion recognition, the lack of a proper formulation may be the reason why the field is still in its infancy. A look at the most recent survey papers shows that there is still no agreement on how the emotion recognition problem should be addressed. This may be due to the lack of a fully established theory of emotions and a general consensus on how to label them [Moo+13; SEA15]. In this sense, some works carry out a categorical classification, while others turn the task into a regression problem. Whereas solutions based on regression use a dimensional approach to label emotions, systems that follow the classification approach commonly employ both categorical and/or dimensional labeling. Moreover, the great variety of the features, classifiers, data sets and evaluation metrics employed across experiments poses an additional problem. These issues, among others, are the basis of the current main challenges in the SER field that are listed below:

- *Formulation.* Although categorical classification over the 'big six' emotions (Happiness, Sadness, Fear, Disgust, Anger, Surprise) is the most common approach, there is no universal method formulated that could be employed across experiments.

- *Technical agreement.* Despite all studies follow the same schema based on feature extraction and classification/regression and despite SVMs can be considered the most commonly employed machine learning algorithm, the set of most relevant features and/or classifiers needs to be defined. Together with the exploitation of common datasets, such definition would allow the comparison of different systems and approaches.

- *Spontaneous speech.* Emotion recognition on spontaneous speech over naturalistic databases containing real-life speech is a hot topic in the community, since most previous work has been carried out on acted speech and on sentences

containing one isolated emotion. Real-life speech has the particularity of containing audio tracks with more dynamic, blended and less pure emotions, which represents a more difficult challenge.

- *Speaker-independent*. Nowadays, the availability of rich acted data of a certain speaker allows training highly accurate speaker-dependent models. Nevertheless, there is still a wide gap between the performance of speaker-dependent and speaker-independent systems.

- *Real-time.* Since one of its main applications is extracting information about the user state in HCI systems, accurate SER technology working in real-time will be needed in order to assess the mood of the conversational partner as fast as possible.

- *Portability.* As more accurate SER solutions with low computational cost are developed, they will need to be adapted to be integrated in small electronic devices.

With regard to real-time speech-driven lip-sync and facial animation, a number of approaches have been explored in the field but it is still a challenging topic in the speech community. First, optimal speech motion requires contextual information from the past and from the future. Despite such kind of information can be used to apply global techniques and find the optimal facial animation at each time in pre-recorded speech, the approach is not feasible in live visual speech synthesis because forthcoming speech is not available at synthesis time. This makes it very difficult, if not impossible, to achieve the same level of realism in offline and real-time scenarios. Second, live speech-driven facial animation algorithms must be highly efficient to ensure good performance in real-time speed. When audio-to-visual conversion is carried out at phoneme level, the speed of the system is set by the delay of the online phoneme recogniser. Obviously, phoneme recognition accuracy will also impact the quality of the final animations. The real-time issue becomes even more challenging when paralinguistic information (e.g. prosody) is used to generate other expressive facial motions. Furthermore, since paralinguistic information is speaker, language and culture dependent, the development of speaker-independent live speech-driven visual animation technology is a great challenge for the future.

On the other hand, dealing with the co-articulation effect can be considered the other main challenge of the field. Although several works have tried to address this problem, modeling how the pronunciation of each phoneme affects that of the adjacent ones is still challenging and becomes even more difficult when emotional states and/or significant prosody variations (e.g. speech rate) are involved. Finally, given the wide spread use of the Internet and the smart electronic devices, developing

technology that integrates expressive avatars and/or talking heads in such kind of environments is a trend that will have to tackle all the challenges described above in computationally limited conditions.

4.4 Main contributions

The main contributions made to the SER field in this PhD project are linked to the principal feature extraction and classification components. Regarding feature extraction, several techniques have been implemented with the aim of selecting the most relevant features for emotion recognition of acted speech in Basque and Spanish. This work was performed over the Rekemozio data set [Lop+07], and the selected features were divided by language and genre.

On the other hand, a new multi-classifier based on the Stacked Generalization method for the fusion of single machine learning algorithms applied to the SER field was proposed and evaluated with different sets of features, data sets and languages. Two sets of features were used in the experiments, including an in-house defined set of 123 features, and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [Eyb+16]. The Rekemozio data set (Basque and Spanish) and the well-known Berlin Emotional Speech database (German) [Bur+05] were employed in the experiments and the performance of the new classifier was compared with other classification methods.

All of the studies described in this dissertation work were focused on the classification of emotions at categorical level, including the six primary emotions plus the neutral one.

With regard to speech-driven facial animation, research work carried out to animate a virtual character in real-time through the application of techniques for lip sync, co-articulation and facial motion generation is presented.

4.4.1 Feature Subset Selection for Speech Emotion Recognition in Basque and Spanish languages

The goal of this work was to take a step forward in the search for the most relevant emotional speech features for Spanish and Basque. The experimentation was divided in three phases, each including a different set of features. Four well-known supervised classification algorithms were applied in each phase, including two Decision Trees, Instance-Based Learning and Naive Bayes classifiers. The results obtained by

each classifier at each phase were then intended to be outperformed by applying a Feature Subset Selection (FSS) technique through the Estimation of Distribution Algorithm (EDA). The final results confirmed the good performance of the EDA based FSS technique in the selection of the most relevant features to obtain the best classification scores.

This work was performed over the Rekemozio data set, which contains instances of both Basque and Spanish. It consists of a multimedia database with video and audio recordings and is described in more detail in [Lop+07]. In these experiments, only the spoken material was employed. Rekemozio uses a categorical model based on the six basic and universal emotions defined by Ekman and Friesen [EFP75], which correspond to Sadness, Fear, Joy, Anger, Surprise and Disgust. The Neutral emotion category was also added to the dataset. Recordings were carried out by skilled actors and actresses using semantically and non-semantically relevant words and sentences. They were then validated by fifty-seven volunteers who selected the audio tracks which most closely reflected the corresponding emotion. The validation concluded that 78% of audio stimuli were valid to express the intended emotion as the recognition accuracy percentage was over 50% [Lóp+06].

Regarding feature extraction, a number of features which had been frequently used in other similar studies in the literature [Iri+00; Nav+04] were selected and checked in the first phase. Information related to prosody, such as the fundamental frequency, energy, intensity and speaking rate was extracted using a 20 ms frame-based analysis with an overlapping of 10 ms, obtaining a total set of 32 features. The computed features are described in more detail in [Álv+06].

In the second phase, 91 new features were extracted following the work presented in [Tat+02], which proposed new interesting formulas to extract information from speech, and also defined a novel technique for signal treatment, not only extracting information by frames, but by regions consisting of more than three consecutive frames, for the analysis of both voiced and unvoiced parts. The features extracted in this phase are described in more detail in [Álv+07a].

In the third phase, the whole set of features of the previous two phases was compiled, obtaining a total of 123 speech features.

Four supervised machine learning algorithms were applied in each phase, including the ID3 [Qui86] and C4.5 [Qui14] decision trees, IB [AKA91] as the Instance-Based Learning paradigm and a Naive-Bayes (NB) [Min61] classifier. Using the 10-fold cross-validation method, accuracies for each of the described classifiers over the three sets of features presented above were first computed separately for the Basque and Spanish audio recordings of the Rekemozio dataset.

With the aim of reducing dimensionality, removing irrelevant or redundant characteristics, selecting the most relevant speech features and improving results, a Feature Subset Selection (FSS) approach was then applied over each machine learning algorithm. The feature selection task can be exposed as a search problem, where each state in the search space identifies a subset of possible features. In this sense, Genetic Algorithms [Hol75] are one of the best known techniques to solve optimization problems, and propose a population-based search method. An interesting adaptation is the Estimation of Distribution Algorithm (EDA) [Inz+00], where the new population is sampled at each iteration from a probability distribution which is estimated from the selected individuals. This way, a randomized, evolutionary, population-based search can be performed using probabilistic information to guide the search. In Figure 4.1, the main computation scheme of the EDA algorithm is presented.

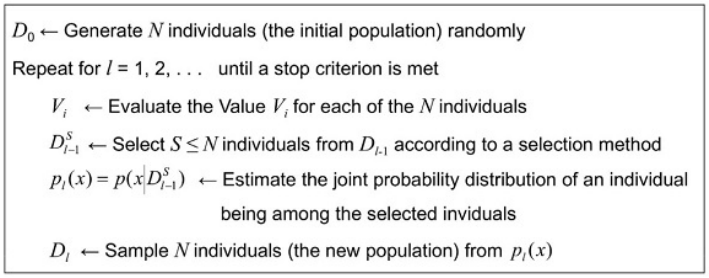


Fig. 4.1: Main scheme of the Estimation of Distribution Algorithms (EDA) approach

All the accuracies obtained through the application of machine learning algorithms with and without applying the EDA-based FSS technique are shown and explained in detail in [Arr+14].

Figures 4.2 and 4.3 show the scores reached for Basque and Spanish in the three phases respectively, for each of the machine learning algorithms without EDA-FSS.

In addition, Figures 4.4 and 4.5 present the scores achieved when the EDA-FSS technique is applied to each of the machine learning algorithms for Basque and Spanish. Results are shown divided by phases. In these figures, the accuracies obtained when applying a greedy FSS search algorithm (FSS-Forward) instead of EDA are also given. This method was tested only for the third phase feature set, and was included for comparison purposes with the aim of emphasizing the goodness of the EDA search process.

If we compare the results in Figures 4.2 and 4.3 with the ones shown in 4.4 and 4.5, the clear improvements achieved when applying the EDA-FSS method can be observed. It is worth emphasizing that the difference between the classification accuracies obtained with the initial set of 32 features without FSS and those obtained

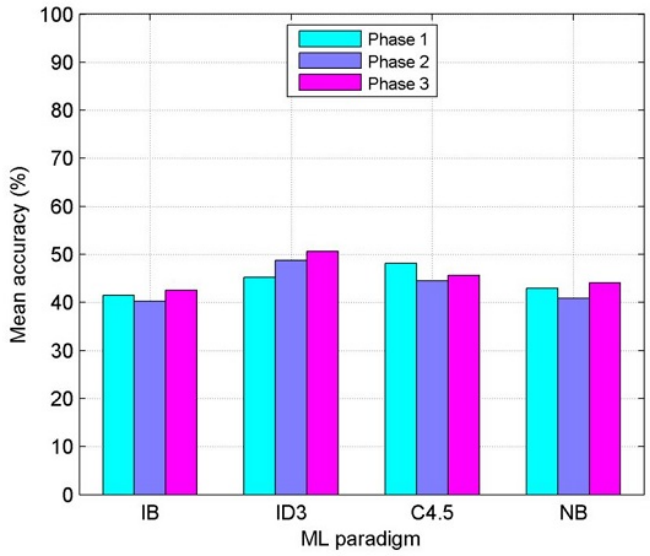


Fig. 4.2: Scores for each of the machine-learning paradigms without EDA-FSS for Basque. Results correspond to the mean accuracy computed over all the actors in the database for each phase of the experiments

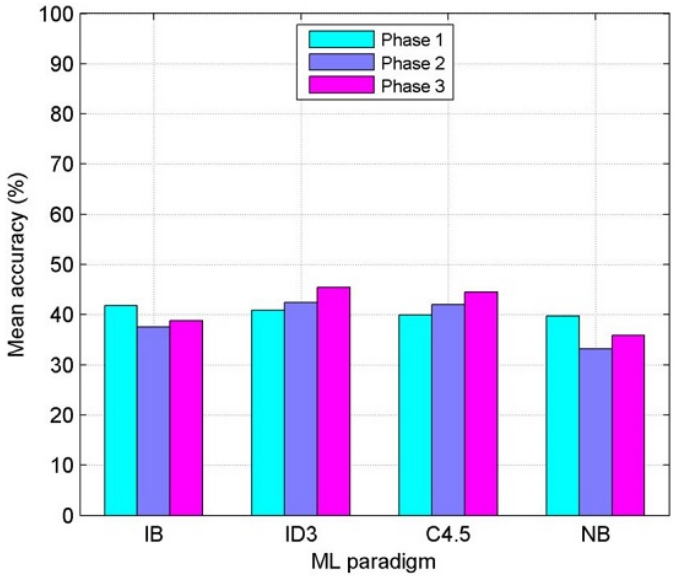


Fig. 4.3: Scores for each of the machine-learning paradigms without EDA-FSS for Spanish. Results correspond to the mean accuracy computed over all the actors in the database for each phase of the experiments

with the whole set of 123 features after applying FSS sum up notable average increases of 30.62 and 30.61 percentage points for Basque and Spanish respectively. Concerning the FSS-Forward method, its best results seem to be obtained with NB classifier for both languages. Nevertheless, the classification performances are disappointing, since they are similar to those obtained using the initial set of 32 features without applying EDA-FSS, and they are thus far from the best scores obtained with the EDA-FSS method.

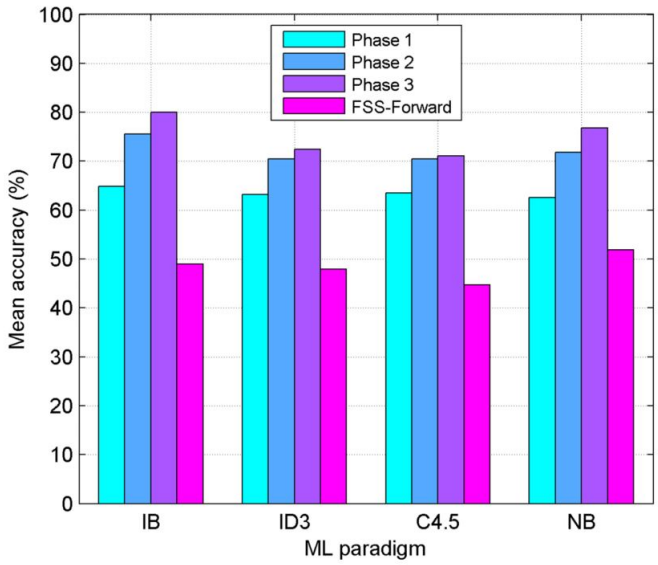


Fig. 4.4: Scores for each of the machine-learning paradigms when EDA-FSS is applied to Basque. Results correspond to the mean accuracy computed over all the actors in the database for each phase of the experiments. Results obtained with a standard FSS-Forward approach are also shown.

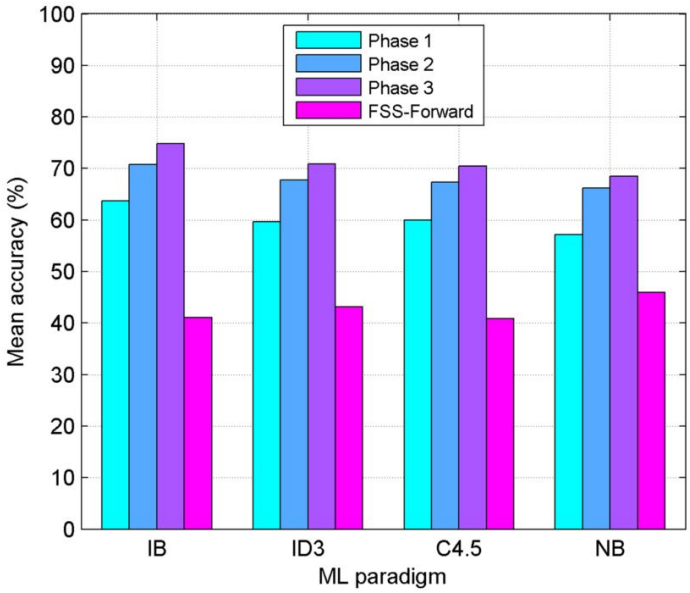


Fig. 4.5: Scores for each of the machine-learning paradigms when EDA-FSS is applied to Spanish. Results correspond to the mean accuracy computed over all the actors in the database for each phase of the experiments. Results obtained with a standard FSS-Forward approach are also shown.

The EDA-FSS method allowed us to select the best set of speech features that achieved the highest accuracies for each language. This selection was performed over the set of features of the third phase, where the best classification rates were reached and the classifiers achieved the best accuracies. In this sense, the IB paradigm was the classifier which obtained the best results for most of the actors.

Given that speaker-dependent models were constructed, each actor may have different relevant features. These features were analyzed by grouping actors per language and gender aiming at obtaining a partial independence of the actor. The purpose of this grouping was to shed more light on the impact that gender and language can have in the final features of each subgroup. The criterion to consider a feature relevant within a subgroup was that more than 50% of the actors had to have that feature selected by the algorithm.

It must be highlighted that several features were common for both languages and genres. These features mainly corresponded to prosodic features, such as the Fundamental Frequency (the mean, variance, the mean square error of the regression coefficient and mean of the pitch means in every voiced region); Energy (maximum, mean and variance); RMS energy (maximum and mean), and Loudness. The voice quality features shared by both languages and genres were less and mostly corresponded to the third formant mean, the first and second formant bandwidths and the activation level of the speech signal; where, the maximum and mean stand out among all the voiced regions.

These shared prosodic and voice quality features can be considered to be the most relevant features for developing speaker- and language-independent systems, at least for Basque and Spanish. The rest of the more relevant features for each language and genre are presented in [Arr+14].

The main publications related to the work described above are listed bellow:

- [Arr+14] Arruti, A., Cearreta, I., Álvarez, A., Lazkano, E., and Sierra, B. (2014). *Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction.* PloS one, 9(10), e108975.

- [Álv+07a] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007, September). *A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms.* In Text, Speech and Dialogue (pp. 423-430). Springer Berlin Heidelberg.

- [Álv+07b] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2007). *Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech.* In Advances in Nonlinear Speech Processing (pp. 273-281). Springer Berlin Heidelberg.

- [Álv+06] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2006, September). *Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in spoken Spanish and Standard Basque Language.* In Text, Speech and Dialogue (pp. 565-572). Springer Berlin Heidelberg.

4.4.2 Classifier Subset Selection for the Stacked Generalization applied to Speech Emotion Recognition

Besides the feature extraction component and the selection of the most relevant features for the SER area, the search for the most suitable classification method is also a hot trend in the field.

With the aim of improving previous results of different studies, a new classification approach was proposed to deal with emotion recognition in speech, following the current tendency of fusing classifiers to take advantage of their individual benefits (see Section 4.2).

In this work, stacked generalization [Wol92] was employed as the main method to construct a multi-classifier system. Whereas ensemble strategies, such as bagging or boosting, obtain the final decision after a vote among the predictions of individual classifiers, Stacked Generalization applies another individual classifier to these initial predictions with the aim of detecting patterns and enhancing the performance of the vote. The Stacked Generalization can be considered a framework for the combination of classifiers, in which each layer of classifiers is used to combine the predictions of the classifiers of its preceding layer. A single classifier at the top-most level outputs the final prediction.

In this work, a two-level system was employed, composed of a first layer (level-0) with several single classifiers, and a second layer (level-1) with a single meta-classifier which returns the final decision. While the data to train the single classifiers of the first layer corresponds to the features extracted from the emotional audio tracks, the data used to build the meta-classifier is obtained after a validation process, where the outputs of the first layer classifiers are taken as attributes, and the class is the real class of the example (emotion). This implies the creation of a new dataset in which the number of predictor variables corresponds to the number of classifiers of the bottom layer, and all of the variables have the same value range as the class variable. Figure 4.6 presents an example of the implemented Stacked Generalization (SG) multi-classifier.

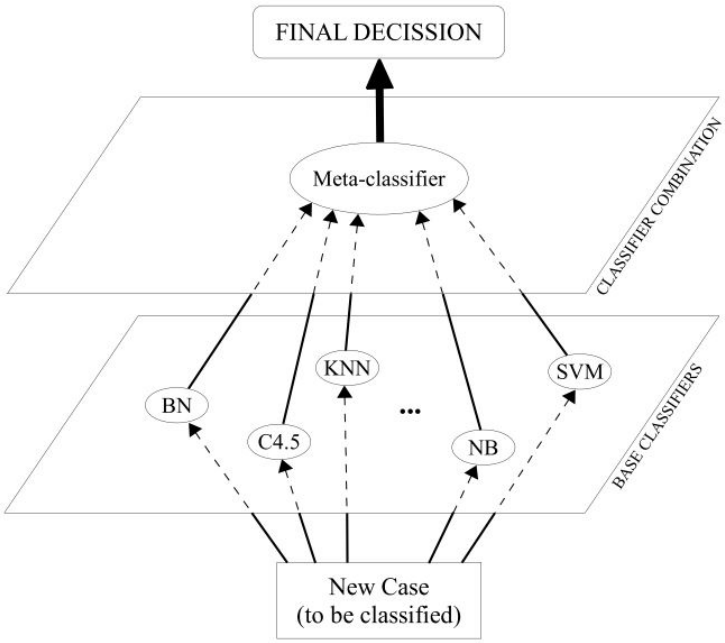


Fig. 4.6: Stacked Generalization schemata

Following a similar approach to that employed with the Feature Subset Selection method, using many classifiers in the first layer of the SG classifier can be very effective, but selecting an optimal subset can reduce the computational cost and improve accuracy, assuming that the selected classifiers are diverse and independent enough. To this end, the Classifier Subset Selection (CSS) concept was included as it is graphically illustrated in Figure 4.7.

As shown in Figure 4.7, an intermediate phase is added to the multi-classifier to select a subset of the best classifiers in the first layer. In this scenario, evolutionary approaches are often employed as the method to make the optimal selection. Following the work in [Inz+00] an Estimation of Bayesian Network Algorithms (EBNA) [EL99] were integrated. EBNA adopt Bayesian networks as the probabilistic model and can be considered one of the most sophisticated algorithms in the Estimation of Distribution Algorithms (EDA).

The focus of the work was to compare the CSS-based SG classifier with single classifiers and other multi-classifier systems, such as Bagging, Boosting and Standard Stacking Generalization. The experimentation was divided into three main phases, each designed to test several classifiers, feature sets and databases.

The first two phases were performed over the Rekemozio dataset. In the first phase, the following ten initial supervised classifiers were selected: Bayesian Network (BN), C4.5, k-Nearest Neighbors (kNN), KStar, Naive Bayes Tree (NBT), Naive Bayes

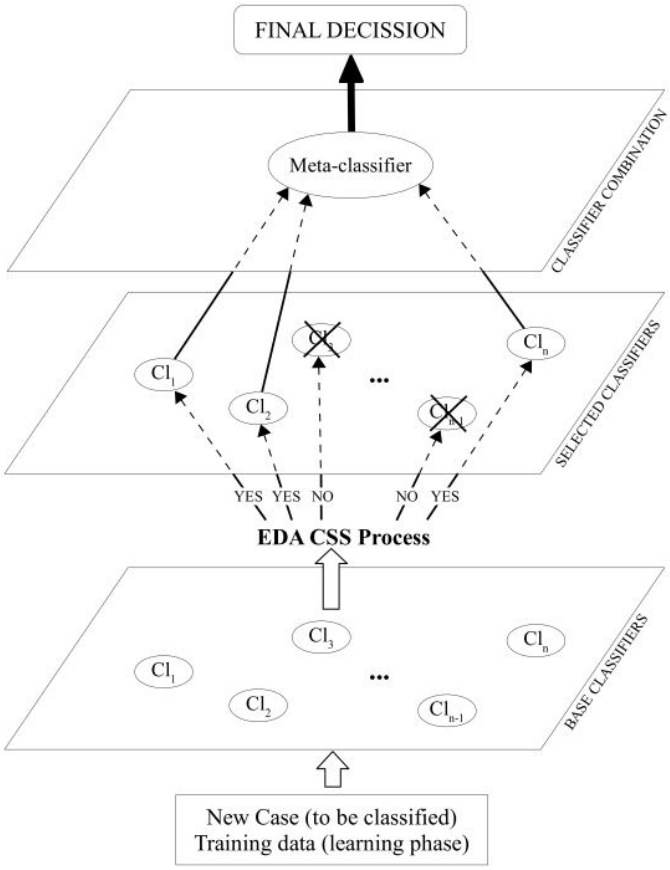


Fig. 4.7: Stacked Generalization with Classifier Subset Selection

(NB), One Rule (OneR), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Random Forest (RandomF) and Support Vector Machines (SVM). These classifiers were also used to construct the multi-classifiers. Besides, the whole set of 123 features employed in our previous work [Arr+14] were used as the principal speech characteristics. Results obtained in this phase concluded that the CSS-based SG classifier achieved higher accuracies than most of the single and multi-classifiers. In fact, results clearly outperformed 14 of the 17 actors in the database. In addition, the SVM classifier was shown to be the meta-classifier which provided the best results.

The aim of the second phase was to verify the efficiency of the CSS-based SG classifier on the RekEmozio dataset using: (1) the well-known set of acoustic parameters called eGeMAPS [Eyb+16], instead of the set of 123 features; and (2) a different configuration of the base classifiers in the first layer. Only the best meta-classifier of the first phase (SVM) was used to construct multi-classifiers in this second phase. Regarding base classifiers, the following two artificial neural networks were added: MultiLayer Perceptron (MLP) and Radial Basis Function (RBF). The average accuracy of the CSS-based SG classifiers in this phase outperformed the mean accuracy of the

first phase by 4.56 percentage points, which demonstrated the effectiveness of the eGeMAPS parameters and the new classifiers included in the first layer.

Once a suitable configuration of the CSS-based SG classifiers concerning the features (eGeMAPS) was obtained, the configuration of the first layer and the optimal meta-classifier (SVM) was tested on the publicly available EmoDB database [Bur+05] and results were compared with those obtained by other studies in the literature. On average, the CSS-based SG classifier with the SVM acting as the meta-classifier reached an average accuracy of 82.45% for all of the actors in the EmoDB database. Considering that the human perception rate for this database was set to 84% [Esp+12], this mean value can be seen as a promising result. Moreover, this score outperforms the results of other works in the literature over the same EmoDB database, such as the scores obtained in [Esp+12; Cos+14], which reached accuracies of 79% and 77%, respectively, although these works analyzed the whole database and used different machine learning algorithms and audio features. Table 4.1 shows the results of the three best classifiers of each system in the third phase. The best result of the three classification systems is highlighted in bold per actor. Interestingly, MLP, RandomF and SVM are the best three classifiers for each of the classification systems.

Tab. 4.1: Third phase. Accuracy percentages per actor for the best three classifiers of each system built on the Berlin Emotional Speech database (Emo-DB). Mean and SD rows represent the average and standard deviation considering all of the actors.

	Single			**Standard Stacking**			**CSS Stacking**		
	MLP	**RandomF**	**SVM**	**MLP**	**RandomF**	**SVM**	**MLP**	**RandomF**	**SVM**
A1	79.59%	73.46%	77.55%	63.26%	71.42%	61.22%	79.59%	**81.63%**	79.59%
A2	94.82%	87.93%	86.20%	79.31%	89.65%	72.41%	93.10%	94.83%	**96.55%**
A3	74.41%	62.79%	67.44%	62.79%	67.44%	62.79%	74.42%	74.42%	**76.74%**
A4	84.21%	84.21%	81.57%	68.42%	71.05%	68.42%	**89.47%**	84.21%	86.84%
A5	63.63%	**80.00%**	72.72%	56.36%	65.45%	54.54%	67.27%	72.73%	78.18%
A6	77.14%	74.28%	80.00%	71.42%	68.57%	68.57%	**82.86%**	**82.86%**	**82.86%**
A7	78.68%	75.40%	72.13%	67.21%	70.49%	65.57%	77.05%	**80.33%**	78.69%
A8	78.26%	75.36%	78.26%	73.91%	76.81%	78.26%	82.61%	**86.96%**	85.51%
A9	67.85%	**82.14%**	66.07%	69.64%	71.42%	64.28%	76.79%	75.00%	75.00%
A10	74.64%	83.09%	76.05%	73.23%	71.83%	76.05%	83.10%	80.28%	**84.51%**
Mean	77.32%	77.87%	75.80%	68.55%	72.41%	67.21%	80.63%	81.32%	**82.45%**
SD	8.52	7.17	6.28	6.56	6.76	7.13	7.41	6.57	6.33

In [Álv+16d], the confusion matrices obtained by the CSS stacking classifiers in the second and third phases are also presented for each actor.

In summary, the overall results demonstrate the good performance of the CSS-based SG classification paradigm and confirm the robustness of this classification system to deal with speech emotion recognition over several conditions and datasets.

The main publication related to the work described above is given bellow:

- [Álv+16d] Álvarez, A., Sierra, B., Arruti, A., López-Gil, J. M., and Garay-Vitoria, N. (2015). *Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech.* Sensors, 16(1), 21.

4.4.3 Speech-driven facial animation

The first contribution is related to the development of a mixed reality prototype, which provided a 3D virtual presenter embedded in a real TV scenario driven by an actor in real-time. The prototype solved some of the gaps regarding the applicability of 3D real-time virtual presenters to the TV environment; such as (1) character flexibility, allowing the reuse of characters that had not been created exclusively for the mixed reality system; (2) equipment need, letting the computer create the real/virtual mix directly without the need of any chroma, (3) mixed reality flexibility, allowing the cameraman to change the camera parameters (zoom, movements, etc.) in real-time, and (4) actor's comfort, integrating voice analysis for the automatic animation of the character instead of using motion capture systems.

The voice analysis module aimed at processing the actor's speech in real-time and obtaining phonemes and prosodic information to generate synchronized lip and facial motions for the virtual character animation. Phonemes were obtained through an HTK/ATK-based online recognition system. Once recognized, phonemes were then mapped to their corresponding visemes in order to carry out lip-synchronization. In parallel, the speech of the actor was processed to extract paralinguistic information related to the pitch and RMS energy of the signal. In order to make the component adaptable to each user, the initial seconds of speech were used to generate a speaker model containing the maximum, minimum and mean values of the pitch and energy features. These values were then employed to shoot facial animations linked to head, eye, eyebrow and mouth movements following heuristically predefined rules. The communication interface between the speech application and the animation platform was developed using sockets based on TCP/IP. This way, the animation module was fed with the recognized unit and the facial movements needed to realistically animate the character through real speech.

On the other hand, a system to learn foreign language pronunciation was developed and integrated within a web application (see Section 2.4.2 for details). It was composed of an expressive talking head, which was able to pronounce any text sentence in Basque and English. Speech was generated through a text-to-speech synthesizer and lip-synchronization was carried out automatically after a forced-alignment process, which allowed obtaining time-tamps at phoneme level. For the lip animation, the co-articulation problem was tackled considering syllables instead

of phonemes during the actual audio-to-visual conversion process. In this sense, the current, previous and next syllables were employed at each time in order to smooth the co-articulation effect. The syllabification module was rule-based and was developed ad-hoc for Basque considering all the pronunciation rules of the language. The facial synthesis was performed following the method in [CM93].

Following the challenge of integrating expressive avatars and/or talking heads in the Internet environment, the pronunciation learning prototype focused on directly using the web browser to render the virtual head exploiting the WebGL standard, so that the animation process was entirely performed in the cloud.

The main publications related to the work performed on the field are listed bellow:

- [Oya+10] Oyarzun, D., Mujika, A., Álvarez, A., Legarretaetxeberria, A., Arrieta, A., and del Puy Carretero, M. (2010). *High-realistic and flexible virtual presenters.* In Articulated Motion and Deformable Objects (pp. 108-117). Springer Berlin Heidelberg.

- [Muj+13] Mujika, A., Diez, H., Alvarez, A., Urteaga, M., and Oyarzun, D. (2013). *Realistic visual speech synthesis in WebGL.* In Proceedings of the 18th International Conference on 3D Web Technology (pp. 207-207). ACM.

4.5 Conclusions and Future Work

This chapter has focused on describing the current status, challenges and main contributions to the fields of speech emotion recognition and speech-driven facial animation, both aimed at enhancing the interaction between computers and humans.

Although significant advances have been done in the last decades in other speech processing fields such as speech recognition or speaker verification, the community is still far from achieving natural human-computer interactions. Apart from the current limitations of dialogue systems, many of which are still rule-based instead of statistical, the inability of machines to understand the users' emotional states is one of the main reasons for such unnatural interactions.

A huge number of studies have presented speech emotion recognition systems, many of them following the same processing pipeline, including a feature extraction module and a subsequent classification process. This dissertation work has attempted to give a step forward in both components. First, a new Feature Subset Selection

method was introduced to obtain the most relevant features for emotion recognition of Basque and Spanish audio tracks. Experiments have shown that there are some speech features common to both languages (Basque and Spanish) and the genres involved. Besides, the presented EDA-FSS approach has shown to be a suitable method to select the most relevant features that improve accuracy and applicable to other languages. Furthermore, a new classification approach based on fusing the benefits from several single classifiers has also been proposed. Such classifier was tested over several features and corpora, outperforming the results obtained by previous studies on the same data sets.

Future work in the field should involve carrying out new experiments on naturalistic databases which include real-life spontaneous speech, in order to test the goodness of the presented feature selection and classification paradigms in other data set conditions and domains. Besides, new standard classifiers should be explored to be included in the first layer and as meta-classifiers, and a combination of data from several databases could be used, with the aim of exploring speaker- and language-independent classification systems.

Regarding speech-driven facial animation, two solutions were presented for different environments. The first solution embedded a 3D virtual presenter in a real TV scenario. The virtual character was animated in real-time through the voice of an actor, both linguistically (lip-synch) and paralinguistically (face motions), combining online phone recognition and speech prosody analysis technologies. The second solution broke the barrier of the Internet and presented a web application where the browser was used to render an expressive talking-head. In addition, a pronunciation evaluation system was implemented and integrated for the Basque and English languages.

5 Transfer of Speech Solutions to Industry

As a non-profit applied research center, Vicomtech-IK4 aims to work in the benefit of enterprises and society. One of its main goals is the transfer of technology to Industry in order to provide companies with high-value technological assets within a collaborative framework, which guarantees that clients can implement their business models exploiting the transferred results and Vicomtech-IK4 can continue to enhance and specialize the technology.

Along its professional path, Vicomtech-IK4 has succeed with a high number of technology transfers to Industry. And several of the developments described in this dissertation work have also provided solutions to various companies, either as core technology or as a means to improve internal workflows and processes.

5.1 Mixer Servicios Audiovisuales S.L.

Mixer is one of the principal dubbing and subtitling service providers of the Basque Country's public broadcast corporation, Euskal Irrati Telebista (EITB, Basque Radio-television). One of their main activities is focused on dubbing TV contents from several languages into Basque and on generating the corresponding subtitles. For many years, they first translated the original script into Basque and adapted it for dubbing purposes. Then, they created the Basque subtitles based on the translated script. These two tasks were carried out manually.

With the aim of improving its workflow, Mixer integrated long-audio alignment technology provided by Vicomtech-IK4. To this end, a standalone application was implemented and adapted to the particular needs of Mixer. The application was designed to be installed locally in different computers. And the technology was adapted to the specific subtitling rules of the company, including features related to segmentation and the persistence time of subtitles on screen. As Mixer reported, they achieved a productivity gain greater than 60%. While previously they needed around 10 hours to generate subtitles for each hour of content, the transferred technology allowed them to reduce this time to less than 4 hours. It needs to be

noted that the resulting automatic subtitles, were revised and post-edited to correct possible errors.

http://www.mixer.com.es/es/mixer/

5.2 Ubertitles S.L.

Ubertitles S.L. was born in 2012 as a spin-off of Vicomtech-IK4 with the main objective of providing a service to subtitle videos rapidly, simply and automatically. The core technology of the company was transferred by Vicomtech-IK4 and is based on the long-audio alignment approach described in subsection 3.4.4. Users can provide contents with their transcriptions and the corresponding subtitles are generated automatically on the spot. Ubertitles also offers a manual transcription service for contents for which transcriptions are not available. Ubertitles' solution adapts to the specific needs of each user, as the number of lines per subtitle or the amount of characters per line can be adjusted for segmentation purposes. Ubertitles also provides functionalities to automatically assign colors to the subtitles of the different characters. This information is particularly useful to generate subtitles for the deaf and the hard of hearing.

Ubertitles provides a SaaS service in the cloud that operates from any computer with Internet acces. The service does not involve any monthly costs or licenses and users only pay for its actual use.

http://www.ubertitles.com/

5.3 Irekia

As briefly introduced in section 2.1.6, Irekia is the Open Government (oGov) portal of the Basque Government and as such, it serves as the direct communication channel between the Basque Administration and the citizens. They disseminate the activities carried out by the Basque Government through their website and across dissemination events that aim to break barriers and keep people aware of all the political activities carried out by their government.

At the moment, Irekia is making a great effort in order to make their contents accessible through the integration of technologies such as ReadSpeaker[1], which includes

[1] http://www.readspeaker.com/

text-to-speech for web reading or the interaction tools offered by Inclusite[2]. Irekia also sought for solutions to generate subtitles automatically, since they do not count with the resources to accomplish such work manually. As a result, Vicomtech-IK4 has integrated a subtitling platform similar to that described in section 3.4.6 in the facilities of Irekia. The platform was installed on a server connected to their internal network, so that any user with permissions can access it. In addition, the transferred automatic subtitling technology has been adapted to Irekia's domain including new models trained on in-domain texts. Irekia is currently using the platform to generate automatic subtitles and publish them online with minor post-editing. Automatically subtitled contents include a message that warns and apologizes about the potential errors.

http://www.irekia.euskadi.eus/

5.4 Serikat Consultoría e Informática S.A.

Serikat is the company responsible for transcribing the Basque Parliament sessions. Serikat's professionals transcribe the sessions manually. Along each session, they need to keep transcribing 15 minute audio segments as they are produced and return the corresponding transcriptions within a 75 minute time frame. In order to meet the required deadline, Serikat's team transcribes shorter segments in parallel and then joins them together before delivery. Transcriptions have to comply with the specific format requested by the Basque Parliament.

With the aim of increasing productivity and competitiveness, Serikat is integrating Vicomtech-IK4's rich transcription technology in its workflow. A platform similar to that detailed in section 3.4.6 is being transferred, adapted to the domain and requirements of the Basque Parliament. Serikat's transcription team is starting to use the platform to create draft transcriptions for post-editing, instead of transcribing the parliament sessions from scratch.

http://www.serikat.es/

[2]http://www.inclusite.com/

6 Conclusions

This dissertation work describes how speech technologies can contribute to the audiovisual and multimedia interaction environments. As technology and computers improve and more data resources and tools are available to build robust statistical components, speech and natural language processing technologies are increasingly being integrated in several Industries and domains. In this sense, the new audiovisual law approved by the European Parliament and the Council in 2010 to guarantee the accessibility of audiovisual contents to all the people prompted the need to develop technology that can automatically generate accessible information in a more effective and productive way. This is also being extended to contents shared through other media channels such as the Internet and to contents linked to public services or related to the political and social life. Besides, it is worth mentioning that subtitles are not just useful for the deaf and hearing impaired but for all those people who are learning new languages or for anyone who cannot hear the audio due to adverse acoustic conditions.

Although an increasing amount of broadcast content is being subtitled, some broadcasters do not yet comply with the European audiovisual law. Producing subtitles of pre-recorded and live contents is expensive and hardly achievable in some cases. Live subtitling requires professionals to type subtitles in extreme time constraints or to re-speak a summary version of the spoken content to a LVCSR engine, which involves a high cognitive effort. The amount of professionals qualified for the live subtitling task is limited and, thus, the related economical rates are high. On the other hand, subtitling pre-recorded contents takes no less than 6-8 hours per hour, depending on the content and the skills of the subtitler. These issues along with the good performance achieved by current LVCSR technology in bounded domains, have positioned automatic subtitling as one of the most promising solutions for the present and the future.

The speech technology employed for automatic subtitling can also be used for automatic rich transcription. There is an emerging demand from parliaments and political institutions, whose sessions need to be transcribed. In these domains, technology helps reduce transcription time and make the task more pleasant, as only automatic transcription errors need to be revised and post-edited.

Other markets such as the e-learning, healthcare, call center, defense or enterprise markets are also critical for the evolution of speech technologies and their integration in Industry and people's common lives. In fact, according to BCC research[1], the speech recognition market will continue to grow from a size of USD 47 Billion in 2011 to USD 113 Billion in 2017, which means a CAGR[2] of 16.2%.

Future work on the speech technologies applied to the audiovisual environment will most probably focus on further improving automatic subtitling and rich transcription systems, as they are two of the most demanded technological resources needed to generate accessible contents and extract their semantic information in a rapid and economic way. These improvements will be centered on enhancing current performance and also on extending them to more languages. The basic resource required to develop robust automatic subtitling and rich transcription systems is annotated corpus. And the size of such corpus has to be large, both at acoustic and textual level. As an example, the SAVAS corpus consisted of 200 hours and texts containing 1,000 million words per language. A look at the corpus resources distributed by the Linguistic Data Consortium (LDC)[3] gives references mainly for languages such as English, Arabic and Mandarin Chinese, in addition to a few others such as Czech, Korean, Turkish and Spanish in the broadcast domain. Such scarcity of available data resources definitely affects the development of technology for more languages in this domain.

Speech technology will also evolve in constructing more robust automatic subtitling and rich transcription solutions. From the speech recognition point of view, DNN-HMM acoustic models have outperformed the widely used GMM-HMM based acoustic approach and have enabled the construction of LVCSR engines that work robustly in speaker-independent and acoustically different conditions. However, adverse and noisy environments still pose a great challenge. And the recognition of spontaneous speech also needs to be improved, especially at language model level. N-gram language models usually require the application of smoothing techniques such as Kneser-Ney [KN95] when big vocabularies and high N orders are involved. New language modelling paradigms based on Recurrent Neural Networks (RNN) have recently emerged to overcome the limits of the traditional N-gram models [Joz+16], such as the use of a short history of previous words to predict the next word or their limitations to model long context dependencies. Nevertheless, the research community needs to continue exploring new paradigms and techniques to overcome these limitations in the coming years.

[1] http://www.bccresearch.com/
[2] Compound Annual Growth Rate (CAGR) is the mean annual growth rate of an investment over a specified period of time longer than one year.
[3] https://catalog.ldc.upenn.edu/

The other components of the automatic subtitling and rich transcription systems also need to be improved, with the aim of reaching human quality performance. Taking current results into account, those components that involve a greater technological challenge are automatic punctuation and speaker diarization/identification. And in particular, comma prediction and dealing with acoustically adverse conditions.

Overall, automatic subtitling and rich transcription technology is expected to increasingly evolve towards domain-, acoustic- and speech type-independent modelling and to reach the same performance levels in batch and live operation modes.

Speech emotion recognition and synthesis (e.g. facial animation) is not as mature as automatic subtitling and rich transcription. As in real life, the interpretation of paralinguistic information is more subjective and its interest has emerged more recently. Although the roots of emotional intelligence can be traced back to Darwin's early work on the importance of emotional expression for survival and adaptation in the 19th century, the term only appeared in 1964 in [Dav+64] and did not gain popularity until 1995 with Goleman's famous book [Gol95]. Emotions have also gained relevance in Human-Computer Interaction applications and are considered essential to enhance communications between computers and humans and to endow machines the ability to understand the emotional state of the user at all times. As for the market, emotion recognition and synthesis is considered an emerging field which is expected to grow from USD 5.66 Billion in 2015 to USD 22.65 Billion in 2020, with a CAGR of 31.9%. Apart from the HCI environment, several other rising sectors such as the marketing, advertising and security sectors and vertical markets such as the banking, defense or the commercial security markets among many others, have contributed to the emergence and growing interest in this technology.

Future directions in the SER field shall center on agreeing the focus of the problem and also on obtaining the definitive set of speech features which more accurately describe the human emotional state. In the same way, new classification paradigms and modelling techniques should be tested, with the aim of evolving the SER technology to make it capable of working efficiently in speaker-independent environments. In this line, pioneer studies that model and use phonetic and speaker variability information have emerged recently. All these possible improvements would bring the SER technology closer to real-life speech, in which more challenging issues such as spontaneous speech, short speech segments, adverse acoustic conditions, blended emotions or the mix of several emotions in the same spoken utterance occur. The lack of naturalistic databases for different languages is a problem that also needs to be tackled in order to advance research in this field. Existing databases of this kind are the Belfast naturalistic emotion database [Dou+03] for English, the Vera am Mittag audio-visual emotional speech database [GKN] for German and the FAU Aibo Emotion Corpus [BSN08] for German also. The sophistication of SER will allow its

integration in a wider set of HCI applications, with the aim of improving the *user experience* and understanding the user needs in a more natural way.

Finally, main future work in Speech-driven Facial Animation will be centered in dealing with the constrains posed by a real-time scenario, naturalizing the lip-sync and co-articulation effects and generating natural and expressive facial motions linked to the paralinguistic information carried in human speech.

The four fields addressed in this work will continue to evolve and improve, thanks to several emerging markets that are betting on these technologies with the aim of enhancing their products, being more competitive and responding to the needs of the new social and digital life.

7 Publications

7.1 APyCA: Towards the automatic subtitling of television content in Spanish

- **Authors:** Aitor Álvarez, Arantza del Pozo, and Andoni Arruti
- **Booktitle:** Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)
- **Year:** 2010
- **Publisher:** IEEE

Proceedings of the International Multiconference on Computer Science and Information Technology pp. 567–574
ISBN 978-83-60810-27-9
ISSN 1896-7094

APyCA: Towards the Automatic Subtitling of Television Content in Spanish

Aitor Álvarez, Arantza del Pozo
Vicomtech Research Centre
Mikeletegi pasealekua, 57
Miramon Teknologia Parkea
20009 Donostia-San Sebastian, Spain
Email: {aalvarez, adelpozo}@vicomtech.org

Andoni Arruti
The University of the Basque Country
Dept. of Computer Architecture and Technology
Manuel de Lardizabal Pasealekua 1
20018 Donostia-San Sebastian, Spain
Email: andoni.arruti@ehu.es

***Abstract*—Automatic subtitling of television content has become an approachable challenge due to the advancement of the technology involved. In addition, it has also become a priority need for many Spanish TV broadcasters, who will have to broadcast up to 90% of subtitled content by 2013 to comply with recently approved national audiovisual policies. APyCA, the prototype system described in this paper, has been developed in an attempt to automate the process of subtitling television content in Spanish through the application of state-of-the-art speech and language technologies. Voice activity detection, automatic speech recognition and alignment, discourse segment detection and speaker diarization have proved to be useful to generate time-coded colour-assigned draft transcriptions for post-editing. The productive benefit of the followed approach heavily depends on the performance of the speech recognition module, which achieves reasonable results on clean read speech but degrades as this becomes more noisy and/or spontaneous.**

I. Introduction

SUBTITLING plays an important role in the increasingly multimedia and globalised world we live in. Its usefulness extends from the enrichment of TV content – in order to make it more accessible for people with hearing difficulties or to facilitate audiovisual information retrieval – to its application in noisy environments such as airports and transit stations, where it is not possible to hear TV broadcasts. In addition, subtitling has also become a priority need for many Spanish TV broadcasters, who will have to broadcast up to 90% of subtitled content by 2013 to comply with recently approved national audiovisual policies[1].

However, subtitling is a labor-intensive and economically costly process. As a general rule, manual production of high-quality subtitles can be assumed to take between 8 and 10 times the length of the video material [1]. Nevertheless, mainly due to the higher demands, the time allotted to production of the subtitled material has decreased in recent years [1], [2].

Experienced professionals currently employ dedicated subtitling software tools to help them generate subtitles faster. However, these tools simply display the subtitles on the computer screen as they will appear on the television or movie screen and facilitate purely mechanical functions, such as cueing the subtitles, spell-checking and other basic text processing functions [3]. Only recently speaker-dependent automatic speech recognition has become popular for live subtitling through re-speaking, a technique in which a professional subtitler is trained to dictate live subtitles as the programme happens. Products such as Protitle Live® (NINSIGHT)[2] and WinCAPS® (Sysmedia)[3] allow trained speakers to dictate live subtitles into trained ASR engines. Nevertheless, there is still no ASR-based system in use for fully automated subtitling.

The application of the following state-of-the-art technologies can also contribute to making the subtitling process more automatic and productive:

A. Voice Activity Detection (VAD)

TV content presents a wide range of acoustic conditions: e.g. music, clean speech, outdoor speech, speech with background music, sound effects, noise, etc. However, only those segments that contain speech are to be subtitled. In addition, the different acoustic conditions might require different kinds of processing.

VAD technology can be used to automatically detect the audio segments containing speech. VAD segmentations can also be used to automatically classify and group audio segments with similar acoustic characteristics for further processing.

B. Automatic Speech Recognition (ASR) and alignment

ASR can be employed to obtain automatic transcriptions of the spoken information. Even though ASR can potentially save a lot of time, it is a difficult task mainly due to the high variability of the spoken environments, speakers and speech types present in TV content. Spoken environments vary from clean (studio recordings) to noisy (outdoor recordings, speech mixed with background music or sound effects). The type of speech may differ from dictation (newsreader) to spontaneous (debate or interview). The combination of these

[1] S. Government, "Spanish Audiovisual Law on Subtitles. http://www.cesya.es/es/normativa/legislacion/Financiacion_Radio_TV," 2008.

[2] http://www.ninsight.fr/FR/

[3] http://www.sysmedia.com/

possibilities seriously challenges ASR technology, which also needs to deal with speaker independence and the uncontrolled vocabulary of TV programs.

The time-stamps output by the ASR system can also be employed to align the recognised transcripts to the audio signals. In cases where the transcripts already exist, forced alignment can be used instead of recognition to obtain more accurate synchronizations between audio and text.

C. Discourse segment detection (DSD)

The detection of entities, relationships or individual events of speech and its segmentation into sentences and phrases is a crucial step for the transition from speech recognition to its full understanding. Unless explicitly dictated, speech recognisers output strings of words without a right segmentation of the output into discursive segments. As a result, ASR transcriptions consist of raw text that is quite difficult to understand for the reader.

DSD techniques can be used to automatically segment ASR transcriptions into segments which contain whole meaning, in order to make them more readable.

D. Speaker diarization (SD)

SD is the task of segmenting a multi-speaker audio signal into homogeneous parts and clustering them into different groups, each containing the voice of a single speaker.

In the context of subtitling, SD can be employed to automatically assign a specific color to the subtitles spoken by each speaker.

APyCA, the prototype system described in this paper, integrates the four technologies described above in a unique application, whose aim is to facilitate the manual production of subtitles by experienced professionals, reducing as a result the high cost of subtitle production.

The paper is structured as follows. Section 2 describes the state-of-the-art of the technologies involved and Section 3 presents the resources and tools developed and integrated within the project. Section 4 then describes the implemented prototype. Evaluation of the different modules is presented in Section 5 and finally, Section 6 discusses the main conclusions and further work.

II. State of the art

Much work has been made on the four main technologies involved in APyCA: Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Discourse Segments Detection (DSD) and Speaker Diarization (SD).

A. Voice Activity Detection (VAD)

With increasing demand for voice interfaces, the ability to distinguish human speech from other sounds is becoming crucial. Many works have attempted to discover characteristic features of human voices that are present only in speech. Since such characteristic features have not yet been discovered, short-time energy, zero crossing rate (ZCR), low-variance spectrum (LVS), spectral entropy (SE), periodicity, and so on have been used instead [4], [5]. While it is true that speech has such characteristics, the problem is that they can also be present in some non-speech sounds. This leads to a high false acceptance rate for specific kinds of noise. For example, loud white noise can also have high energy and ZCR.

For these reasons, statistical pattern classification approaches such as Gaussian Mixture Models (GMMs) have gained wider acceptance [6], [7]. In statistical VAD methods, both speech and noise models are trained via corresponding training data. Then, log likelihood ratio tests are applied to input data for speech and noise discrimination. These VAD methods have been shown to exhibit superior performance than the previous approach.

B. Automatic Speech Recognition (ASR)

There have been several projects focused on the development of ASR technology for the automatic transcription of Broadcast News (BN). However, most of them were developed for languages other than Spanish, such as English [8], French [9], Portuguese [10] or German [11]. As a result, there is not much data available in Spanish to train a robust speech recogniser for the automatic transcription of broadcast content. Several studies [9], [12] state that at least 100 hours of annotated and transcribed data is required for the adequate training of BN ASR engines and practical development works tend to use as much data as possible. For example, [13] uses up to 1000 hours of training speech data for Persian while [14] employs 81 hours for English, 52 for Portuguese and, in comparison, only 15 for Spanish. This lack of data is the main reason why we decided to use a commercial ASR engine within the APyCA prototype, and to explore adaptation of its default models to improve performance.

Despite the improvement of automatic speech recognisers, developing a system for the automatic transcription of content broadcasted in radio or television is still a challenge for many research groups. A system aimed at the automatic transcription of Portuguese BN, working in a real application scenario currently is [10]. It is based on a hybrid acoustic modelling approach that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multilayer Perceptrons (MLPs). Such acoustic modelling combines phoneme probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. The feature extraction methods are PLP, Log-RASTA and MSG. The training of the language model is done using both, Portuguese newspaper texts combined with the transcriptions used for acoustic model training.

With regard to the performance of the ASR systems developed for the different languages on the broadcast domain, the resulting error rates reflect in general the varying level of the acoustic and linguistic complexity of the recordings [11]. WERs range from 16.1% to 64.5%. For Spanish, [14]

achieved a mean WER of 18.9%, while [12] managed to decrease it up to 10% by restricting the recognition domain.

The vast majority of the previous studies consider classifying, labeling and structuring the acoustic signal into homogeneous segments essential to optimise the training of the acoustic models of the recogniser, for its subsequent proper operation [15].

C. Discourse Segments Detection (DSD)

The most widely investigated two sources of information to resolve the problem of detecting discursive segment boundaries are word transcriptions (what the speakers say) and prosody (how they say it). It is common to use two statistical models: language and prosodic models. In general, the language model gives the probability of a segmental boundary occurring in a context, while the prosodic model expresses the relationship between prosodic features and segmental boundaries.

Most previous works on discourse segment detection, e.g. [16], are based on the combination of these two information sources. [17] presents a system for punctuation generation which combines both prosodic and linguistic information, in addition to acoustic models. [18] and [19] use a general HMM framework that allows the combination of lexical and prosodic information to recover punctuation marks. A similar approach was used to detect sentence boundaries in [20] and [21].

D. Speaker Diarization (SD)

The varied and wide applicability of speaker diarization technology has led different research groups to develop several systems. The SD process often consists of three main phases: front-end acoustic processing, initial segmentation and final speaker clustering and refinement. The pre-processing step has two main goals. The first one is to normalise the signal in order to remove corrupting noise. In [22], for example, Wiener filtering is applied on each audio channel with that purpose. The second one is to parameterise the signal. Mel Frequency Cepstum Coefficients (MFCC) and Linear Frequency Cepstrum Coefficients (LFCC) are commonly used parameter features, in vectors of several dimensions which often include deltas and/or deltas-deltas.

The initial segmentation phase aims to provide an approximate speaker turn labeling to initialise and speed-up the subsequent segmentation and clustering stages. Several distance criterions can be used in this step. While [23] applies a classical GLR speaker turn detection criteria, [24] uses a segmentation similar to the KL2 metric, measuring the maxima of a local Gaussian divergence between two adjacent sliding windows of five seconds.

The most common clustering method employed in the speaker clustering and refinement phase is the Bayesian Information Criteria (BIC) or a variation called ΔBIC [22, 24]. Initial clusters are generally modelled by single Gaussians with full covariance matrices estimated on the acoustic frames of each segment output by the initial segmentation step. The BIC or ΔBIC metrics are commonly used both, to measure inter-cluster distances and as stop criterions.

III. Resources, tools and applications developed

The main tools integrated in APyCA are: (1) a VAD module; (2) a large vocabulary continuous speech recognition module for recognition and alignment and modules for (3) the detection of discursive segment boundaries and (4) speaker diarization.

A. Voice Activity Detection (VAD)

In order to feed the speech recogniser with audio segments containing speech, a previous segmentation and classification of the audio signal is required. This classification should be as comprehensive as possible, to ensure that no misclassified speech segments are lost.

APyCA segments and classifies the input audio into four different acoustic types: speech, speech plus noise, noise and silence, based on the speech detection functionality of the open source LIUM_SpkDiarization tool [25]. Such segmentation is obtained through Viterbi decoding of one-state Hidden Markov Models (HMMs) trained for the different acoustic conditions on the ESTER broadcast news corpus.

B. Automatic Speech Recognition (ASR) and alignment

APyCA employs the Windows Speech Recogniser (WSR) 8.0 as its ASR engine, integrated through the SAPI 5.3 functionality on the .NET Framework 3.5 environment.

In order to improve its performance, default models have been adapted with acoustically similar (i.e. clean speech vs. noisy speech) and/or TV genre-specific data by feeding the system with the corresponding audio recordings and text transcripts.

As well as for generating textual transcriptions of the spoken information, the ASR module is also used to obtain word-level time-stamps to align the audio and the text. In those cases where transcriptions already exist and the recognition step is not required, audio and text synchronization is computed by an alignment module developed using the HTK Toolkit [26]. The alignment module is a monophone recogniser trained on the Albayzin corpus [27], which extracts 39-dimensional feature vectors containing MFCC, delta and delta-delta coefficients on 25ms windows every 10ms and uses the Spanish version of SAMPA as its phoneme set, plus silence and short pause models. Each monophone (except from the short pause model) consists of non-emitting start and end states plus three emitting states, connected left-to-right with no skips and modelled by a single Gaussian. Viterbi is used for decoding.

C. Discourse Segment Detection (DSD)

Any subtitling platform integrating a speech recognition engine requires the development of algorithms for the automatic segmentation of the recognised output into dircursive segments.

APyCA has four different ways to automatically predict discourse segment boundaries: two of them are related to the acoustic and prosodic processing of the speech signal, another one is based on the linguistic analysis of the transcribed text and the last one combines the previous three approaches. The different techniques employed are presented in more detail in the following sections.

1) DSD based on Acoustic Information

Acoustic pauses are detected by analysing word start and end time-stamps produced by the recogniser or the alignment module during the recognition and alignment processes respectively. Whatever the difference, any non-coincidence in time between the end of a word and the start of the next has been taken as a potential acoustic pause.

It is important to emphasise at this point that even if acoustic pauses do not always correspond to discursive breaks, their relationship is evident in many cases.

2) DSD based on Prosodic Information

Acoustic pauses are not always grammatically correct as they may coincide with breathings, stops or speech difluencies that are not always related to true discursive boundaries. Discourse segment detection based on prosodic information can help resolve this problem.

The implemented algorithm detects discursive segment boundaries based on CART classifiers trained with the Waikato Environment for Knowledge Analysis (WEKA) tool [28]. Three different classes have been used: "silence", "question" and "nothing" - corresponding to the cases where a word is followed by silence, question mark or nothing, respectively. Each class is trained on prosodic features extracted for each word of the Multext Prosody corpus [29] using the Purdue Prosodic Feature Extraction Tool (PPFE) [30]. 232 prosodic features are extracted around each word. These features are mainly related to:

•***Duration:*** the duration and normalised duration of each word and word boundary are extracted. In addition, the duration and normalised duration of the last vowel and rhyme before a word boundary are also measured.

•***Pitch:*** several different types of F0 features are computed, based on the stylized pitch contour.

◦ *Range features:* these include the minimum, maximum, mean, and last F0 values of each word and reflect its pitch range.

◦ *Movement features:* measure the movement of the F0 contour within the voiced regions of the words preceding and following a boundary. The minimum, maximum, mean, first and last stylized F0 values of each word are computed and compared to those of the following word, using log differences and ratios.

◦ *Slope features:* the last slope value of a word preceding a boundary and the first slope value of a word following a boundary are also calculated.

•***Energy:*** similar to the F0 features, a variety of energy related range features, movement features, and slope features are computed, using various normalization methods.

Each word in the training corpus was manually labeled to belong to one of the three classes defined above.

3) DSD based on Linguistic Information

The linguistic algorithm has the same purpose as the prosodic and acoustic algorithms, i.e. estimating discourse segmentations of the transcribed text. The philosophy used to develop this module has been based on two types of heuristics: grammatical and structural.

On the one hand, a probabilistic part-of-speech (PoS) tagger based on Hidden Markov Models (HMM) has been developed in order to grammatically categorise each word. It has been trained on a proprietary lexical database that includes thousands of grammatical categories of words. In addition, heuristic rules have been developed to detect combinations of grammatical categories, before or after which it is more likely to have segment discourse boundaries.

On the other hand, the most frequent and meaningful structural elements present in the Multext Prosody corpus before or after which it is highly likely to have a discourse segment boundary have been identified. Based on this information, heuristic rules to detect discursive boundaries have been designed.

The latter approach has been found to be more robust than the former one in the automatic subtitling scenario, since recognition errors can lead to grammatical miscategorisations which weaken the designed grammatical heuristic rules.

4) DSD based on Combined Information

The global APyCA DSD system is modular in nature. This means that the input text can be independently segmented into discourse segments using any of the modules designed: acoustic, prosodic and/or linguistic.

A combined model has also been developed, which takes the predictions and confidence measures provided by the three modules described above and gives a final result based on their weighted combination. In general, if two modules detect a pause in a word boundary with enough confidence, we take it as a real pause. This approach exploits the complementarity of the three very different sources of information used for the detection of discourse segments.

D. Speaker Diarization (SD)

This module aims at segmenting the acoustic signal according to the speaker identities, so that each speaker can be assigned a different subtitle colour.

APyCA uses the LIUM_SpkDiarization open source tool [25] to solve the task of speaker diarization. Signals are parameterised using 13 Mel Frequency Cepstral Coefficients (MFCC) including coefficient C0 as energy, computed with the Sphinx 4 tools [31]. 20 ms windows are employed with an overlap of 10 ms. Cepstral Mean Normalisation (CMN) is

not applied, due to its tendency to increase the error rate of the diarization task.

The diarization process consists of three main phases. Instantatenous signal change points corresponding to segment boundaries are detected first, using distance-based segmentation metrics which combine the Generalised Likelihood Ratio (GLR) and Bayesian Information Criterion (BIC). GLR is computed using full covariance Gaussians estimated on sliding windows of five seconds and followed by a second ΔBIC pass, which also uses full covariance Gaussians, to fuse consecutive segments of the same speaker.

Then, nonadjacent segments of the same nature and speaker are brought together in clusters using a hierarchical agglomerative clustering algorithm with a stopping criterion based on the ΔBIC metric.

Finally, Viterbi decoding is performed to generate improved segmentations. Each cluster is modeled by a one-state HMM, represented by a GMM with 8 components and a diagonal covariance matrix learned by Expectation-Maximization Maximum-Likelihood (EM-ML) over the segments of the cluster. The log-penalty between two HMMs is fixed experimentally.

IV. Integration of Components into a Demo Application. Description of the Prototype

APyCA is a prototype oriented towards the automatic transcription of TV content in Spanish which integrates the technologies described in the previous sections of the paper and aims to serve the professional subtitlers as a tool to facilitate the creation and editing of subtitles. The following sections describe the main features and architecture of the developed demo application.

A. Features

Its input is TV content in the form of video or audio. It supports many different formats, including the main standards used by television producers, such as *mpeg2, h.264, aac or wav.*

Its output is a well-formed STL (binary) or SRT subtitle file, which respects the maximum number of characters allowed per line and includes colours to differentiate speakers. Time-spotting is based on the estimated word time-stamps and discourse segment boundaries. If needed, these subtitle files can be easily edited further using commercial software for subtitle generation, e.g. WinCAPS, FAB Teletext and Subtitling, Subtitle Workshop, etc.

The technologies involved have been grouped into three automatic functionalities: transcription, time-spotting and

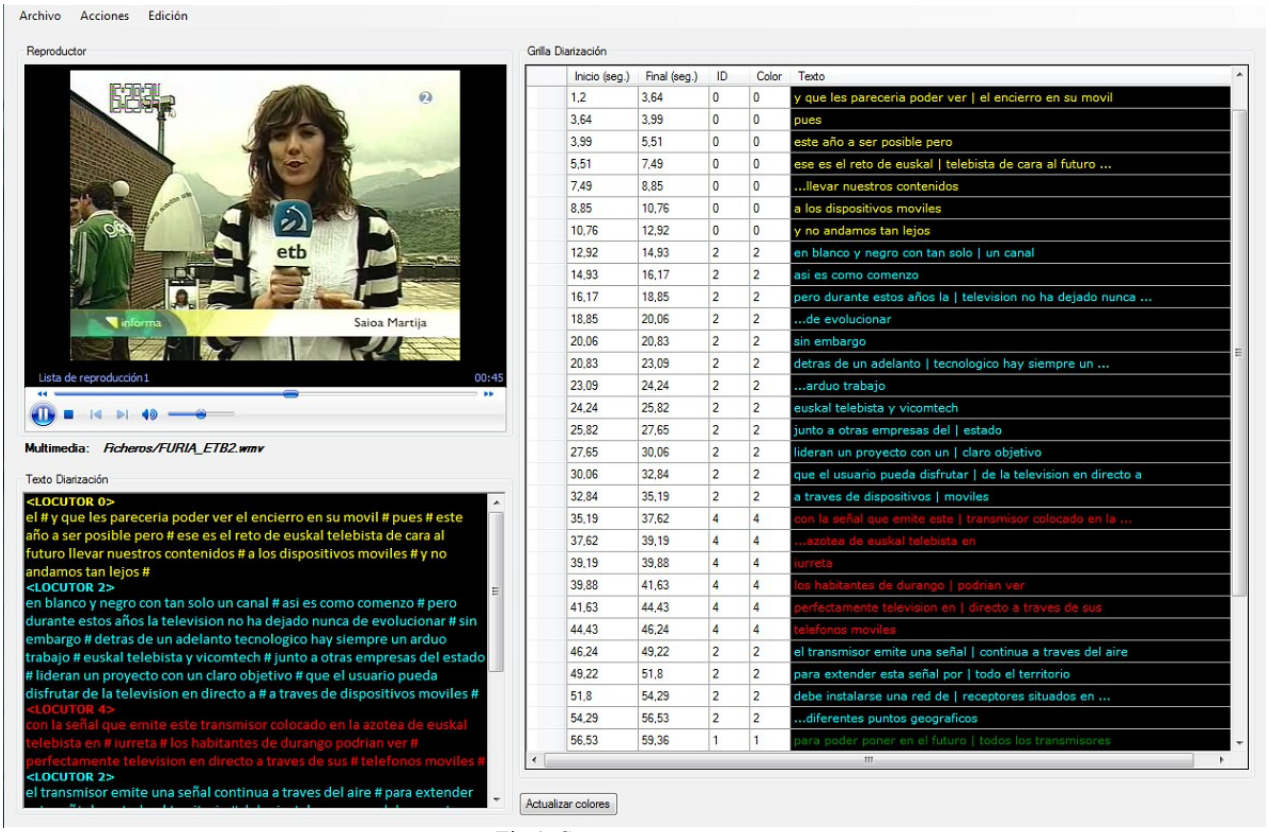


Fig 1: System screen capture

speaker diarization – which can be applied and edited independently through dedicated graphical user interfaces.

•The **Automatic Transcription** screen shows the raw text returned by the ASR engine for those segments that contain speech. It also allows playback and editing of the transcriptions, so that subtitlers can manually correct the errors of the recogniser. The fewer the transcription mistakes, the better the time-spotting will be.

•The **Automatic Time-Spotting** functionality chunks the transcribed text into discoursive segments and aligns them with the audio. The start times, end times and text of each subtitle can be edited manually.

•The **Speaker Diarization** screen automatically assigns different colours to the subtitles spoken by different speakers, also allowing their manual edition.

These three functionalities can be combined to suit the needs of the professional subtitlers. It is possible, for example, to skip the automatic transcription step and upload already transcribed audiovisual content. Or to generate the subtitle files without speaker diarization information.

The prototype has been developed entirely using the Microsoft .NET platform, the C# programming language and several Perl scripts for text processing.

A screen capture of the system is shown in Fig. 1.

B. *Architecture*

Fig. 2 illustrates how the different modules interact within the system and with the user.

The system supports the input of TV content in video or audio formats, as well as with or without its corresponding textual transcription. The FFmpeg [32] tool is used to extract the audio from the video in different formats and configurations. If the transcription does not exist, the audio will be recognised. If the transcription exists, forced alignment will be applied instead to obtain word time-stamps. In any case, time-stamps are required for the discourse segment detection and speaker diarization modules. Output subtitle files can be generated after recognition/alignment, after discourse segment detection or after speaker diarization.

The modular architecture of the system will allow simple integration of additional modules providing new functionalities in the future.

V. Evaluation

A. *Data*

The main modules of the APyCA prototype have been evaluated individually on two different genres of TV content: weather forecasts and political interviews.

Weather forecasts do not present difficulties related to the the spoken environment, the type of speech used, the number of speakers involved or the quantity and quality of their interventions. In fact, they contain just one anchor presenter following a previously written and rehearsed script in a noise-free recording studio. However, the employed vocabulary is very specific of the meteorological domain. Political interviews involve many different types of spoken environments (studio, parliament, street), types of speakers and speech (presenters following pre-prepared scripts and/or spontaneous interviewees) and a very domain specific vocabulary, with many mentions to names of politicians.

Ten programs of each type were recorded and used for training (8) and testing (2) the different modules of the system. Their reference transcripts were obtained by manually correcting the transcriptions output by the WSR 8.0 with default models.

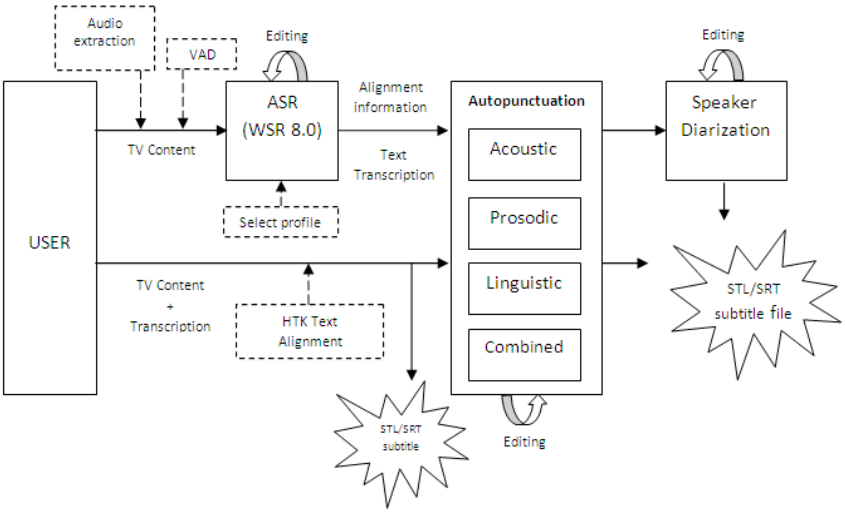


Fig 2: System architecture

B. Automatic Speech Recognition (ASR)

The performance of the WSR 8.0 recogniser was tested in three different conditions: (i) using the default recogniser models, (ii) using clean and noisy speech profiles adapted to the clean and noisy acoustic conditions found in each corpus, and (iii) using TV genre-specific profiles trained for the weather forecast and political interview domains.

Results for *the weather forecast* corpus are shown in Table 1. The average percentage of words correctly recognised overall is especially promising. The column labeled *Baseline* shows the performance of the default profile of the commercial WSR 8.0 engine. The column labeled *TV-genre profile* corresponds to the recognition rate achieved using a profile trained with all the training content of the weather forecast corpus. The *Acoustic profiles* column shows the results obtained after applying the recognition profiles trained with clean and noisy speech. Contrary to expectations, acoustic profiling does not achieve the best results probably due to the loss of context caused by the more detailed audio segmentation involved.

TABLE I.
AVERAGE RECOGNITION RATE IN THE WEATHER FORECAST CORPUS

Baseline	*TV-genre profile*	*Acoustic profiles*
81.3 %	96.65 %	92.34 %

Results for the *political interview* corpus are shown in Table 2. Less satisfactory recognition rates were obtained with this corpus overall, due to the inherent difficulty of the content type. It is remarkable that the application of acoustic profiling did not improve the results obtained by the default recognition profiles. On the other hand, TV-genre profiling only improves baseline results slightly.

TABLE II.
AVERAGE RECOGNITION RATE IN THE POLITICAL INTERVIEW CORPUS

Baseline	*TV-genre profile*	*Acoustic profiles*
79.54 %	79.80 %	78.60 %

C. Discourse segment detection (DSD)

Results concerning the evaluation of the different DSD modules are shown in Tables III, IV and V.

Each module was evaluated individually, against manually labeled reference test files. Acoustic labels take into account acoustic silences and short pauses. Prosodic labels are based on the intonation of the related sound files. Linguistic labels consider syntactic information of the associated text.

According to the followed evaluation methodology, "*Matching breaks*" refers to the percentage of breaks that match the reference file, while "*Unassigned breaks*" relates to the percentage of breaks present in the reference labels which have not been assigned by the different modules. The percentage of extra breaks assigned by the DSD modules that do not appear in the reference files is counted as "*Extra breaks*".

TABLE III.
RESULTS OF THE ACOUSTIC MODULE

Matching breaks	*Unassigned breaks*	*Extra breaks*
92.67 %	7.33 %	16.60 %

TABLE IV.
RESULTS OF THE PROSODIC MODULE

Matching breaks	*Unassigned breaks*	*Extra breaks*
64.49 %	35.51 %	63.64 %

TABLE V.
RESULTS OF THE LINGUISTIC MODULE

Matching breaks	*Unassigned breaks*	*Extra breaks*
51.92 %	48.08%	1.44 %

Results show that in 92.67% of the cases, acoustic segmental boundaries were assigned correctly, 7.33% of the acoustic pauses were not detected and 16.80% were wrongly assigned, particularly those matching breathing stops and speech disfluencies. Spontaneous speech was the main enemy of the prosodic module, mainly trained under a database of read speech. Nevertheless, it achieved a non negligible 64.49% accuracy rate. As for the linguistic module, its performance was penalised by the recognition errors which affect the designed heuristic rules. Overall, the acoustic module has proved to be the most efficient to detect discourse segment boundaries, due to its high speed and hit rate.

D. Speaker Diarization (SD)

The speaker diarization module achieved very good performance. Even in the rich acoustic environment of the political interview corpus, results achieved 87% success rate. Errors were mainly due to background acoustic changes, which caused the same speaker to be classified as two in some cases where the background acoustic environment was different, since the BIC criterion employed in APyCA for speaker diarization was actually designed to classify those segments as different.

VI. CONCLUSIONS AND FURTHER WORK

Voice activity detection, automatic speech recognition and alignment, discourse segments detection and speaker diarization technologies have been developed, customized and integrated in a prototype to support the subtitle generation process of Spanish TV content.

Objective evalutation of the different modules has shown that the proposed approach is feasible and applicable to generate automatically time-coded and colour-assigned draft transcriptions for post-editing. The commercial WSR 8.0 engine has shown adequate performance for the task. Adaptation of the default profiles to each TV-genre has shown to improve recognition accuracy. However, transcription performance degrades overall as the input speech becomes more noisy and/or spontaneous. Acoustic discourse segment detection has been found to be very efficient in terms of high speed and hit rate for time-spotting. The LIUM_SpkDiariza-

tion tool has also shown good results in the colour assignment task.

However, there is still quite a lot of room for improvement. Techniques to enhance ASR accuracy in noisy and/or spontaneous environments could be integrated. The prosodic DSD module could be trained on spontaneous and/or emotional speech corpora to better match intonation patterns of certain TV content. Finally, feature normalization techniques could be added to the speaker diarization module to obtain one-to-one relationships between clusters and speakers.
Although some positive informal usability tests have been done with professional subtitlers, a more comprehensive assesment should be carried out in order to verify the feasibility and quantify the time and money savings which could be provided by a software tool similar to the developed prototype.

Acknowledgment

This work has been partially funded by the Basque Government. The authors would like to thank Mixer, Irusoin and ETB for providing the corpora and giving usability feedback.

References

[1] M. Flanagan, “Human Evaluation of Example-Based MT of subtitles for DVD,” Dublin City University, 2009.

[2] M. Carroll, “Subtitling: Changing standards for new media? LISA Newsletter Global Insider, XIII, 3.5. 2004. http://www.lisa.org/globalizationinsider/2004/09/subtitling_chan.htm,”

[3] L. Bowker, Computer-aided Translation Technology: A Practical Introduction, Ottawa: University of Ottawa Press, 2002.

[4] J.L. Shen, J.W. Hung, and L.S. Lee, “Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Environments”, Proc. Int. Conf. Spoken Language Process., paper 0232, 1998.

[5] I.D. Lee, H.P. Stern, S.A. Mahmoud, “A Voice Activity Detection Algorithm for Communication Systems with Dynamically Varying Background Acoustic Noises,” Proc. Veh. Technol. Conf., 1998.

[6] J. Sohn, N.S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection”, IEEE Signal Process. Lett., vol. 6, no. 1, pp. 1-3, 1999.

[7] A. Davis, S. Nordholm, R. Togneri, “Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold” , IEEE Trans. on Signal Proc., vol 14, no 2, pp. 412-424, 2006.

[8] J.S. Garofolo, J.G. Fiscus, W.M. Fisher, “Design and preparation of the 1996 hub-4 broadcast news benchmark test corpora,” in Proceedings of the DARPA Speech Recognition Workshop., pp. 15–21, 1997.

[9] S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, K. Choukri. "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News". In Proceedings of the 5th International Conference on Language Resources and Evaluation 2006.

[10] H. Meinedo, D. Caseiro, J. Neto, I. Trancoso. “AUDIMUS.MEDIA: a broadcast news speech recognition system for the European Portuguese language”. In Proceedings of PROPOR 2003, Portugal, 2003.

[11] D. Baum, B. Samlowski, T. Winkler, R. Bardeli, Schneider: “DiSCo - a speaker and speech recognition evaluation corpus for challenging problems in the broadcast domain”. Proceedings of the GSCL Symposium 'Sprachtechnologie und eHumanities' 2009.

[12] J. Loof, Ch. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, Ch. Plahl, D. Rybach R. Schluter and H. Ney. "The RWTH 2007 TC-STAR Evaluation System for European English and Spanish". Interspech 2007.

[13] C. Gollan, H. Ney, “Towards automatic learning in LVCSR: Rapid development of a Persian broadcast transcription system,” Interspeech' 08.

[14] F. Batista, I. Trancoso, N. J. Mamede. "Comparing Automatic Rich Transcription for Portuguese, Spanish and English Broadcast News". In Automatic Speech Recognition and Understanding Workshop, 2009.

[15] J.-L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. de Kercadio, “The Limsi SDR systemfor TREC-9,” in Proc. 9th Text Retrieval Conference, TREC-9, pp. 335–341, Gaithersburg, Md, USA, 2000.

[16] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, B. Peskin, and M. Harper. “The ICSI-SRI-UW Metadata Extraction System”. ICSLP 2004, International Conf. on Spoken Language Processing, Korea. 2004.

[17] J.H. Yim. “Named Entity Recognition from Speech and Its Use in the Generation of Enhanced Speech Recognition Output”. Darwin College, University of Cambridge and Cambridge University Engineering Department. 2001.

[18] H. Christensen, Y. Gotoh, and S. Renals, “Punctuation annotation using statistical prosody models,” in Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 35–40, 2001.

[19] J. Kim, P. C. Woodland, “The use of prosody in a combined system for punctuation generation and speech recognition,” Proc. Eurospeech' 01.

[20] Y. Gotoh and S. Renals, “Sentence boundary detection in broadcast speech transcripts,” in Proc. of the ISCA Workshop: ASR-2000.

[21] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, “Prosody based automatic segmentation of speech into sentences and topics,” Speech Communications, vol. 32, no. 1-2, pp. 127–154, 2000.

[22] T. L. Nwe, H. Sun, H. Li, S. Rahardja, "Speaker Diarization in Meeting Audio", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, April 19-24, 2009.

[23] J. Huang, E. Marcheret, K. Visewswariah, G. Potamianos, “The IBM RT07 Evaluation Systems for Speaker Diarization on Lecture Meetings”, in Multimodal Technologies for Perception of Humans, Springer, 2008.

[24] C.Wooters, M. Huijbregts. “The ICSI RT07s Speaker Diarization System”. In Rich Transcription 2007 Meeting Recognition Workshop.

[25] S. Meignier, T. Merlin. “LIUM_SpkDiarization: An Open Source Toolkit For Diarization”. CMU Sphinx Workshop 2010, Dallas, 2010.

[26] Hidden Markov Model Toolkit (HTK) 3.2, Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/, 2002.

[27] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio: “Development of Spanish Corpora for Speech Research (Albayzin)”. Workshop on International Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods, Italy, 199.1

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. “The WEKA Data Mining Software: An Update”; SIGKDD Explorations, Volume 11, Issue 1. 2009.

[29] E. Campione, (Ed.) Multext-Prosody. A multilingual prosodic database. CD-ROM Distributed by ELRA/ELDA. 1999.

[30] Z. Huang, L. Chen, M. Harper. “Purdue Prosodic Feature Extraction Toolkit on Praat”. Spoken Language Processing Lab, Purdue University. 2006.

[31] Sphinx-4. “A speech recognizer written entirely in the Java programming language”. http://cmusphinx.sourceforge.net/sphinx4/

[32] FFmpeg. “A complete, cross-platform solution to record, convert and stream audio and video”. http://www.ffmpeg.org/

7.2 Automating live and batch subtitling of multimedia contents for several European languages

- **Authors:** Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, Tiago Luís, Sérgio Paulo, Nicola Piccinini, Haritz Arzelus, João Neto, Carlo Aliprandi and Arantza del Pozo
- **Journal:** Multimedia Tools and Applications
- **Pages:** 1–31
- **Year:** 2015
- **Publisher:** Springer

Multimed Tools Appl
DOI 10.1007/s11042-015-2794-z

Automating live and batch subtitling of multimedia contents for several European languages

Aitor Álvarez[1] · Carlos Mendes[2] · Matteo Raffaelli[3] · Tiago Luís[2] · Sérgio Paulo[2] · Nicola Piccinini[3] · Haritz Arzelus[1] · João Neto[2] · Carlo Aliprandi[3] · Arantza del Pozo[1]

Received: 22 December 2014 / Revised: 22 June 2015 / Accepted: 29 June 2015

Abstract The subtitling demand of multimedia content has grown quickly over the last years, especially after the adoption of the new European audiovisual legislation, which forces to make multimedia content accessible to all. As a result, TV channels have been moved to produce subtitles for a high percentage of their broadcast content. Consequently, the market has been seeking subtitling alternatives more productive than the traditional manual process. The large effort dedicated by the research community to the development of Large Vocabulary Continuous Speech Recognition (LVCSR) over the last decade has resulted in significant improvements on multimedia transcription, becoming the most powerful technology for automatic intralingual subtitling. This article contains a detailed description of the live and batch automatic subtitling applications developed by the SAVAS consortium for several European languages based on proprietary LVCSR technology specifically tailored to the subtitling needs, together with results of their quality evaluation.

Keywords Multimedia communication · Multimedia systems · Automatic speech recognition · Automatic subtitling · Subtitling quality · Access services

✉ Aitor Álvarez
aalvarez@vicomtech.org

[1] Department of Human Speech and Language Technologies, Vicomtech-IK4 Foundation, San Sebastian-Donostia, Spain

[2] VoiceInteraction-Speech Processing Technologies, SA, Lisbon, Portugal

[3] Synthema-Language and Semantic Technologies, Pisa, Italy

Published online: 11 July 2015

1 Introduction

The subtitling demand of multimedia content has grown quickly over the last years, especially after the adoption of the new European audiovisual legislation (Article 7 of the Audiovisual Media Services Directive). This law regulates the right of persons with disabilities and elderly people to participate and be integrated in the social and cultural life of the Community, through accessible multimedia services including aspects such as sign-language, subtitling, audiodescription and easily understandable menu navigation.

As a result of this new legal framework, public and private TV channels have been moved to produce subtitles for a high percentage of their broadcast content. However, the subtitling process is traditionally based on the manual production of time-aligned transcriptions of audiovisual content, a task which requires considerable effort. Manual production of high-quality subtitles has been reported to take between 8 to 10 times the length of the multimedia material [14]. Hence, broadcasters and subtitling companies are looking for solutions that can help them cope with the increasing subtitling volumes and demand.

The large effort in research and development of Large Vocabulary Continuous Speech Recognition (LVCSR) over the last decade has resulted in significant improvements on multimedia data transcription, retrieval and indexation [16, 23, 43], making it the most powerful technology available to increase productivity in several automated intralingual subtitling tasks. In the last few years, respeaking - a technique in which a professional listens to the source audio and dictates it so that his/her voice input can be processed by a speech recognition engine which transcribes it, thus producing subtitles - has consolidated as the most widely adopted live subtitling technique. Another trend in use today is the application of LVCSR to automatically generate transcripts from the source audio as the basis for subtitles. The main advantage of this method compared to respeaking is that it can actually produce similar results in bounded domains without the need of a respeaker, which reduces costs.

In order to comply with the new legal requirements, broadcasters started focusing their increased subtitling effort on quantity, considering quality a secondary issue. However, an increasing demand to improve the quality of automatic subtitles has arised recently. The quality of subtitles involves several features linked to subtitle layout, duration and text editing. Layout parameters include: the position of subtitles on screen; the number of lines and the amount of characters contained in each line; the typeface, distribution and alignment of the text; the front and background colors; speaker colors; and transmission modes, i.e. blocks or scrolling/word-by-word. Duration features involve delay in live subtitling and the persistence of subtitles on screen. Finally, text editing parameters are related to capitalization and punctuation issues, segmentation or the use of acronyms, apostrophes and numerals.

This article contains a detailed description of the automatic live and batch subtitling applications developed by the SAVAS consortium[1] for several European languages based on proprietary LVCSR technology tailored to the specific needs of the subtitling industry, together with results of their quality evaluation. Applications have been developed for Portuguese, Spanish, Basque, Italian, French, German and the Swiss variants of the latter three, and trained and tested on several domains such as broadcast news, sports, interviews and debates. Their performance has been evaluated against a variety of metrics, including

[1] http://fp7-savas.eu

both standard LVCSR and subtitle quality metrics. In Section 2, an overview of the state-of-the-art of LVCSR and the existing assisted subtitling applications is presented. Section 3 details the SAVAS technology and the developed live and batch subtitling applications. The methodology followed to evaluate them is then described in Section 4. Finally, Section 5 presents the evaluation results and the main conclusions are summarized in Section 6.

2 Related work in automatic speech recognition and assisted subtitling applications

LVCSR technology is employed to transcribe speech into text for further linguistic processing. Despite progress in the last decade, LVCSR is still highly task- and domain-dependent due to its statistical nature. In terms of accuracy, LVCSR system performance varies with the task [3]: clean read speech transcription achieves better performance than TV and radio news broadcasts, telephone conversations, lectures or plenary sessions of the European Parliament. Although comparable performance has been achieved for several languages, English is still the most developed one today.

LVCSR technology has been exploited commercially, mainly for dictation and command-based interaction applications in specific domains, like the HealthCare or the Parliamentary domain [20, 30, 32]. The main LVCSR engines of this type available are IBM ViaVoice [20], now discontinued from the market, Microsoft Windows Speech Recognition [30] and Nuance Dragon Naturally Speaking [32]. Many subtitling tools currently employed by the industry (e.g. WINCAPS Q-Live [35], WINCAPS Qu4ntum [36], FAB Subtitler Live Edition [12], Grass Valley captioning and subtitling solution [19], Starfish Isis [37]) support the Nuance Dragon Naturally Speaking dictation engine for respeaking purposes. However, there are no tools on the market that allow generating automatic intralingual subtitles from the source audio without respeaking.

This has been limited by the unsuitability of the available dictation engines for audio transcription [33] and by the absence of more sophisticated LVCSR technology for transcription of multimedia contents. Dictation engines have several limitations. First, they are speaker dependent; that is, they have to be adapted to each user. Second, they do not perform well on multimedia material containing complex acoustic conditions (e.g. background music or noise) or spontaneous speech, because they have been designed for dictation purposes. Finally, they have only been developed for languages with a high market potential (e.g. English and Spanish) and are not available for many other languages. On the other hand, training high-quality LVCSR transcription engines requires huge amounts of audio and text per language and specific domain. Several studies (see [15] and [24]) state that at least 100 hours of annotated and transcribed audio are necessary to adequately train the acoustic models of such kind of LVCSR engines. Regarding language modelling, [29] have estimated that ideally one billion words of texts are required.

Recently, few internet services offering the automatic generation of time-aligned subtitles from a source audio and its transcript have arisen. Ubertitles [39], eCaption [11] or SyncWords [38] offer such kind of service in different languages, based on proprietary audio and text alignment technology. However, they are not capable of producing subtitles automatically from the audio source and carry out the transcription step manually.

On the other hand, Koemei [21], SailLabs [34], Vecsys [40] or Verbio [41] are companies that commercialise automated transcription solutions for varying pools of languages and application scenarios such as lectures, open source intelligence or media, but do not produce subtitles. Since recently, Google [17] supports the automatic generation of time-aligned

draft transcriptions and subtitles from the videos uploaded to Youtube - serving multiple languages and allowing their automatic translation through Google Translate [18]. Nevertheless, for the moment Youtube transcriptions do not include punctuation and capitalization nor follow the standard professional subtitling practices (see Section 4).

In the more specific subtitling field, VoiceInteraction pioneered a transcription solution [27] capable of generating subtitles for Portuguese broadcast news, which was adopted and is currently in daily use by the public Portuguese broadcaster RTP. The SAVAS automatic live and batch subtitling applications described in the next section have followed this approach, being specifically designed for subtitling purposes and taking into account the most relevant features of quality subtitles.

3 SAVAS technology, development and applications

In this article, full systems for automatic subtitling and transcription of audiovisual contents are presented. The systems were trained over Broadcast News, sports and interview/debate domains, which contain a great variety of speakers (journalists and citizens), topics (economy, politics, sports), acoustic conditions (studio and outside news), and types of speech (planned and spontaneous). These aspects thus turned this type of content into an optimal resource to train the more robust and flexible LVCSR systems as possible.

The systems were developed for several European languages, including Portuguese, Spanish, Basque, Italian, French, German and the Swiss variants of the latter three. All the systems were initially trained over Broadcast News contents. In the case of Portuguese, this system was then extended and adapted to a more complex domain, i.e. interview and debate domain, containing repetitions, hesitations, disfluences, unfinished sentences, overlapping speech etc. All of these issues make the recognition process more difficult. Moreover, the Basque and Italian dictation systems were also adapted to the Sports domain for Respeaking purposes.

In addition, three type of applications were built over the systems described above for several subtitling and transcription purposes. The first application is a batch Speaker Independent Transcription and Subtitling application (S.Scribe!), capable of automatically transcribe pre-recorded audio and video files into time-aligned enriched subtitles. The second application corresponds to an Online Subtitling System (S.Live!) and it is able to automatically transcribing live audio into configurable and well-formatted subtitles. The final application involves a Respeaking engine (S.Respeak!) for dictation, which can be easily integrated into any commercial subtitling solution and capable of producing subtitles with an acceptable delay. Table 1 summarizes the LVCSR based SAVAS systems and applications that we developed per language and domain.

As it can be seen in Table 1, the three applications were developed for all the languages involved. Unlike the rest of the languages, Portuguese S.Scribe! and S.Live! applications were developed for both Broadcast News and Interview/debate domains. It implied an adaptation process of the base Portuguese models trained on Broadcast News contents to a more specific domain, in which more complicated issues related to spontaneous speech had to be considered.

Regarding S.Respeak! dictation system, all the languages were covered by engines trained on the news domain. In addition, these engines were also adapted to the Sports domain in the case of Basque and Italian languages. The S.Respeak! application is composed by an engine to produce transcriptions for dictation and respeaking purposes. It is not a complete solution for respeaking, since it has to be integrated with a subtitling software to

Table 1 LVCSR based SAVAS systems and applications per language

Language	S.Scribe!	S.Live!	S.Respeak!
Portuguese	Broadcast News	Broadcast News	News
	Interview/debate	Interview/debate	–
Spanish	Broadcast News	Broadcast News	News
Basque	Broadcast News	Broadcast News	Sports and News
Italian	Broadcast News	Broadcast News	Sports and News
Swiss Italian	Broadcast News	Broadcast News	News
French	Broadcast News	Broadcast News	News
Swiss French	Broadcast News	Broadcast News	News
German	Broadcast News	Broadcast News	News
Swiss German	Broadcast News	Broadcast News	News

create subtitles. Synthema's Voice Subtitle [4] and SysMedia's SpeakTitle [22] are examples of these subtitling solutions.

In the next subsections, the data resources we compiled to train the systems, the development of the main technological components of the systems, in addition to the type of applications, are described in more detail.

3.1 Compiled data resources

The development of robust LVCSR systems for automatic transcription and subtitling in the audiovisual domain requires considerably large audio and text corpora for the acoustic and language modeling. Based on previous experience [26, 29], we estimated that the development of good performance transcription systems would ideally require at least 200 hours of audio and 1000 million words of text. The same data could also be exploited to develop dictation systems. Besides, the adaptation of an already existing transcription or dictation system to a new domain was estimated to be achievable with 20 hours of audio with at least 500k words.

With regard to audio data for acoustic modeling, in this work the most of the data were gathered from programs produced by broadcasters. The TV programs were then manually annotated using the Transcriber tool,[2] following internal transcription conventions. Since manually annotating 200 hours of audio data is a highly costly task, we followed an incremental automation approach. The first 50 hours per language were annotated manually from scratch, followed by several stages of manual annotation with draft automatic transcriptions. At each stage, new acoustic models were trained to decrease the amount of errors produced by automatic recognition and thus to speed up the manual transcription process.

On the other hand, the text sources for language modeling and vocabulary creation were a mix of autocue scripts and subtitles provided by the broadcasters and subtitling companies, plus newswire and sports text crawled from the Internet. In addition, the transcriptions of the collected audio content were also used as text data. Table 2 shows the final amounts of audio and text corpora collected for each language.

As it can be seen from the Table 2, most of the targeted amounts were almost reached, except for Basque text corpora in the broadcast news and Portuguese in the interview/debate

[2] http://trans.sourceforge.net/en/presentation.php

Table 2 Collected audio and text corpora per language and domain

Language	Variant	Domain	Audio	Text
Portuguese	European	Broadcast News	113H	1012M
		Interview/debate (adaptation)	20H	200K
Spanish	European	Broadcast News	200H	1009M
Basque	Standard Basque	Broadcast News	200H	329M
		Sports (adaptation)	20H	500K
Italian	Italian	Broadcast News	162H	950M
		Sports (adaptation)	–	500K
	Swiss Italian	Broadcast News	50H	100M
French	European	Broadcast News	150H	932M
	Swiss French	Broadcast News	50H	100M
German	European	Broadcast News	151H	808M
	Swiss German	Broadcast News	51H	100M

domain for adaptation purposes. As a minority language, the availability of Basque text corpora in the news domain was limited. With Portuguese, the difficulty was to find text resources containing the type of spoken information common in the interview/debate domain.

Although an exact 1000 million word text corpus was not achieved for all languages, this cannot be considered as critical, since the purpose of having such large text corpora was to use pruning techniques in the final language models in order to reduce noise and texts particularly out of domain. With regard to Italian, French, German and their Swiss variants, the originally targeted amounts were distributed according to previous experience on dialect adaption [26].

All the audio and text data collected from broadcasters and subtitling companies were shared through the META-SHARE[3] repository, once the corresponding permissions were granted by to the contents owners. In addition to the raw audio and text, three transcribed audio test sets were also shared per language. These test sets will allow other LVCSR technology developers to compare the performance of their systems with that of the SAVAS engines. The commercial license established for the sharable resources is the META-SHARE Commercial-NoReDistribution-For-a-Fee (C-NoReD-FF) license. On the other hand, the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license was established for research purposes. All the data were shared through the SAVAS META-SHARE specific repository, which has become one of the biggest available audio and data sources exploitable for LVCSR development.

Further information about the collected data resources can be found in the work presented in [9].

3.2 Development of the systems

Automatic Subtitling and Transcription of audiovisual contents is a highly complex task that requires several modules of functionalities to provide useful operational capabilities. Although the LVCSR is the most important component, there are other technologies

[3] http://www.meta-net.eu/meta-share

involved in this process. The SAVAS systems can thus be represented as a pipeline of processing blocks, which represent the different components, as shown in Fig. 1.

In global terms, the Audio Pre-Processing block receives the program audio, discriminates between Speech and Non-speech and sends the audio to the Large Vocabulary Continuous Speech Recognition in case of speech. Additionally, it gives information on speaker clustering, speaker gender and speaker identification in case of relevant speakers. The Large Vocabulary Continuous Speech Recognition block transcribes the audio input stream according to a vocabulary and a language model. This component is the most important and critical one since its performance will be reflected directly in the final result. The Output Normalization block converts sequences of words representing digits, connected digits, and numerals into numbers. It also capitalizes the names and introduces the punctuation marks. Finally, the Subtitling Generation block creates the subtitles according to each broadcaster subtitling rules and specifications.

The overall system works in a pipeline and asynchronous operation mode, where each block is responsible for fulfilling its own task and send the results to the next block. In the following subsections, we give a more detailed description of the system components.

3.2.1 Audio pre-processing (APP)

The full operation of the APP block is intended to provide a complete description of the input audio, including speech/non-speech segmentation (SNS), gender classification (male/female), background classification (clean, noise, music) and speaker diarization, which performs speaker clustering and speaker identification (in case of relevant speakers as pivots).

The technology integrated for acoustic change detection and classification, background conditions classification and gender classification is described in the work presented in [28]. However, new algorithms were developed in this work for a more efficient Speaker Clustering and Speaker Identification, described in the following two subsections.

Speaker clustering Our previous work [45] in speaker clustering based on the Bayesian Information Criteria (BIC) obtained low performance mainly caused by the audio segmentation component (SNS), which sometimes produced small Speech segments. The new

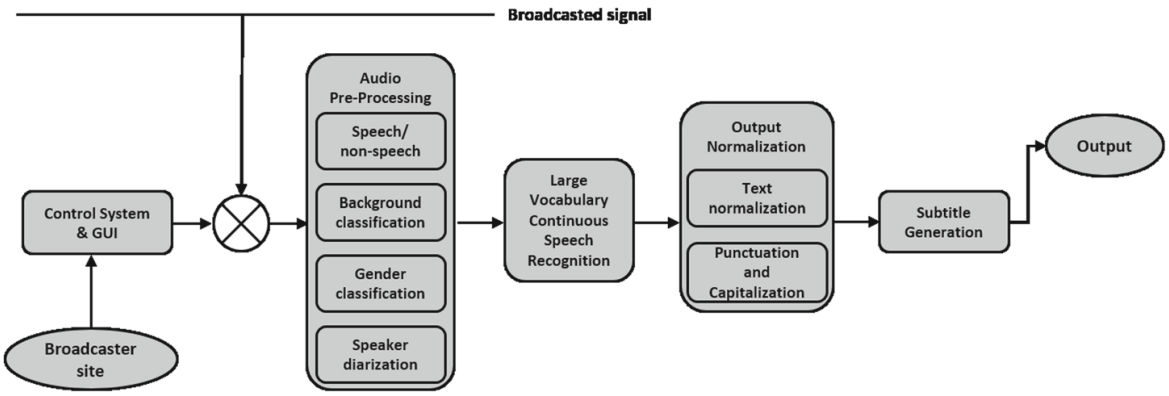


Fig. 1 Pipeline of the SAVAS subtitling systems

Table 3 Improvement on DER with the new algorithm

Language	Previous Algorithm	Improved Algorithm
Portuguese	29.06 %	24.04 %
Basque	41.30 %	32.60 %
Italian	30.07 %	16.84 %

algorithm adds a BIC based speaker turn detection before the BIC clustering to overcome this problem.

The algorithm starts by detecting speaker turns using BIC, where change points are detected through generalized likelihood ratio (GLR), using Gaussians with full covariance matrices. SNS segments are also modeled with Full Gaussian and compared with the current speaker Full Gaussian.

$$BIC_{i,j} = \frac{n_i + n_j}{2} log\,|\Sigma| - \frac{n_i}{2} log\,|\Sigma_i| - \frac{n_j}{2} log\,\left|\Sigma_j\right| - \lambda P \tag{1}$$

Equation (1) gives the BIC score of the similarity of two segments/clusters, where $|\Sigma_i|$ and $\left|\Sigma_j\right|$ are the determinants of the Gaussian associated to segment/cluster i and j respectively, $|\Sigma|$ is the determinant of the Gaussian associated to segment/cluster i plus j and P is a penalty factor. If the BIC score is lower than 0, then the two segments/clusters are merged together as one, otherwise a new speaker is detected and new statistical information is gathered for the new speaker Full Gaussian.

The hierarchical clustering algorithm is an adaptation of the algorithm described in [25]. In our hierarchical clustering algorithm, the current speaker cluster, provided by turn detection and modeled with Full Gaussian, is compared with the clusters obtained so far. This comparison differs from the implementation in [25], where all clusters are compared. This difference allows on-line processing of the clusters.

Table 3 presents results obtained for 3 languages, where significant reduction in Diarization Error Rate (DER) can be observed comparing the old BIC algorithm with the new one. DER metric defines the ratio of incorrectly detected speaker time to total speaker time, as it is described in [13].

Speaker identification Total Variability has emerged as one of the most powerful approaches to the problem of speaker verification. This technique jointly models speaker and channel variabilities as a single low rank space. Our Speaker Identification component uses the low-dimensionality total variability factors, known as identity-vectors (i-vectors), produced by the Total Variability technique to model known speaker identities. The i-vectors are extracted, as depicted in Fig. 2, using an Universal Background Model (UBM) and

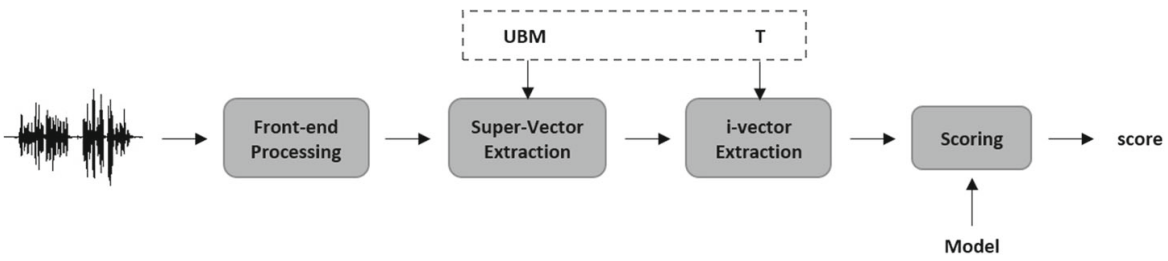


Fig. 2 Total Variability i-vector extraction

the Total Variability matrix (T). This technology is based on previous work conducted on Language Identification [1].

The Speaker Identification component works after the Speaker Clustering, once all the speakers have been grouped into clusters. Since this component works on-line, every time an unseen speaker starts talking, the component is not able to know the speaker identity immediately. To overcome this problem, we produce a first estimate for its identity (if it is a known speaker) after 10 seconds of speech and a final identity estimation after 30 seconds. Since the zero- and first-order sufficient statistics (and the respective i-vectors) from Total Variability are associated with the cluster, the speaker information is immediately available whenever a cluster with a known identity appears.

For this work, we trained Total Variability models with the same acoustic features and parameters used in [1], but modeling speakers instead of languages. Regarding the scoring, we also use Linear Logistic Regression (LLR), but only using the information from the i-vectors.

3.2.2 Large vocabulary continuous speech recognition (LVCSR)

The LVCSR engine named Audimus [31] is based on a hybrid speech recognition structure combining the temporal modeling capabilities of Hidden Markov Models (HMMs), with the pattern discriminative classification capabilities of Multilayer Perceptrons (MLPs). The processing stages are represented in Fig. 3.

The system uses and combines phone probabilities generated by several MLPs trained on distinct feature sets, resulting from different feature extraction processes in order to better model the acoustic diversity. This is relevant in the recognition of TV programs and multimedia contents, with a high diversity of speakers and environments. These probabilities are taken at the output of each MLP classifier and combined using an appropriate algorithm.

The decoder is based on the Weighted Finite-State Transducer (WFST) approach [7]. In Audimus systems, the search space is a large WFST, which results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model. This decoder uses a specialized WFST composition algorithm of the lexicon and language model components in a single step. Furthermore it supports lazy implementations, where only the fragment of the search space required in runtime is computed. Besides the recognized words, the decoder outputs a series of values describing the recognition process. In order to generate a word confidence measure, these features are combined through a maximum

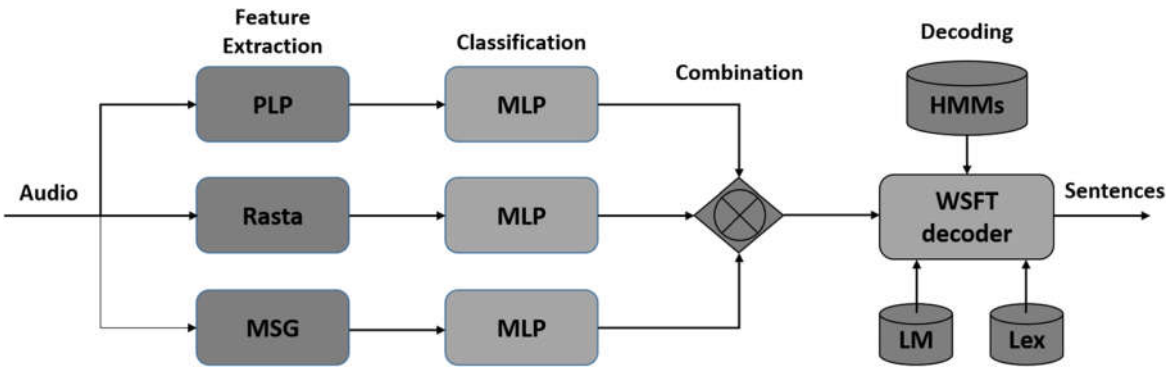


Fig. 3 Audimus processing pipeline

entropy classifier, whose output represents the probability of each word being correct [7]. Confidence measures for the recognized text are necessary to filter the output text in the subtitling composition stage.

Acoustic modelling The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches. Different feature extraction and classification branches effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, commonly present in multimedia content. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the third uses 28 MSG (Modulation SpectroGram) coefficients for each audio frame. Each MLP classifier incorporates local acoustic temporal context through an input window of 13 frames (the MSG branch uses 15 frames) and two fully connected non-linear hidden layers. The number of units on each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characteristic phonetic set of each language.

The acoustic model training was carried out with the same techniques used in [26]. For the first stages of the annotation process, monophone based systems were trained for each language. The acoustic models were trained in a series of stages to aid the process of manual annotation. Figure 4 presents the Word Error Rate (WER) against the amount of annotated material for the initial stages of the monophone based systems. As it can be observed in Fig. 4, not all languages have the same level of WER, but they all exhibit the same exponential decay behavior with the amount of training material.

Once a language reached all the training material, diphone systems were built using alignments and labels generated by the monophone systems. Since there is much less data for the Swiss variations, the Italian, French, German and their Swiss variations were trained with the combined data of both counter parts to enrich the models. Specifically, Italian was trained in conjunction with Swiss Italian, and the same for the other pair of languages.

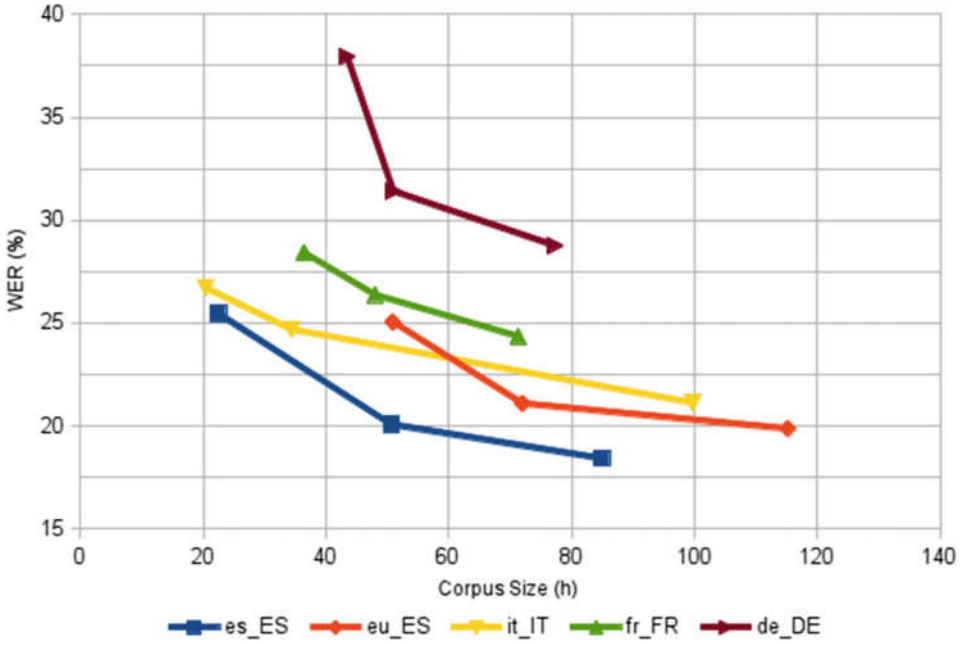


Fig. 4 WER vs annotation material

Language modelling The techniques used in the construction of the language models are the same as in [26]. The language models were created using statistical backed-off N-gram models. N-grams are probabilistic models which exploit the ordering of words predicting the next word from the previous N-1 words. In a bit of terminological ambiguity, the term N-gram is usually used to refer to either the word sequence or the predictive model.

In SAVAS systems, N-gram models resulted from the interpolation of several specific language models. The number of these specific language models varies across languages and depends on the availability of text data from different sources. For instance, in the case of the Spanish system, the first specific language model (LM) was a 4-gram LM trained on data from several online ES newspapers texts ranging from 2000 to 2014 totalizing 900M words. The second one was a 3-gram LM estimated on the BN training transcriptions which has 1.2M words. The third model was a 4-gram LM estimated on autocue scripts toting up 200M words. These three language models were then linearly interpolated with optimization of the weights on the automatic transcription texts. The final interpolated LM for Spanish was a 4-gram LM, with Kneser-Ney modified smoothing, with 100k words (1-gram), 8.0M 2-gram, 14.9M 3-gram, 9.8M 4-gram, with a perplexity value of 88.

In Table 4, the obtained number of N-grams (N-gram counts) and perplexity (PPL) values are presented for each model and language. Perplexity is the most common evaluation metric for N-gram language models and it is a function of the probability that the language model assigns to a data set. The smaller perplexity is, the better is the model.

Pronunciation lexicons In TV and multimedia contents large variety of topics are discussed over time. Additionally, in order to guarantee the performance of the LVCSR systems, the vocabularies have to be limited to a fixed amount of words, typically to the most frequent ones. Both constrains imply that Out-Of-Vocabulary (OOV) words cannot be avoided. The regular approach is to use a vocabulary containing at least 60K words for English and even more for other inflectional languages.

With the aim of having a reasonable coverage of the languages, while maintaining performance, we used the 100K most common words for the vocabularies. Even if agglutinative languages, like Basque and German, may require larger vocabularies, increasing the language coverage beyond the 100K word vocabulary can dramatically degrade performance, mainly in on-line systems.

For the pronunciation lexicons we used lexica developed for each language. For unavailable words we used grapheme-to-phoneme systems to generate the corresponding

Table 4 Language models n-gram counts and perplexities (PPL)

Language	2-gram	3-gram	4-gram	PPL
Portuguese	8.1 M	11.2 M	7.3 M	144
Spanish	8.0 M	14.9 M	9.8 M	88
Basque	11.5 M	13.3 M	5.3 M	261
Italian	10.4 M	14.1 M	9.2 M	151
French	9.2 M	12.1 M	6.8 M	88
German	10.2 M	12.9 M	7.5 M	174
Swiss Italian	10.4 M	14.1 M	9.2 M	198
Swiss French	9.2 M	12.1 M	6.8 M	130
Swiss German	10.2 M	12.9 M	7.5 M	214

pronunciations. Table 5 presents the number of pronunciations obtained for each language with the 100k vocabularies.

3.2.3 Output normalization

This component involves a set of actions to convert the raw output of the LVCSR into normalized text suitable for subtitles. These actions aim to reduce the dimensionality of the text and to improve the readability of subtitles.

The normalization operation is two-fold. First, numbers, numerals, dates and amounts (e.g. money and percentage) are converted to their digit representation through rule-based functions developed for each language. Besides, this block also punctuates the text and capitalizes the acronyms and proper names. The most common acronyms were included in the vocabularies, and techniques based on maximum entropy models are used for the automatic Punctuation and Capitalization of named entities [6]. These techniques are based on the information provided by the preceding APP and LVCSR components, such as pauses, speaker changes, Part-Of-Speech (POS) information of the present, previous and following words, in addition to the confidence measure associated to each word.

3.2.4 Subtitle generation

After the normalization operation, words are organized with the aim of composing the final subtitles, according to a series of options configurable by the user.

The SAVAS systems allow the configuration of the most common layout features related to subtitling, such as the position of subtitles on screen, the number of lines per subtitle, the amount of characters per line, the typeface, the distribution and alignment of the text, the transmission modes (i.e. blocks or scrolling), and the colors linked to different speakers. For this last feature, the information given by the APP component about the speaker gender is used to change the colors of subtitles.

Regarding features related to the duration of subtitles on screen, automatic pre-recorded subtitles created by the S.Scribe! application can be configured either to be synchronized to the audio or to follow minimum and maximum duration and speed rules to improve readability.

3.3 SAVAS applications

As mentioned before, three type of applications were developed for different subtitling purposes: batch transcription and subtitling system (S.Scribe!), online subtitling system

Table 5 Amount of pronunciations in the lexicons

Language	Pronunciations
Portuguese	120 K
Spanish	104 K
Basque	155 K
Italian	138 K
French	159 K
German	169 K

(S.Live!) and a respeaking and dictation engine (S.Respeak!). In the following subsections, these applications are described in more detail.

3.3.1 S.Scribe!

S.Scribe! is a client/server system, working offline: it is capable of processing a file of previously recorded audio or video and transcribe it, producing a subtitle file.

The application has an interface for administration and usage; it receives an audio/video file, adds it to a processing list and notifies the user upon completion, so that he/she can download the result. The most common and standard subtitling formats, like TTML or SRT, are supported. The results can also be downloaded in text (TXT) and meta-data (XML) formats. S.Scribe! includes 2 operation modes:

- HTML Interface: the system is available at a given web address (URL). The user has to log in and then he/she can submit audio/video files to be processed.
- Webservice interface (SOAP/WSDL): the system is invoked through a webservice. The user specifies a URL where the audio/video file is expected to be available for downloading and processing.

3.3.2 S.Live!

S.Live! is a speech transcription system which operates in both online and real time modes. It receives input streaming (audio or video in digital or analogue format) from the broadcaster or a multimedia content from the web, producing live captioning and broadcasting live subtitles in both analogue or digital formats.

The system can be used for captioning television programmes as well as available multimedia content on the web. It offers a Language Model Adaptation Service, which allows daily adaptation to new and different topics. It works in conjunction with an Acoustic Segmentation Module for separating the relevant acoustic areas for captioning. The S.Live! application was prepared to be integrated with the most commercial subtitling softwares, such as Screen or FAB subtitling systems, thorugh an IP based protocol.

Nowadays, the S.Live! system is used in several television channels in a number of countries like Portugal and Brazil.

3.3.3 S.Respeak!

The main engine of the S.Respeak! application is VOXControl, a software for dictation and re-speaking. Its use is foreseen in audiovisual contents with a high degree of background noise, music and spontaneous speech, where S.Live! or S.Scribe! applications underperform and a human operator is required to re-speak the relevant information. With this operation, the difficulties associated to the automatic transcription of those kinds of programs are overcome. This application allows to adapt the acoustic model to a specific user, in order to increase performance.

VoxControl can be integrated with re-speaking applications in two different modes:

- The first solution is to put the cursor on top of a window where the output text is needed. This is the normal operation of a dictation system. However, this implies that the user is not allowed to move the cursor and that only one user can operate the application. For re-speaking, this option could be limiting.

– Due to the limitations posed by the first solution, a second mode was implemented in which the application runs in the background. In this mode, the software communicates with the commercial subtitling software through the same mechanism and protocols of the S.Live! application.

4 Evaluation methodology

The evaluation methodology had the objective of measuring the performance of the developed SAVAS applications regarding the subtitling quality features accepted by the industry. With this aim, a varied set of metrics were employed, some of which were designed specifically for this work.

A number of guidelines and good practice codes for subtitling have been published over the last years. Among well-known ones are: Ofcoms Guidance on Standards for Subtitling;[4] BBCs Online Subtitling Editorial Guidelines;[5] ESISTs Guidelines for Production and Layout of TV Subtitles;[6] the Spanish UNE 153010 norm [2] on subtitling for the deaf and hard of hearing and a reference textbook on generally accepted subtitling practice published in 2007 by Jorge Diaz-Cintas et al. [10]. Standard guidelines cover the various aspects of subtitle quality, such as subtitle layout, duration and text editing, which are shared among subtitling companies and broadcasters.

Concerning layout features, the most widely accepted subtitling practice uses two centered lines at the bottom of the screen, of 37 characters each one, with white font over a black background and in block mode, which is much easier to read than scrolling. In order to highlight different speakers, colours such as yellow, cyan or green are usually employed.

With regard to duration, recommendations span features related to the delay or the persistence of subtitles on screen. While high latencies have a negative impact on the perceived quality of subtitles, short persistence on screen has shown to decrease their readability. Standard guidelines recommend a maximum delay of 3 seconds and a maximum speed of 160-180 words per minute. Although these recommendations are generally followed for pre-recorded programs, the difficulties posed by live subtitling currently result in median latencies of around 6 seconds, with spikes of up to 24 seconds.

Finally, the standard subtitling practice related to text editing employs mixed letter case as in printed material, splits subtitled text at the highest possible syntactic nodes and makes use of the most common and recognizable acronyms, apostrophes and numerals to save character space.

With the aim of measuring all these parameters, we employed and defined several metrics, which are described in more detail in the following subsections.

[4] http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/standards_for_subtitling/subtitling_1.asp.html

[5] http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf

[6] http://www.translationjournal.net/journal/04stndrd.htm

4.1 Metrics

4.1.1 Word error rate (WER)

WER is a common metric used to measure the performance of speech recognition systems. It is computed by comparing reference annotations against automatic transcriptions, which in our case correspond to automatic subtitles. WER is calculated through the following formula:

$$WER = \frac{Substitutions + Deletions + Insertions}{Words\ in\ the\ reference} \times 100 \tag{2}$$

Substitutions refer to words which are replaced. *Deletions* are related to words which are missed out and *Insertions* are words incorrectly added by the recognizer. *Words in the reference* is the number of total words in the reference annotation.

4.1.2 Speaker change detection (SD)

The SD metric was computed thorough the F1-measure metric, which combines the harmonic mean of Precision and Recall metrics as follows:

$$F1 - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision refers to the fraction of retrieved instances that are correct, while the Recall metric describes the fraction of correct instances that are retrieved.

4.1.3 Capitalization and punctuation

The performance of the capitalization and punctuation features was measured using the Precision, Recall and F1-measure metrics.

4.1.4 Timing

As mentioned before, two duration features are linked to subtitle timing: delay and persistence.

Delay The delay of automatic live subtitles is composed by (1) the latency of the LVCSR technology, (2) the time needed to compose each subtitle and (3) the time length to insert it into the audiovisual signal to be transmitted. This feature was measured by the broadcasters, once the SAVAS technology for automatic subtitling was integrated in their premises. The delay was computed comparing the time of the broadcasted subtitles against word-level time-codes synchronized to the audio.

The technology for Basque and Spanish was evaluated at Euskal Telebista (ETB, Basque Television), the Basque Countrys public broadcast service, while the Portuguese Radio and Television (RTP) integrated the technology for Portuguese. The Italian Public Service Broadcaster (RAI, Radio Televisione Italiana) tested Italian and SWISS Teletext (SWISS TXT) was in charge of evaluating the rest of the languages. Each of the broadcasters used a different insertion software

Persistence Subtitle persistence on screen was measured by subtracting the time-in from the time-out of each subtitle and averaging across entries. Because the number of words or characters per subtitle also has an impact in their readability, metrics such as words per minute (wpm) or characters per second (cps) were employed.

4.1.5 Splitting

It is difficult to measure subtitle splitting objectively, since there is generally more than one way of segmenting correctly a particular subtitle. Thus, the approach adopted to measure splitting quality involves asking subtitling experts to manually rank the quality of both inter- and cross-subtitle splitting.

4.1.6 Overall quality

The NER model[7] has been used since some years to measure live respoken subtitle errors. The model uses the following formula to determine the quality of live respoken subtitles:

$$NERvalue = \frac{N - E - R}{N} \times 100 \tag{4}$$

where N is the number of words in the respoken text, E corresponds to the edition errors caused by the respeaker's strategies, and R is the errors committed by the recognizer. Computing the formula, a NER value of 100 indicates that the content was subtitled entirely correctly. Good quality live subtitles are expected to go beyond 98 % accuracy according to this.

Since the NER model was devised for quality assessment of respoken subtitles, it considers only recognition and respeakerṣ edition errors. However, speaker colour, timing and splitting information is also relevant in automatic subtitling and helps establishing the quality of subtitles. In this work, the NER model was extended to also consider errors related to those features.

The extended eNER formula is as follows:

$$eNERvalue = \frac{(N \times P) - \sum_{i=1}^{N}(R + SD + T + S)}{(N \times P)} \times 100 \tag{5}$$

where N is the number of test subtitles, P is the number of parameters to be evaluated, R corresponds to the recognition errors, SD represents the speaker change errors, T is the timing persistence errors scoring 0 (no error) or 1 (error) values, and S represents the splitting errors, scoring 0 (no error), 0.5 (inter- or cross- subtitle error) or 1 (inter- and cross-subtitle errors). All the parameters has a maximum value of 1.

The recognition (R) and speaker change (SD) errors were calculated using the WER and F1-measure metrics respectively. The timing (T) and splitting (S) parameters were evaluated manually by subtitling experts.

The evaluation of the S.Respeak! applications for Basque and Italian was carried out using the NER model through the NERstar tool.

[7] http://www.speedchill.com/nerstar/

4.1.7 Productivity gain

The aim of the productivity gain evaluation was to test whether post-editing automatic subtitles is faster than creating them manually from scratch. Subtitling professionals in all languages were asked to post-edit automatic subtitles and to create them from scratch, using their usual subtitle editing software and quality standards. The productivity gain was measured using the Subtitles per minute (spm) metric.

5 Evaluation and results

5.1 Test Set description

The Test Set for the evaluation of the systems was composed of a total amount of 30 hours of news, interview/debate and sports TV programs broadcasted in 2014. This material was annotated to compare it against the outputs generated by each application. Table 6 details the amount of data collected for each application type per language.

For Basque and Spanish, 5 hours were compiled per language. In Basque, 2 hours were used to test the S.Live! subtitling application and other 2 hours to evaluate the S.Scribe! application in the news domain. Another hour was compiled to evaluate the S.Respeak! application in the sports domain. In Spanish, the test set was divided in two parts and employed to evaluate S.Live! and S.Scribe! applications. The data for Basque and Spanish were gathered from news and sport programs broadcasted by ETB.

For Portuguese, 2 hours were compiled from a debate program to test the Live subtitling application in the interview/debate domain. Overall, 6 hours were compiled for the Italian Test Set, including 4 hours for Italian and 2 hours for Swiss Italian. The Italian data were gathered from news programs broadcasted by RAI.

Regarding French, German and their Swiss variations, they follow the same distribution as Italian. The French contents were gathered from news programs broadcasted by Euronews, France24 and the Swiss French television. The German contents were collected from news programs broadcasted by DasErste, ZDF and the Swiss German television.

5.2 Word error rate

As shown in Fig. 5, WERs follow similar distribution for Basque, Spanish, Italian, French and German, achieving performances around 15 %. The Swiss variants of French and German languages goes up to 20 % and the more challenging Portuguese case reaches 30 %. There is no significant difference between live and pre-recorded mode and performance variations are most probably due to the use of different testing contents.

Table 6 Test Set data amounts per language

Application	PT	ES	EU	IT	FR	DE	IT_CH	FR_CH	DE_CH
S.Live!	2 H	2.5 H	2 H	1.5 H	2 H	2 H	1 H	1 H	1 H
S.Scribe!	-	2.5 H	2 H	1.5 H	2 H	2 H	1 H	1 H	1 H
S.Respeak!	-	-	1 H	1 H	-	-	-	-	-

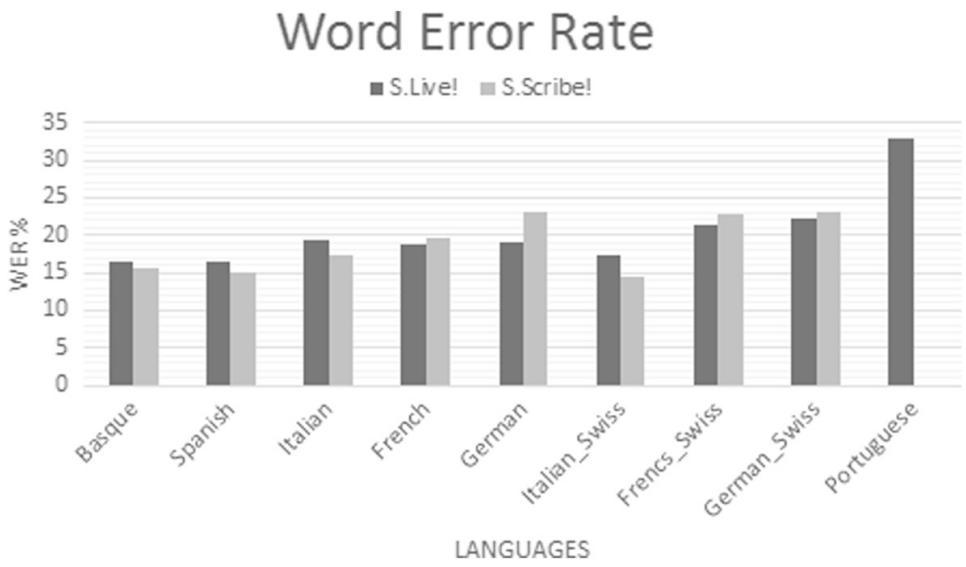


Fig. 5 WERs per application and language

5.3 Speaker change detection

Speaker Change Detection (SD) performance is shown in Fig. 6 per language and application type. The results are around 80 % with slightly worse performance going down to 60 % for Swiss French due to the acoustic conditions of these language on the test set. In the case of Portuguese, the particularity of the test material, which was composed by debate programs including many overlapped turns, degrades the accuracy for this language.

5.4 Capitalization and punctuation

Regarding Capitalization and Punctuation, Figs. 7 and 8 show that average F1-measures are around 85 % and 50 % respectively. Even if the results for Capitalization are promising, the accuracy obtained on automatic Punctuation reflects the difficulty posed by these type of contents, containing topic and speaker-dependent emotional pronunciations and intonations, in which acoustic pauses usually do not correspond to the real ends of phrases and sentences.

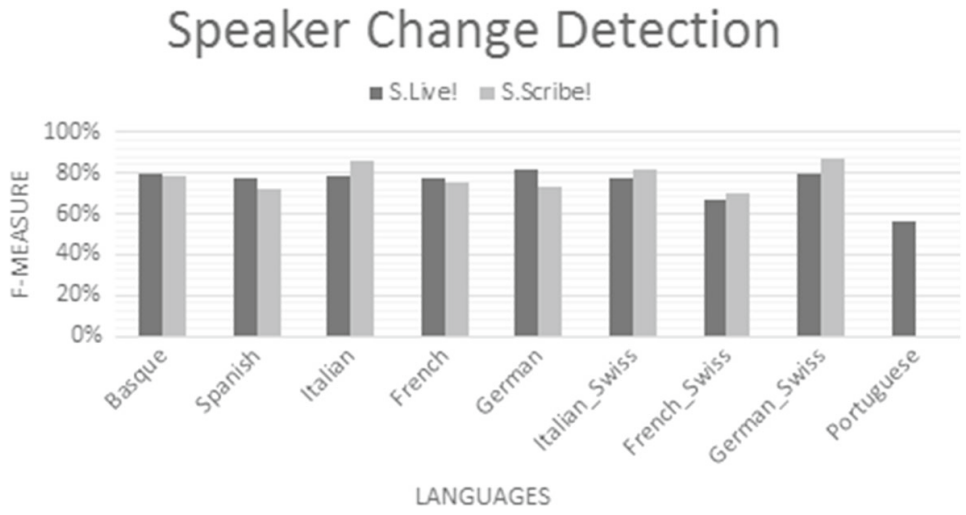


Fig. 6 Speaker Change Detection accuracies per application and language

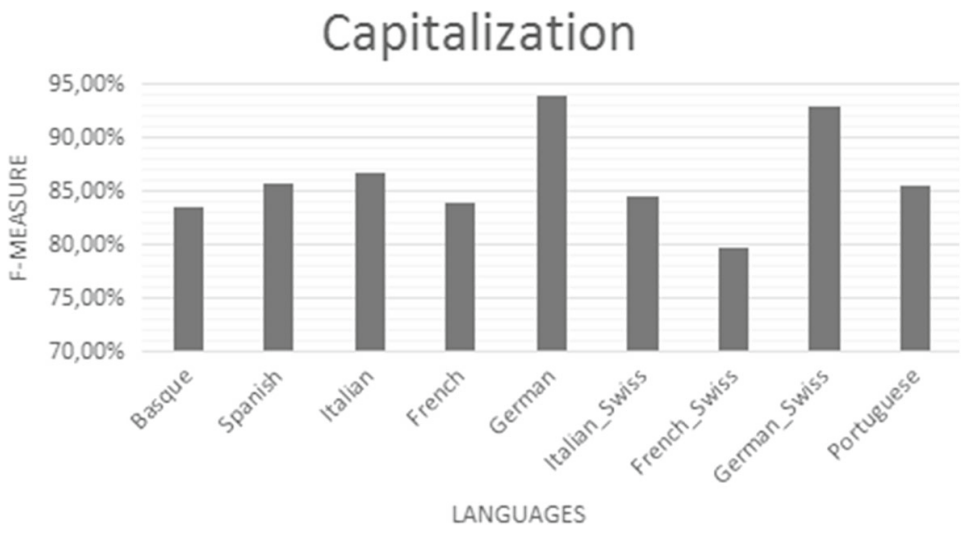


Fig. 7 Capitalization accuracies per language

5.5 Timing

5.5.1 Delay

The delay of the S.Live! applications is presented in Fig. 9. The results suggest that there could be a clear dependence between the delay and the software employed by each of the broadcasters to compose and insert the automatic subtitles into the broacasted signal. Each of the broadcasters used a different subtitling software: ETB employed WinCAPS for Basque and Spanish, RTP and SWISS TXT integrated FAB for Portuguese and Swiss variants respectively, and RAI used Speech Title for Italian. Nevertheless, if we leave the Italian outlier aside, the average delay results in 7 seconds which can be considered state-of-the-art performance of live respoken subtitling.

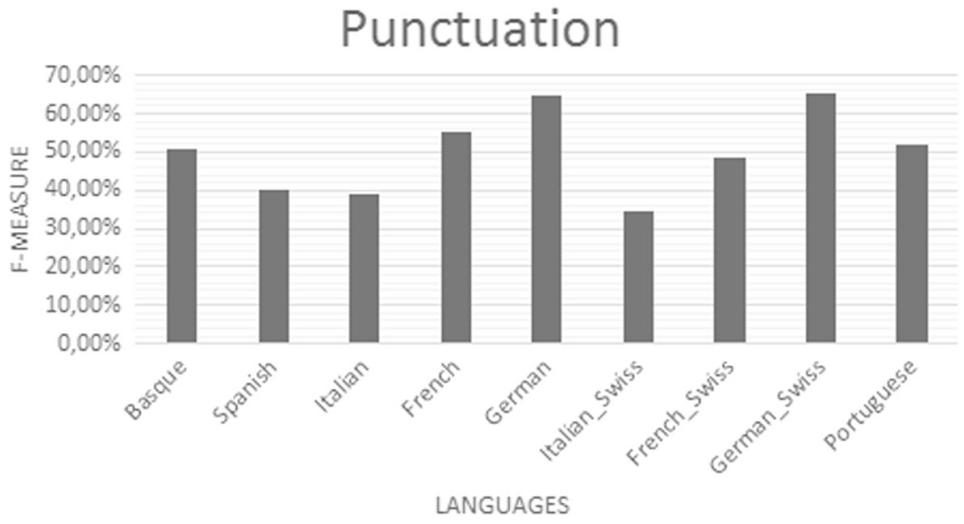


Fig. 8 Punctuation accuracies per language

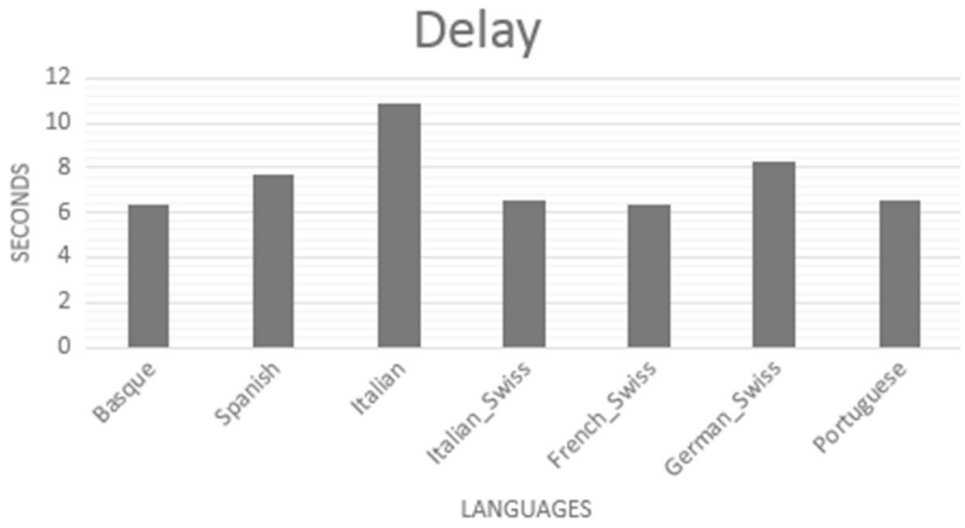


Fig. 9 Delay in seconds of the S.Live! applications

5.5.2 Persistence

Regarding persistence, Fig. 10 shows the average characters per second (CPS) computed per language in live and pre-recorded mode. Overall, average automatic subtitle persistence is below the maximum thresholds of 17 and 19 accepted by the subtitling community for pre-recorded and live programs respectively.

Results suggest persistence to be language specific, with Basque and Spanish automatic subtitles having the slowest speed, followed by Portuguese, Italian and the Swiss variants of Italian, French and German. Further inspection of the content also showed positive correlations between the amount of spontaneous sections in the audiovisual material, in which speech rate is usually faster, and their higher CPS values. The Swiss news programs in particular have higher proportions of interviews and spontaneous interactions than the rest. The reason why the live persistence of automatic Basque and Spanish subtitles is lower than their pre-recorded counterparts is that the WinCAPS insertion software employed at ETB had been configured to force subtitle speed output to 10 cps. Similarly, the persistence of the Swiss languages was also configured in the insertion software during live operation while

Fig. 10 The persistence in CPS per language and applications

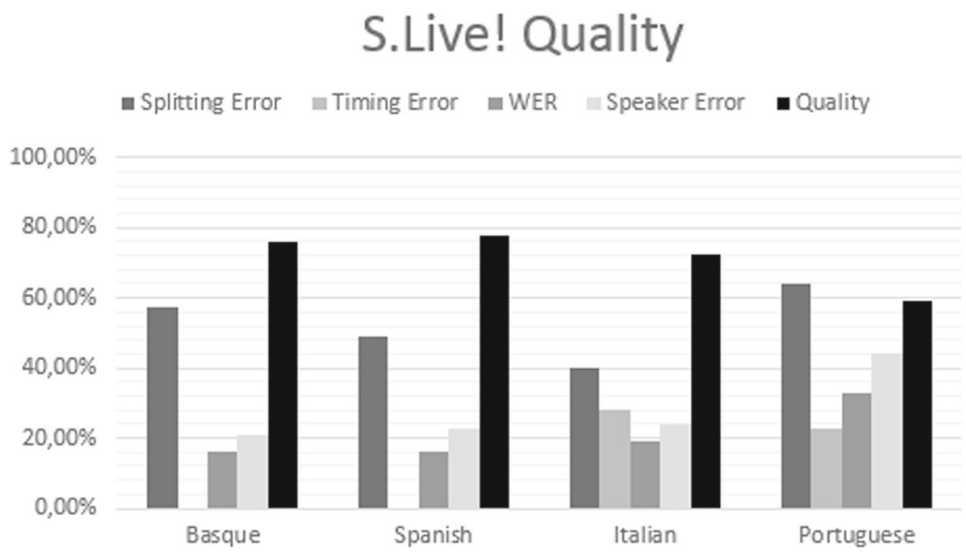


Fig. 11 S.Live! overall quality results per language

pre-recorded timing was simply set to be synchronized with the audio. These results show that configuring automatic subtitles just to be synchronized to the audio reduces delay but worsens persistence and, as a consequence, readability.

5.6 Overall quality

5.6.1 eNER of S.Live! and S.Scribe!

Figures 11 and 12 show the overall quality of the S.Live! and S.Scribe! applications. It was computed for Basque, Spanish, Italian and Portuguese languages. As it can be appreciated, eNER values are around 75 % on average for the broadcast news domain in Basque, Spanish and Italian without significant performance differences across operation modes and 60 % for the interview/debate domain in Portuguese. Although these values are far from the 98 % NER values considered to correspond to good quality subtitles, eNER results are expected

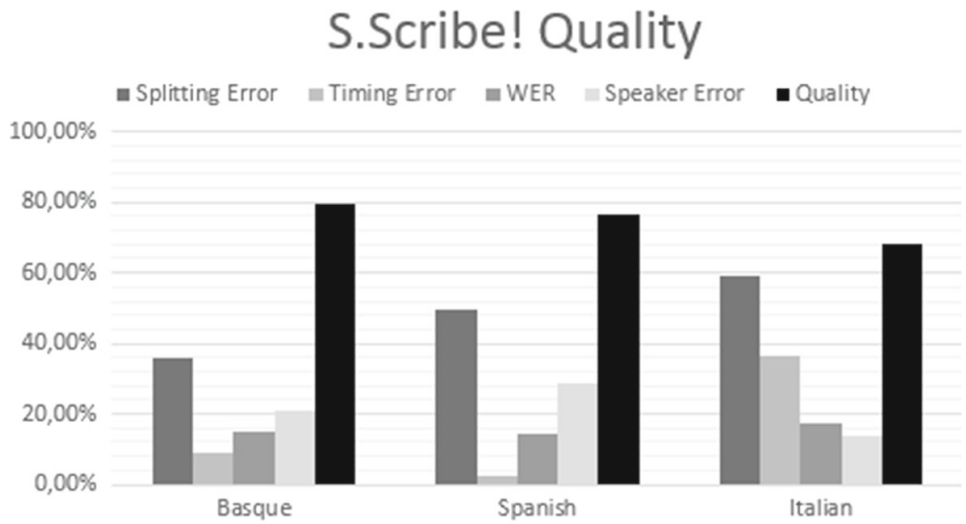


Fig. 12 S.Scribe! overall quality results per language

to reach relatively lower values because the extended formula considers a higher amount of quality features.

If we look into the specific weight of each one of the considered quality features on the overall eNER metric, we can see that splitting errors are the most frequent ones for all the languages.

Generally speaking, only 10 percent of the examined subtitles were free of splitting errors, about half of them contained at least one of the two possible splitting errors, a third both of them. And as far as the error-free subtitles were concerned, one needed in most cases to link them with other subtitles in order to guarantee readability. This can be considered as reasonable, since technology for automatic splitting was not trained and developed within this work. The splitting of subtitles was done just counting up the characters and controlling not to exceed the maximum length of each subtitle line.

Manual evaluation of timing errors resulted zero for S.Live! in Basque and Spanish because the WinCAPS insertion subtitle software was configured to force subtitle speed output to 10 cps during broadcasting.

In Italian, timing errors outweight those related to speaker change and WER for both applications. Even if the average persistence achieved for Italian was in the 13 cps range, a further study of the results demonstrated a high fluctuation between low and high cps values. Regarding timing errors for S.Live! in Italian, the high delay presented above was the main reason for these discrete results.

5.6.2 NER of S.Respeak!

The S.Respeak! applications were evaluated for Basque and Italian in the sports domain. Due to the effort required by the only respeaker available, the respeaking task for Basque was divided into two parts of 20 and 30 minutes. The NER values achieved for each part were 86.55 % and 85.05 % respectively. Most of the errors were due to minor edition mistakes related to the speakers' strategies, while recognition errors were mostly caused by substitutions.

In Italian, a 84,64 % NER value was obtained. In this case, the great majority of errors was due to minor recognition mistakes mainly classified as deletions. Edition errors committed by the respeaker were smaller and minor.

We believe that the presented results could be improved by more proficient respeakers for both languages.

5.7 Productivity gain

Finally, Figure 13 summarizes the productivity gains achieved in the post-editing task. All but one subtitler have managed to increase their productivity post-editing automatic pre-recorded subtitles when compared to creating them from scratch. Gains are highly subtitler dependent, ranging between 33 % to 2 % across post-editors. We believe post-editing training and practice should help increase them.

For Italian, post-editing S.Scribe! output has also been compared against post-editing stenotype output. As it can be appreciated in the Fig. 13, the latter has achieved higher productivity gains. Stenotypists probably generate less text editing errors than state-of-the-art LVCSR technology, particularly in what capitalization and punctuation features are concerned and, thus, the time devoted to correcting such kind of errors is reduced. However, stenotyping requires personnel resources that automatic transcription does not, which is an important factor to be considered.

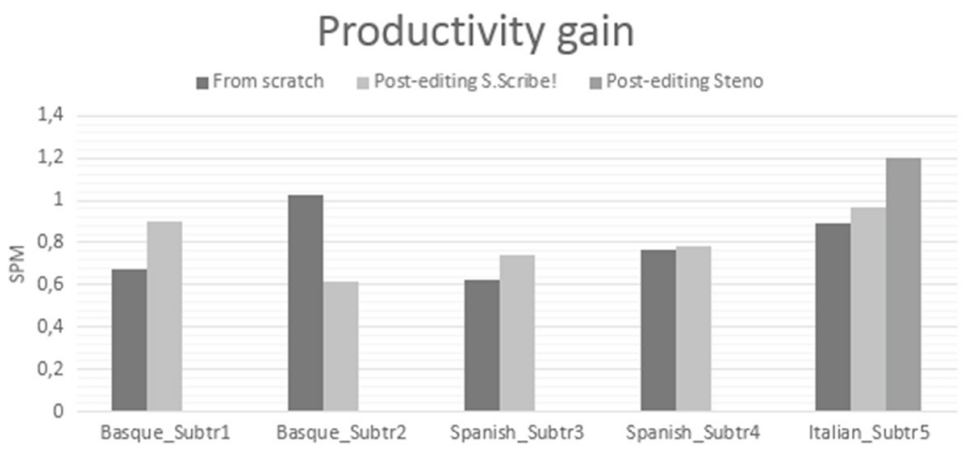


Fig. 13 Productivity gain results

5.8 Comparative evaluation

As we detailed in Section 2, Google is currently the only company which provides publicly the automatic generation of time-coded transcriptions and subtitles. This is a service available to the videos uploaded by the users to Youtube platform and it works in batch mode. However, the automatically generated transcriptions and subtitles from Youtube do not include punctuation and capitalization at the moment. Moreover, subtitles do not fulfill the standard subtitling practices defined by professionals. Thus, SAVAS systems could be compared to Google technology only at word error rate level.

For this comparative evaluation, we employed contents from 4 languages of the SAVAS project, including Spanish, Italian, French and German. We uploaded contents to Youtube and obtained the related transcriptions. These transcriptions were then compared to the results obtained with the SAVAS S.Scribe! application for the same audiovisual contents. The comparison was done using the WER metric explained in Section 4. Table 7 shows the results obtained for each language using Youtube and SAVAS S.Scribe! application.

As it can be seen in Table 7, S.Scribe! application outperforms the results obtained by Google technology through Youtube platform for all the languages in the test set. However, it could be expectable since the domain of these test contents was the same employed to train SAVAS systems; that is, broadcast news domain. Google offers technology which contains models trained on general domain data. The most remarkable differences are appreciated in Italian and French languages with improvements of around 10 and 12 points respectively.

Table 7 WER metrics for Youtube and S.Scribe! application

Language	Duration	Youtube	S.Scribe!
Spanish	2 H	24.91 %	16.50 %
Italian	2.5 H	27.80 %	17.81 %
French	2 H	37.61 %	25.81 %
German	2 H	31.10 %	26.92 %

6 Conclusions

In this article, new LVCSR based subtitling systems for both batch and live multimedia contents have been presented for several European languages. As it was described in Section 2, SAVAS systems can be considered pioneers offering a solution for full transcriptions and subtitles generation for different types of applications and languages. A survey of the literature in the field actually provides no references on such full automatic subtitling systems. Besides supporting different European languages, SAVAS systems include LVCSR engines trained over a huge amount of annotated data, technology for automatic punctuation and capitalization, speaker clustering and identification, in addition to modules for text normalization and subtitles generation following configurable standard subtitling rules. The systems were developed considering the needs of the subtitling companies and market, including the integration of the systems with the main subtitling softwares. Furthermore, SAVAS systems include three types of applications per language; S.Scribe!, a batch Speaker Independent Transcription system for offline subtitling; S.Live!, a Speaker Independent Transcription System, with real-time performances for live subtitling; and S.Respeak!, a dictation engine for live and batch production of subtitles.

The SAVAS systems were further evaluated using several metrics related to LVCSR technology and subtitling quality specific features. Although the developed automatic subtitling applications do not perform as well as professional subtitlers, they have achieved favorable results. The WER both in live and pre-recorded mode can be considered promising since it performed much better than other reference systems like Google for this specific domain. The delay of S.Live! subtitles is perfectly consistent with the recommendations. The speaker change detection technology works well even if it can be refined to work better with spontaneous speech. Automatic punctuation is error-prone considerably often due to the difficulty posed by the TV contents, and the splitting algorithm shall look into making use of syntactic information to achieve better results. Finally, productivity gain experiments suggest that post-editing automatic subtitles is faster than creating them from scratch.

The future work will be focused on the improvement of the technologies involved in the SAVAS systems. Regarding LVCSR technology, recent studies have demonstrated that Deep Neural Networks (DNNs) models have driven significant improvements on a variety of speech recognition benchmarks and data sets [42, 44]. With regard to SAVAS systems, further work will include research and development of a hybrid DNN-HMM recognition system for a more efficient offline and specially online subtitling. With the aim of improving the automatic punctuation module, new features will be included for classification, including prosodic and speaker related information. Furthermore, recent advances in Natural Language Processing with Neural Networks [8] has shown promising results that could be useful in sentence boundaries detection and punctuation marks prediction. The automatic splitting of subtitles should be also improved considering syntactic information to create linguistically coherent line-breaks, which is the preferred and most adopted solution in the community. To this end, the literature offers solutions based on the use of machine learning algorithms [5]. Finally, future work will also include the expansion of the SAVAS systems to more European and Asian languages.

Acknowledgments This work was funded by the FP7-ICT-2011-SME-DCL project 296371 - SAVAS (Sharing Audiovisual contents for Automatic Subtitling). http://www.fp7-savas.eu

References

1. Abad A (2007) The L2F language recognition system for NIST LRE 2011. In: The 2011 NIST language recognition evaluation (LRE11) workshop
2. AENOR (2003) Spanish Technical Standards. Standard UNE 153010:2003: Subtitled Through Teletext. http://www.aenor.es
3. Ajot J, Fiscus J (2009) The rich transcription 2009 speech-to-text (STT) and speaker attributed STT results. Tech. rep., NIST - National Institute of Standards and Technology, Rich Transcription Evaluation Workshop, Melbourne, Florida
4. Aliprandi C, et al. (2003) RAI voice subtitle: how the lexical approach can improve quality in Speech Recognition Systems. https://www.voiceproject.eu/
5. Álvarez A, Arzelus H, Etchegoyhen T (2014) Towards customized automatic segmentation of subtitles. In: Advances in speech and language technologies for Iberian languages. Springer, pp 229–238
6. Batista F, Caseiro D, Mamede N, Trancoso I (2008) Recovering capitalization and punctuation marks for automatic speech recognition: case study for Portuguese broadcast news. Speech Comm 50(10):847–862
7. Caseiro D, Trancoso I (2006) A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Trans Audio Speech Lang Process 14(4):1281–1291
8. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537
9. Del Pozo A, Aliprandi C, Álvarez A, Mendes C, Neto J, Paulo S, Piccinini N, Raffaelli M (2014) SAVAS: collecting, annotating and sharing audiovisual language resources for automatic subtitling. In: LREC 2014. Proceedings of the 9th international conference on language resources and evaluation
10. Díaz-Cintas J, Orero P, Remael A (2007) Media for all: subtitling for the deaf, audio description, and sign language, vol 30. Rodopi
11. eCaption: http://www.ecaption.eu/
12. FAB - Teletext & Subtitling Systems: FAB Subtitler Live Edition. http://www.fab-online.com/eng/subtitling/production/subtlive.htm
13. Fiscus J, Garofolo J, Ajot J, Michet M (2006) Rt-06s speaker diarization results and speech activity detection results. In: NIST 2006 spring rich transcrition evaluation workshop, Washington DC
14. Flanagan M (2009) Recycling texts: human evaluation of example-based machine translation subtitles for DVD. Ph.D. thesis, School of applied language and intercultural studies. Dublin City University, Dublin
15. Galliano S, Geoffrois E, Gravier G, Bonastre JF, Mostefa D, Choukri K (2006) Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In: Proceedings of LREC, vol 6, pp 315–320
16. Gauvain JL, Lamel L, Adda G (2001) Audio partitioning and transcription for broadcast data indexation. Multimedia Tools Appl 14(2):187–200
17. Google: Automatic captions in youtube. https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html (2009)
18. Google: Translate youtube captions. https://www.mattcutts.com/blog/youtube-subtitle-captions/ (2009)
19. Grass Valeey: Subtitle and Caption Creation. http://www.grassvalley.com/products/subcat-subtitle_and_caption_creation
20. IBM: Viavoice. http://www-01.ibm.com/software/pervasive/viavoice.html
21. Koemei: https://www.koemei.com/
22. Lambourne A, Hewitt J, Lyon C, Warren S (2004) Speech-based real-time subtitling services. Int J Speech Technol 7(4):269–279
23. Lan ZZ, Bao L, Yu SI, Liu W, Hauptmann AG (2013) Multimedia classification and event detection using double fusion. Multimedia Tools Appl 1–15
24. Lööf J, Gollan C, Hahn S, Heigold G, Hoffmeister B, Plahl C, Rybach D, Schlüter R, Ney H (2007) The RWTH 2007 TC-STAR evaluation system for european English and Spanish. In: INTERSPEECH, pp 2145–2148
25. Meignier S, Merlin T (2010) LIUM SpkDiarization: an open source toolkit for diarization. In: CMU SPUD workshop, vol 2010, Dallas
26. Meinedo H, Abad A, Pellegrini T, Trancoso I, Neto J (2010) The L2F broadcast news speech recognition system. Proc Fala 93–96
27. Meinedo H, Caseiro D, Neto J, Trancoso I (2003) Audimus.media: a broadcast news speech recognition system for the european portuguese language. In: Computational Processing of the Portuguese Language. Springer, pp 9–17

28. Meinedo H, Neto JP (2005) A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models. In: INTERSPEECH. Citeseer, pp 237–240
29. Meinedo H, Viveiros M, Neto JP (2008) Evaluation of a live broadcast news subtitling system for portuguese. In: INTERSPEECH, pp 508–511
30. Microsoft: windows speech recognition. http://www.windows.microsoft.com/en-us/windows7/dictate-text-using-speech-recognition
31. Neto J, Meinedo H, Viveiros M, Cassaca R, Martins C, Caseiro D (2008) Broadcast news subtitling system in portuguese. In: IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE, pp 1561–1564
32. Nuance: Dragon Naturally Speaking. http://www.nuance.com/index.htm
33. Obach M, Lehr M, Arruti A (2007) Automatic speech recognition for live TV subtitling for hearing-impaired people. Challenges for Assistive Technology: AAATE 07 20:286
34. Sail Labs: http://www.sail-labs.com/
35. Screen Systems: WinCAPS Q-live for live and news subtitling and captioning. http://www.screensystems.tv/products/wincaps-q-live/
36. Screen Systems: WINCAPS QU4NTUM subtitling software. http://www.screensystems.tv/products/wincaps-subtitling-software/
37. Starfish Technologies: Subtitling and closed captioning systems. http://www.starfish.tv/captioning-and-subtitling/
38. SyncWords: https://www.syncwords.com/
39. Ubertitles: http://www.ubertitles.com/
40. Vecsys: http://www.vecsys-technologies.fr/en/
41. Verbio: https://www.verbio.com/
42. Vu NT, Imseng D, Povey D, Motlicek P, Schultz T, Bourlard H (2014) Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 7639–7643
43. Woodland PC (2002) The development of the HTK broadcast news transcription system: an overview. Speech Comm 37(1):47–67
44. Zhang X, Trmal J, Povey D, Khudanpur S (2014) Improving deep neural network acoustic models using generalized maxout networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 215–219
45. Zibert J, Mihelic F, Martens JP, Meinedo H, Neto J, Docio L, García-Mateo C, David P, Zdansky J, Pleva M et al (2005) The COST278 broadcast news segmentation and speaker clustering evaluation: overview, methodology, systems, results. In: 6th Annual conference of the international speech communication association (Interspeech 2005); 9th European conference on speech communication and technology (Eurospeech), vol 2005. International Speech Communication Association (ISCA), pp 629–632

Aitor Álvarez He works as staff Researcher of the Human Speech and Language Technologies group at VICOMTECH. He studied Computer Science at the University of the Basque Country (2005). He carried out his final year project at the Department of Architecture and Computer Technology of the same university, where he continued working as a scholar. He is currently a PhD student, has completed the Diploma of Advanced Studies and is working on his thesis on advanced audio and speech processing techniques

7.3 SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling

- **Authors:** Arantza del Pozo, Carlo Aliprandi, Aitor Álvarez, Carlos Mendes, Joao P. Neto, Sérgio Paulo, Nicola Piccinini, and Matteo Raffaelli
- **Booktitle:** Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)
- **Year:** 2014

SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling

Arantza del Pozo[1], Carlo Aliprandi[2], Aitor Álvarez[1], Carlos Mendes[3], Joao P. Neto[3], Sérgio Paulo[3], Nicola Piccinini[2], Matteo Raffaelli[2]

[1]Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
[2]Synthema, Pisa, Italy
[3]VoiceInteraction, Lisbon, Portugal

[1]{adelpozo, aalvarez}@vicomtech.org, [2]{carlo.aliprandi, nicola.piccinini, matteo.raffaelli}@synthema.it,[3]{carlos.mendes, joao.neto, sergio.paulo}@voiceinteraction.pt

Abstract

This paper describes the data collection, annotation and sharing activities carried out within the FP7 EU-funded SAVAS project. The project aims to collect, share and reuse audiovisual language resources from broadcasters and subtitling companies to develop large vocabulary continuous speech recognisers in specific domains and new languages, with the purpose of solving the automated subtitling needs of the media industry.

Keywords: audio and text corpora, speech recognition, automatic subtitling

1. Introduction

Due to recently approved European and National directives and laws, the subtitling demand has grown fast in the past few years throughout Europe. As a result, broadcasters and subtitling companies are seeking for subtitling alternatives more productive than the traditional manual process.

Large Vocabulary Continuous Speech Recognition (LVCSR) is proving to be a useful technology for such a purpose. Respeaking is consolidating as the main subtitling technique employed for live and pre-recorded broadcast productions. Another trend in use today is the application of speech recognition to automatically generate a transcript of a programme's soundtrack without the need of a respeaker [1], and to use this as the basis of subtitles in order to increase subtitling efficiency and reduce costs. Unfortunately, the high expenses associated to the collection and annotation of the audio and text corpora required to train each LVCSR system for respeaking or automatic transcription has hindered the development of new languages and application domains.

Within the SAVAS project[1], data is being collected and annotated and LVCSR technology for automated subtitling is being developed for the languages and domains shown in Table 1.

The consortium is formed by a mix of broadcasters, subtitling companies and technology developers with expertise in the targeted languages. Given the participation of a Swiss partner, Swiss Italian, Swiss French and Swiss German data is also being collected and annotated and the development of Swiss variants of the Italian, French and German systems is being explored.

In addition, the consortium is striving to share the collected language resources for which intellectual property rights can be cleared, without compromising the business plan of the main results of the project. In this paper, we describe the data collection, annotation and sharing activities carried out within the project.

System type	Domain	Language(s)
Transcription	Broadcast news	Basque, Spanish, Italian, French and German
	Interview/debate	Portuguese
Dictation	Broadcast news	Italian, French, German
	Sports domain	Basque, Italian

Table 1. SAVAS systems, domains and languages

2. Data collection

The development of robust LVCSR systems for automatic subtitling, capable of producing transcriptions with word error rates (WER) below 20%, requires considerably large audio and text corpora for acoustic (AM) and language modeling (LM).

2.1 Data targets

Table 2 summarizes the targeted amounts of audio and text data aimed to be collected within the project.

Based in previous experience [2,3], we estimated that the development of transcription systems for automatic subtitling in new languages within the broadcast news domain would ideally require at least 200 hours of audio and one billion words of text. The same data could also be exploited to develop dictation systems in the same domain for the considered languages. On the other hand, the adaptation of an already existing transcription system for automatic subtitling to a new domain was estimated to be achievable with 20 hours of audio and 500k words. Finally, 20 hours and 500k words of audio and text were deemed enough to adapt existing transcription systems for automatic subtitling to a different dictation domain.

[1] http://www.fp7-savas.eu/

System type	Data	
	Audio for AM	**Text for LM**
Transcription/Dictation in the broadcast news domain	200 hours	1B words
Transcription in the interview/debate domain	20 hours	500k words
Dictation in the sports domain	[20 hours]	500k words

Table 2. Targeted audio and text data

For existing dictation systems with robust enough acoustic models, the text alone was estimated sufficient for adaptation.

2.2 Collected data

Most of the required audio data has been gathered from programs produced by the broadcasters in the consortium. The text sources are a mix of autocue scripts and subtitles provided by the broadcasters and subtitling companies in the consortium, plus newswire and sports text crawled from the Internet. In addition, the transcriptions of the collected audio content have also been used as text data for language modeling. Table 3 shows the final amounts of audio and text corpora collected for each language and domain.

Language	Domain	Audio	Text
Basque	Broadcast news	200h	350M
	Sports	20h	200k
Spanish	Broadcast news	200h	1B
Portuguese	Interview/debate	20h	200k
Italian	Broadcast news	150h	1B
	Sports	--	500k
Swiss Italian	Broadcast news	50h	100M
French	Broadcast news	150h	1B
Swiss French	Broadcast news	50h	100M
German	Broadcast news	150h	1B
Swiss German	Broadcast news	50h	100M

Table 3. Collected audio and text corpora per language and domain

As it can be seen from the table above, most of the targeted amounts have been reached, except for Basque and Portuguese text corpora in the broadcast news, sports and interview/debate domains, respectively.

As a minority language, the availability of Basque text corpora in the news and sports domains is limited.

With Portuguese, the difficulty has been to find text resources containing the type of spoken information common in the interview/debate domain: repetitions, hesitations, disfluencies, unfinished sentences, etc. Thus, the text corpus from the conversational domain has been compiled based on the transcriptions of the corresponding 20 audio hours.

For Italian, French, German and their Swiss variants, the originally targeted amounts have been distributed according to previous experience on dialect adaptation [3].

3. Data annotation

The annotations used in the SAVAS project are composed of spoken utterance transcriptions combined with speaker turn and background noise segmentations.

3.1 Tools & methodology

Transcriber 1.5.1 [4] has been chosen as annotation tool, since it has been developed for the creation and management of speech corpora closely following the Linguistic Data Consortium's [2] annotation conventions and recommendations which SAVAS follows.

The methodology employed has aimed at making the annotation process as productive as possible, following an incremental automation approach. The first 50 hours per language have been annotated manually from scratch, with the support of autocue scripts as a basis for transcription when available. Such manual annotations have then been used to develop a set of automation tools described in Section 0, for the automatic generation of transcriptions and annotations that can be imported into Transcriber. From then on, annotators only needed to correct the errors produced by the automatic tools instead of transcribing and annotating audio content from scratch.

3.2 Quality assurance

The consistency and accuracy of the annotations has been ensured through personalized training and a centralized review methodology. An annotation core team has been established per language, responsible for training the rest of annotators and reviewing the consistency and quality of their annotations.

Training courses have been organised for annotators to learn the SAVAS annotation guidelines and carry out their first annotation tasks in a supervised manner. Each annotator's initial set of annotations were then thoroughly reviewed by the core annotation team in each language. This intensive review-and-feedback process was repeated with each annotator until the core team considered that the quality of their annotations was good and consistent. After that, core teams kept reviewing all annotations and reporting on repeated mistakes annotators may have produced.

3.3 Automation tools

3.3.1 For transcriptions

The first batch of manually annotated 50 hours was used, together with the text material available at the time, to develop full LVCSR systems for each language. Annotators started employing the output of these systems as draft timed-transcriptions to be post-edited with Transcriber. As more annotated audio and text became available, updated versions of the LVCSR systems were

[2] http://www.ldc.upenn.edu/

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	AVG	STD
P1	45	50	60	50	70	30	60	30	50	28	60	48.45	14.08
P2	32	40	50	35	65	20	45	24	40	21	48	38.18	13.72
P3	28	35	40	30	55	18	35	18	28	18	32	30.63	11.02
P4	26	30	40	25	43	16	26	18	18	16	22	25.45	9.18

Table 4. Annotators' effort across phases: P1=first annotations from scratch; P2=use of autocue scripts; P3=experienced; P4=use of automation tools

trained with more data. This kept improving transcription accuracy.

Automatic punctuation and capitalization modules were also iteratively trained with the data available in each cycle, which further improved the quality of the automatic transcriptions for post-edition.

3.3.2 For annotations

Audio segmentation and speaker diarization modules were implemented to allow the creation of draft background noise and speaker turn labels.

In addition, already annotated named entity labels were exploited to allow their automatic tagging in the remaining data.

The development of these modules was also carried out in an iterative manner, based on the annotated data available per language and cycle.

3.3.3 Achieved productivity gain

Table 4 shows the effort, measured in terms of reported average hours needed to annotate one hour of content, required by a control group of 11 annotators, A1-A11, across several languages and annotation phases.

Annotation effort has resulted to be heavily linked to each particular annotator. As shown in the table, the quickest annotator, A10, is almost three times faster than the slowest one, A5.

The numbers in the table also show that all annotators managed to reduce the required effort with automation and experience. On average, productivity increased 21% from P1 to P2, which suggests that the use of autocue scripts as draft transcriptions markedly speed up the annotation process. The experience gained between P2 and P3 further improves productivity in 16% on average, so we can conclude that the more content an annotator annotates, the more productive he/she will become. Finally, the use of automatic tools for transcription and annotation also increased productivity from P3 to P4 in 11% on average. The overall productivity gain achieved on average from P1 to P4 was thus a considerable 48%.

4. Data sharing

Most of the resources present in the existing data infrastructures and repositories such as ELDA[3], LDC[2] or META-SHARE[4] are textual (i.e. written corpuses, lexical resources and treebanks). In comparison, oral resources such as those required to build LVCSR systems are less common [5], since their collection is in general more costly.

In order to leverage the audio and text data compilation work carried out within the project, the consortium has made an effort to clear intellectual property rights (IPR) and maximize the amount of resources to be shared with the rest of the community without compromising its business plan. It is worth to note that the call[5] under which the SAVAS project is funded seeks among other things for the financed consortia, in particular SMEs, to commercially exploit project results.

A SAVAS META-SHARE repository that hosts the project data has been developed and can be accessed through http://metashare.synthema.it:8000. The legal status of the shared audiovisual resources has been cleared and their licensing foundations have been established.

Table 5 summarizes the type, amounts and license schemes of the data shared for each language. As it can be seen, all audio and text data collected from the consortium broadcasters and subtitling companies is shared. In addition to raw audio and text, two transcribed audio test sets are also shared per language. These test sets will allow other LVCSR technology developers compare the performance of their systems with that of the SAVAS engines, which we plan to publish in other relevant conferences.

Because the consortium SMEs are looking into the market exploitation of automated subtitling and transcription applications trained on the compiled annotations, these will not be made available in the repository.

In those cases in which data sources external to the consortium such as French and German audio or Internet text crawls have had to be exploited, data sharing permission has been requested to the respective owners. Unfortunately, this process has resulted highly time-consuming and little productive. More than 20 broadcasters and newspapers have been contacted following an approach based in [6]. Among those, only the main Basque newspaper, Berria[6], has agreed to clear its copyright for sharing purposes. Our experience has shown that most data owners fear undue competition and brand damaging derived from misuse of their data and, in general, rather not risk. Although data sharing negotiations with some IPR owners are still pending, we do not expect big changes from the figures reported in Table 5 by the end of the project.

The commercial license established for the SAVAS sharable resources is the META-SHARE C_NoReD_FF[7]

[3] http://www.elda.org/

[4] http://www.meta-share.eu/

[5] FP7-ICT-2011-SME-DCL

[6] http://www.berria.info/

[7] Commercial_NoReDistribution_For-a-Fee

Basque	Language Resource		Amount	Commercial license	Research license
	Audio	Broadcast news	200h	C-NoReD-FF	CC BY NC SA
		Sports	20h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	143M	C-NoReD-FF	CC BY NC SA
		Crawled news	176M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	5h+5h	C-NoReD-FF	CC BY NC SA
Spanish	Audio	Broadcast news	200 h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	178M	C-NoReD-FF	CC BY NC SA
		Crawled news	17M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	5h+5h	C-NoReD-FF	CC BY NC SA
Portuguese	Audio	Interview/debate	20h	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Italian	Audio	Broadcast news	150h	C-NoReD-FF	C-NoReD-FF
	Text	Crawled news	9M	C-NoReD-FF	C-NoReD-FF
	Transcribed audio	Test sets I and II	4h+4h	C-NoReD-FF	C-NoReD-FF
Swiss Italian	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	6M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Swiss French	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	31M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA
Swiss German	Audio	Broadcast news	50h	C-NoReD-FF	CC BY NC SA
	Text	Autocues, scripts and subtitles	32M	C-NoReD-FF	CC BY NC SA
	Transcribed audio	Test sets I and II	2h+2h	C-NoReD-FF	CC BY NC SA

Table 5. SAVAS META-SHARE repository data

license[8]. On the other hand, the Creative Commons CC BY NC SA license has been established mainly for research purposes. In a nutshell, language resources with the C-NoReD-FF licensing schema cannot be redistributed, require the payment of a fee and allow derivatives which can be used for commercial purposes. Resources with CC BY NC SA schemes cannot be redistributed either, require attribution, are free and allow derivatives which can only be used for non-commercial purposes and need to be shared under the same terms.

5. Conclusions and future work

This paper has described the data collection, annotation and sharing activities of the SAVAS project. A considerable amount of audio and text data has been collected for each of the targeted languages. In addition, the followed annotation methodology has managed to ensure high quality annotations and improve the productivity of the task. Finally, the consortium has worked to share the greatest number of compiled resources with the rest of the community. The compiled SAVAS META-SHARE repository can be currently considered one of the biggest available multilingual audio and text data sources exploitable for LVCSR development.

The final types, amounts and license schemes of the shared resources do not match the total collected for two main reasons. On the one hand, data can be considered a commercial asset by owners looking into its exploitation. On the other hand, data owners fear competition and damage from its misuse. We believe considerable work remains to be done to inculcate the data sharing culture among data owners. Future data collection, annotation and sharing approaches of this kind aiming to achieve higher impact in data compilation for open technology research and development should devote special efforts to negotiate and clear copyright issues with the corresponding data owners.

[8] http://www.meta-net.eu/meta-share/licenses

6. Acknowledgements

The authors wish to thank all the data owners who have agreed to share the resources described in this article. SAVAS is funded through the EU FP7 SME-DCL Programme (2012-2014), under grant agreement 296371.

7. References

[1] C. Aliprandi, C. Scudellari, I. Gallucci, N. Piccinini, M. Raffaelli, A. del Pozo, A. Álvarez, H. Arzelus, R. Cassaca, T. Luis, J. Neto, C. Mendes, S. Paulo and M. Viveiros, "Automatic Live Subtitling: state of the art, expectations and current trends", in Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies, Las Vegas, April 2014

[2] H. Meinedo, M. Viveiros and J. Neto, "Evaluation of a Live Broadcast News Subtitling System for Portuguese", in Proceedings of Interspeech, Brisbane, Australia, 2008.

[3] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso and J. Neto, "The L2F Broadcast News Speech Recognition System", in Proceedings of Fala, Vigo, Spain, 2010

[4] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production", Speech Communication special issue on Speech Annotation and Corpus Tools, vol. 33, no. 1-2, January 2000

[5] C. Parra, N. Bel, V. Quochi, "Survey and assessment of methods for the automatic construction of LRs", deliverable D6.1a, FlaReNEt, September 2009

[6] O. De Clercq and M. Montero Perez, "Data Collection and IPR in Multilingual Parallel Corpora. Dutch Parallel Corpus", in Proceedings of LREC, Malta, 2010

7.4 Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks

- **Authors:** Germán Bordel, Mikel Peñagarikano, Luís Rodríguez-Fuentes, Aitor Álvarez, and Amparo Varona
- **Journal:** IEEE Signal Processing Letters
- **Year:** 2016
- **Publisher:** IEEE
- **DOI:** 10.1109/LSP.2015.2505140

Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks

Germán Bordel, *Member, IEEE*, Mikel Penagarikano, *Member, IEEE*, Luis Javier Rodríguez-Fuentes, *Member, IEEE*, Aitor Álvarez, and Amparo Varona, *Member, IEEE*

***Abstract*—The synchronization of text transcripts with audio tracks is typically solved by forced alignment at the phonetic level. However, when dealing with either very long audio tracks or acoustically inaccurate text transcripts, more complex methods are needed, usually based on heavy and costly ASR systems. In a previous work, we showed that a simple and lightweight method could be effectively applied, based on a free phonetic decoding of the speech signal and the alignment of the free and reference phonetic sequences, allowing the transfer of timestamps from the former to the latter. This method has yielded competitive results on the Hub4-97 dataset and is currently applied to synchronize the videos and minutes of the Basque Parliament plenary sessions. In this paper, probabilistic kernels (similarity functions) are applied, based on the hypothesis that a confusion matrix computed from a large corpus of speech conveys key information about the behavior of the phonetic decoder, and that the probabilistic interpretation of this information may help design informative kernels leading to improved alignments. The probabilistic kernels proposed in this work outperform our baseline kernels and other alternatives, including a reference ASR-based approach and a knowledge-based kernel, in experiments on the Hub4-97 dataset.**

***Index Terms*—Long audio tracks, probabilistic kernel, text-to-speech alignment.**

I. Introduction

THE need to technically address accessibility issues is making video captioning an increasingly demanding area for speech and language technologies. In 2007, the European Parliament established new rules for the audiovisual industry that posed a significant pressure to the video producers and broadcasters. At that time, there was no economically reasonable solutions to fit the new exigence levels: manual processing would be too intensive and costly, whereas automatic technology was not applicable due to the low quality of automatic speech recognition (ASR) techniques when applied without restrictions. In these situations, a mixed approach can be applied by manually producing a text transcript and then automatically aligning audio and text. In fact, this problem had been already addressed in the late nineties by several authors [1][2]. Shortly afterwards, the method proposed in [3] achieved good quality results, provided that relatively clean speech segments and accurate text transcripts were available. The approach worked by first detecting fairly good matches between the text and the output of an ASR system customized to the transcript, and then recursively processing smaller portions of the problem. More recently, other related works have been reported which address different objectives, such as correcting human transcripts [4], or dealing with highly imperfect transcripts [5]. In [6], an open source toolkit was presented based on a complex strategy similar to [3]. More recently, different approaches aiming to also obtain simple alignment procedures have been proposed [7], [8], [9], mostly related to the generation of resources for training ASR systems.

The method proposed in [3] was really successful but required a large amount of resources and was computationally costly. In [10], the same author presented a different approach, where the segmentation of long sequences was performed at the acoustic level, rejecting noisy regions, and focusing on the nearly-forced alignment of the selected portions. In [11][12], we applied an unconstrained phonetic decoder to a long speech signal and aligned the recognized sequence of phones to the phonetic transcript derived from the reference text (see Fig. 1). The accuracy figures obtained in word-level alignment experiments on the widely known Hub4-97 dataset [13] were only slightly lower than those reported in [3]. Being fast and comparatively light in terms of both computational costs and required resources, we have been using this system since 2010 to add subtitles to the videos of plenary sessions of the Basque Parliament. These sessions include speech in both Spanish and Basque, and speakers switch frequently from one to the other. Transcripts come from the official minutes, which are clean versions of the speech actually uttered (sometimes disfluent or grammatically incorrect). Note that the lack of correspondence between text transcripts and audio recordings increases the difficulty of the alignment task.

The two alignment strategies applied in [11] were: (1) *maximum number of matches* (*MaxMatch*), which is related to the longest common subsequence problem [14]; and (2) the *minimum edit (Levenshtein) distance* (*MinDist*) [15]. Taken as an optimization problem, the first strategy evaluates exact symbol matches as positive events and tries to maximize the number of them, whereas the second strategy evaluates mismatches (substitutions, deletions and insertions) as negative events and tries to minimize the number of them. When comparing both

Manuscript received June 26, 2015; revised November 19, 2015; accepted November 27, 2015. Date of publication December 03, 2015; date of current version December 10, 2015. This work was supported by the University of the Basque Country, Spain, under Grant GIU13/28 . The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Frederic Bechet.

G. Bordel, M. Penagarikano, L. J. Rodríguez-Fuentes, and A. Varona are with the Department of Electricity and Electronics, University of the Basque Country, UPV/EHU, 48940 Leioa, Spain (e-mail: german.bordel@ehu.es; mikel.penagarikano@ehu.es; luisjavier.rodriguez@ehu.es; amparo.varona@ehu.es).

A. Álvarez is with Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain (e-mail: aalvarez@vicomtech.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2505140

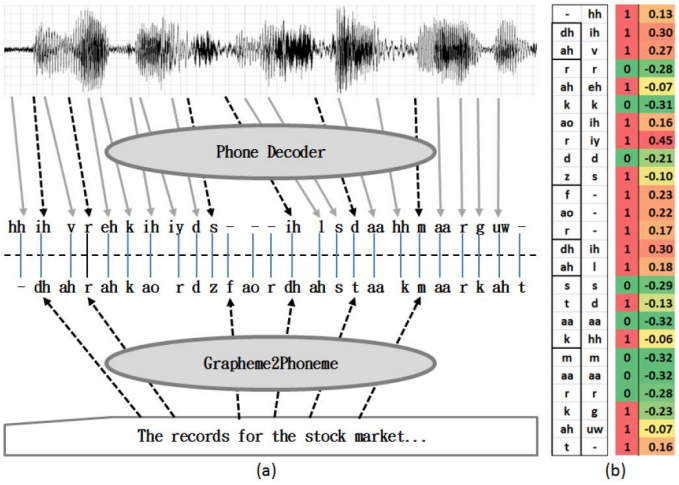


Fig. 1. Speech-to-text alignment is performed in this work by aligning a recognized phone sequence to the phonetic transcript of the reference text (a). A straightforward approach consists in minimizing a distance metric like Levehnstein's, but more informative metrics may lead to better results (b).

strategies in terms of time deviations with regard to a reference ground-truth, we observed that the *MinDist* approach provided slightly better performance than the *MaxMatch* approach.

As suggested in [11], a potential way for improvement is the use of information about the phone decoder confusion probabilities (see Fig. 1). In [16], this information is used along with two knowledge-based probability functions.

In this paper, we present an alternative approach, where the information provided by the phone decoder confusions is transformed according to probabilistic criteria. Three baseline systems will be considered for comparison: the *MaxMatch* and *MinDist* systems, and the ASR-based system described in [3].

II. Database and Phonetic Decoder

The Hub4-97 database contains about 3 hours of transcribed broadcast audio, classified into six categories according to the acoustic conditions, plus a seventh category (*Other*) and an additional set of *Unclassified* segments (see Table I).

The phonetic decoder used in this work is based on a 40-phone set for English. Acoustic models were initialized on the TIMIT database [17] and then re-estimated on the Wall Street Journal database [18]. Left-to-right monophone continuous Hidden Markov Models, with three states and 64 Gaussian distributions per state, were used. It must be noted that, to keep things simple, the decoder was not particularly suited to the Hub4-97 dataset, that is, no adaptation was performed. Phone error rates attained on the six main Hub4-97 categories are shown in Table I. We would like to emphasize, at this point, that our alignment approach is able to achieve competitive results even though a low performance phone decoder is used.

III. Confusion Matrix and Kernels

Let us consider a speech corpus for which text transcripts are available. Then, phonetic transcripts can be easily extracted by applying either a pronunciation dictionary or a set of rules (or both). Let Σ be the set of phonetic symbols, $\mathcal{R}$ the reference phonetic transcript of a speech utterance and $\mathcal{H}$ the phonetic sequence hypothesized by a phonetic decoder when processing the speech utterance. Then, by aligning all the pairs of sequences

TABLE I
Hub4-97 Data Distribution per Categories, and Phone Error Rates on the Six Main Categories Using Our Phonetic Decoder

Cat.	Description	Time (Secs.)	Words (count)	PhonERR %
F0	Baseline Broadcast Speech	4405	13040	41.4
F1	Spontaneous Broadcast Speech	1890	6331	48.6
F2	Speech Over Telephone Channels	1495	4672	57.3
F3	Speech in the Presence of Background Music	498	1554	53.3
F4	Speech Under Degraded Acoustic Conditions	1048	3263	48.6
F5	Speech from Non-Native Speakers	232	663	43.5
FX	All other speech (61 transcribed: other languages + 71 non-speech: silence, music, noise...)	1220	2528	—
?	Unclassified (not transcribed: mostly overlapped speech)	140	0	—

$\mathcal{R}$ and $\mathcal{H}$ in a dataset, we can get the counts c_{rh} representing the number of times that symbols $r, h \in \Sigma$ (r coming from $\mathcal{R}$ and h from $\mathcal{H}$) have been paired together in those alignments. Similarly, we can get the counts c_r representing the number of times that a symbol r from $\mathcal{R}$ has not been paired with any symbol (i.e. *deleted*), and the counts c_h representing the number of times that a symbol h from $\mathcal{H}$ has no counterpart in $\mathcal{R}$ (i.e. it has been *inserted*).

The pairings (r, h) can be *matches* ($r = h$) or *substitutions* ($r \neq h$), but this fact is irrelevant for the rest of the work. We introduce the symbol ϵ in order to represent all these counts under a single structure commonly known as a *confusion matrix* ($c : \Sigma \cup \{\epsilon\} \times \Sigma \cup \{\epsilon\} \rightarrow \mathbb{N}$) considering $c_r \equiv c_{r\epsilon}$ and $c_h \equiv c_{\epsilon h}$.

A perfect decoder processing a perfectly labeled corpus should output a sequence of symbols identical to the reference transcript, so that all counts in the confusion matrix would be zero except for the diagonal elements. Any deviation from this unrealistic situation is somehow *informed* by the non-zero counts outside the diagonal.

To take advantage from the information provided by the confusion matrix in the alignment procedure, we must interpret and transform it into a kernel. Here we use the term "kernel" in the sense it is used in pattern analysis, meaning a similarity function over pairs of data. This kernel is the piece of information used by a sequence alignment algorithm to evaluate different possible subsequence alignments. The simplest one will evaluate different ways of confronting just two symbols, each one coming from one of the aligned sequences. This can be done by considering the primitive editing operations: matches, substitutions, deletions and insertions; or by considering different (particularized) evaluations for each symbol, in a similar manner as that considered in the confusion matrix.

When confronting a symbol r from the reference and a symbol h from the hypothesis, the aligner will need three kernel values: K_{rh}, for the case of pairing them, K_r, for the case of deleting r (not taking into account the presence of h), and K_h, for the case of considering h as an insertion (not taking into account the presence of r).

IV. Probabilistic Kernels

The kernels proposed in this work are based on the probabilities of the three alternatives: P_{pair}, the probability of pairing r and h; P_{del}, the probability of deleting r; and P_{ins}, the probability of inserting h. They can be estimated as the frequency of occurrence of each event in the master alignment:

$$P_{pair} = P(pair|r,h) = c_{rh} / \left(c_{rh} + \frac{c_{r\epsilon}}{N} + \frac{c_{\epsilon h}}{N} \right) \quad (1)$$

$$P_{del} = P(del|r) = c_{r\epsilon} / \left(c_{r\epsilon} + \sum_{\forall h} \left(c_{rh} + \frac{c_{\epsilon h}}{N} \right) \right) \quad (2)$$

$$P_{ins} = P(ins|h) = c_{\epsilon h} / \left(c_{\epsilon h} + \sum_{\forall r} \left(c_{rh} + \frac{c_{r\epsilon}}{N} \right) \right) \quad (3)$$

where $N \equiv |\Sigma|$ is the number of phonetic symbols. N is used to normalize the counts in cases where all possible contexts (represented by ϵ) have been added up. Note that these $P_{(\cdot)}$ values do not sum up to one, that is, we are not calculating the probability to take one of the three possible actions given a confronted pair (r, h), since the deletion and insertion probabilities do not depend on both symbols (*i.e.* $P_r \neq P(del|r,h)$, $P_h \neq P(ins|r,h)$). This fact will result in probabilistic kernels where the deletion and insertion values, K_{del} and K_{ins}, will not depend on their corresponding context symbols, h and r, respectively. Hereinafter, any kernel will be defined as:

$$K = \{K_{pair}, K_{del}, K_{ins}\} \quad (4)$$

The sequence alignment algorithm aims to find the path that minimizes or maximizes the cumulated kernel value, depending on the meaning of their figures, that can be *costs* or *benefits* respectively. Regardless of which of the two representations is considered more natural, we can switch from one to the other by inverting the sign of the values. For example, the two baseline kernels considered in this work can be expressed as:

$$K_{MaxMatch} = \{\delta_{rh}, 0, 0\}_{r,h \in \Sigma} \quad (5)$$

$$K^{\text{cost}}_{MinDist} = \{1 - \delta_{rh}, 1, 1\}_{r,h \in \Sigma} \quad (6)$$

where K is a benefit kernel, K^{cost} is a cost kernel, and δ_{rh} is the Kronecker delta ($\delta_{rh} = 1$ if $r = h$; otherwise $\delta_{rh} = 0$).

In this work, we adopt the benefit maximization representation, where the minimum distance kernel is given by:

$$K_{MinDist} = \{\delta_{rh} - 1, -1, -1\}_{r,h \in \Sigma} \quad (7)$$

Next we define the three kernels that will be evaluated in this work.

A. Expected Match Kernel

The *MaxMatch* kernel can be directly generalized to a probabilistic kernel just by using the probability P_{pair} (*i.e.* the probability of r and h being paired in an alignment) instead of δ_{rh}:

$$K_{ExpectedMatch} = \{P_{pair}, 0, 0\} \quad (8)$$

B. Expected Edit Distance Kernel

Similarly to the *MaxMatch* generalization, the *MinDist* kernel can be directly converted to a probabilistic one that minimizes the expected number of editing operations (the expected edit distance). Given the probability P_e of an event (a pairing, a deletion or an insertion), the probability of an edit error is $1 - P_e$, and therefore, the benefits-based probabilistic kernel is:

$$K_{ExpectedDist} = \{P_{pair} - 1, P_{del} - 1, P_{ins} - 1\} \quad (9)$$

TABLE II
ALIGNMENT ACCURACY AT THE WORD LEVEL FOR DIFFERENT KERNELS. THE FIGURES REPRESENT THE PERCENTAGE OF WORDS WITH HYPOTHESIZED BOUNDARIES CLOSER THAN A GIVEN DISTANCE $|x|$ (IN SECONDS) TO THE TRUE BOUNDARIES

$\lvert x\rvert <$ seconds	Max Match	Min Dist	Kondrak	ASR[3]	Expected Match	Expected Dist	Logit
$\lvert x\rvert < 0.1$	78.31	80.58	85.2		87.76	84.91	**89.02**
$\lvert x\rvert < 0.2$	86.75	89.70	92.23		93.64	91.83	**94.4**
$\lvert x\rvert < 0.3$	90.65	93.37	95.04		95.85	94.58	**96.39**
$\lvert x\rvert < 0.4$	93.50	95.74	96.88		97.39	96.51	**97.79**
$\lvert x\rvert < 0.5$	95.24	97.16	97.92	98.5	98.25	97.64	**98.54**
$\lvert x\rvert < 1.0$	98.48	99.31	99.63		99.58	99.58	**99.71**
$\lvert x\rvert < 1.5$	99.35	99.78	99.91		99.82	99.94	**99.94**
$\lvert x\rvert < 2.0$	99.70	99.91	99.97	99.75	99.87	**99.98**	**99.98**

C. Logit Kernel

The direct translation of probabilities to benefits (or costs) is a way of introducing some information that could improve the alignment results, but a kind of logarithmic transformation converting the multiplicative probabilities into additive benefits could better match the nature of the algorithm. Moreover, given the symmetry of costs and benefits, a symmetric function mapping the [0, 1] domain to $(-\infty, \infty)$ seems a good candidate to be tried. The logit function: $\text{logit}(p) = \log(p(1-p))$ fulfills these requirements, so we propose a logit kernel:

$$K_{Logit} = \{\text{logit}(P_{pair}), \text{logit}(P_{del}), \text{logit}(P_{ins})\} \quad (10)$$

.

V. EXPERIMENTAL RESULTS

The different approaches presented in Section IV have been tested on the Hub4-97 dataset, using a confusion matrix estimated on the WSJ training corpus. The alignment effectiveness is evaluated in terms of the time deviation of each word hypothesized boundary with respect to the reference timestamps. This reference was built by carefully performing forced alignment in small pieces and checking the results by both hearing and visualizing them. The linguistic knowledge-based *Kondrak* kernel [16][19] was also tested in order to evaluate the amount of phonetic information provided by the probabilistic kernels.

Table II presents the results of the tested kernels for different tolerance intervals. Both the *ExpectedMatch* and the *ExpectedDist* kernels clearly beat their non-probabilistic versions (*MaxMatch* and *MinDist*), whereas the results of the *Kondrak* kernel are halfway between them. The *Logit* kernel gets the most competitive results, outperforming all the previous ones. Furthermore, at a much lower computational cost and required resources, the *Logit* kernel attains the same performance than the reference method [3] for both 0.5 and 2.0 seconds tolerance intervals.

The improvement provided by probabilistic kernels with respect to their non-probabilistic counterparts can be easily explained by the fact that they include information about the confusability of the phone decoder. If two different symbols r and h present a high confusion rate, they should be considered closer to a match than another pair with a lower confusion rate, so the benefit of accepting a substitution should be higher.

On the other hand, given that the simple *ExpectedMatch* kernel beats the *Kondrak* kernel, we can hypothesize that the

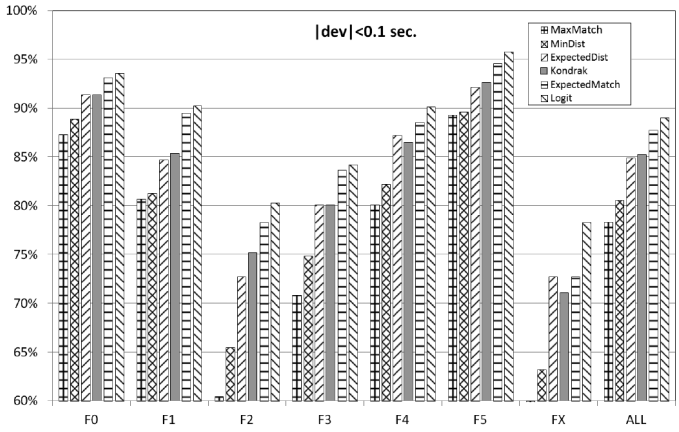


Fig. 2. Alignment accuracy (tolerance interval: 0.1 seconds) for the six kernels considered in this work, broken down by the acoustic conditions of Hub4-97.

information extracted from the confusion matrix is rich enough to avoid the need of any other source of linguistic information.

The high performance of the *Logit* kernel is in accordance with the criteria we adopted to propose it. Given the probability estimate P_e of an event, all the other kernels produce bounded values whereas the *logit* kernel produces values in the range $(-\infty, +\infty)$. This means (for the bounded kernels) that even in the case that $P_e = 0$, this event could be inside the selected path, because the contributions of other pairings can compensate for it. A similar (symmetric) argument could be used for the case $P_e = 1$. On the contrary, the *Logit* kernel value for $P_e = 0$ is $-\infty$ and, therefore, this event will never belong to the best path alignment, whereas for the case where $P_e = 1$, the kernel value is $+\infty$ and, therefore, this event must be part of the best path alignment.

Fig. 2 shows the results for the six considered kernels and for the highest precision ($|dev| < 0.1$ seconds), broken down by the acoustic condition categories in Hub4-97. As expected, relative differences among kernels are consistent across acoustic conditions. Moreover, the quality of the alignments does not degrade to a higher degree than the quality of the decoded phonetic transcripts. The phone error rate ranges in [41.4, 57.3] whereas the corresponding alignment error ranges in [4.2, 19.7] (for the logit kernel), the size of the interval being approximately 16 in both cases. This shows that the alignment procedure performs reasonable well in all conditions. We should expect, however, that if the acoustic conditions were extremely adverse, the decoded phonetic transcript may become almost random, and the alignment procedure would probably fail.

VI. Conclusion

In this work, we have shown that the characterization of the phonetic decoder by means of a confusion matrix can be effectively used to estimate an informative kernel able to achieve remarkable performance gains in text-to-speech alignment of long audio tracks. The results obtained with a kernel based on phonetic knowledge reveal that the information conveyed by the confusion matrix is more precise and/or more specific to the particular task, leading to better performance.

As a potential line for future research, after using the system to align a speech signal to its text transcript for the first time, a new confusion matrix may be estimated and used to realign the sequences. In this way, the second matrix would be adapted to the specific problem being treated. The new matrix would carry information about specific phonetic confusability due to the particular acoustic conditions of the signal, and about any specificity of the text sequence, which may eventually lead to a better result in the new realignment. It would be interesting to check whether this iterative procedure would effectively converge to a stable point with maximum alignment performance.

References

[1] C. W. Wightman and D. T. Talkin, "The aligner: Text-to-speech alignment using Markov models," in *Progress in Speech Synthesis*. Berlin, Germany: Springer, 1997, pp. 313–323.

[2] J. Robert-Ribes and R. Mukhtar, "Automatic generation of hyperlinks between audio and transcript," in *Fifth Eur. Conf. Speech Communication and Technology*, 1997, pp. 903–906.

[3] P. Moreno, C. Joerg, J. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth Int. Conf. Spoken Language Processing*, 1998.

[4] T. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. Interspeech*, 2006, pp. 1606–1609.

[5] A. Haubold and J. Kender, "Alignment of speech to highly imperfect text transcriptions," in *2007 IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 224–227.

[6] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. Workshop on New Tools and Methods for Very Large Scale Research in Phonetic Sciences*, Philadelphia, PA, USA, Jan. 2011.

[7] S. Hoffmann and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," in *Proc. Interspeech 2013*, Lyon, France, Aug. 2013, pp. 1520–1524.

[8] I. Ahmed and S. K. Kopparapu, "Technique for automatic sentence level alignment of long speech and transcripts," in *Proc. Interspeech*, 2013, pp. 1516–1519.

[9] X. Anguera, J. Luque, and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources," in *Proc. Interspeech 2014*, Singapore, Sep. 2014, pp. 1405–1409.

[10] P. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. IEEE ICASSP*, Apr. 2009, pp. 4869–4872.

[11] G. Bordel, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *Proc. Interspeech 2012*, Portland, OR, USA, Sep. 2012.

[12] G. Bordel, S. Nieto, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "Automatic subtitling of the basque parliament plenary sessions videos," in *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 1613–1616.

[13] D. Graff, J. Fiscus, and J. Garofolo, *1997 HUB4 English evaluation speech and transcripts*. Philadelphia, PA, USA: Linguistic Data Consortium, 2002.

[14] D. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Commun. ACM*, vol. 18, no. 6, pp. 341–343, 1975.

[15] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[16] A. Álvarez, H. Arzelus, and P. Ruiz, "Long audio alignment for automatic subtitling using different phone-relatedness measures," in *Proc. IEEE ICASSP*, 2014, pp. 6280–6284.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[18] J. S. Garofolo, D. Graff, D. Paul, and D. S. Pallett, *CSR-I (WSJ0) Complete*. Philadelphia, PA, USA: Linguistic Data Consortium, 2007.

[19] G. Kondrak, "Algorithms for language reconstruction," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2002.

7.5 Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque

- **Authors:** Aitor Álvarez, Pablo Ruiz, and Haritz Arzelus
- **Booktitle:** Proceedings of the 17th International Conference on Text, Speech and Dialogue (TSD)
- **Year:** 2014
- **Publisher:** Springer

Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque

Aitor Álvarez, Pablo Ruiz, and Haritz Arzelus

Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
{aalvarez,pruiz,harzelus}@vicomtech.org

Abstract. A multilingual long audio alignment system is presented in the automatic subtitling domain, supporting English, Spanish and Basque. Pre-recorded contents are recognized at phoneme level through language-dependent triphone-based decoders. In addition, the transcripts are phonetically translated using grapheme-to-phoneme transcriptors. An optimized version of Hirschberg's algorithm performs an alignment between both phoneme sequences to find matches. The correctly aligned phonemes and their time-codes obtained in the recognition step are used as the reference to obtain near-perfectly aligned subtitles. The performance of the alignment algorithm is evaluated using different non-binary scoring matrices based on phone confusion-pairs from each decoder, on phonological similarity and on human perception errors. This system is an evolution of our previous successful system for long audio alignment.

Keywords: Long audio alignment, automatic subtitling, phonological similarity matrices, perceptual confusion matrices.

1 Introduction

Subtitling is one of the most important means to make audiovisual content accessible. To promote accessibility, current European audiovisual law is forcing TV channels to subtitle a huge proportion of their contents. To address this increased demand, broadcasters and subtitlers are seeking alternatives more productive than manual subtitling. Speech recognition technologies have proved useful in this respect. One efficient approach, when the script for the content exists, is speech-text alignment, which relies on aligning audio with its script to automatically recover time stamps. Forced-alignment is challenging with long signals, because of the widely-used Viterbi algorithm, which forms very large lattices during decoding, requiring a lot of memory.

In this work, the system presented in [1] for long audio alignment in an automatic subtitling scenario has been improved and extended. Phone-decoder accuracy was improved using context-dependent acoustic models, besides implementing an adaptation of the generic language models to the script of the contents to subtitle. The system was also extended to Basque, its original languages being English and Spanish, and additional linguistic resources were created for the Spanish aligner.

The paper is structured as follows. Section 2 looks at related work in long audio alignment and in phone-relatedness measures. Section 3 describes our speech-text alignment system, and Section 4 presents the phoneme similarity matrices created.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 473–480, 2014.

Section 5 discusses the evaluation method and results. Section 6 presents conclusions and suggestions for further work.

2 Related Work

The reference for many of the related studies is the work done in [2], where the forced alignment was turned into a recursive and iteratively adapted speech recognition process. They used dynamic programming to align the hypothesis text and the reference transcript at word level. Subsequent works proposed improvements of this system, to deal with scenarios in which transcripts are not exact. In [3] a Driven Decoding Algorithm (DDA) was proposed to simultaneously align and correct the imperfect transcripts. At a new generated assumption of the speech recognizer in the lattice, DDA aligned it with the approximated transcript and a new matching score was computed and integrated with the language model for linguistic rescoring. An efficient, and simpler, long audio alignment approach was presented in [4]. They developed a system based on Hirschberg's dynamic programming algorithm [5] to align the phone decoder output with the transcription at phoneme level. They used a binary matrix to score alignment operations, with a cost of one for insertions, deletions and substitutions, and a cost of 0 for matches. Inspired on [4], for our experiments in [1] we created several scoring matrices, based on criteria like phonological similarity, phone-decoder confusion and phone confusion in human perception.

Concerning literature relevant for the creation of our scoring matrices, our phonological similarity metric is based on [6], where Kondrak constructed a metric that outperformed previously available ones, evaluating it with cognate alignment tasks. The metric was also successfully employed in spoken document retrieval in [7]. Regarding phone confusion in human perception, our American English matrices rely on perceptual error data reported in [8], who used a phoneset that closely corresponds to our phone-decoder's phoneset. Our Spanish data are based on the corpus of misperceptions developed by [9], which provides data covering our entire phoneset.

3 Long Speech-Text Alignment System

The goal of any speech-text alignment system is to obtain a perfect timing synchronization between the source audio and related text recovering the time codes for each word in the transcript. Our multilingual long speech-text alignment system is trained to align long audios and related transcripts for English, Spanish and Basque. For each language, a language-dependent phone decoder was developed, in addition to a grapheme-to-phoneme transcriptor. The aim of the alignment algorithm is to find matches between the phones recognized by the phone-decoder and the reference phoneme transcription. Only the time-codes of the correctly aligned phones will be used as reference times for further synchronization.

However, all the phonemes are not always correctly aligned during alignment; substitutions, deletions and insertions may occur. In fact, using the evaluation contents presented in Section 5, only 34% of the phonemes were correctly aligned for English,

while 57% and 48% of the phonemes were matched for Spanish and Basque respectively in the best-performing configuration. These time-codes at phoneme level are then used to estimate the start time of each word and thus of each subtitle. The promising results presented in this paper prove that the time-codes recovered by the aligner are good enough to generate near-perfectly aligned subtitles.

3.1 Context-Dependent Phone Decoders

The phone-decoders have been improved from the last version of the system presented in [1], in which monophone models were employed. For this study, cross-word triphone models were built for each language to deal with coarticulation effects. With the aim of reducing linguistic variability, the language model consisted of an interpolation of the generic language model and a specific model created for each transcript. The interpolated models were bigram triphone models. The triphone-based phone decoders were trained using the HTK[1] tool. The parametrization of the signal consisted of 18 Mel-Frequency Cepstral Coefficients plus the energy and their delta and delta-delta coefficients, using 16-bit PCM audios sampled at 16 KHz.

The English triphone-based decoder system was built using the TIMIT database [11], which is composed by 5 hours and 23 minutes of clean speech data. Texts totaling 369 million words, gathered from digital newspapers, were used to train the generic language model. The Phone Error Rate (PER) of this decoder was 24.71%.

The Spanish triphone-based decoder system was based on 20 hours of clean-speech from three databases; Albayzin [12], Multext [13], and records of broadcast news contents from the SAVAS corpus [14]. The generic language model was trained with texts crawled from national newspapers, toting up 45 million words. The PER of the Spanish decoder was 31.79%.

The Basque triphone-based decoder system was generated using 36 hours of clean speech records of broadcast news contents. The generic language model was built using texts crawled from national newspaper, totaling 91 million words. The PER of the Basque decoder was 20.92%.

For all three languages, the corpora were split between training and test sets containing 70% and 30% of the data, respectively.

3.2 Grapheme-to-Phoneme Transcriptors

The grapheme-to-phoneme (G2P) transcriptors used for English and Spanish were the same used in the previous work [1]. The Spanish G2P was ruled based and inspired on the tool provided by Lopez[2]. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary[3] using Phonetisaurus[4] tool. The Basque G2P transcriptor was based on manually created heuristic rules. The phonesets for all the languages are available on our project's website[5].

[1] http://htk.eng.cam.ac.uk/

[2] http://www.aucel.com/pln/

[3] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

[4] http://code.google.com/p/phonetisaurus/

[5] http://sites.google.com/site/similaritymatrices/

3.3 Algorithm for Alignment of Phoneme Sequences

Our alignment algorithm is a slightly modified version of the well-known divide-and-conquer Hirschberg's algorithm. These modifications were established once their effectiveness in the alignment process was tested.

Given the two phoneme sequences $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ to be aligned, the algorithm forces them to be recursively divided at indexes x_{mid} and y_{mid} respectively. Hirschberg defined x_{mid} as *round(length(x)/2)*. Nevertheless, following the procedure several candidates can arise for y_{mid}. In our algorithm, y_{mid} always corresponds to the candidate-index closest to the middle of Y. The other modification relies on forcing a substitution operation, even if the phonemes do not match, when the recursive algorithm only has sequences of one symbol left to align.

Four edit-operations are allowed in the alignment algorithm: matches, substitutions, deletions and insertions. The scores for matches and substitutions are defined by the scoring matrices (See Section 4), while deletions and insertions incur a gap penalty. Since each matrix-type tested has a different range of values, the gap penalties are also different for each matrix-type. In our binary matrix, the gap penalty was 2. For all other matrices, the penalty was a quarter of the matrix' maximum value, following one of the practices for gap penalties referenced in [6].

4 Phoneme-Relatedness Scoring Matrices

The phoneme-relatedness matrices provide information to the aligner about how likely it is for an alignment between two phonemes to be correct. The matrices favour aligning similar phonemes, by giving such alignments higher scores than to alignments between less similar phonemes. The matrices give the lowest scores to alignments between highly dissimilar phones, which are unlikely to be correct.

We created different scoring matrices for each language, applying different phoneme-relatedness criteria. The first scoring matrix is decoder-dependent, based on errors made by the phone decoder. The second matrix is decoder-independent, and based on phonological similarity, assessed by comparing largely articulatory features. The final matrix is also decoder-independent, and relies on phoneme confusion in human perception. Samples for all types of matrices are available on our project's website.

4.1 Matrices Based on Phone-Decoding Errors

The matrices were created based on HTK's HResults logs, when aligning the phone-decoding output and the G2P transcription for sequences of approx. 200,000 phonemes in English, 1,000,000 in Spanish and 2,000,000 in Basque. For each phone in the phoneset, the matrices contain the percentages of misrecognitions and correct recognitions by the decoder, normalized to a 1–1000 integer range. For instance, if 4% of the occurences of /ɲ/ were misrecognized as /n/, the matrix shows a score of 40 for the [ɲ,n] phoneme pair. In order to prevent substitutions between phonemes never mistaken by the decoder, a score of −500 was entered in the matrix for such phoneme-pairs. This score corresponds to $1/2 \times (0 - \max(\{\text{Score Range}\}))$.

4.2 Matrices Based on Phonological Similarity

Our phonological similarity scores are based on the metric devised by Kondrak in [6], as part of the ALINE cognate alignment system[6]. Phonemes are described with Ladefoged's [14] multivalued features, and a *salience* factor weights each feature according to its impact for phoneme similarity. The features, values and saliences employed for each language are available on our project's website.

$$\sigma_{\text{sub}}(p, q) = (C_{\text{sub}} - \delta(p, q) - V(p) - V(q))/100$$
$$\text{where } V(p) = \begin{cases} 0 & \text{if } p \text{ is a consonant or } p = q \\ C_{\text{vwl}} & \text{otherwise} \end{cases}$$
$$\delta(p, q) = \sum_{f \in R} diff(p, q, f) \times salience(f)$$
$$\sigma_{skip}(p) = ceiling(|C_{\text{sub}}/400|)$$

Fig. 1. Similarity function, based on Kondrak (2002)

Fig. 1 shows equations with our scoring function. $\sigma_{\text{sub}}(p, q)$ returns the similarity score for phonemes p and q, $C_{\text{sub}}/100$ being the maximum possible similarity score. C_{vwl} represents the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically as described in [15]. The function diff(p, q, f) yields the similarity score between phonemes p and q for feature f, and the feature-set R is configurable. Last, $\sigma_{\text{skip}}(p)$ returns the penalty for insertions and deletions used in the aligner. We defined heuristically a C_{sub} value of 3,500 (i.e. a maximum similarity score of 35), and a gap penalty of 9 for alignment, which corresponds to ceiling($|C_{\text{sub}}/400|$).

Kondrak's original function was designed for cognate alignment. We modified the function, for coherence with our audio aligner, and to adapt it to audio alignment tasks, achieving better results with the modified version than with the original. Details about the modifications are discussed in [1] and in the project's website.

4.3 Matrices Based on Perceptual Errors

The English matrices were based on human perceptual error data from [8]. They performed a phoneme identification study with native speakers of American English, asking them to identify the initial or final phoneme of 645 syllables of types CV (ConsonantVowel) and VC, at signal-to-noise ratios (SNR) of 0, 8 and 16. The noise type was multi-speaker babble. Participants chose a response among several possibilities presented to them visually. The phoneme-set in the study covers all of our decoder's phoneset except schwa. We only used the SNR 16 data, since a matrix based exclusively on this subset of the data yielded better alignment results than when considering data at other SNR for building the matrix.

The Spanish matrix was based on an extended version, provided by the authors directly, of the corpus of human misperceptions in noise developed in [9]. The

[6] ALINE is available at
`http://webdocs.cs.ualberta.ca/~kondrak/#Resources`

methodology involved presenting 69 native speakers of Spanish with over 20,000 single-word stimuli, under different masking-noise conditions, and asking the speakers to write the word they had heard. Only stimuli for which certain agreement thresholds were reached among participants' responses were kept for the final misperception corpus, which consists of 3,294 stimuli and their associated responses. The study is thus a free-response error-elicitation task, not a closed-response task like [8]. However, we chose [9] as our data source, since, unlike other Spanish perception studies, it provides data for all phonemes in our decoder's phoneset. For coherence with our English data, we based our matrices on the 1,838 stimuli where multi-speaker babble was used as the masker. SNR in these stimuli ranged between -8 and $+1$. For computing our confusion matrix, we compared the corpus' stimulus and responses in cases where the response involved a single-phoneme error. We recorded the percentage of matches and mismatches between each stimulus and each response in the stimulus' response-set (a maximum of 15 responses were available per stimulus). Match and mismatch percentages were normalized to a 1–1000 range. For phoneme pairs where no confusion had taken place, a score of -500 (i.e. $1/2 \times (0 - max(\{Score\ Range\})$ was entered in the matrix. The matrix was based on 6,807 stimulus-response pairings.

Perceptual-relatedness matrices were not created for Basque, since we are not aware of appropriate data that could be exploited for their creation.

5 Evaluation and Results

The English test-set totaled 21,310 phonemes, 4,732 words and 471 subtitles, and contained non-clean speech from television audios. Its reference subtitles contained some stretches where transcription was imperfect, with subtitles missing for some parts of the audio. The Spanish test-set consisted of 47,480 phonemes, 8,774 words and 1,249 subtitles, and was composed of clean speech from documentaries. The Basque test-set totaled 26,712 phonemes, 4,331 words and 726 subtitles, containing a concatenation of a documentary and a film, and included noisy-speech.

Long audio alignment accuracies using different phone-relatedness matrices for English, Spanish and Basque are presented in Table 1. The results present the percentage or words and subtitles correctly aligned within the specified deviation range from the reference. The real time-codes at word level were obtained applying a forced-alignment algorithm for each subtitle in the reference material, which was composed of time-coded subtitles manually created by professional subtitlers. For subtitle-level evaluation, the deviation of the first and last words of the subtitles were measured.

The results show the effectiveness of our long audio alignment system, even with contents containing noisy-speech and imperfect transcriptions. Besides, the improvements using non-binary matrices are clearly proved comparing to the accuracies obtained with the binary matrix. Considering that a maximum deviation of 1 second is not long enough for listeners to have difficulties associating the subtitle and the audio, near-perfectly aligned subtitles were obtained for all three languages. In fact, alignment accuracies of 91.30%, 96.72% and 95.18% were obtained for English, Spanish and Basque respectively at this maximum deviation time.

Regarding non-binary matrices performance, the PDE matrices achieve the most accurate alignment results for English and Basque. It was expectable since these

Table 1. Alignment accuracy at word and subtitle level. **PDE**: Phone-decoder-error based matrix, **PHS**: Phonological similarity, **PCE**: Perceptual error matrix.

		Word-level deviation (seconds)					Subtitle-level deviation (seconds)					
	Matrix	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**	**Matrix**
English	Binary	0.25	8.13	28.12	40.36	56.14	0.42	4.46	38.64	83.65	100	Binary
	PDE	**1.02**	**29.88**	**60.17**	**72.94**	**84.52**	**0.85**	**14.86**	**54.35**	**91.30**	100	PDE
	PHS	0.87	25.41	56.10	69.55	79.73	0.64	11.25	53.50	88.54	100	PHS
	PCE	0.72	26.66	57.26	70.31	82.57	0.64	14.65	53.29	90.02	100	PCE
Spanish	Binary	2.47	47.70	69.11	75.55	80.21	0.48	21.06	63.49	92.47	100	Binary
	PDE	**5.55**	77.42	**92.21**	**94.39**	95.93	1.12	40.19	**80.22**	96.64	100	PDE
	PHS	5.44	**77.45**	92.17	94.31	**95.97**	**1.20**	**40.67**	80.14	96.64	100	PHS
	PCE	5.22	74.83	92.03	94.48	96.35	1.28	38.83	78.78	**96.72**	100	PCE
Basque	Binary	1.55	34.98	56.63	61.06	65.25	0.83	24.24	64.05	92.29	100	Binary
	PDE	2.34	48.91	76.21	80.91	85.05	**1.65**	**44.63**	**75.76**	**95.18**	100	PDE
	PHS	**2.49**	**49.97**	**77.00**	**82.10**	**86.45**	1.38	35.54	74.24	95.04	100	PHS

matrices were based on each phone-decoder phone confusion-pairs. However, the improvements with the PDE matrix comparing to improvements with the other non-binary matrices are not relevant. For Spanish, the PCE matrix obtained the best results, although the PDE and PHS matrices achieved very similar accuracies.

6 Conclusions and Further Work

The adequate performance of our multilingual long audio alignment system in the automatic subtitling scenario was presented in this work. We established the effectiveness of a customized version of the well-known Hirschberg algorithm, and proved that using several scoring matrices based on different phoneme-relatedness criteria obtains well-performed alignments.

Since the current system works with triphone-based phone decoders, ongoing work is focused on the development of context-dependent phoneme scoring matrices. The goal behind this approach will be to improve the alignment process considering not only phones, but also biphones and triphones, to deal with coarticulation effects.

References

1. Álvarez, A., Arzelus, H., Ruiz, P.: Long audio alignment for automatic subtitling using different phone-relatedness measures. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Florence, Italy (2014)
2. Moreno, P.J., Joerg, C., Van Thong, J.-M., Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP, Sydney, Australia (1998)
3. Lecouteux, B., Linàres, G., Nocéra, P., Bonastre, J.: Imperfect transcript driven speech recognition. In: Proceedings of INTERSPEECH, pp. 1626–1629 (2006)
4. Bordel, G., Nieto, S., Peñagarikano, M., Rodríguez-Fuentes, L.J., Varona, A.: A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In: Proceedings of INTERSPEECH, Portland, Oregon (2012)

5. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM 18(6), 341–343 (1975)
6. Kondrak, G.: Algorithms for Language Reconstruction. PhD Thesis. University of Toronto (2002)
7. Comas, P.: Factoid Question Answering for Spoken Documents. PhD Thesis. Universitat Politècnica de Catalunya (2012)
8. Cutler, A., Weber, A., Smits, R., Cooper, N.: Patterns of English phoneme confusions by native and non-native listeners. Journal of the Acoustical Society of America 116(6), 3668–3678 (2004)
9. García Lecumberri, M.L., Toth, A.M., Tang, Y., Cooke, M.: Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish. In: Proceedings of INTERSPEECH, pp. 2807–2811 (2013)
10. Garafolo, J.S.L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia (1993)
11. Díaz, J.E., Peinado, A., Rubio, A., Segarra, E., Prieto, N., Casacuberta, F.: Albayzín: a task-oriented Spanish speech corpus. In: Proceedings of LREC, Granada, Spain (1998)
12. Campione, E., Véronis, J.: A multilingual prosodic database. In: Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP, Sydney, Australia (1998)
13. Del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J.P., Paulo, S., Piccinini, N., Rafaelli, M.: SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. In: Proceedings of LREC, Reykjavik, Iceland (2014)
14. Ladefoged, P.: A Course in Phonetics. Harcourt Brace Jovanovich, New York (1995)
15. Ruiz, P., Álvarez, A., Arzelus, H.: Phoneme similarity matrices to improve long audio alignment for automatic subtitling. In: Proceedings of LREC, Reykjavik, Iceland (2014)

7.6 Long audio alignment for automatic subtitling using different phone-relatedness measures

- **Authors:** Aitor Álvarez, Haritz Arzelus, and Pablo Ruiz
- **Booktitle:** Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- **Year:** 2014
- **Organization:** IEEE

LONG AUDIO ALIGNMENT FOR AUTOMATIC SUBTITLING USING DIFFERENT PHONE-RELATEDNESS MEASURES

Aitor Álvarez, Haritz Arzelus, Pablo Ruiz

Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
{aalvarez,harzelus,pruiz}@vicomtech.org

ABSTRACT

In this work, long audio alignment systems for Spanish and English are presented in an automatic subtitling scenario. Pre-recorded contents are automatically recognized at phoneme level by language-dependent phone decoders. A dynamic-programming alignment algorithm finds matches between the automatically decoded phones and the ones in the phonetic transcription for the content's script. The accuracy of the alignment algorithm is evaluated when applying three non-binary scoring matrices based on phone confusion-pairs from each phone decoder, on phonological similarity and on human perception errors. Alignment results with the three continuous-score matrices are compared to results with a baseline binary matrix, at word and subtitle levels. The non-binary matrices achieved clearly better results. Matrix samples are given in the project's website.

Index Terms— Long audio alignment, phonological similarity matrices, perceptual confusion matrices, automatic subtitling.

1. INTRODUCTION

Due to the huge subtitling demand generated by current accessibility policies, broadcasters and subtitling companies are looking for solutions to automate subtitling.

Speech processing technologies are proving helpful in speeding up the subtitling process. A widespread approach for subtitling pre-recorded contents exploits existing text transcriptions for the content (scripts). Under this approach, automatic speech-text alignment systems recover word-level time-codes from the audio for the scripts. Although speech-text alignment is an interesting approach for automatic subtitling, aligning long audio signals is challenging, given memory demands, processing time, and the decreased reliability of the commonly employed Viterbi search algorithm when aligning long sequences.

For the present work, the successful system for long audio alignment described in [1] was taken as the basis. Their alignment method was based on Hirschberg's algorithm [2], using a binary matrix for scoring alignment operations. Our study deployed a similar algorithm, but using three types of non-binary scoring matrices, based on different phoneme-relatedness criteria. Alignment systems were developed for Spanish and English.

The paper is structured as follows. Section 2 looks at related work in long audio alignment and in phoneme-relatedness measures. Section 3 presents our system, and Section 4 describes the similarity matrices created. Section 5 discusses evaluation methods and results. Section 6 presents conclusions and suggestions for future work.

2. RELATED WORK

Speech-text alignment has been extensively studied. Many studies follow work by [3], where forced alignment was turned into a recursive speech recognition process, iteratively adapted to the content. Dynamic programming was used to align the hypothesis text and the reference transcript at word level. Subsequent works proposed improvements to this system: [4], [5].

A different approach, which does not require adapting the models and vocabulary, is in [1]. They developed an aligner based on Hirschberg's algorithm, a dynamic programming algorithm used in bioinformatics for genetic sequence alignment. They used a binary matrix to score alignment operations: insertions, deletions and substitutions had a cost of 1, while matches received a score of 0.

Whereas [1] used binary matrices, our study tested non-binary scoring matrices, based on phone-confusion ratios in our phone decoder, on phonological similarity, and on phone confusion in human perception. Our phonological similarity metric was based on [6], where a metric was presented that outperformed previously existing measures, applied to the task of cognate alignment. The metric was also successfully employed in spoken document retrieval [7]. Regarding phone confusion in human perception, [8] provided phone confusion results for American English, using a phoneset that is very close to our aligner's phoneset.

3. LONG AUDIO ALIGNMENT SYSTEM

The goal of an audio alignment system is to recover time-codes from the audio for words in the audio's script. To this end, our speech-text alignment system aligns two sequences of phonemes obtained from different sources. A language-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

dependent phone-decoder recognizes the phones and their time-codes from the audio. The decoder's output usually contains mistakes due to common recognition errors. Besides, a grapheme-to-phoneme module translates the input transcript into the reference phoneme transcription. An alignment algorithm finds phoneme matches between the phones recognized by the phone-decoder and the reference phoneme transcription. Correctly aligned phonemes will receive the time-codes obtained by the phone-decoder. Phonemes are not always correctly aligned; substitutions, deletions and insertion errors may occur. Nonetheless, the results of this study suggest that the number of matching phonemes found by our aligner is sufficient to recover enough time-codes to create near-perfectly aligned subtitles.

3.1. Phone recognition system

The phone recognition systems were trained using HTK, a toolkit for building hidden Markov models. The acoustic models were based on a monophone model, with three left-to-right emitting states using 32 Gaussian mixture components. The language models were bigram phoneme models. The parametrization of the signal consisted of 18 Mel-Frequency Cepstral Coefficients plus the energy and their delta and delta-delta coefficients, using 16-bit PCM audios sampled at 16 KHz.

The Spanish phone recognition system was trained and tested with 20 hours of audios from three databases; Albayzín [9], Multext [10] and records of clean-speech broadcast news contents. The contents were mixed and divided into training (70%) and test (30%) sets. Texts totaling 45 million words were crawled from national newspapers to train the language model. The Phone Error Rate (PER) for the Spanish phone-decoder was 40.65%.

The English phone recognition system was trained and tested on the TIMIT database [11], which consists of 5 hours and 23 minutes of speech data. 70% of the database was used for training, leaving the rest for testing. Texts totaling 369 million words, collected from digital newspapers, were used to train the language model. The PER for the English phone-decoder was 35.52%.

3.2. Grapheme-to-phoneme transcriptors

Two language-dependent grapheme-to-phoneme (G2P) transcriptors were developed for Spanish and English. The Spanish transcriptor was rule-based. It was inspired on the tool provided by López (*www.aucel.com/pln/*), and adapted to our phonelist. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary (*svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/*) using Phonetisaurus (*code.google.com/p/phonetisaurus/*), a grapheme-to-phoneme framework driven by Weighted Finite State Transducers (WFST). The Spanish and English phonesets are available on our project's website (see *sites.google.com/site/similaritymatrices/*).

3.3. Algorithm for long sequences alignment

For our study, Hirschberg's algorithm was modified in two respects. (1) Given the sequences $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ which have to be recursively divided at indexes x_{mid} and y_{mid} respectively, Hirschberg defined that x_{mid} will always correspond to X's middle index, i.e. *round(length(X)/2)*. However, for sequence Y, when determining the y_{mid} index with Hirschberg's function, several candidate-indexes can arise. Our modification consists in always choosing as y_{mid} the candidate-index that is closest to the middle of Y. (2) During the recursive application of the algorithm, when both sequences have a length of 1 and their phones do not match, a substitution operation is forced, in order to avoid excessive gaps. Both modifications were applied once their effectiveness was established.

In our alignment algorithm, four edit-operations are allowed: matches, substitutions, deletions and insertions. The scores for the first two operations are defined by the scoring matrices (See Section 4), while deletions and insertions incur a gap penalty. Since each matrix-type tested has a different range of values, the gap penalties are also different for each matrix-type. In our binary matrix, the gap penalty was 2. For all other matrices, the penalty was a quarter of the matrix' maximum value, following one of the practices for gap penalties referenced in [6].

4. SIMILARITY MATRICES

Our scoring matrices provide the alignment algorithm with phoneme-similarity information. Depending on the scoring matrix, the alignment will be different, since scores for phoneme matches and mismatches (substitutions) differ across matrices. We developed three types of matrices, applying different phoneme-relatedness criteria. These matrices can help the alignment algorithm consider mismatches between similar phonemes as possible correct substitutions. They also prevent substitutions between very dissimilar phonemes, which are unlikely to be correct. The phone-decoder error based matrix achieves these ends by providing information about the phone decoder's phone confusion-pairs. The phonological similarity matrix is decoder-independent and estimates similarity based on common articulatory characteristics between phonemes. The perceptual matrix, also decoder-independent, reflects phone confusion-pairs in human perception. Its use is justified by the way the signal was parameterized: The frequency warping scale used for filter spacing in MFCC computation is the Mel scale, which was originally created through human perception experiments. Our project's website provides samples for the three types of matrices.

Alignment results with the three matrix-types were compared to results with a baseline binary matrix where matches had a score of 1 and mismatches had a score of 0.

4. 1. Phone-decoder error-based matrices

Our phone-decoder error-based matrix was computed from the output of HTK's HResults tool, when aligning the phonetic recognition and G2P transcription for sequences of ca. 25000 phonemes in Spanish and ca. 12700 phonemes in English. The matrix represents the percentage of times each phone in the phoneset was recognized correctly or misrecognized. Percentages were normalized to a 1-1000 integer range. E.g. if 8.5% of the occurrences of /θ/ were misrecognized as /f/, the matrix shows a score of 85 for phoneme pair [θ, f]. For pairs where phonemes were never mistaken for each other, we stipulated a score of –500, i.e. *½ × (0 – max({Score Range})*, preventing substitutions between members of such phoneme pairs.

4.2. Phonological similarity matrices

The phonological similarity scores were based on the metric described by Kondrak [6] as part of a cognate alignment system. Phonemes are described with Ladefoged's [12] multivalued features, and weighted according to their *salience*: the feature's impact for similarity. Our feature set, feature and salience values are on our project's website (see *sites.google.com/site/similaritymatrices/*).

$$\sigma_{sub}(p,q) = (C_{sub} - \delta(p,q) - V(p) - V(q)) / 100$$

where

$$V(p) = \begin{cases} 0 \text{ if p is a consonant or p = q} \\ C_{vwl} \text{otherwise} \end{cases}$$

$$\delta(p,q) = \sum_{f \in R} \text{diff}(p,q,f) \times \text{salience}(f)$$

$$\sigma_{skip}(p) = \text{ceiling}(|C_{sub} / 400|)$$

Figure 1: Similarity function, based on Kondrak (2002)

Figure 1 shows our scoring function: $\sigma_{sub}(p, q)$ yields the similarity score for segments p and q. $C_{sub}/100$ is the maximum possible similarity score. C_{vwl} defines the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically. Function diff(p, q, f) returns the difference between segments p and q for feature f. Feature-set R is configurable. Finally, $\sigma_{skip}(p)$ returns the penalty for insertions and deletions used in the aligner (see Section 3.3).

We modified Kondrak's original function, making it more suitable for audio alignment, and for coherence with our aligner. First, the definition of *V(p)* was modified. Note that, as C_{vwl}, approaches 0, scores for vowel matches become closer to scores for consonant matches, increasing the weight of vowels in alignment. Kondrak mentions that prioritizing consonant matches is desirable in cognate alignment. Nonetheless, for audio alignment we obtained slightly better results by assigning the same weight to all matches. This can be achieved in the original function by setting $C_{vwl} = 0$. However, as C_{vwl} approaches 0, substitutions between vowels and consonants become less clearly penalized by the matrix, which is undesirable. By adding the *or*-clause "*or p = q"* in the definition of *V(p)*, we can give equal weight to all matches, while still setting $C_{vwl} > 0$, and thus still applying an extra penalty to vowel/consonant substitutions that is not applied to vowel/vowel substitutions.

Further modifications were the following. First, adding a denominator of 100 to σ_{sub}, in order to keep Kondrak's output range, but using integer feature values and avoiding decimals to reduce memory use. Second, redefining σ_{skip}, for coherence with the way the gap penalty is calculated (see Section 3.3) when aligning with the perceptual and decoding-error based matrices.

The final modification was omitting a clause from the original function, which evaluates two-to-one phoneme alignments. These are not implemented in our audio aligner.

We defined heuristically a C_{sub} value of 3500, yielding a maximum possible similarity score of 35 ($C_{sub}/100$), and a gap penalty of 9 for alignment: *ceiling(|C_{sub}/400|)*.

4.3. Perceptual similarity matrices

We created these matrices for English only. The scores were based on perceptual confusion matrices from [8]. They asked native speakers of English to identify 645 CV (ConsonantVowel) and VC syllables containing a phoneme from a 39-phoneme set (covering all of our phoneset but schwa), at signal-to-noise ratios (SNR) of 0, 8 and 16. Our scores reflect confusion percentages at SNR 16; the scoring matrix thus obtained yielded better results on our test-set than data at other SNR.

We normalized the confusion percentages for each phoneme-pair into a 1-1000 range. For phoneme-pairs where no confusion had taken place, we stipulated a score of –500, i.e. *½ × (0 – max({Score Range})*.

5. EVALUATION AND RESULTS

Our long audio alignment system was evaluated at word and subtitle level. The Spanish test-set (47,480 phonemes; 8,774 words and 1,249 subtitles) was composed of clean-speech audios from films. By contrast, the English test-set (21,310 phonemes; 4,732 words and 471 subtitles) consisted of non-clean speech from television audios, containing disfluencies, music, noise and overlapping speech. In addition, the English reference contained segments with imperfect transcriptions, missing subtitles for certain parts of the audio. Due to these difficulties, lower accuracy in English was expected and observed at all evaluation levels.

Matrix type	*Eval Level*	*0*	*≤0.1*	*≤0.5*	*≤1.0*	*≤2.0*
Binary Baseline	WL	14.15	57.82	72.68	76.20	79.02
	SL	10.57	45.24	73.34	94.96	100
Phone-Decoder Error-Based *(PDE)*	WL	**23.34**	**82.28**	**94.42**	**95.99**	**97.21**
	SL	**18.01**	**66.85**	**87.99**	**98.96**	**100**
Phonological Similarity *(PHS)*	WL	23.00	82.16	93.65	95.58	96.92
	SL	17.85	66.53	**87.99**	98.80	100

Table 1: **Spanish** word-level (**WL**) and subtitle-level (**SL**) alignment accuracy. Percentage of words and subtitles aligned within each deviation range from reference.

Matrix type	*Eval Level*	*0*	*≤0.1*	*≤0.5*	*≤1.0*	*≤2.0*
Binary Baseline	WL	0.28	4.81	19.02	29.67	43.20
	SL	0.21	4.03	36.94	84.08	100
Phone-Decoder Error-Based *(PDE)*	WL	**2.20**	**25.73**	**52.53**	**65.35**	76.31
	SL	**0.42**	**11.46**	**48.83**	**88.32**	**100**
Phonological Similarity *(PHS)*	WL	1.97	24.23	49.81	62.71	73.89
	SL	0.42	8.28	43.10	85.99	100
Perceptual Error-Based *(PCE)*	WL	1.89	23.76	50.54	63.93	**76.44**
	SL	0.21	8.92	47.98	**88.32**	100

Table 2: **English** word-level (**WL**) and subtitle-level (**SL**) alignment accuracy. Percentage of words and subtitles aligned within each deviation range from reference.

Tables 1 and 2 present the alignment accuracy at word and subtitle level for Spanish and English. We adopted the evaluation method from [3] and [1]. The cumulative percentage of correctly aligned words within several deviation ranges was recorded: Column 0 presents the percentage of perfectly aligned words, column ≤0.1 means the percentage of words correctly aligned within a maximum deviation of 0.1 sec, etc. To obtain the actual word-level time-codes for the reference files, a forced alignment was done subtitle by subtitle using the reference material, which contained subtitles manually created by professional subtitlers, as well as their time-codes. For subtitle-level evaluation, the deviation of the subtitle's first and last word compared to the reference was measured.

The most salient conclusion supported by the results is that using matrices based on phone-decoder errors (PDE), phonological similarity (PHS) or perceptual errors (PCE) significantly improved alignment accuracy compared to using the binary matrices. The improvements are noticeable at all deviation ranges.

For both languages, the best alignment accuracy was obtained using the PDE matrix. This was expectable, since the matrix is based on phone confusion-pairs from the phone-decoder used by the aligner.

In Spanish, with the PDE matrix, accuracy gains of 21 and 14 percentage points (ptp) were obtained at subtitle-level compared to the binary matrix, at 0.1 and 0.5 second deviations respectively. Considering that 0.1 and 0.5 sec. are acceptable deviations for subtitling applications, these gains represent a positive impact at an actual application level. Improvements were even higher in word-alignment accuracy: 37 ptp and 21 ptp at the same deviation ranges.

For English, alignment accuracies are lower, given difficulties posed by the test-set. However, the improvements with the PDE matrix compared to the binary matrix are also clear: 7 ptp and 12 ptp at subtitle level with 0.1 and 0.5 second deviations respectively, while at word level accuracies reached improvements of 21 ptp and 33 ptp for the same deviation ranges.

Regarding the PHS and PCE matrices, their alignment accuracy was close to the accuracy obtained with the PDE matrix. This finding suggests that the performance of decoder-independent matrices can get close to the performance of decoder-dependent matrices. Also note that, even if both PHS and PCE matrices obtained similar results, there was a small trend for the PCE matrix to be more accurate. The nature of the PCE matrix, more closely-related to the signal parametrization than the PHS matrix, could explain these minimal accuracy differences.

6. CONCLUSIONS AND FUTURE WORK

Several scoring matrices, based on different phoneme-relatedness criteria, were tested for aligning long audios with Hirschberg algorithm, and found to improve alignment accuracy compared to a binary matrix. As expectable, the matrix based on phone-decoder errors achieved the most accurate alignment results, while the matrices based on phonological similarity and perception errors also obtained clear improvements in alignment accuracy. Improvements were observed at word and subtitle level for Spanish and English, even if the English test-set posed serious difficulties. Thus, the effectiveness of the matrices under adverse conditions was also established.

Accuracy does not only depend on the alignment algorithm and the scoring matrices, but also on the performance of the phone decoders. Improving their robustness, by using a larger training-set or adapting models to the content or domain, will increase alignment accuracy.

Other future work would be extending the system to more languages. For this study, a matrix based on perceptual errors was created for English, but not for Spanish.

Regarding Hirschberg's algorithm, it sometimes offers more than one possible optimal alignment. In this study, we forced the algorithm to compute just one solution. Considering the different possible solutions, and defining criteria to choose among them could be an interesting study.

7. REFERENCES

[1] G. Bordel, S. Nieto, M. Peñagarikano, L. J. Rodríguez-Fuentes, A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *13th Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, Portland, Oregon, 2012.

[2] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences.," *Communications of the ACM,* vol. 18, no. 6, pp. 341-343, 1975.

[3] P. J. Moreno, C. Joerg, J-M Van Thong and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proceedings of the 5th International Conference on Spoken Language Processing*, *ICSLP*, Sydney, Australia, 1998.

[4] P. J. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Taipei, Taiwan, 2009.

[5] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein and S. Narayannan, "SailAlign: Robust long speech-text alignment," in *Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[6] G. Kondrak, Algorithms for Language Reconstruction, PhD Thesis. University of Toronto, 2002.

[7] P. Comas, Factoid Question Answering for Spoken Documents, PhD Thesis. Universitat Politècnica de Catalunya, 2012.

[8] A. Cutler, A. Weber, R. Smits and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *Journal of the Acoustical Society of America,* vol. 116, no. 6, pp. 3668-3678, 2004.

[9] J. E. Díaz, A. Peinado, A. Rubio, E, Segarra, N. Prieto and F. Casacubierta, "Albayzín: a task-oriented Spanish speech corpus," in *Proceedings of the First International Conference on Language Resources and Evaluation*, *LREC,* Granada, Spain, 1998.

[10] E. Campione and J. Véronis, "A multilingual prosodic database," in *Proceedings of the 5th International Conference on Spoken Language Processing*, *ICSLP,* Sydney, Australia, 1998.

[11] J. S. L. Garafolo, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[12] P. Ladefoged, A Course in Phonetics, New York: Harcourt Brace Jovanovich, 1995.

7.7 Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

- **Authors:** Pablo Ruiz, Aitor Álvarez, and Haritz Arzelus
- **Booktitle:** Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)
- **Year:** 2014

Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

Pablo Ruiz, Aitor Álvarez and Haritz Arzelus
Vicomtech-IK4
Mikeletegi Pasealekua, 57
20009 Donostia/San Sebastián, Spain
{pruiz,aalvarez,harzelus}@vicomtech.org

1 Introduction

Accessibility needs, and policies addressing them, are stimulating a large demand for subtitling in the broadcast industry. Manual subtitling being time and labour-intensive, automatic subtitling is an attractive option, as it saves time and resources.

Our approach to automatic subtitling aligns the audio signal with a human transcript for the audio. Aligning long audio signals is challenging, given memory demands, processing time and error-proneness of algorithms when aligning long sequences.

A successful system for long audio alignment is Bordel et al. (2012). They report alignment results for 3-hour long audios. Their alignment method is based on Hirschberg's algorithm (1975), originally used for genetic sequence alignment. The scoring matrix for alignment operations in Bordel et al. is binary: insertions, deletions and substitutions bear a cost of 1, while matches bear a cost of 0.

In this paper, we follow Bordel et al.'s long audio alignment approach, improving one aspect: We show that, as compared with results for a binary matrix, scoring alignment operations with a matrix based on phoneme-similarity improves alignment results at phoneme level, word level and subtitle level. We present results for the alignment of long audios in Spanish and English.

Our similarity scores follow Kondrak's metric (2002), based on multivalued phonological features weighted by salience. The metric has been successfully employed in cognate alignment and spoken document retrieval (Comas, 2012).

The paper is structured as follows. Section 2 presents our long audio alignment system, and Section 3 describes the similarity matrices created. Section 4 discusses evaluation methods and results. Section 5 contains conclusions and suggestions for future work.

2 Speech-text alignment system

The speech-text alignment system aligns two sequences of phonemes obtained from different sources. Given the audio and the transcript of the content to be automatically subtitled, a language-dependent phone decoding is used to recognize phonemes and their time-codes from the audio. In addition, a grapheme-to-phoneme module translates the input transcript into the reference phoneme transcription. An alignment algorithm finds phoneme correspondences between the reference phoneme transcription and the phonemes recognized by the phone-decoder, which usually contain common recognition errors. Aligned phonemes are assigned the time-codes obtained by the phone-decoder. Phoneme alignment may present substitutions, deletions and insertion errors. However, the number of phone correspondences found generally provides enough time-codes to create subtitles with near-perfect alignment at word-level.

2.1 Phone decoding module

The phone decoding module was trained using HTK[1], a hidden Markov model toolkit. The acoustic model was based on a monophone model, with three left-to-right emitting states using 32 Gaussian mixture components. The language model was a bigram phoneme model. The parametrization of the signal consisted of 18 Mel-Frequency Cepstral Coefficients plus the energy and their delta and delta-delta coefficients, using 16-bit PCM audios sampled at 16 KHz.

The Spanish phone-decoder was trained and tested with 20 hours of audios from three databases; Albayzín (Díaz et al., 1998), Multext (Campione and Véronis, 1998) and records of clean-speech broadcast news. The contents were

[1] http://htk.eng.cam.ac.uk/

mixed and divided into training (70%) and test (30%) sets. Texts totaling 45 million words were crawled from a national newspaper to train the language model. The Spanish phone-decoder yielded a Phone Error Rate (PER) of 40.65%.

The English phone-decoder was trained and tested on the TIMIT database (Garafolo et al, 1993), which consists of 5 hours and 23 minutes. 70% of the database was used for training, leaving the rest for testing. Texts totaling 369 million words, collected from digital newspapers, were used to train the language model. The English phone-decoder yielded a PER of 35.52%.

2.2 Grapheme-to-phoneme transcriptors

Grapheme-to-phoneme (G2P) transcriptors were developed for Spanish and English. The Spanish transcriptor was rule-based, inspired on an open-source tool[2], and adapted to our phonelist. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary[3] (CMUdict) using Phonetisaurus[4], a G2P framework based on weighted finite state transducers.

The Spanish and English phonesets are available on our project's website.[5]

2.3 Algorithm for long sequence alignment

We used Hirschberg's (1975) algorithm, an optimization of Needleman and Wunsch's (1970) algorithm to calculate the optimal alignment of two sequences of length n and m in $n \times m$ steps.

With Hirschberg's algorithm, each alignment operation receives a score, and the alignment obtaining the best score is chosen. Substitutions are evaluated with a scoring matrix. Gaps (insertions and deletions) incur a penalty. When aligning with the binary scoring matrix, our gap penalty was 2. When using the phoneme-similarity based matrices, our gap penalty was 10, based on parameter C_{skip} from our similarity function (Section 3).

Bordel et. al (2012) based their audio alignment system on Hirschberg's algorithm, showing its suitability. Nevertheless, they used a binary scoring matrix, while in this study matrices based on phoneme similarity were developed.

[2] http://www.aucel.com/pln/

[3] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

[4] http://code.google.com/p/phonetisaurus/

[5] https://sites.google.com/site/similaritymatrices/

3 Phoneme similarity matrices

Our similarity scores are based on the metric in Kondrak's (2002) ALINE cognate alignment system.[6] Phonemes are described with Ladefoged's (1995) multivalued features. Features are weighted according to their *salience*: the feature's impact for similarity. Features *place* and *manner* need to bear significantly higher salience than the rest.

$$\sigma_{sub}(p,q) = (C_{sub} - \delta(p,q) - V(p) - V(q)) / 100$$

where

$$V(p) = \begin{cases} 0 \text{ if } p \text{ is a consonant} \\ C_{vwl} \text{ otherwise} \end{cases}$$

$$\delta(p,q) = \sum_{f \in \boldsymbol{R}} \text{diff}(p,q,f) \times \text{salience}(f)$$

$$\sigma_{skip}(p) = C_{skip} / 100$$

Figure 1: Similarity Function

The phoneme and feature set, feature values and salience weights need to be adapted to each language. For each phone in our Spanish and English phonesets, we created feature specifications, available on our project's website (see footnote 5). Samples are shown in Table 5. Salience weights are in Table 6.

The scoring function[7] is in Figure 1: $\boldsymbol{\sigma_{sub}(p, q)}$ returns the similarity score for segments p and q. $\boldsymbol{C_{sub}}/100$ is the maximum similarity score attainable. $\boldsymbol{C_{vwl}}$ determines the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically. The function $\textbf{diff}(\boldsymbol{p, q, f})$ outputs the difference between segments p and q for feature f. The set of features $\boldsymbol{R}$ is configurable. Finally, $\boldsymbol{\sigma_{skip}(p)}$ returns $\boldsymbol{C_{skip}/100}$, which is used to define the penalty for insertions and deletions employed in the aligner (see Section 2.3).

We created different matrices, varying the settings for elements (1) through (3) below. Table 1 and Table 2 show a summary of the settings for each matrix. Table 3 and Table 4 show matrix samples.

For all matrices, C_{sub} was 3500, yielding a maximum possible similarity score of 35 ($C_{sub}/100$), and C_{skip} was −1000, yielding a gap penalty of 10 when aligning ($|C_{skip}/100|$).

[6] http://webdocs.cs.ualberta.ca/~kondrak/#Resources for Kondrak's ALINE. A Python implementation (PyAline) by Huff (2010) is at http://sourceforge.net/projects/pyaline/

[7] The original function contains an additional clause, not used by us, which evaluates two-to-one phoneme alignments. We added a denominator of 100 to σsub and $\sigma skip$, to use integer feature values (not decimals) while keeping our similarity scores in the range reported by Kondrak.

***(1)* C_{vwl}: 0 vs. 1000**. A desirable outcome of setting $C_{vwl} > 0$ is that substitutions between vowels and consonants are more clearly penalized by the matrix, getting lower scores than when $C_{vwl} = 0$. However, with $C_{vwl} > 0$, vowel matches get a lower similarity score than consonant matches, decreasing the weight of vowels in alignment. This is useful for cognate alignment (Kondrak, 2002, p. 48). The question arises whether this is also beneficial when aligning decoded phonemes and a G2P output. We tested this by defining *V(p)* in the scoring function differently.

***(2) V(p)*: original vs. alternative definition**. The alternative definition of V(p) in Figure 2 allows us to give equal scores to vowel matches and consonant matches, while still setting $C_{vwl} > 0$, and thus still obtaining the beneficial effect of penalizing consonant/vowel substitutions more than consonant/consonant ones.

With parameters p, q from $\sigma_{sub}(p, q)$

$$V(p) = \begin{cases} 0 \text{ if } p \text{ or } q \text{ is a consonant or } p = q \\ C_{vwl} \text{ otherwise} \end{cases}$$

Figure 2: Alternative definition for *V(p)*

***(3)* Diphthongs: binary vs. continuous scores**. This applies only to our English matrices. Our English phoneset treats diphthongs as single phones, but in Kondrak they are two-phoneme sequences. To score diphthong substitutions with Kondrak's function, we assigned them features and values heuristically. For comparison, we created matrices where diphthong scores were binary (match *vs*. mismatch).

Matrix Name	C_{vwl}	Definition of *V(p)*
C_v0_VpO	0	original
C_v1K_VpO	1000	original
C_v1K_VpA	1000	alternative

Table 1: Spanish Similarity Matrices and their settings

Matrix Name	C_{vwl}	Definition of *V(p)*	Diphthong Scores
C_v0_VpO_DB	0	original	binary
C_v0_VpO_DC	0	original	continuous
C_v1K_VpO_DB	1000	original	binary
C_v1K_VpA_DB	1000	alternative	binary
C_v1K_VpO_DC	1000	original	continuous
C_v1K_VpA_DC	1000	alternative	continuous

Table 2: English Similarity Matrices and their settings

IPA	a	i	n	p	r	s	j
a	35	7	−50	−56	−30	−50	2
i	7	35	−26	−32	−6	−26	10
n	−50	−26	35	9	−5	−5	−21
p	−56	−32	9	35	−11	9	−27
r	−30	−6	5	−11	35	−5	9
s	−50	−26	5	9	−5	35	−21
j	2	10	−21	−27	9	−21	35

Table 3: Sample from Spanish Matrix C_v*1K_VpA*

IPA	æ	i:	n	p	ɹ	s	aj
æ	35	9	−46	−57	−16	−36	10
i:	9	35	−26	−37	4	−16	−46
n	−46	−26	35	4	5	5	−46
p	−57	−37	4	35	−6	14	−46
ɹ	−16	4	5	−6	35	15	−46
s	−36	−16	5	14	15	35	−46
aj	10	−46	−46	−46	−46	−46	35

Table 4: Sample from English Matrix C_v*1K_VpA_DC*

4 Evaluation and Results

We evaluated alignment at phoneme, word, and subtitle level, aligning long audios containing spontaneous speech, with disfluencies. The Spanish test-set was clean speech. The English test-set was non-clean speech, with music, noise and overlapping utterances. Accordingly, lower accuracy in English was expected and observed, at all evaluation levels. Another difficulty with English subtitles, which also led to lower accuracy, is that they represent a less literal transcription of the audio than the Spanish subtitles, due to a different subtitling approach in each language.

The test-sets are different to the ones used to evaluate the phone-decoder, and consist of television audios, providing results that are more indicative of alignment quality in a real application scenario.

The Spanish test-set contained 47,480 phonemes, 8,774 words and 1,249 subtitles. The English test-set contained 21,310 phonemes, 4,732 words and 471 subtitles.

4.1 Evaluation at phoneme level

The number of correctly aligned phonemes, based on the number of matches during the alignment process, increased when using the phoneme-similarity based matrices. Improvements were around 11 percentage points in Spanish, from 38.14% with the binary matrix to 49.69% with the best-performing phoneme-similarity based matrix. Improvements in English were around 12 percentage points (15.57% with the binary matrix vs. 27.91% with the best phoneme-similarity based matrix).

SPANISH																	
IPA	**Vic**	**Place[1]**		**Manner[1]**		**V**	**Syl**	**Voi**	**Nas**	**Lat**	**Tri**	**High[1]**		**Back[1]**		**Ro[1]**	***E.g.***
a	**a**	velar	60	low vowel	0	1	100	100	0			low	0	front	100	0	v***a***
i	**i**	palatal	70	high vowel	40	1	100	100	0			high	100	front	100	0	d***i***
n	**n**	alveolar	85	stop	100	0	0	100	100	0	0						***n***o
p	**p**	bilabial	100	stop	100	0	0	0	0	0	0						***p***an
r	**R**	alveolar	85	approximant	60	0	0	100	0	0	100						pe***rr***o
s	**s**	alveolar	85	fricative	80	0	0	0	0	0	0						***s***on
j	**j**	palatal	70	high vowel	40	1	0	100	0	0	0	high	100	front	100	0	ho***y***
ENGLISH																	
IPA	**Vic**	**Place[1]**		**Manner[1]**		**V**	**Syl**	**Voi**	**Nas**	**Lat**	**Asp**	**High[1]**		**Back[1]**		**Ro[1]**	**Lo[1]**	***E.g.***
æ	**ae**	palatal	70	low vowel	0	1	100	100	0			low	0	front	100	0	0	c***a***t
i:	**iy**	palatal	70	high vowel	40	1	100	100	0			high	100	front	100	0	100	f***ee***l
n	**n**	alveolar	85	stop	100	0	0	100	100	0	0							***n***od
p	**p**	bilabial	100	stop	100	0	0	0	0	0	100							***p***od
ɹ	**r**	alveolar	85	approximant	60	0	0	100	0	0	0							***r***ed
s	**s**	alveolar	85	fricative	80	0	0	0	0	0	0							***s***et
aj	**ay**	palatal	70	low vowel+ high vowel	16	1	100	100	0			low+ high	40	central+ front	70	0	100	s***id***e

[1]To compare with each other phonemes where **V=1**, *Place* and *Manner* are replaced with *High*, *Back*, *Round*, and, if available, *Long*.

Shaded cells indicate features that are not used to define similarity for the segment in the language

Abbreviations	**V**: Vowel, **Syl**: Syllabic, **Voi**: Voice, **Nas**: Nasal, **Lat**: Lateral, **Asp**: Aspirated, **Tri**: Trill **Ro**: Round, **Lo**: Long, **Vic:** ASCII-based phone code

Table 5: Samples from the Phoneset, Features and Feature Values for Spanish and English

Place	40	Nasal	10	High	5
Manner	50	Lateral	10	Back	5
Syllabic	5	Aspirated	5	Round	5
Voice	10	Trill	10	Long	1

Table 6: Salience Weights for each feature

4.2 Evaluation at word level

We adopted Moreno et al.'s (1998) measure of word-level alignment, also used by Bordel et al. As Table 7 and Table 8 show, we record the cumulative percentage of correctly aligned words within a given deviation range: Column *0* shows the percentage of perfectly aligned words, column *≤0.1* means words whose misalignment goes up to 0.1*sec*, and so on. In the tables, we highlighted the best and worst results at 0, ≤0.1, ≤0.5 and ≤2 seconds.

Improvements with the phoneme-similarity based matrices were observed. In Spanish, exactly aligned words increased by ca. 9 percentage points, while improvement at a ≤0.5 deviation range was 20.85 percentage points. In English, improvements between ca. 20 and 30 percentage points were observed for each deviation range.

4.3 Evaluation at subtitle level

This is the most important evaluation, since it is indicative of the system's alignment quality in its application scenario: automatic subtitling. Reference subtitles were created manually by subtitling professionals.

For subtitle-level evaluation, we measured the deviation, compared to the reference, of the beginning of the subtitle's first word and of the end of the subtitle's last word. Cumulative percentages are given in Table 9 and Table 10.

In Spanish, when using the best-performing phoneme similarity based matrix, exactly aligned subtitles increased by 7.44 percentage points compared to results with the binary matrix. At the ≤0.5 deviation range, gains were 14.57 percentage points. In English, alignment improved at each deviation range, e.g. gains of 4.03 percentage points at ≤0.1 seconds and 8.92 percentage points at ≤0.5 seconds.

sec	0	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**14.17**	**57.71**	66.80	70.09	71.66	**72.65**	76.21	78.04	**79.02**
C_v0_VpO	23.17	81.08	87.87	90.27	91.42	92.17	93.96	95.01	95.64
C_v1K_VpO	22.77	80.78	87.96	90.65	91.62	92.41	94.43	95.23	95.89
C_v1K_VpA	**23.01**	**82.21**	89.31	91.83	92.87	**93.50**	95.50	96.21	**96.83**

Table 7: Spanish word alignment accuracy. Percentage of words aligned within each deviation range from reference

sec	0	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**10.57**	**45.08**	56.20	61.97	67.57	**73.26**	95.12	**99.76**	100
C_v0_VpO	18.33	65.73	76.62	82.23	85.27	87.35	98.56	**100**	100
C_v1K_VpO	17.93	65.57	76.94	82.95	85.35	87.43	98.56	99.84	100
C_v1K_VpA	**18.01**	**66.45**	77.50	82.71	85.59	**87.83**	98.80	99.84	100

Table 9: Spanish subtitle alignment accuracy. Percentage of subtitles aligned within each deviation range from reference

sec	0	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**0.28**	**4.81**	9.21	12.73	16.31	**19.04**	29.69	37.38	**43.24**
C_v0_VpO_DB	1.59	19.57	28.72	34.94	40.66	45.00	58.22	65.46	70.31
C_v0_VpO_DC	1.59	19.86	28.97	34.33	38.78	42.78	56.61	63.64	68.72
C_v1K_VpO_DB	1.67	20.31	29.82	36.26	41.25	45.55	58.79	66.14	70.84
C_v1K_VpO_DC	1.67	20.03	28.84	33.91	38.29	42.23	54.76	61.10	64.85
C_v1K_VpA_DB	1.80	22.87	33.38	39.50	44.60	48.73	61.56	68.64	**73.08**
C_v1K_VpA_DC	**1.93**	**23.76**	34.20	40.03	44.79	**48.90**	61.58	68.34	72.72

Table 8: English word alignment accuracy. Percentage of words aligned within each deviation range from reference

sec	0	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤1.0	≤1.5	≤2.0
Binary	0.21	**4.25**	9.77	18.26	27.18	**37.15**	84.29	**98.73**	100
C_v0_VpO_DB	0.42	7.64	16.99	25.05	33.55	43.10	86.41	98.30	100
C_v0_VpO_DC	0.42	7.43	16.14	25.48	34.39	42.46	87.47	98.73	100
C_v1K_VpO_DB	0.42	8.92	16.56	26.54	33.33	40.55	87.47	98.51	100
C_v1K_VpO_DC	0.42	9.13	17.20	27.18	35.88	42.25	87.05	98.51	100
C_v1K_VpA_DB	0.42	**11.04**	19.53	30.79	39.70	45.86	87.47	98.94	100
C_v1K_VpA_DC	0.42	8.28	16.99	26.96	40.13	**46.07**	86.84	**99.58**	100

Table 10: English subtitle alignment accuracy. Percentage of subtitles aligned within each deviation range from reference

5 Conclusions and Future Work

This study shows that long audio alignment using Hirschberg's algorithm can be improved by using, instead of a binary scoring matrix, a scoring matrix based on phoneme similarity defined via phonological features. Improvements were observed at phoneme, word and subtitle level, when aligning both clean speech (Spanish tests) and non-clean speech (English tests).

As expectable, improvements at word level were higher than at subtitle level. At subtitle level, we only assess the position of the first and last word of each subtitle. This restricts the set of word-alignments that can contribute to a subtitle-level improvement.

Regarding the different matrices tested, we obtained slightly better results with the matrices created using a modified scoring function, that gives equal weight to consonant matches and vowel matches.

As future work, several approaches to improve alignment could be tested. First, our phoneme decoding applied MFCC coefficients, based on a perceptually motivated Mel frequency scale. However, our phoneme-similarity metric relied on phonological features that follow articulatory criteria. Using MFCC parametrization together with matrices based on perceptual similarity could be tested. The converse approach is also possible: Keeping a similarity metric based on articulatory criteria, but using an acoustic parametrization that provides a good description of the speech articulators, e.g. linear predictive coding (LPC). Finally, since alignment quality depends on phone-decoder accuracy, similarity matrices based on phone-decoding confusion matrices could be tested.

6 References

Bordel, G., S. Nieto, M. Peñagarikano, L. J. Rodríguez-Fuentes, A. Varona. 2012. A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*. Portland, Oregon.

Campione, E. and J. Véronis. 1998. A multilingual prosodic database. In *ICSLP 1998, Proceedings of the 5th International Conference on Spoken Language Processing.* Sydney, Australia.

Comas, P. 2012. *Factoid Question Answering for Spoken Documents*. PhD Thesis. Universitat Politècnica de Catalunya.

Díaz, J. E., A. Peinado, A. Rubio, E, Segarra, N. Prieto and F. Casacubierta. 1998. Albayzín: a task-oriented Spanish speech corpus. In *LREC 1998. Proceedings of the First International Conference on Language Resources and Evaluation.* Granada, Spain.

Garafolo, J. S, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia.

Hirschberg, D. S. 1975. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM* 18 (6): 341-343.

Huff, P. 2010. *Automatically Growing Language Family Trees Using the ALINE Distance*. M.A. Thesis. Brigham Young University.

Kondrak, G. 2002. *Algorithms for Language Reconstruction*. PhD Thesis. University of Toronto.

Ladefoged, P. 1995. *A Course in Phonetics*. Harcourt Brace Jovanovich. New York

Moreno, P. J, C. Joerg, J-M Van Thong, O. Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *ICSLP 1998, Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney, Australia.

Neeedleman, S. B, and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3): 443-53

7.8 Improving the Automatic Segmentation of Subtitles through Conditional Random Field

- **Authors:** Aitor Álvarez, Carlos-D. Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, Arantza del Pozo
- **Journal:** Speech Communication
- **Publisher:** Elsevier
- **Status:** Major revision

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2016) 1–18

Speech Communication

www.elsevier.com/locate/procedia

Improving the Automatic Segmentation of Subtitles through Conditional Random Field

Aitor Álvarez[a], Carlos-D. Martínez-Hinarejos[b], Haritz Arzelus[a], Marina Balenciaga[a], Arantza del Pozo[a]

[a] *Human Speech and Language Technology Group, Vicomtech-IK4, San Sebastian, Spain*
[b] *Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Spain*

Abstract

Automatic segmentation of subtitles is a novel research field which has not been studied extensively to date. However, quality automatic subtitling is a real need for broadcasters which seek for automatic solutions given the demanding European audiovisual legislation. In this article, a method based on Conditional Random Fields is presented to deal with the automatic subtitling segmentation. This is a continuation of a previous work in the field, which proposed a method based on Support Vector Machine classifier to generate possible candidates for breaks. For this study, two corpora in Basque and Spanish were used for experiments, and the performance of the current method was tested and compared with the previous solution through several evaluation metrics. Finally, an experiment with human evaluators was carried out with the aim of measuring the productivity gain in post-editing automatic subtitles generated with the new method presented.

Keywords: automatic subtitling, subtitle segmentation, pattern recognition, machine learning

1. Introduction

Subtitles have acquired great relevance within the audiovisual community during the last years, mainly after the adoption of the new audiovisual directives (Article 7 of the Audiovisual Media Services Directive[1]) of the European Parliament and of the Council in March of 2010. This legislation regulates the rights of people with a visual or hearing disability, and moved member states to take the necessary measures to guarantee that the services of audiovisual providers under their jurisdiction are gradually more and more accessible by means of sign-language, audio-description, easily menu navigation and subtitling.

Given the new audiovisual legislation, broadcasters and subtitling companies are seeking automatic solutions to be more productive than with the traditional manual subtitling. At the same time, disability organisations are pushing for both quantity and quality of subtitles, in order to not only increment the percentage of subtitling in the TV and the Internet, but also request quality

[1] `http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32010L0013\&from=EN`

subtitles. As a result, the demand of automatic solutions for quality subtitling has grown fast in the audiovisual community.

Several parameters take part in the definition of what the quality of subtitles is [1]. Apart from features related to subtitle layout, duration and text editing, subtitling segmentation is one of the most relevant, as it was demonstrated in [2], a study whose aim was to verify whether a correct text chunking in subtitles had an impact on both comprehension and reading speed using human evaluators. Even though important differences were not found in terms of comprehension, they demonstrated that a correct segmentation by phrase or by sentence significantly reduced the time needed to read subtitles. Furthermore, the strong need for proper segmentation is supported by the psycholinguistic literature on reading [3], where the consensual view is that subtitle lines should end at natural linguistic breaks to improve readability and reduce cognitive effort produced by poorly segmented text lines [4].

In this article, a new method based on the probabilistic Conditional Random Field is applied to the field of automatic subtitling segmentation for Basque and Spanish languages. This work is a continuation of the previous research presented in [5], in which Support Vector Machine and Logistic Regression classifiers were employed for the subtitling segmentation task in the Basque language. In the present study, the same Basque corpus was used in order to compare the performance using the new classification method. In addition, the work has been extended to the Spanish language. It allowed us to confirm that the new classification method employed was valid for different types of corpora and languages. Given that the results obtained in [5] by the Support Vector Machine and Logistic Regression classifiers were very close due to its similar nature, in this work the performance of Support Vector Machine and Conditional Random Field were compared for both languages, leaving out the Logistic Regression classifier. The results obtained proved that the classification method based on Conditional Random Field outperformed clearly the results obtained by Support Vector Machine in terms of accuracy and computation time for both languages.

The article is structured as follows. Section 2 describes existing work on automatic subtitling and segmentation. Section 3 looks at Conditional Random Field method and how it fits subtitle segmentation task. Section 4 describes the methodology we designed and implemented to build the new classification approach. Section 5 presents the experimental framework and the evaluation metrics. Section 6 summarizes the evaluation results and the performance comparison between the methods based on Conditional Random Field and Support Vector Machine. Section 7 shows the human performance results in segmentation correction for two options of obtaining draft segmentations. Finally, Section 8 draws conclusions and describes future work.

2. Related work in Subtitle Segmentation

Automatic segmentation of subtitles is a novel line of research which has not been studied extensively up to the present. To date, most of the automatic subtitling solutions have not been capable of generating syntactic and semantically coherent breaks for quality segmentation and, thus, segmentation is mainly performed considering the maximum number of characters allowed per line or through manual intervention. A survey of the literature actually provides only one reference in the field of automatic segmentation; the study presented in [5], where automatic subtitle segmentation was treated as a machine learning problem. In this work, Support Vector Machine and Logistic Regression classifiers were built over a Basque corpus consisting of TV cartoon programs and subtitles generated by professional subtitlers. With the help of an iterative algorithm, all the possible candidates for a line break were generated at each iteration. These candidates were then measured against the machine-learned classifiers and optimal candidates selected according to the obtained score. They reached promising results with an average F1-measure score of around 75% considering both classifiers. However, they estimated only the possible segmentation points, without distinguishing between different types of breaks. This implies considering line-breaks and subtitle-breaks, which is a critical information to automatically generate the final subtitles correctly. It has to be noted that not all the subtitles have to have the same number of lines; there can be

subtitles with just one line combined with others with two lines depending on the content and the segmentation rules. It is therefore critical to differentiate between line-breaks and subtitle-breaks. Finally, the computation time needed to generate all the candidates and select the optimal ones in the method presented in [5] was inefficient for a real application. Computing the iterative algorithm for one hour of content with 900 subtitles it took four hours of processing time (16 seconds per subtitle) on an Intel(R) Xeon(R) 2.00GHz and 32GB based server.

The rest of works in the literature regarding subtitles segmentation are focused on comparing the comprehension and reading speech in live-respoken subtitles segmented in a correctly and poorly manner [2], measuring the impact of arbitrary segmented subtitles on readers [4] and on studying the way line-breaking is commonly performed [6]. None of these three last works include technology to automatically create and segment subtitles.

3. Conditional Random Field for Segmentation

Conditional Random Fields (CRF) have been applied in different domains and applications, such as computer vision [7], bioinformatics [8], and specially in Natural Language Processing (NLP). In the NLP field, applications go from recognition and classification in text and speech [9, 10, 11] to segmentation and labelling of text [12, 13, 14]. These last applications inspired this work on the application of CRF to the subtitle segmentation problem.

Segmentation of subtitles can be seen as a label assignment to the sequence of words to be segmented, where the labels will basically indicate if a word pertains to the extreme (beginning or end) of a segmentation unit. Following a statistical approximation, the objective is obtaining the optimal assignment from a sequence of words. If the sequence of words is $W = w_1^n = w_1 w_2 \cdots w_n$, and the sequence of labels is $L = l_1^n = l_1 l_2 \cdots l_n$, the problem can be statistically stated as:

$$\hat{L} = \underset{l_1^n}{\operatorname{argmax}} \Pr(l_1^n | w_1^n) \tag{1}$$

The problem can be solved by defining a model to estimate $\Pr(l_1^n|w_1^n)$ from training data (training process) and applying a searching algorithm on that model for a given sequence of words (decoding process). CRF offer an appropriate framework for modeling conditional probability between input-output sequences, as well as searching algorithms that allow to obtain the decoding results.

Following a notation similar to that employed in [15], a linear chain Conditional Random Field can be formulated as:

$$\Pr(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{\tau=1}^{\mathcal{T}} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_\tau, y_{\tau-1}, \vec{x}_\tau) \right) \tag{2}$$

In Equation (2), the meaning of the different terms is the following:

- $\vec{x}$ and $\vec{y}$ represent input and output sequences, respectively (both of size $\mathcal{T}$).
- $Z(\vec{x})$ is a normalization factor in order to ensure a proper probability distribution.
- f_k (with $k = 1, \ldots, K$) is the set of features functions; these feature functions establish the correspondence between input and/or output elements, and they actually form the probability distribution; in this formulation, they are said to be *bigram* models, since output in time $\tau - 1$ ($y_{\tau-1}$) is related to output in time τ (y_τ).
- θ_k (with $k = 1, \ldots, K$) is the set of weights associated to each feature function f_k.

In the case of subtitle segmentation, input is a sequence of feature vectors derived from the sequence of words to be segmented, whereas output is a sequence of labels that represent for each word its situation inside the segmentation. Details on the specific input features and output labels are described in Section 4. Feature functions f_k and weights θ_k will be obtained in the training process.

According to this formulation, the final CRF model for subtitle segmentation can be stated as:

$$\hat{L} = \underset{l_1^n}{\operatorname{argmax}} \Pr(l_1^n | w_1^n) = \underset{l_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} \exp\left(\sum_{k=1}^{K} \theta_k f_k(l_i, l_{i-1}, \vec{w}_i) \right) \tag{3}$$

As it can be seen, the maximization allows to avoid the normalization term $Z(\vec{x})$. Notice that $\vec{w}_i$ is the feature vector derived from word w_i in the input.

4. Methodology

4.1. Important considerations in subtitles' segmentation

Automatic segmentation of subtitles can be treated as a text sequence labeling problem. However, it has some particularities which have to be considered carefully. Firstly, there are several features that have to be taken into account at the same time. Apart from the text analysis, a correct segmentation of subtitles depends on other characteristics like (1) the amount of characters allowed per line, (2) timing issues related to long pauses and speech rhythm, (3) speaker changes, (4) the preceding and posterior words to select the most appropriate break type and point, and (5) the subtitle persistence on screen, which has a real impact in the readability. Moreover, it has to be noted that although there are some standard guidelines for a correct subtitling, such as Ofcom's Guidance on Standards for Subtitling[2]; BBC's Online Subtitling Editorial Guidelines[3]; and ESIST's Guidelines for Production and Layout of TV Subtitles[4], each subtitling company tends to have its own subtitling rules which may differ with each others in some specific points.

Secondly, besides looking upon the characteristics described above, the segmentation should be done including a syntactic analysis to create linguistically coherent breaks. It is the preferred and most adopted solution in the subtitling community and it follows from experiments and conclusions in psycholinguistic research, which show that readers analyze texts considering syntactic information [16], grouping words corresponding to syntactic phrases and clauses [17]. Therefore, with the aim of facilitating readability, subtitle lines should thus be split according to coherent linguistic breaks and considering the highest possible syntactic node as possible.

Finally, the demand of automatic solutions for subtitling comes from the need of tools to operate fast and provide quality results. Within an automatic subtitling solution which includes speech recognition technology, it is expected an output with well formatted and segmented subtitles, and few recognition mistakes. Besides, it should be executed in the shortest time as possible, requiring optimal solutions with high performance and low processing cost.

4.2. Conditional Random Field's configuration

Before constructing a CRF graphical model for any application, it has to be previously defined a dependence structure, which will be obeyed by the class labels given the observed data. This structure defines the transitions between the class labels at the graph node. In a Markov dependence structure, each class label and its corresponding feature vectors depend on the neighboring class labels and their features in the predefined neighborhood distance.

[2] http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/standards_for_subtitling/subtitling_1.asp.html

[3] http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf

[4] http://www.translationjournal.net/journal/04stndrd.htm

With the aim of defining a dependence structure for automatic segmentation through a CRF graphical model, eight class labels were created to define the function of each word within the subtitle, as listed below:

- B-SU *(Begin-Subtitle):* For each first word in subtitles.
- I-LI *(In-Line):* For each word in subtitle which is not the first or last word of a line or subtitle.
- E-LI *(End-Line):* For each word which represents the last word of a line which does not correspond to the end of a subtitle (e.g. last word of the first line in a subtitle with two lines).
- B-LI *(Begin-Line):* For each first word in a line that is not the first word in a subtitle (i.e., first word in second line for subtitles with more than one line).
- E-SU *(End-Subtitle):* Each final word of a subtitle composed by one or more lines.
- BE-SU *(BeginSubtitle-EndSubtitle):* For words in an one-word subtitle.
- BS-EL *(BeginSubtitle-EndLine):* For words in the first one-word line for subtitles with more than one line.
- BL-ES *(BeginLine-EndSubtitle):* For words in the second one-word line for subtitles with more than one line.

00:00,166 - 00:05,333
Come here,
Mum come here

Example 1: Subtitle example composed by 6 words and 2 lines

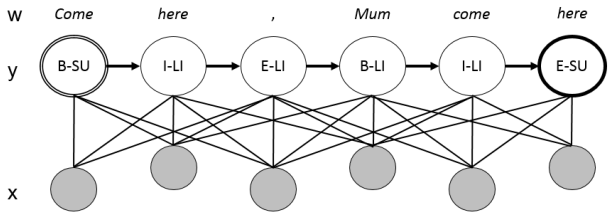


Fig. 1: Graphical model of the executed CRF over the Example 1. Transition factors depend on the surrounding two observations.

In Figure 1, a practical execution of the defined Markov dependence structure is presented, given the example subtitle shown in Example 1, which is composed by 6 tokens and formatted in 2 lines. Given this input example, the target of the CRF model would be to predict an output vector $y = \{y_0,y_1,..., y_N\}$ through the observed feature vectors $\{x_0,x_1,..., x_N\}$ extracted from the sequence of words $\{w_0,w_1,..., w_N\}$. In the CRF graphical models constructed for this work, each variable y_j corresponds to one of the class labels described above for each word at position j. For its part, each x_j contains the feature vector values about the word at position j. The transition factors of our CRF models depend on the surrounding two observations. The features used to describe each of the words at position j are described in Subsection 4.3.

4.3. Conditional Random Field's feature vectors

The feature vectors which describe the information for each word were composed of a total of 15 characteristics. They can be divided into the following subsets:

- Words: The current word and the surrounding 2 words. *(5 features)*
- Part-Of-Speech: The current word's Part-Of-Speech and the surrounding 2 words' Part-Of-Speech information. *(5 features)*
- Amount of characters per line and subtitle: A boolean value to control if the amount of characters per line and subtitle has been exceeded. *(2 features)*
- Speaker Change: A boolean value to control if there is a speaker change in the current word or not. *(1 feature)*
- Time difference between the current and the neighboring words: Two parameters to compute the time difference between the current word and the previous and next word. We used 5 different discrete values for these parameters, including the value 0 for time differences lower than 100 milliseconds, 0.1 for differences lower than 500 milliseconds, 0.5 for differences between 500 and 1000 milliseconds, the value 1 for differences in the range of 1000 and 1500 milliseconds, the value 1.5 for differences higher than 1500 milliseconds and lower than 2000 milliseconds, and the value 2 for differences higher than 2000 milliseconds. The reference values were fixed looking at the training corpus, once all the time differences at the breaks were computed and analyzed. *(2 features)*

5. Experiments

5.1. Corpora description and processing

The new automatic segmentation method based on CRF graphical models was tested over two languages, each with a particular corpus. For the Basque language, we used the same corpus of that employed in [5]. It was composed of TV cartoon programs in Basque with manually generated subtitles by professional subtitlers, for a total amount of 109,006 subtitles. The subtitle files were provided in SRT format, indicating start and end time-codes for each subtitle and presented in blocks of a maximum of two lines. The subtitles were carefully generated and segmented maintaining a linguistic coherence and splitting subtitles according to the highest possible syntactic node.

With regard to the Spanish language, the new corpus was composed of 98 episodes of the TV Spanish series "Mi querido Klikowsky", with a total amount of 81,802 subtitles. The subtitle files, which were provided also in SRT format, were created manually by professionals, and the segmentation was performed following specific predefined rules based on keeping a linguistic and syntactic coherence. The contents include many segments with spontaneous speech, grammatically incorrect sentences, and some words and expressions pronounced in several Spanish dialects, such as Argentinian and Andalusian. This issue triggers the Part-Of-Speech technology to make more mistakes than desired.

Since the Spanish 98 episodes do not include speaker change information, an additional subcorpora was also created to test the impact of this feature on the segmentation task. We generated two additional subcorpora with 23 episodes from the original 98 ones, one containing speaker changes, which were included manually by a professional, and the other without speaker change information.

With regard to the feature vectors, the computation of the POS information was performed using the Eustagger toolkit [18] and ixa-pipe-pos [19] for the Basque and Spanish languages respectively. In addition, the time-codes at word level were obtained through the audio forced-alignment algorithms presented in [20] for both languages. This last information allowed us to obtain the differences in time between neighboring tokens.

	Basque Corpus			Spanish Large Corpus			Spanish Small Corpora		
	CV	Training	Test	CV	Training	Test	CV	Training	Test
Programs	358	283	14	98	96	2	23	22	1
Subtitles	109006	86656	5307	81802	80058	1744	20154	19150	1004
Lines	166986	132832	8337	149774	146618	3156	37209	35356	1853
Words	768394	610471	37579	857648	839917	17731	211317	200964	10353
Lines/Subt	1.53	1.53	1.57	1.83	1.83	1.81	1.85	1.85	1.85
Words/Lines	4.60	4.60	4.51	5.73	5.73	5.62	5.68	5.68	5.59
Words/Subt	7.05	7.04	7.08	10.48	10.49	10.17	10.49	10.49	10.31

Table 1: Distinctive features of the different corpora (CV for cross-validation corpus, Training and Test for comparative corpus).

	Basque Corpus			Spanish Large Corpus			Spanish Small Corpora		
Label	CV	Training	Test	CV	Training	Test	CV	Training	Test
B-SU	14.06%	14.06%	13.99%	9.50%	9.50%	9.82%	9.53%	9.52%	9.68%
I-LI	56.69%	56.64%	55.79%	65.12%	65.13%	64.43%	64.80%	64.82%	64.22%
E-LI	7.47%	7.49%	7.97%	7.91%	7.91%	7.95%	8.06%	8.06%	8.18%
B-LI	7.52%	7.54%	8.03%	7.91%	7.91%	7.95%	8.06%	8.06%	8.20%
E-SU	14.11%	14.12%	14.05%	9.51%	9.50%	9.82%	9.53%	9.52%	9.70%
BE-SU	0.05%	0.05%	0.04%	0.02%	0.02%	0.01%	0.00%	0.00%	0.00%
BS-EL	0.08%	0.08%	0.10%	0.02%	0.02%	0.01%	0.01%	0.01%	0.02%
BL-ES	0.02%	0.02%	0.03%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%

Table 2: Proportion of the different labels in the different corpora (CV for cross-validation corpus, Training and Test for comparative corpus).

Tables 1 and 2 describe the distinctive features and proportion of labels in all the corpora respectively.

5.2. Experiments setup

The Basque and Spanish CRF models were built and evaluated in two ways. Initially, the whole corpus for each language was used to train and evaluate models applying 10-fold cross validation technique. This evaluation was performed at class label and segmentation levels for both languages.

Each corpus was then split in train and test sets, and results were compared with the ones obtained with the SVM based classification method. In the case of Basque, the division followed the partition made in [5] to evaluate the SVM based classification method. In this previous work, about 80% (86,656 subtitles) of the corpus was used to train the SVM models, 15% to evaluate them, and the rest (final-test) to evaluate the complete method including the iterative algorithm. For this work, we used the train and final-test partitions in order to compare both methods with the same size of corpus. Thus, 86,656 subtitles were used to train the basque CRF models and 5,307 subtitles to test them. For the Spanish Large corpus without the speaker change information (98 episodes), the distribution was carried out keeping 80,058 subtitles for training and the rest (1,744 subtitles) for testing. Finally, for the two Spanish Small subcorpora (23 episodes) with and without speaker information, 19,150 of the subtitles were used to train models, and 1,004 subtitles for testing purposes. The procedure followed to create segmentation breaks using the SVM based classification method was the same explained in the work [5], as it was briefly summarized in the previous Section 2.

All the experiments were performed using the CRF++ toolkit [21].

5.3. Evaluation metrics for segmentation

Apart from the classical metrics for label assignment (Precision, Recall, and F1-Score), since the problem to study was the subtitle segmentation, segmentation evaluation metrics had to be be used. Four main evaluation metrics were used to test the performance of the CRF models and SVM based classification method, as they are described in the following subsections.

5.3.1. F1-LINE

It is the evaluation metric proposed in [5] and it was only used in this work to compare the performance of both CRF and SVM based classification method in testing mode. It measures

segmentation errors (false negatives and false positives) and correct segmentations (true negatives and true positives), and computes the accuracy through the F1-Score. It does not distinguish between line and subtitle breaks. The `conlleval` script[5] (which is the one used for the CoNLL-2000 shared task) was employed for measuring this metric, as well as for the Precision, Recall, and F1-Score calculations presented in Section 6.

5.3.2. NIST-SU

This well-known metric was provided by NIST for the Rich Transcription Fall evaluations [22], and it computes the number of segmentation errors (missed segments and false alarm segments) divided by the number of segments in the reference. Its limitation is that it does not consider position substitutions. For this work, it was computed at line level (NIST-SU-LI), which included both line-breaks and subtitle-breaks, and at subtitle level (NIST-SU-SUB).

5.3.3. DSER

It is computed dividing the number of incorrectly segmented portions in the reference by the total of segments in the reference. This is a more greedy metric if comparing with the NIST-SU, and its limitation lies in that it takes segments as whole sequence, and not as limits. For this work, it was computed at line level (DSER-LI), composed by line-breaks and subtitle-breaks, and at subtitle level (DSER-SUB).

5.3.4. SegER

It was proposed in [14] as an alternative evaluation measure to overcome the limitations posed by the previous NIST-SU and DSER metrics. SegER is computed as the edit distance between sequences of reference positions and hypothesis positions (those obtained automatically by the classifiers), using Insertion, Deletion, and Substitution as edition operations. As for the previous two metrics, it was also computed at line level (SegER-LI), which included both line-breaks and subtitle-breaks, and at subtitle level (SegER-SUB).

In Table 3 an example is given on how these metrics are computed taking as input the reference and the hypothesis, both composed of the class labels defined for the segmentation task. The computation scores of the segmentation measures are presented in Table 4.

Table 3: An example of how the different metrics are computed given a reference and the hypothesis estimated by the classifiers. For the *F1-LINE* metric calculation, *TN* means *True Negative*, *TP* corresponds to *True Positive*, *FP* is *False Positive* and *FN* means *False Negative*. The sign x corresponds to an error and ✓means correct. Finally, Correct and Substitution are represented by the C and S symbols respectively.

Segmentation measures											
Reference:	B-SU	I-LI	E-LI	B-LI	I-LI	E-SU	B-SU	E-LI	B-LI	I-LI	E-SU
Hypothesis:	B-SU	I-LI	E-LI	B-LI	E-SU	B-SU	I-LI	E-LI	B-LI	I-LI	E-SU
F1-LINE	*TN*	*TN*	*TP*	*TN*	*FP*	*FN*	*TN*	*TP*	*TN*	*TN*	*TP*
NIST-SU-SUB					x	x					✓
NIST-SU-LI			✓		x	x		✓			✓
DSER-SUB						x					x
DSER-LI			✓			x		x			✓
SegER-SUB					S_1	S_1					C
SegER-LI			C		S_1	S_1		C			C

[5] `http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt`

Table 4: Computation scores of the example given in Table 3. It has to be noted that the scores can be positive (Acc, which means Accuracy) or negative (Err, which denotes Error).

Segmentation scores		
Metric	Computation	Score (Acc/Err)
F1-LINE	(2*TP) / (2*TP+FP+FN)	75% (Acc)
NIST-SU-SUB	2 Err / 3 Ref	66.67% (Err)
NIST-SU-LI	2 Err / 5 Ref	40% (Err)
DSER-SUB	2 Err / 2 Ref	100% (Err)
DSER-LI	2 Err / 4 Ref	50% (Err)
SegER-SUB	(1S) / (1C+1S)	50% (Err)
SegER-LI	(1S) / (3C+1S)	25 % (Err)

6. Results and discussion

6.1. Basque Corpus

6.1.1. Training and evaluation

This subsection describes the results obtained during the training and evaluation through the 10-fold cross-validation technique of the Basque CRF model using the whole corpus of this language. Table 5 presents the results for each class label, whilst Table 6 shows the accuracy reached and the number of tokens correctly tagged by the classifier.

Class labels evaluation				
	#	Precision	Recall	F1-Score
I-LI	435,575	93.5%	95.7%	94.6%
B-SU	108,014	87.3%	87.2%	87.2%
E-SU	108,441	87.4%	87.2%	87.3%
B-LI	57,819	69.3%	63.5%	66.3%
E-LI	57,392	69.0%	63.3%	66.1%
BL-ES	161	47.0%	5.0%	9.0%
BS-EL	589	90.1%	52.5%	66.3%
BE-SU	403	92.4%	84.9%	88.5%

Table 5: Precision, Recall and F1-Score values for each class label applying 10-fold cross-validation in the Basque corpus.

#Correct	679,347
#Labels	768,394
Accuracy	88.4%

Table 6: 10-fold cross-validation accuracy at class label level in the Basque corpus.

As it is shown in Table 5, among the most common labels in the Basque corpus, the labels representing the I-LI, B-SU, and E-SU labels reached the best results, obtaining a F1-Score of 94.6%, 87.2%, and 87.3% respectively. It means that the CRF classifiers modeled accurately subtitles boundaries and in-line words. However, the scores obtained at line-breaks level through the E-LI and B-LI labels are not as precise as at subtitle-breaks. It is due to the fact that there are more features at subtitle-level which could stand for a subtitle break, such as speaker changes, full stops or long silences, than at line-level, which usually depends exclusively on the syntactic information to predict a correct line break. Nevertheless, the F1-Score for the E-LI label achieved an interesting 66.1%. Besides, it has to be considered that the performance of the B-SU and B-LI labels are entirely dependent on the E-SU and E-LI labels respectively. However, as it can be appreciated in Table 6, if we consider the whole set of labels to be predicted (768,394 labels), an accuracy of 88.4% was achieved, given that 679,347 labels were correctly tagged.

Table 7: 10-fold cross-validation scores at segmentation level in the Basque corpus.

Segmentation evaluation					
NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
25.3	15.8	44.4	27.6	21.6	12.8

On the other hand, Table 7 presents the results for the NIST-SU, DSER and SegER evaluation metrics over the 10-fold cross-validation technique applied during the training of the CRF Basque

model on the whole corpus. As it can be seen, the segmentation scores follow the same tendency as the example given in Table 4, where the segmentation errors at line-level are lower than at subtitle-level for any case. The CRF model achieved promising performance for Basque. The interesting low error rates presented in Table 7 demonstrated the good performance of the labels, as it was shown in Table 5.

6.1.2. Testing and comparison

In this subsection, the Basque CRF model is compared at segmentation level with the Basque SVM based classification method through the metrics described in Section 5.3, and using the train and test distributions described previously.

Initially, Table 8 presents the Precision, Recall and F1-Score values achieved with the Basque CRF model over the Basque test data set. The CRF model was built on the train data set of the Basque corpus. Since the amount of the BL-ES, BS-EL and BE-SU labels was insignificant in the Basque test, we did not include their scores. As it can be seen in Table 8, the subtitles boundaries and inline words reached high accuracies, obtaining 90.6%, 86.6%, and 86.7% F1-Scores for I-LI, B-SU, and E-SU labels respectively. On the contrary, the performance of the labels related to the line boundaries was not as precise as the ones related to subtitle boundaries. The labels B-LI and E-LI, which correspond to begin-line and end-line words, achieved 44.4% and 44.5% F1-Scores values respectively. However, 31,200 of the 37,579 labels were correctly classified in overall, obtaining a global accuracy of 83.0%, as it is shown in Table 9.

Table 8: Precision, Recall and F1-Score values of each class label for the Basque test data set.

Class labels evaluation				
	#	Precision	Recall	F1-Score
I-LI	22,466	87.6%	93.9%	90.6%
B-SU	5,326	86.0%	87.2%	86.6%
E-SU	5,324	86.4%	87.1%	86.7%
B-LI	2,224	52.3%	38.5%	44.4%
E-LI	2,224	52.2%	38.7%	44.5%

Table 9: Accuracy at class label level for the Basque test data set.

#Correct	31,200
#Labels	37,579
Accuracy	83.0%

Table 10 presents the segmentation scores of each classification method for the Basque test set. As it can be seen, the low performance of the B-LI and E-LI labels presented in Table 8 had a real impact on the segmentation scores for the CRF model. For the NIST-SU and DSER metrics, the error rate at line level reached a higher error than the metrics related to the subtitle level. If we compared both classification methods, the CRF model outperformed clearly the results obtained by the SVM based classification method for all cases. The difference is even higher for the metrics related to measure the subtitle boundaries.

Table 10: Segmentation scores of the CRF and SVM models for the Basque test set.

	Segmentation score						
	F1-LINE	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
CRF	83.0	26.5	28.3	47.1	47.4	22.6	21.6
SVM	74.7	81.6	56.1	110.5	79.1	59.0	33.7

6.2. Spanish Large Corpus

6.2.1. Training and evaluation

The results obtained during the training and evaluation of Spanish CRF models with the Spanish Large Corpus (98 episodes) and applying 10-fold cross-validation technique are presented in this subsection. This corpus did not include information about speaker changes. Table 11 describes the results at class label, and the accuracy along with the number of correctly tagged labels are shown in Table 12. Finally, the results at segmentation level are presented in Table 13.

Table 11: Precision, Recall and F1-Score values for each class label applying 10-fold cross-validation in the Spanish Large corpus (without speaker change information).

Class labels evaluation				
	#	Precision	Recall	F1-Score
I-LI	558,500	90.2%	92.6%	91.4%
B-SU	81,513	98.8%	93.6%	96.1%
E-SU	81,543	98.9%	93.6%	96.2%
B-LI	67,861	60.5%	57.7%	59.1%
E-LI	67,831	60.5%	57.8%	59.1%
BL-ES	111	75.0%	24.3%	36.7%
BS-EL	141	25.0%	0.7%	1.4%
BE-SU	148	100.0%	92.6%	96.1%

Table 12: 10-fold cross-validation accuracy at class label level in the Spanish Large corpus (without speaker change information).

#Correct	748,270
#Labels	857,648
Accuracy	87.3%

Table 13: 10-fold cross-validation scores at segmentation level in the Spanish Large corpus (without speaker change information).

Segmentation evaluation					
NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
7.4	34.6	11.8	57.9	7.2	25.9

In the case of the Spanish Large Corpus evaluation, the low accuracy of the labels at line level particularly affects the segmentation scores for all the metrics. As it can be seen in Table 13, the error rates of the metrics related to line boundaries are specially higher than the rates of subtitle boundaries. It can be explained by the really good performance of the B-SU, E-SU, and BE-SU labels involved in specifying the subtitles boundaries, achieving F1-Score of 96.1%, 96.2%, and 96.1% respectively. On the contrary, the B-LI, E-LI, and BS-EL labels scored accuracies of 59.1%, 59.1%, and 1.4% respectively.

6.2.2. Testing and comparison

In this subsection, the Spanish CRF model and SVM based classification method, both built over the Spanish Large Corpus, which does not contain speaker changes, are compared at segmentation level. Firstly, the results reached at label level through the CRF model are presented in Table 14 and Table 15. In this case, the impact of not having speaker changes marks is clearly appreciated in all the labels related to describe the subtitles and lines boundaries. The performance of B-SU and E-SU labels at subtitle level has decreased clearly when comparing with the Basque corpus. Besides, the precision of the B-LI and E-LI labels does not reach 45%. The only label which has kept a good performance is the I-LI label, achieving a F1-Score value of 90.4%.

Table 14: Precision, Recall and F1-Score values for each class label for the Spanish Large corpus test set.

Class labels evaluation				
	#	Precision	Recall	F1-Score
I-LI	11,966	88.3%	92.5%	90.4%
B-SU	1,094	98.2%	61.7%	75.7%
E-SU	1,093	98.2%	61.7%	75.7%
B-LI	1,788	44.7%	56.7%	50.0%
E-LI	1,789	44.8%	56.8%	50.1%

Table 15: Accuracy at class label level for the Spanish Large corpus test set.

#Correct	14,316
#Labels	17,731
Accuracy	80.7%

The results obtained at segmentation level over the test data set of the Spanish Large corpus are presented in Table 16 for CRF and SVM based classification method. Naturally, the lower performance of the previously described labels should affect directly all the metrics which measured segmentation of the CRF model. However, the differences at line level are not very low when comparing with the results obtained in the training and evaluation phase, where the performance

of the B-SU/E-SU and B-LI/E-LI labels was better. Even though, the error rates grew notably at subtitle level if we compared with the results in Table 13.

Table 16: Segmentation scores of CRF and SVM models for the Spanish test data, without including speaker change information

	Segmentation score - No Speaker Change						
	F1-LINE	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
CRF	80.7	39.4	38.2	58.0	61.6	38.8	28.4
SVM	41.4	146.6	119.2	159.4	140.0	82.7	66.6

As in Table 10, both the accuracy of F1-LINE and the error rates of the other metrics are outperformed by the CRF model when comparing with the SVM based classification method, which obtained error rates higher than the 100% for the NIST-SU and DSER metrics. The main reason for these high error rates is that the SVM based classification method generate candidates for any type of break, without distinguishing between line and subtitle breaks. Using the break points proposed by the SVM based classification method, we assigned automatically labels to each word of the test contents, generating two-lines subtitles consecutively from the beginning of each content. This was the only way to create subtitles using the SVM based classification method's output, since no more information was provided by this method. This procedure could therefore generate multiple errors in tagging words with incorrect labels, and mainly in differentiating between the E-LI and E-SU labels. In addition, it has to be considered that the Spanish Large corpus contains multiple segments with spontaneous speech, unfinished sentences and words, and expressions from Spanish dialects such as Andalusian and Argentine. One of the main parameters in the SVM based classification method, which is described in detail in [5], was the perplexity given by a language model (LM) built on the train data and using Part-Of-Speech (POS) tags as units. The difficulties posed by these type of contents produced mistakes in the POS information extraction, and thus in the high perplexities given by the LM. This also affected segmentation error rates to be extremely high for the SVM based classification method.

6.3. Spanish Small Subcorpora

6.3.1. Training and evaluation

The results reached during the training and evaluation of Spanish CRF models for the two subcorpora (23 episodes) with and without speaker change information and applying 10-fold cross-validation technique are presented in this subsection. This evaluation was focused on checking the impact of having speaker change information in the accuracy of the subtitle segmentation. Table 17 describes the results at class label, and the accuracy along with the number of correctly tagged labels are shown in Table 18. The BL-ES, BS-EL and BE-SU labels are not shown because of their low count. Finally, the results at segmentation level are presented in Table 19.

Table 17: Precision, Recall and F1-Score values for each class label applying 10-fold cross-validation in the Spanish Small corpora, with and without speaker change information.

		Class labels evaluation					
		With Speaker Change			Without Speaker Change		
	#	Precision	Recall	F1-Score	Precision	Recall	F1-Score
I-LI	136,933	89.9%	93.4%	91.6%	89.0%	92.4%	90.7%
B-SU	20,135	85.7%	71.6%	78.0%	85.4%	70.9%	77.5%
E-SU	20,137	85.8%	71.6%	78.0%	85.4%	70.9%	77.5%
B-LI	17,040	55.2%	57.3%	56.2%	53.0%	55.6%	54.3%
E-LI	17,038	55.2%	57.3%	56.2%	53.0%	55.6%	54.3%

Although the hypothesis was that the speaker change information should help improving the results, this issue was not clearly demonstrated for the Spanish Small corpus in 10-fold cross-validation technique. As it can be seen in Table 17, the differences of the B-SU and E-SU labels

Table 18: 10-fold cross-validation accuracy at class label level in the Spanish Small corpora, with and without speaker change information.

	With speaker change	Without speaker change
#Correct	176,301	174,025
#Labels	211,317	
Accuracy	83.4%	82.3%

in terms of F1-Score are minimum between the two corpora, with and without speaker changes. The improvement is only 0.5 percentage points. The labels which represent the lines boundaries have a similar behavior, achieving improvements of almost 2 percentage points on average. These small improvements are also present in Table 19. Even if all the error rates for the contents with speaker change were lower, the differences with the contents without speaker changes are not as clear as expected. The main reason could be related to the small size of the corpus used for these experiments. As it was described in Table 1, the Spanish Small Corpus was composed by a total amount of 20,154 subtitles, which corresponds to a quarter of the Spanish Large Corpus.

Table 19: 10-fold cross-validation scores at segmentation level in the Spanish Small corpora, with and without speaker change information.

	Segmentation evaluation					
	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
With spk ch	40.3	32.5	64.2	54.4	33.8	25.3
Without spk ch	41.2	36.1	65.4	60.0	34.5	28.0

6.3.2. Testing and comparison

In this last subsection, the CRF model and the SVM based classification method are compared for the two Spanish Small Corpora. Tables 20 and 21 present the scores for each label, whilst Table 22 shows the segmentation score and error rates for each corpus and classification method.

Table 20: Precision, Recall and F1-Score values for each class label applying 10-fold cross-validation in the Spanish Small corpora, with and without speaker change information.

		Class labels evaluation					
		With Speaker Change			Without Speaker Change		
	#	Precision	Recall	F1-Score	Precision	Recall	F1-Score
I-LI	6,833	89.9%	92.4%	91.1%	88.6%	90.6%	89.6%
B-SU	818	86.1%	70.3%	77.4%	84.0%	69.2%	75.9%
E-SU	818	86.3%	70.3%	77.5%	84.2%	69.2%	76.0%
B-LI	942	52.4%	58.2%	55.2%	48.9%	54.8%	51.7%
E-LI	942	52.4%	58.3%	55.2%	48.9%	54.9%	51.7%

The differences between the scores obtained with and without speaker information using separated train and test partitions are more significant than when 10-fold cross-validation technique was applied. The improvements at subtitle and line levels are around 1.5% and 4% percentage points respectively in Table 20. The better performance of the CRF method against the SVM based classification method is demonstrated again in Table 22 for all the metrics. It is interesting to observe how the results of the SVM based classification method are better in this case comparing with the rates obtained with the whole Spanish corpus given in Table 16. For instance, the F1-LINE metric scores 45.4% of accuracy over the Spanish Small Corpus which does not include speaker changes, whilst an accuracy of 41.4% was obtained on the Spanish Large Corpus. The same tendency is kept for the rest of metrics. It can be explained by the fact that in the SVM based classification method the labels are almost randomly assigned just following the break marks given in the output. Hence, it seems that this method is prone to generate more errors as more subtitles are given to test.

Table 21: 10-fold cross-validation accuracy at class label level in the Spanish Small corpora, with and without speaker change information.

	With speaker change	Without speaker change
#Correct	8,539	8,342
#Labels	10,353	
Accuracy	82.5%	80.6%

Table 22: Segmentation scores of CRF and SVM models for the Spanish test data, including or not speaker change information

		Segmentation score						
		F1-LINE	NIST-SU-SUB	NIST-SU-LI	DSER-SUB	DSER-LI	SegER-SUB	SegER-LI
With	CRF	82.5	40.8	33.2	64.5	56.2	34.4	25.8
spk ch	SVM	47.5	115.8	90.8	136.6	119.7	74.2	59.5
Without	CRF	80.6	43.7	38.5	68.2	64.2	36.2	29.6
spk ch	SVM	45.4	118.0	93.6	138.6	120.0	76.1	61.1

On the contrary, the results of the CRF model are more consistent with the size of the corpus used to train and test models. In comparison with the Table 16, the metrics achieved higher error rates for the Spanish Small Corpus. Finally, the impact of the speaker change parameter is demonstrated in Table 22. Although the experiments were carried out with a small corpus, the error rates were lower for all the metrics in the corpus with speaker changes. The higher difference is given by the DSER-LI metric with a difference of 8 percentage points, reaching an error rate of 56.2% and 64.2% for the corpus with and without speaker changes respectively.

6.4. General discussion

Comparing all results from a general point of view, the first remarkable issue is that using CRF for assigning the different subtitle labels to the words is a valid alternative, since in all cases average accuracy is higher than 80%. Comparing with the previous SVM approximation employed in [5], it supposes a large impact on subtitling quality, specially for Spanish language (where SVM results present an accuracy lower than 50%), although Basque language presents a significant improvement as well.

Examining the results, the general tendency is having a more accurate subtitle segmentation than in-line segmentation. This is reasonable since begin and end of subtitles present more specific clues to detect its presence (e.g., punctuation marks, silences, speaker changes, etc.) than line breaks. When looking at the whole subtitle segmentation with respect to line segmentation (i.e., the one that includes all lines as units, independently if they are starting or end lines for subtitles), the general tendency is that line segmentation presents lower error than subtitle segmentation, which is reasonable since subtitle boundaries are a subset of line boundaries, and subtitle segmentation accuracy affects line segmentation accuracy.

However, in a few cases (Basque comparison, Table 10, and Spanish Large comparison, Table 16), differences show an irregular behavior, and even in the Spanish Large Corpus cross-validation experiments (Table 13), the tendency is the opposite. This can be explained by the nature of the corpora and the behavior of the classifier: Spanish Large Corpus presents a high proportion of lines in each subtitle (around 1.8 lines per subtitle, in contrast to what occurs with the Basque corpus with around 1.5 lines per subtitle), and presents a much lower relative accuracy of line boundaries labels (B-LI and E-LI) than the other cases (relative F1-Measure difference is about 60%, in contrast to about 30% in the Basque cross-validation and about 40% in the Spanish Small cross-validation). The combination of the two factors (higher number of lines and lower accuracy for detecting line boundaries with respect to subtitle boundaries) explains the different behavior, since there are more line boundaries to detect and they are detected with less precision, making the whole line segmentation error higher than the simple subtitle segmentation error. Similar arguments explain the irregular behavior of Basque comparison (less lines per subtitle but much lower

detection of line boundaries) and Spanish Large comparison (same number of lines per subtitle but not so low performance on the detection of line boundaries).

These issues allow us to suppose that, given the nature of the corpus (specially proportion of lines by subtitle) and the classifier (accuracy in detecting line boundaries with respect to subtitle boundaries), different performances can be expected at the two levels (subtitle- and line-level) and decisions can be taken on the use of more specialized models for the nature of the corpus, which will allow to obtain more accurate results for the subtitle segmentation task.

In any case, CRF represents a new milestone in this task since results are in all cases much better than the current alternative (SVM based classification method) and the decoding time is really fast (less than 0.1 milliseconds per subtitle in an Intel(R) Core(TM) i7 computer at 3.4 GHz with 16 GB of RAM) with respect to that provided by the SVM based classification method.

7. Productivity gain evaluation

With the aim of testing the efficiency of the CRF classifier, an experiment was carried out with human evaluators. The experiment consisted of measuring the effort of post-editing the segmentation of subtitles generated using two techniques: (1) subtitles segmented considering the maximum number of characters permitted per line, which is the main technique employed in most automatic subtitling systems currently, and (2) subtitles segmented using the information provided by the CRF classifier. Results from both techniques were finally compared to evaluate whether using the CRF-based classification method was more productive and facilitates the process of generating quality subtitles.

Nine students of the Subtitling Module included in the UAB's (Universitat Autònoma de Barcelona) METAV[6] and MTAV[7] Masters Programs volunteered to participate in the evaluation. In addition to the subtitling practice acquired through the masters program, they all had further subtitling expertise varying from one month to three years.

The experiment was performed over the Spanish corpus containing the information related to speaker change, composed of a total amount of 20,154 subtitles, 1,004 of which were used for testing purposes. This test set was first divided into smaller sets of 50 subtitles each, which were generated using both the CRF-based classification method and counting characters method. Each participant was then asked to post-edit the segmentation of two sets, each of which had been segmented using one of the two techniques. In order not to influence the post-editing task, the evaluation sets assigned to each post-editor contained different subtitles. The participants received some previous guidelines on the manner they had to post-edit and correct the subtitles, including some specific and reference rules for a proper segmentation. Subtitling Workshop[8] and the Toggl[9] tools were employed as subtitling and time tracking software, respectively. After finishing the task, participants generated a Toggl report including the time required to complete it.

Figure 2 shows the time in minutes per subtitle (mps) needed by each participant to post-edit the 50 subtitles in the two sets segmented with the described two methods.

As it can be seen in Figure 2, all post-editors needed more time to post-edit a subtitle in the test set segmented with the counting characters technique. On average, it took them 0.3 minutes to post-edit a subtitle segmented with the CRF-based method and 0.88 minutes to post-edit a subtitle segmented with the counting characters method, which is almost 3 times longer overall. These differences are more noticeable in some cases. For instance, P8 needed, on average, only 0.3 minutes to post-edit a subtitle segmented with the CRF classifier and 1.35 minutes to post-edit a subtitle segmented with the other method under evaluation. It is worth mentioning that P8 was one of the most experimented participant in the manual generation of subtitles.

[6] http://metav.uab.cat
[7] http://pagines.uab.cat/mtav
[8] http://subworkshop.sourceforge.net/
[9] https://toggl.com/

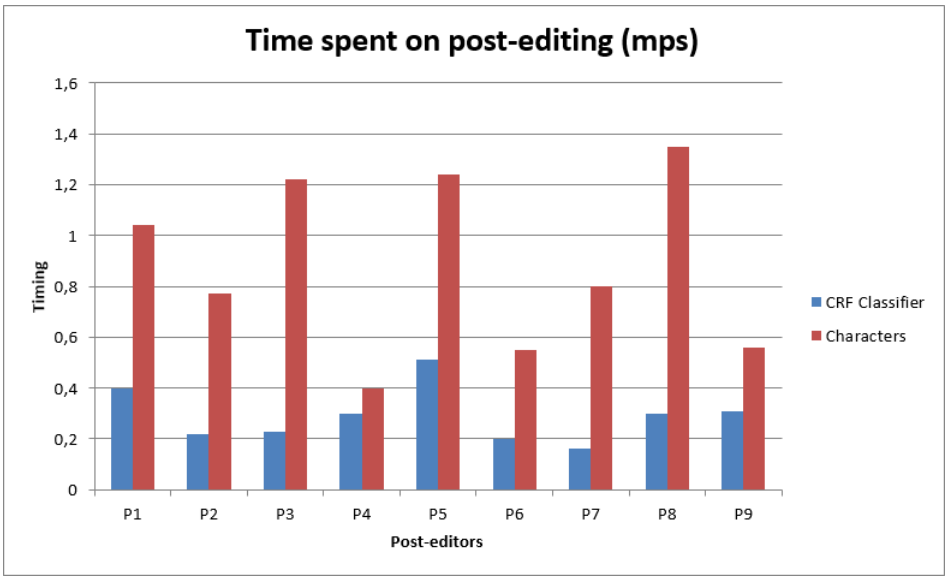


Fig. 2: Productivity evaluation results

The results demonstrated that it was much faster to post-edit the subtitles segmented with the CRF classifier, making the post-editing task an easier and more pleasant activity.

8. Conclusions and Future Work

The use of CRF for automatic segmentation of subtitles allowed us improving the results obtained in [5] in the following points: (1) differing between the type of breaks (line- and subtitle-breaks), (2) obtaining much better scores and thus generating more and better segmented subtitles and (3) faster processing time. The first point is given by the methodology we employed to construct the CRF models, which was focused on modeling transitions between labels corresponding to each word and its function within the subtitle. The second point was demonstrated in Section 6, in which we showed how the CRF models outperformed the results obtained by the SVM based classification method for different types of corpora in Basque and Spanish. For the third point, we presented computation times at subtitle level for each classification method, making clear that CRF model (less than 0.1 milliseconds per subtitle) needed much less decoding time than the SVM classification method (16 seconds per subtitle) on similar computers. Finally, a productivity study was presented with human evaluators, which allowed us to show that post-editing subtitles created through the CRF model took less time than to generate them from those obtained by using a more naive method.

The future work will involve experimentation with Recurrent Neural Networks (RNNs) for the task of automatic segmentation. RNNs have been proven to be useful for sequence labeling due to their several and attractive properties, including that they are able to make use of the past and future contextual information, and that they are robust to possible local distorsions of the input sequence [23]. In addition, more parameters should be explored to test their impact in this labeling task, such as stop words, syntactic functions or grammatical relations of the different clauses within a sentence. Finally, given that automatic subtitling is an alternative for live broadcasts, for which traditional manual subtitling is less effective, a solution for real-time automatic segmentation should be developed. Considering their computing and decoding time, CRF graphical models would be an interesting solution, but features like POS information should be removed because of the time

needed for their computation. Hence, new CRF models should be built including new combinations of different feature sets for the live broadcast environment.

References

[1] A. Álvarez, C. Mendes, M. Raffaelli, T. Luís, S. Paulo, N. Piccinini, H. Arzelus, J. Neto, C. Aliprandi, A. del Pozo, Automating live and batch subtitling of multimedia contents for several european languages, Multimedia Tools and Applications (2015) 1–31.

[2] D. J. Rajendran, A. T. Duchowski, P. Orero, J. Martínez, P. Romero-Fresco, Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination, Perspectives 21 (1) (2013) 5–21.

[3] G. D'Ydewalle, J. V. Rensbergen, 13 Developmental Studies of Text-Picture Interactions in the Perception of Animated Cartoons with Text, Advances in Psychology 58 (1989) 233–248.

[4] E. Perego, F. Del Missier, M. Porta, M. Mosconi, The Cognitive Effectiveness of Subtitle Processing, Media Psychology 13 (3) (2010) 243–272.

[5] A. Álvarez, H. Arzelus, T. Etchegoyhen, Towards customized automatic segmentation of subtitles, in: Advances in Speech and Language Technologies for Iberian Languages, Vol. 8854 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 229–238.

[6] E. Perego, Subtitles and line-breaks: Towards improved readability, Vol. 78, John Benjamins Publishing, 2008, pp. 211–223.

[7] S. Nowozin, C. H. Lampert, Structured learning and prediction in computer vision, Found. Trends. Comput. Graph. Vis. 6 (3-4) (2011) 185–365. doi:10.1561/0600000033.
URL http://dx.doi.org/10.1561/0600000033

[8] Y. Liu, J. Carbonell, P. Weigele, V. Gopalakrishnan, Protein fold recognition using segmentation conditional random fields (scrfs), Journal of Computational Biology 13 (2) (2006) 394–406.

[9] A. Gunawardana, M. Mahajan, A. Acero, J. C. Platt, Hidden conditional random fields for phone classification, in: Interspeech, 2005, pp. 1117–1120.

[10] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 188–191. doi:10.3115/1119176.1119206.
URL http://dx.doi.org/10.3115/1119176.1119206

[11] D. Roth, W.-t. Yih, Integer linear programming inference for conditional random fields, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, ACM, New York, NY, USA, 2005, pp. 736–743. doi:10.1145/1102351.1102444.
URL http://doi.acm.org/10.1145/1102351.1102444

[12] F. Sha, F. Pereira, Shallow parsing with conditional random fields, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 134–141. doi:10.3115/1073445.1073473.
URL http://dx.doi.org/10.3115/1073445.1073473

[13] F. Peng, F. Feng, A. McCallum, Chinese segmentation and new word detection using conditional random fields, in: Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 562–568. doi:10.3115/1220355.1220436.
URL http://dx.doi.org/10.3115/1220355.1220436

[14] C.-D. Martínez-Hinarejos, J.-M. Benedí, V. Tamarit, Unsegmented dialogue act annotation and decoding with n-gram transducers, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (1) (2015) 198–211.

[15] C. Sutton, A. McCallum, An introduction to conditional random fields, Foundations and Trends in Machine Learning 4 (4) (2012) 267–373.

[16] G. B. Flores d'Arcais, Syntactic processing during reading for comprehension., Lawrence Erlbaum Associates, Inc, 1987.

[17] M. E. Coltheart, Attention and performance 12: The psychology of reading., Lawrence Erlbaum Associates, Inc, 1987.

[18] N. Ezeiza, I. Alegria, J. M. Arriola, R. Urizar, I. Aduriz, Combining stochastic and rule-based methods for disambiguation in agglutinative languages, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 1998, pp. 380–384.

[19] R. Agerri, J. Bermudez, G. Rigau, Multilingual, Efficient and Easy NLP Processing with IXA Pipeline, in: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 5–8.

[20] A. Álvarez, P. Ruiz, H. Arzelus, Improving a long audio aligner through phone-relatedness matrices for english, spanish and basque, in: P. Sojka, A. Horák, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue, Vol. 8655 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 473–480.

[21] T. Kudo, Crf++: Yet another crf toolkit, Software available at http://crfpp. sourceforge. net.

[22] NIST, Nist website: Rt-03 fall rich transcription, http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/index.html (2003).
[23] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, Vol. 385 of Studies in Computational Intelligence, Springer, 2012. doi:10.1007/978-3-642-24797-2.
URL http://dx.doi.org/10.1007/978-3-642-24797-2

7.9 Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles

- **Authors:** Aitor Álvarez, Marina Balenciaga, Arantza del Pozo, Haritz Arzelus, Anna Matamala, Carlos-D. Martínez-Hinarejos
- **Booktitle:** Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)
- **Year:** 2016

Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles

Aitor Álvarez[1], Marina Balenciaga[1], Arantza del Pozo[1], Haritz Arzelus[1], Anna Matamala[2], Carlos-D. Martínez-Hinarejos[3]

[1] Human Speech and Language Technology Group, Vicomtech-IK4, San Sebastian, Spain

[2] Department of Translation, Interpreting and East Asian Studies, UAB, Barcelona, Spain

[3] Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Spain

[1]{aalvarez, mbalenciaga, adelpozo, harzelus}@vicomtech.org, [2]anna.matamala@uab.cat, [3]cmartine@dsic.upv.es

Abstract

This paper describes the evaluation methodology followed to measure the impact of using a machine learning algorithm to automatically segment intralingual subtitles. The segmentation quality, productivity and self-reported post-editing effort achieved with such approach are shown to improve those obtained by the technique based in counting characters, mainly employed for automatic subtitle segmentation currently. The corpus used to train and test the proposed automated segmentation method is also described and shared with the community, in order to foster further research in this area.

Keywords: automatic subtitling, subtitle segmentation, machine learning

1. Introduction

Society and the governments are increasingly requesting larger amounts of subtitled TV content (Neto et al., 2008), since subtitles are the most practical technique to guarantee the accessibility of audiovisual material to those who cannot access the audio (AENOR, 2012).

The current demand and promising future of intralingual subtitling has accelerated research into more productive methods that help to cover challenging subtitling situations such as, for example, live broadcasts. In recent years, technological advances in speech recognition have enabled automatic intralingual subtitling to be a reality (Álvarez et al., 2015). However, automatic subtitling technology has limitations, a major one being its inability to segment subtitle text in a logical way. Within the subtitling field, segmentation refers to the division of the original text into sections that viewers can understand immediately (Díaz-Cintas and Remael, 2007), playing a fundamental role in the creation of quality subtitles.

This work analyses the application of a machine learning algorithm to automatically segment the text contained in intralingual subtitles and compares its performance against that of the main technique based in counting characters employed in most automatic subtitling systems currently. Its impact is measured in terms of subtitle quality and regarding the productivity achieved and self-reported effort when integrated in the subtitling process through post-editing. Also, given the little work carried out so far in automatic subtitle segmentation, the corpus employed to train and test the presented machine learning algorithm is described and shared with the aim of fostering further research and technology development.

2. Background

2.1. From traditional to automatic subtitling

Traditional subtitling is carried out by professionals who aim to reproduce in text on screen the original dialogues, the discursive elements in the image and, when addressed to those who cannot hear the original audio, the information contained in the soundtrack of audiovisual contents (Díaz-Cintas and Remael, 2007).

From a linguistic perspective, subtitles can be classified as intralingual, interlingual or bilingual. Depending on the time available for their preparation, they can be prerecorded, live or semi-live. And according to the recipient, subtitles can either be for the hearing or for the deaf and hard-of-hearing, the latter containing additional information to facilitate comprehension, such as contextual information or sound effects (Díaz-Cintas and Remael, 2007).

Several studies collect good subtitling practices (Karamitroglou, 1998; Ivarsson and Carroll, 1998; Díaz-Cintas and Remael, 2007; Ford Williams, 2009; Ofcom, 2015). Also, there are some regulations governing quality subtitling standards such as the UNE 153010 (AENOR, 2012). The main features of good subtitles can be classified into:

- **Spacing features:** distributing the text into one or two lines between 4 and 43 characters.
- **Timing features**: showing subtitles at a speed of 130-170 words per minute keeping them on screen between 1 and 6 seconds, while synchronizing with the audio and inserting short pauses between consecutive subtitles.
- **Linguistic features**: keeping the original terms and avoiding more than two sentences per subtitles, one per line.
- **Orthotypographic features**: following the general guidelines of printed text.

All of the above features impact subtitle segmentation and are taken into account for manual segmentation by

professional subtitlers.

Automatic subtitling was born in response to a high subtitling demand, as a more productive alternative that enabled subtitling in challenging situations, such as live broadcasts, where traditional subtitling was not directly applicable. However, at present, automatic subtitling is yet not capable of creating subtitles that equal human quality and, thus, its focus is on facilitating the generation or post-editing of automatic subtitles by professional subtitlers, both in live and pre-recorded settings. In this context, post-editing can be defined as the process by which a professional edits, modifies and/or corrects the output of an automatic subtitling system. Post-editing is increasingly gaining relevance as it is proving to help achieve subtitles of high quality in a more productive way.
The different types of technology that can be employed for automatic intralingual subtitling are:

- **Stenotyping:** It involves using a shorthand typewriter representing syllables, words, punctuation signs and/or phrases phonetically. Allowing the generation of subtitles at speeds of 220 and 300 words per minute, it is generally used for live subtitling. Its precision reaches 97-98%, the generated delay is low and the severity of errors medium. However, learning this technique requires a long time (around three years) and the cost is high (Romero-Fresco 2011).

- **Respeaking:** This technique involves producing real-time subtitles by means of speech recognition software transcribing a simultaneous reformulation of the source text dictated by the respeaker to the computer (Eugeni, 2008). Being easier to master than stenotyping and similiar in average performance, Respeaking has recently become the most widely used method for live subtitling in countries such as the UK, France or Germany (Mikul, 2014).

- **Automatic transcription:** This technology involves the generation of subtitles directly from the source audio, without the need of human intervention. In state-of-the-art systems, after a pre-processing step to normalize the audio and select the segments with contain speech, a speech recognition software transcribes the speech detected and synchronizes it with the audio. A posterior linguistic processing normalizes the numbers, abbreviations and acronyms, and capitalization and punctuation marks are automatically included. Finally, subtitle segmentation is performed and subtitles are generated in the required format. Various subtitling solutions based on automatic transcription systems have been developed for several languages and domain in recent years (Neto et al., 2008; Ortega et al., 2009; Álvarez et al., 2015). Although they do not perform as well as professional subtitlers, they have achieved promising results with productivity gain experiments suggesting that post-editing automatic subtitles is faster than creating them from scratch (Álvarez et al., 2015).

2.2. Subtitle segmentation

While good segmentation facilitates reading and understanding subtitles (Ford Williams, 2009), bad segmentation can interrupt the natural reading flow, make the audience lose concentration and obscure the subtitle message (Perego et al., 2010).
Good subtitle segmentation involves making each subtitle constitute a complete linguistic unit, according to the main rules of syntax and semantics. In traditional subtitling, the concept of the "highest syntactic node" (Karamitroglou, 1998) is widely employed, establishing that each subtitle line should contain the highest possible level of syntactic information.
The guidelines governing segmentation quality involve the following most relevant criteria (AENOR, 2012; Díaz-Cintas and Remael, 2007):

- Take advantage of silences, grammatical pauses and punctuation signs.
- Do not divide noun-, verb- or prepositional-phrases.
- Do not split compound verb forms, and words.
- In subtitles consisting of two sentences, place each sentence in one line. If compound sentences do not fit into one line, use a line per proposition. Write conjunctions and nexus in the bottom line. If simple sentences require division, put the subject in the top line and the predicate in the bottom line. With question-answers, place the question in the top line and the answer in the bottom line, unless information is exposed too soon this way.

To date, most of the automatic subtitling solutions have not been able to discriminate the natural pauses, syntactic and semantic information relevant for quality segmentation and, thus, automatic segmentation in stenotyping, respeaking or audio transcription applications is mainly performed considering only the maximum number of characters allowed per line or through manual intervention. Machine learning has only recently started to be applied for automatic segmentation. The first reference in the field (Álvarez et al., 2014) focused in the development of Support Vector Machines (SVM) and Logistic Regression (LR) classifiers to automatically segment subtitle text considering the good practices of traditional subtitling.

3. Technical approach

With the aim of improving the results in (Álvarez et al., 2014), a new classification approach was developed in this study. Given that subtitle segmentation can be treated as a text labeling problem, in which each of the words in subtitles carries a specific function, Conditional Random Fields (CRFs) were used as the main machine learning algorithm for the automatic segmentation task. CRFs are often employed for labeling and parsing sequential natural language text (Lafferty et al., 2001) and unlike other classifiers, such as SVM or LR, CRFs consider the surrounding observations to predict the current label. This is an important feature, since predicting the optimal segmentation point depends not only on the current word, but also on the surrounding context.

The feature vectors, which describe the information related to each word during classification, were composed by the following characteristics: (1) the current an the surrounding 2 words, (2) the Part-Of-Speech (POS) information of the current and the surrounding 2 words, (3) the amount of characters per line and subtitle, (4) speaker change information, and (5) the time differences between the current, previous and next words.
The CRFs were trained and evaluated with the CRFSuite software (Naoaki Okazaki, 2007), and the POS information was extracted using the ixa-pipe-pos toolkit (Agerri et al., 2014).

4. Corpus characteristics

The corpus was composed by 23 episodes of the Spanish "Mi Querido Klikowsky" TV series, containing 1,150 minutes and a total amount of 20,154 subtitles, 90% of which was used to train the CRF model, and the rest was kept for testing purposes. The subtitle files, provided in SRT format, were created manually by professionals, and their segmentation was performed following specific predefined rules to keep linguistic and syntactic coherence. The contents include many segments with spontaneous speech, grammatically incorrect sentences, and some words and expressions pronounced in several Spanish dialects, such as Argentinian and Andalusian. Because of this, the POS tagger made more mistakes than desired.
The corpus described above (EiTB_Subt_Corp) will be made available to the research community through the META-SHARE repository[1] under the Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA) license.

5. Evaluation methodology

In order to analyze the impact of the proposed CRF-based segmentation approach described in section 3., we have carried out evaluation at three different levels. First, segmentation quality has been measured through objective metrics. Second, a post-editing experiment has been conducted to test whether the application of the developed algorithm affects productivity. Third, subjective feedback regarding the post-editing task has been collected measuring the self-reported effort of post-editors through a Likert scale questionnaire.

5.1. Quality evaluation

The segmentation quality of subtitles was calculated in terms of precision, recall and F1-score as follows:

$$Precision = \frac{correct\ segmentations}{total\ number\ of\ segmentations}$$

$$Recall = \frac{correct\ segmentations}{total\ number\ of\ correct\ segmentations}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

The test set described in section 4. was segmented using both, the proposed machine learning algorithm and the technique based in counting characters. Then, the precision, recall and F1-score achieved with both methods were compared.

5.2. Productivity evaluation

Whether the developed segmentation approach facilitates and streamlines the process of generating quality subtitles was evaluated through a post-editing task.
Nine students of the Subtitling Module included in the UAB's METAV[2] and MTAV[3] Masters Programs volunteered to participate in the test. Eight of them were Translation and Interpretation graduates, while one had a degree in Business Administration. In addition to the subtitling practice acquired through the masters program, several participants had further experience: three of them had been subtitling between one and three years and other four between one and six months. One participant worked as a professional subtitler at the time of the experiment and another participant had a six-month post-editing experience.
The post-editing task was arranged as follows. First, the test set described in section 4. was divided into smaller sets of 50 subtitles, which were segmented using both, the proposed algorithm and the counting characters method. Then, each participant was asked to post-edit the segmentation of two of these sets, each of which had been segmented using one of the two techniques under evaluation. In order not to influence the post-editing task, the evaluation sets assigned to each post-editor contained different subtitles. Subtitling Workshop[4] and the Toggl[5] tools were employed as subtitling and time tracking softwares, respectively. After finishing the task, participants generated a Toggl report including the time required to complete it.

5.3. Self-reported effort

In order to gather additional subjective information regarding the post-editing experience of the volunteers, they were asked to rate:

- the self-reported effort expended post-editing each of the subtitle sets on a 1 to 5 scale (1 being the lowest and 5 the highest)
- their level of agreement/disagreement on a 1 to 5 scale (1 being "strongly disagree" and 5 being "strongly agree") with the statements shown in Table 2.

6. Results

6.1. Quality evaluation

Table 1 shows results of the quality evaluation. As it can be seen, results achieved by the CRF model outperform those of the counting character technique. With the machine learning method, 85.08% of the retrieved cuts are correct and 80.30% of the correct cuts that should be retrieved are generated. Such results go down to 22.08% and 15.43% in the case of the counting character segmentation technique.

[1] http://www.meta-share.eu/

[2] http://metav.uab.cat

[3] http://pagines.uab.cat/mtav

[4] http://subworkshop.sourceforge.net/

[5] http://toggl.com

Algorithm	Precision	Recall	F1-score
Counting Characters	22.08%	15.43%	18.17%
CRF model	85.08%	80.30%	82.62%

Table 1: Quality evaluation results

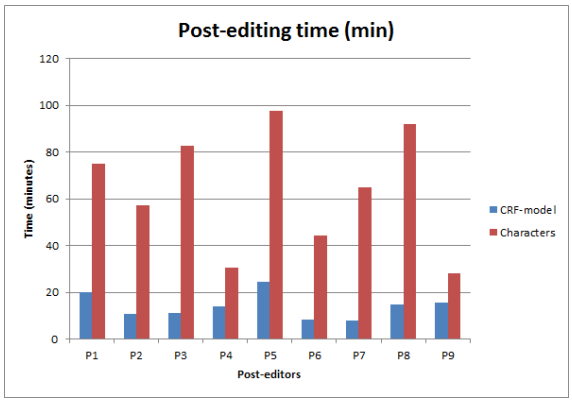


Figure 1: Productivity evaluation results in minutes

6.2. Productivity evaluation

The time in minutes required by each participant to post-edit the subtitle files is presented in Figure 1 for both segmentation methods. This time only includes the post-edition task; that is, the time needed to automatically segment the subtitles was not considered. The participants are ordered considering their previous subtitling experience; P1 being the most experienced post-editor, and P9 the participant with less experience.

As it can be appreciated, all participants needed more time to post-edit a subtitle in the test set segmented with the counting characters technique. Participants needed 14.2 minutes on average to post-edit a subtitle file segmented with the CRF-model, whilst it took them 63.7 minutes to post-edit the subtitles splitted with the counting characters method. In other words, participants needed 49 minutes more to post-edit the same number of subtitles (50) segmented with the counting characters method.

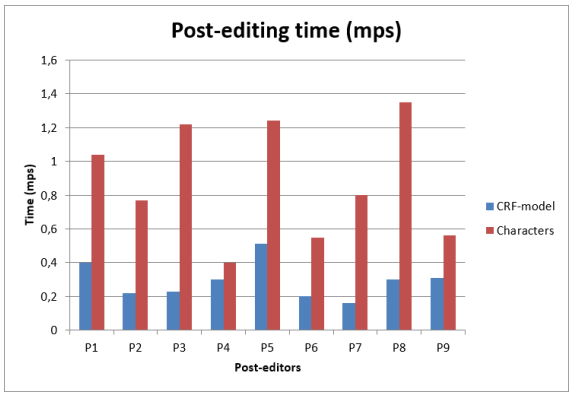


Figure 2: Productivity results in minutes per subtitles (mps)

With the aim of comparing the average time needed per each of participants to post-edit a subtitle in the two sets segmented with the methods under evaluation, the time measured in minutes per subtitle (mps) was computed and presented in Figure 2. On average, it took them 0.3 minutes to post-edit a subtitle segmented with the CRF-model and 0.88 minutes to post-edit a subtitle segmented with the counting characters method, which is 3 times longer overall. Thus, the presented machine learning algorithm allows post-editing segmentation faster, increasing productivity, and making the post-editing task more pleasurable.

6.3. Self-reported effort

Figure 3 shows the self-reported effort results of post-editing the segmentation of the two subtitle sets processed with the methods under evaluation, in a 1 to 5 scale (1 being the lowest and 5 the highest).

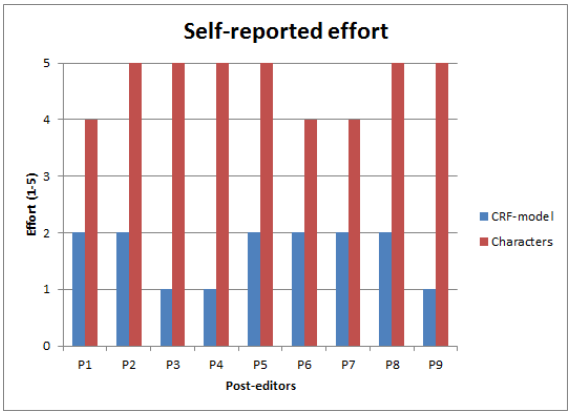


Figure 3: Self-reported post-editing effort results

As it can be seen, there is a clear difference between the self-reported post-editing effort of the two subtitle sets. That segmented using the machine learning algorithm received scores of 1 or 2, with an average of 1.66. On the other hand, the subtitle set segmented using the counting character technique was assessed with scores of 4 or 5, and 4.66 in average. Thus, according to the participants' assessment, post-editing subtitles segmented with the method based in counting characters took them more effort.

Participants also gave their opinion regarding the statements shown in Table 2, in a 1 to 5 scale (1 being "strongly disagree" and 5 being "strongly agree"). From the gathered results, it can be observed that in general most participants found it more enjoyable to post-edit subtitles automatically segmented by the machine learning algorithm, scoring its related statements higher. In particular, participants found it easier and less boring to post-edit, respect and use the guidelines of good segmentation with the machine learning algorithm on average. In addition, segmentations produced by such algorithm were perceived to be of better quality. And the resulting post-edited subtitles were also thought to be better segmented.

7. Conclusions and Future Work

This work has measured the impact of using a machine learning algorithm to automatically segment intralingual

Statement	Counting characters	Machine learning
I found it difficult to post-edit subtitle segmentation	3.88	1.44
I have been able to respect and use the guidelines of good segmentation	3.66	4.55
I found it boring to post-edit this subtitle file	3	2
Subtitles in this file were well segmented before being post-edited	1.11	3.22
I managed to achieve subtitles with good segmentation quality after post-editing this file	3.55	4.33

Table 2: Average subjective assessment results

subtitles in terms of quality, productivity and self-reported post-editing effort. Quality has been evaluated objectively through precision, recall and F1-score metrics; a post-editing task has been carried out to obtain objective measures of productivity; and the self-reported effort has been assessed subjectively through a ranking questionnaire. All evaluations have been performed through comparison with the main technique employed for automatic subtitle segmentation nowadays, which is based in counting characters. The quality achieved by the proposed CRF-based classifier has been shown to outperform that of the counting character technique by far. Post-editing productivity has shown to increase up to three times and the self-reported effort of the post-editing task to decrease three points. In addition, post-editors have found subtitle segmentations generated by the machine learning method to be of better quality, easier and less boring to post-edit respecting the guidelines of good segmentation and their post-edited versions thought to be better segmented. These successful results show the potential of machine learning to model the segmentation rules employed in traditional subtitling from a relatively small corpus of already segmented subtitles.

Future work should involve testing the proposed automatic segmentation approach more extensively on bigger datasets, different languages and speech recognition output. In addition, new classification methods should be tested for the automatic segmentation task. Among others, Recurrent Neural Networks (RNNs) have been proven to be useful for sequence labeling due to their explotable properties, including that they can make use of the past and future contextual information, and that they are robust to possible local distorsions of the input features (Graves, 2012). Finally, more parameters like stop words, syntactic functions or grammatical relations could be explored in order to check their impact on this task. The authors encourage researchers to look into these and other challenges, exploiting the released corpus of quality segmented subtitles as necessary.

8. Acknowledgements

Anna Matamala is member of transMedia Catalonia, which is a research group funded by the Catalan government (2014SGR027).

9. References

AENOR. (2012). Subtitulado para personas sordas y personas con discapacidad auditiva. UNE 153010:2012. Technical report, Madrid.

Agerri, R., Bermudez, J., and Rigau, G. (2014). Multilingual, Efficient and Easy NLP Processing with IXA Pipeline. *EACL 2014*, page 5.

Álvarez, A., Arzelus, H., and Etchegoyhen, T. (2014). Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238. Springer.

Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., and del Pozo, A. (2015). Automating live and batch subtitling of multimedia contents for several european languages. *Multimedia Tools and Applications*, pages 1–31.

Díaz-Cintas, J. and Remael, A. (2007). *Audiovisual Translation, Subtitling*. St. Jerome Publishing.

Eugeni, C. (2008). Respeaking the news for the deaf: for a real special needs-oriented subtitling. *Studies in English Language and Literature*, 21.

Ford Williams, G. (2009). Online subtitling editorial guidelines v1.1. Technical report, BBC.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer, Heidelberg, New York.

Ivarsson, J. and Carroll, M. (1998). Code of good subtitling practice. Technical report, ESIST (European Association for Studies in Screen Translation), Berlin.

Karamitroglou, F. (1998). A proposed set of subtitling standards in europe. *Translation Journal*, 2.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Mikul, C. (2014). Caption quality: International approaches to standards and measurement. Technical report, Media Access Australia.

Naoaki Okazaki. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast news subtitling system in portuguese. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 1561–1564, Las Vegas, Nevada. IEEE.

Ofcom. (2015). Code on television access services. Technical report.

Ortega, A., García, J., Miguel, A., and Lleida, E. (2009). Real-time live broadcast news subtitling system for Spanish. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton.

Perego, E., del Missier, F., Porta, M., and Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, 13(3):243–272.

7.10 Towards Customized Automatic Segmentation of Subtitles

- **Authors:** Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen
- **Booktitle:** Advances in Speech and Language Technologies for Iberian Languages (IberSPEECH)
- **Year:** 2014
- **Publisher:** Springer

Towards Customized Automatic Segmentation of Subtitles

Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen

Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
{aalvarez,harzelus,tetchegoyhen}@vicomtech.org
http://www.vicomtech.org/

Abstract. Automatic subtitling through speech recognition technology has become an important topic in recent years, where the effort has mostly centered on improving core speech technology to obtain better recognition results. However, subtitling quality also depends on other parameters aimed at favoring the readability and quick understanding of subtitles, like correct subtitle line segmentation. In this work, we present an approach to automate the segmentation of subtitles through machine learning techniques, allowing the creation of customized models adapted to the specific segmentation rules of subtitling companies. Support Vector Machines and Logistic Regression classifiers were trained over a reference corpus of subtitles manually created by professionals and used to segment the output of speech recognition engines. We describe the performance of both classifiers and discuss the merits of the approach for the automatic segmentation of subtitles.

Keywords: automatic subtitling, subtitle segmentation, machine learning.

1 Introduction

Automatic subtitling has recently attracted the interest of the speech and natural language processing research communities, notably after the adoption of new audiovisual legislation by the European Parliament in 2007. This legislation regulates the rights of people with disabilities to be integrated in the social and cultural life of the Community, through accessible audiovisual contents by means of sign-language, audio-description and subtitling. As a result, the demand for automatic subtitling has grown rapidly, with public and private TV channels moving to produce subtitles for larger volumes of their content. The effort has focused on quantity in a first step, in order to match legislative requirements, but there is an increasing demand for an improvement in the quality of automatically generated subtitles as well.

The quality of subtitles involves several parameters linked to subtitle layout, duration and text editing. Layout parameters include: the position of subtitles on screen; the number of lines and amount of characters contained in each line; typeface, distribution and alignment of the text; colors for front and

J.L. Navarro Mesa et al. (Eds.): IberSPEECH 2014, LNAI 8854, pp. 229–238, 2014.

background; different colors per speaker; and transmission modes, i.e. blocks or scrolling/word-by-word. Duration parameters involve delay in live subtitling and the persistence of subtitles on screen. Finally, text editing parameters are related to capitalization and punctuation issues, segmentation and the use of acronyms, apostrophes and numerals.

Among these quality features, the strong need for proper segmentation is supported by the psycholinguistic literature on reading [11], where the consensual view is that subtitle lines should end at natural linguistic breaks in order to favor readability and minimize the cognitive effort produced by poorly segmented text lines [21].

In order to address the need to tackle subtitle quality aspects beyond bare speech recognition and to provide solutions adaptable to the standard guidelines and specific rules of companies, we explored a flexible approach based on machine learning techniques and tested it on the automated segmentation task. Specifically, we trained Support Vector Machines and Logistic Regression classifiers on subtitle corpora created by professional subtitlers and used the resulting models to filter and select optimal segmentation candidates. The results we present involve the use of these two classifiers for the automatic segmentation of subtitles in Basque, although the approach is not language-specific as it only requires properly segmented training material of the type created by subtitling companies under their own guidelines.

This processing pipeline for segmentation has been integrated into the automatic subtitling system described in [3], taking the output of speech processing engines to provide customized segmented subtitles.

The paper is structured as follows. Section 2 describes existing solutions and studies regarding automatic subtitling and segmentation. Section 3 looks at standard issues and considerations for the segmentation of subtitles. Section 4 describes the machine learning approach we implemented. Section 5 presents the experiments and evaluation results. Finally, Section 6 draws conclusions and describes future work.

2 Related Work in Automatic Subtitling

There is extensive research focused on automatic subtitling, mainly through the using of Automatic Speech Recognition technology for the recognition and alignment tasks [2,6,13]. Most of the work in the area has centered on improving recognition accuracy and producing well-synchronized subtitles. Audimus [17] is a reference system in the field, as it provides a complete framework for automatic subtitling of broadcast news contents with low error rates in both batch and live modes. It includes an automatic module for subtitle generation and normalization aimed at improving readability, but segmentation is performed minimally, using only information about the maximum amount of characters permitted per line. The Audimus system was improved and extended to several languages within the European project SAVAS[1]. Many quality features were considered in

[1] http://www.fp7-savas.eu/

the development of the new systems, but technology for automatic segmentation was not included.

Although considerable importance is commonly placed on most of the quality parameters described above, proper line-breaking has generally been disregarded [20]. A survey of the literature in the field actually provides no references on the topic of automatically segmenting subtitles. A few studies were carried out on related topics, as can be found for instance in [20], which explores the way line-breaking is commonly performed, and in [21] which studies the impact of arbitrary segmented subtitles on readers. The importance of segmentation has been noted by [23], a study whose aim was to verify whether text chunking over live re-spoken subtitles had an impact on both comprehension and reading speed. They concluded that even though significant differences were not found in terms of comprehension, a correct segmentation by phrase or by sentence significantly reduced the time spent reading subtitles.

3 Subtitle Segmentation

3.1 Standard Guidelines

A number of guidelines for subtitling have been published over the years. Among well-known ones are: Ofcom's Guidance on Standards for Subtitling[2]; BBC's Online Subtitling Editorial Guidelines[3] ; ESIST's Guidelines for Production and Layout of TV Subtitles[4], the Spanish UNE 153010 norm [1] on subtitling for the deaf and hard of hearing and a reference textbook on generally accepted subtitling practice published in 2007 by Jorge Diaz-Cintas and Aline Remael [10]. Standard guidelines cover the various aspects of subtitle quality, such as subtitle segmentation, and standard practices along these recommendations are shared among subtitling companies and broadcasters.

In terms of segmentation, all standard recommendations conclude that it must benefit and improve readability. For this purpose, considering syntactic information to create linguistically coherent line-breaks is the preferred and most adopted solution in the community. This follows from results in psycholinguistic research, which show that readers analyze texts in terms of syntactic information [9], grouping words corresponding to syntactic phrases and clauses [8]. Reading subtitles is a similar task and subtitles for which segmentation is not based on coherent syntactic groups can thus be assumed to trigger sub-optimal reading [15]. In order to facilitate readability, subtitle lines should thus be split according to coherent linguistic breaks, and the generally accepted solution is to operate the splits at the highest possible syntactic node. This ensures that fragments split along these lines encompass the largest possible amount of related semantic information.

[2] http://www.ofcom.org.uk/static/archive/itc/itc_publications/codes_guidance/standards_for_subtitling/subtitling_1.asp.html

[3] http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf

[4] http://www.translationjournal.net/journal/04stndrd.htm

3.2 Issues in Automatic Segmentation

Although the strong need for proper segmentation and the general constraints that apply to it are clear, there are issues regarding the implementation of automated segmentation.

First, as the previously described guidelines are fairly general in terms of what constitutes a proper subtitle split, there is actual variation among professional subtitlers when it comes to executing the actual segmentation. These variants are usually reflected as distinct sets of company-specific rules, which makes a generic automated solution all the more difficult to achieve as such a solution would have to either disregard company-specific rules or require resource-consuming adaptation of syntactic rule sets on a case by case basis.

Secondly, the automatic detection of the highest syntactic node requires language processing tools for sentence analysis. For major languages like Spanish or English, several such tools are available, e.g. Freeling [18], OpenNLP [4] or the parsers developed by the Berkeley [22] and Stanford [5] groups. For other languages, particularly under-resourced ones, there can be a lack of robust natural language analyzers, which would limit the possibilities of using a syntax-based approach for segmentation.

Finally, a correct syntactic analysis and detection of the highest nodes in subtitles does not guarantee proper segmentation. Several other features have to be considered simultaneously, such as the amount of characters, timing issues and, as previously mentioned, the specific splitting rules used by each subtitling company. All these features have a clear impact on proper subtitle segmentation and need to be taken into account for each specific subtitle.

An ideal solution for the automatic segmentation of subtitles would thus have to (1) correspond to the specific rules used by each subtitling company, and (2) simultaneously consider all relevant information like character sequence length and timing.

In the remainder of the paper, we present a possible solution that involves the use of machine learning classifiers to create segmentation models adapted to each company's needs, thus providing a highly customizable and language-independent solution. This approach has the additional advantage of allowing the simultaneous integration of different features to reach optimal segmentation.

4 Machine Learning for Automatic Segmentation

This section describes the core components of the machine learning approach we followed. We define the automatic segmentation problem as a binary classification task, where subtitles with correct or incorrect segmentation are split into two classes. Positive (correct) feature vectors were extracted from professionally-created subtitle data and contain the segmentation marks found in the corpus; negative (incorrect) vectors were generated by automatically inserting improper segmentation marks. Classifiers were then trained on balanced sets formed with these two types of vectors and used for the segmentation task.

4.1 Corpus Characteristics

The corpus used to train and test the classifiers was composed of subtitles that were manually created by professional subtitlers for TV cartoon programs in Basque, for a total amount of 158,011 subtitles. The files were provided in SRT format, indicating start and end time codes for each subtitle and presented in blocks of a maximum of two lines. The corpus was split between training and test sets containing 80% and 20% of the data, respectively. The subtitles in the corpus were manually generated considering subtitle layout, duration and text editing features. In particular, the segmentation rules followed by the subtitlers focused on maintaining linguistic coherence, splitting subtitles according to the highest possible syntactic node.

4.2 Corpus Processing

In order to train classifiers, both positive and negative examples are necessary, from which to extract feature vectors suitable for the task. We thus prepared a balanced set of positive and negative sets by transforming the original subtitles into a task-specific format. Positive examples were generated by merging consecutive lines in reference subtitles into a single sentence containing the original segmentation mark. Each such transformed sentence was then used as a basis to generate a set of negative examples, by moving the original correct segmentation symbol to other positions in the sentence. All possible negative training examples were generated in a first step, each with a different segmentation point, and a randomly selected subset of these possible incorrect examples was used to balance the amount of positive and negative training elements.

Table 1 provides examples of transformed subtitles, including positive and negative candidates.

Table 1. Training data. The #S# mark denotes a segmentation symbol. The <S1> and <S2> marks correspond to speaker information marks.

Reference subtitles	*Transformed data (examples)*	*Label*
1 00:00:47,430 → 00:00:49,448 <S1>Lapitza eta erregela. 2 00:00:51,283 → 00:00:54,660 <S2>Ez dira berdinak, ezta pentsatu ere.	Lapitza eta erregela. #S# Ez dira berdinak	Correct
	Ez dira berdinak, #S# ezta pentsatu ere.	Correct
	Lapitza #S# eta erregela. Ez dira berdinak	Incorrect
	Lapitza eta #S# erregela. Ez dira berdinak	Incorrect
	Lapitza eta erregela. Ez #S# dira berdinak	Incorrect
	...	...

Training sentences were thus composed of two parts corresponding to each line in a subtitle and divided by the #S# symbol. Features were computed on each of the parts and on the entire sentence as well. Each feature vector was then categorized with the corresponding label.

4.3 Feature Vectors

The features extracted from the transformed data can be divided into four types of characteristics related to (1) timing, (2) number of characters, (3) speaker change and (4) perplexity as given by a language model built over the training data. A feature vector was calculated for each of the sentences in the transformed data and used to train the classifiers.

The timing feature involved the time difference between the first and second parts of each sentence in the transformed data. It was calculated from the start time of the first word of the second part and the end time of the last word of the first part. Since the reference subtitles provided just the time-codes of the first and last words at subtitle level, the forced alignment system for Basque presented in [3] was used to obtain the start and end times-codes for all the words.

To characterize aspects related to the number of characters, three features were calculated from the transformed data. The first two contained the amount of characters of the first part and second part of each sentence, respectively, and the third feature indicated the total number of characters in the entire (bi-)sentence.

Speaker change information was available in the reference subtitles and converted into a boolean value: speaker changes were defined to have value 1 if true and 0 otherwise.

The last feature indicated the perplexity value given by a language model built on the correct sentences in the transformed data. Given that Basque is a morphologically rich language and considering the scarcity of the training data, the language model was built using part-of-speech (POS) information. For this end, the Eustagger [12] toolkit was used, which includes a morphological analyser and a POS tagger for Basque. An unpruned 9-gram language model was estimated using the KenLM toolkit [14], with modified Kneser-Ney smoothing [16]. The average perplexity value was 24.25 on the test set.

4.4 Segmentation Algorithm

As previously mentioned, the automatic segmentation module was integrated into our automatic subtitling system for Basque. This system produces alignments between audio signal and transcripts, thus producing time-codes for each word and providing a basis for the complete generation of subtitles. The segmentation module benefits from the automatically aligned and time-coded words to create candidates for segmentation. These candidates are then measured against the machine-learned models and optimal candidates selected according to the score obtained by their feature vectors. The algorithm for candidate generation and selection is described below.

```
INITIALIZE start_index to zero;
INITIALIZE end_index to one;
SET max_length to maximum length of characters per subtitle;
CALL get_words() RETURNING words;
while end_index is less than length of words do
    COMPUTE generate_candidates(start_index,end_index);
    if exist_valid_candidates() then
        CALL get_best_candidate() RETURNING cut_index;
        COMPUTE insert_cut (cut_index);
        SET previous_maxindex to end_index + 1;
        SET start_index to cut_index + 1;
        SET end_index to start_index + 1;
        SET right_block to words (start_index...end_index);
    else
        if length of right_block is greater than one then
            COMPUTE insert_cut (previous_maxindex);
            SET start_index to previous_maxindex + 1;
            SET end_index to start_index + 1;
        else
            SET start_index to start_index + 1;
            SET end_index to start_index + 1;
        end
    end
end
```

Algorithm 1: Segmentation procedure

The procedure for the generation and selection of segmentation candidates is shown in Algorithm 1. Processing of the text to be segmented is iterative, with validated insertion points taken as new starting points for further processing of the remainder of the text. In other words, we compute segmentation points through short windows of text and repeat the process on the yet unprocessed text after an optimal segmentation has been found for the current window.

Potential points of segmentation are inserted between sequences of consecutive words, where the sole constraint is a maximum allowed sequence length before and after segmentation points. That is, neither sequence on either side of a potential segmentation point can have more characters than this fixed value, which is computed at the beginning of the process and comes from the maximum length in characters for a subtitle line, as observed in the training data. The candidates are created through the subroutine generate_candidates(). The initial candidates correspond to all combinations of the current sequence of words and segmentation points. Each candidate thus includes only one segmentation point and the set of all candidates covers the space of potential segmentation points for the current window of words. Feature vectors are then extracted for each candidate and classified according to the previously trained models. In order to reduce the list of

current candidates into a more manageable set, we only retain candidates with a model-predicted probability above a fixed threshold. Empirical determination for the task at hand yielded a fixed value of 0.7 for this threshold.

The sub-function exist_valid_candidates() checks for the existence of any valid candidate in the filtered set. If the test is positive, the best candidate is selected through the sub-routine get_best_candidate(), which returns the candidate with the highest probability according to the model. In case of a tie, the longest candidate is selected. The sequence of words to the right of the last segmentation point is then stored for the next iteration. As this sequence has already been determined to be a an autonomous sequence at this point, it is taken as an indivisible block in the next iteration, i.e. no new segmentation points can be set between the words that compose it.

If exist_valid_candidates() indicates no candidates at all, the last stored sequence is considered. If this sequence consists of more than one word, a segmentation point is inserted by default; if it contains just one word, the case is taken to be identical to processing the first word of the text: new sequences and candidates are generated from the sequences that include this word and the next ones.

5 Experiments and Evaluation

Several experiments were carried out using Support Vector Machines (SVM) and Logistic Regression (LR) classifiers using the LibSVM [7] and Scikit-learn [19] toolkits respectively. For the SVM classifier, after testing and comparing different combinations of Kernel functions and methods to perform multi-class classification, the Radial Basis Function (RBF) kernel with nu-support vector classification (nu-SVC) algorithm was selected, as this setup gave the best results. The LR classifier was trained through Stochastic Gradient Descent (SGD). Fig. 1 presents results in terms of F-1 measure for each of the test files for both classifiers.

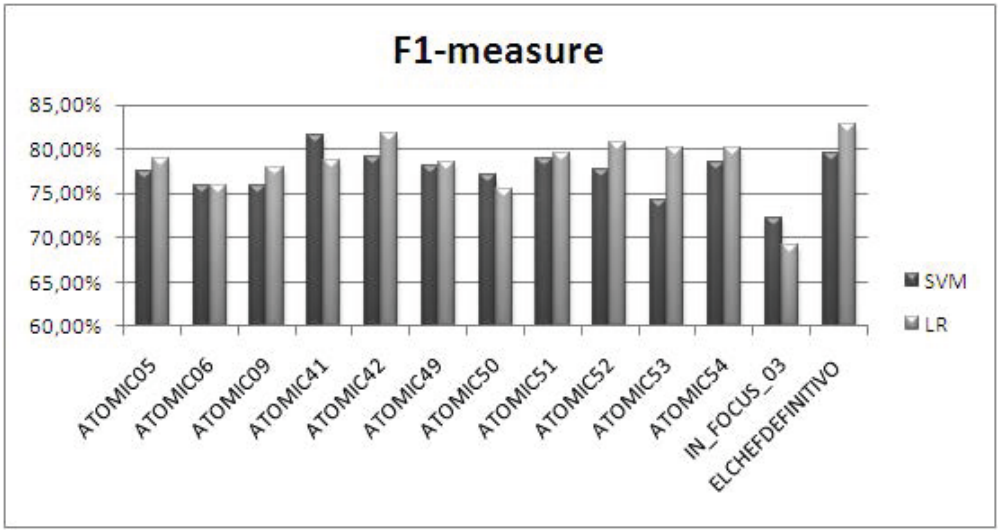


Fig. 1. Segmentation accuracy using SVM and LR classifiers

The results demonstrated similar performance for the two classifiers, with an average score of 74.71% and 76.12% for the SVM and LR classifiers, respectively. In terms of precision and recall, the SVM classifier obtained scores of 82% and 69%. In contrast, the LR classifier achieved a precision of 85% and a recall of 69%. Interestingly, both classifiers reached identical recall, showing that on average almost seven out of ten segmentation points are correctly identified in this approach. Combining this result with precision scores above 80%, the general approach can be seen as promising.

6 Conclusions and Future Work

We presented a novel approach to automatic subtitle segmentation which generates and selects optimal segmentation points according to the predictions made by machine-learned classifiers. This method provides a customized solution to company-specific segmentation guidelines and rules, as the models are strictly induced from existing segmented corpora and generate similar segmentation on new input. Additionally, the approach fills a void as far as generating quality subtitles is concerned, given that automatic subtitle segmentation, which is a crucial quality feature, has been somewhat neglected within the research community. Finally, the method offers a versatile solution as it permits the addition of new features to further tune and improve classification models and subsequent segmentation accuracy.

The preliminary results we have presented are quite satisfactory, with an average recall of nearly 70% and precision above 80% on the test set of a professionally-created corpus of TV cartoon programs in Basque.

In future work, we will pursue experiments on additional corpora, to further evaluate the approach with different domains. More languages will also be tested, as different linguistic characteristics can have an impact on segmentation results, notably in terms of language-dependent infelicitous line endings. We will also explore the impact of including additional features to train classifiers, and evaluate the performance of different feature sets. For instance, incorporating perplexity scores from additional language models trained on surface forms and morphemes might prove beneficiary, as the models would thus include a measure of superficial linguistic knowledge which can be assumed to further improve the proper segmentation of subtitles.

References

1. AENOR: Spanish Technical Standards. Standard UNE 153010:2003: Subtitled Through Teletext, `http://www.aenor.es`
2. Álvarez, A., del Pozo, A., Arruti, A.: APyCA: Towards the Automatic Subtitling of Television Content in Spanish. In: Proceedings of IMCSIT, pp. 567–574. IEEE, Wisla (2010)
3. Álvarez, A., Ruiz, P., Arzelus, H.: Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS (LNAI), vol. 8655, pp. 473–480. Springer, Heidelberg (2014)

4. Baldridge, J.: The OpenNLP Project (2005), http://opennlp.sourceforge.net/
5. Baldridge, J.: Stanford Parser 1.6 (2007), http://nlp.stanford.edu/software/lex-parser.shtml
6. Bordel, G., Peñagarikano, M., Rodríguez-Fuentes, L.J., Varona, A.: A Simple and Efficient Method to Align Very Long Speech Signals to Acoustically Imperfect Transcriptions. In: Proceedings of INTERSPEECH, Portland (2012)
7. Chang, C.C., Lin, C.J.: Libsvm: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27:1–27:27 (2011)
8. Coltheart, M.: What Would We Read Best? Attention and Performance II: The Psychology of Reading. Lawrence Erlbaum Associates, London (1987)
9. D'Arcais, F., Giovanni, B.: Syntactic Processing during Reading for Comprehension. Attention and Performance II: The Psychology of Reading, pp. 619–633. Lawrence Erlbaum Associates, London (1987)
10. Díaz-Cintas, J., Orero, P., Remael, A.: Media for All: Subtitling for the Deaf, Audio Description, and Sign Language, vol. 30. Rodopi (2007)
11. D'Ydewalle, G., Rensbergen, J.V.: Developmental Studies of Text-Picture Interactions in the Perception of Animated Cartoons with Text. Advances in Psychology, vol. 58, pp. 233–248. Elsevier, Amsterdam (1989)
12. Ezeiza, N., Alegria, I., Arriola, J.M., Urizar, R., Aduriz, I.: Combining Stochastic and Rule-based Methods for Disambiguation in Agglutinative Languages. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 380–384, Montreal (1998)
13. Automatic Captions in YouTube (2009), http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html
14. Heafield, K.: KenLM: Faster and Smaller Language Model Queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187–197, Edinburgh (2011)
15. Karamitroglou, F.: A Proposed Set of Subtitling Standards in Europe. Translation Journal 2(2), 1–15 (1998)
16. Kneser, R., Ney, H.: Improved Backing-off for n-gram Language Modeling. In: Proceedings of ICASSP, pp. 181–184, Detroit (1995)
17. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast News Subtitling System in Portuguese. In: Proceedings of ICASSP, pp. 1561–1564, Las Vegas (2008)
18. Padró, L.: Stanilovsky. E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the 8th Language Resources and Evaluation Conference, Istanbul (2012)
19. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
20. Perego, E.: Subtitles and line-breaks: Towards improved readability. In: Between Text and Image: Updating Research in Screen Translation, vol. 78, pp. 211–223. John Benjamins Publishing (2008)
21. Perego, E., Del Missier, F., Porta, M., Mosconi, M.: The Cognitive Effectiveness of Subtitle Processing. Media Psychology 13(3), 243–272 (2010)
22. Petrov, S., Klein, D.: Improved Inference for Unlexicalized Parsing. In: Proceedings of HLT-NAACL, pp. 404–411, Rochester (2007)
23. Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J., Romero-Fresco, P.: Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination. Perspectives 21(1), 5–21 (2013)

7.11 Rich Transcription and Automatic Subtitling for Basque and Spanish

- **Authors:** Aitor Álvarez, Haritz Arzelus, Santiago Prieto and Arantza del Pozo
- **Booktitle:** Advances in Speech and Language Technologies for Iberian Languages (IberSPEECH)
- **Year:** 2016
- **Publisher:** Springer
- **Status:** Submitted

Rich Transcription and Automatic Subtitling for Basque and Spanish

Aitor Álvarez, Haritz Arzelus, Santiago Prieto, and Arantza del Pozo

Human Speech and Language Technology department,
Vicomtech-IK4, San Sebastián, Spain
{aalvarez,harzelus,sprieto,adelpozo}@vicomtech.org
http://www.vicomtech.org/

Abstract. In this paper, complete rich transcription and automatic subtitling systems for Basque and Spanish are described. They enable the automatic transcription and/or subtitling of bilingual contents, through the integration of a language tracker that discriminates between segments spoken in Basque and Spanish. The technology is accessible through a web platform hosted on the Internet. The paper details the architecture of the systems and focuses on the description and evaluation of each technological component. Performance results are reported for the parliamentary domain.

1 Introduction

The new Digital Era has driven a huge increase of the amount of contents that are created and publicly shared on a daily basis. These contents may include text, images, video and/or audio. The generation of such vast amount of contents has led to the progress of technology for their optimal analysis and for the automatic extraction of semantic information in several domains, such as security, surveillance, information retrieval, the audiovisual sector and forensics, among others.

Concerning audio content analysis, both rich transcription and automatic subtitling systems based on Large Vocabulary Continuous Speech Recognition (LVCSR) technology have been turned into promising solutions for many applications in different fields. On the one hand, rich transcription systems have become widely used for tasks such as spoken document retrieval, spoken term detection, summarization, semantic navigation, speech data mining or annotated automatic transcription. These systems allow the automatic generation of a transcript from an audio clip along with metadata to enrich the word stream with useful information such as punctuation, capitalization, speaker identification, sentence units or proper names. On the other hand, the increasing use of multimedia and the accessibility policies promoting the subtitling of broadcast contents at European and national levels have increased the subtitling demand in recent years and LVCSR technology-aid solutions such as re-speaking or automatic subtitling have arisen as alternatives more productive than manual subtitling [2].

In this paper, our proprietary systems for the rich transcription and automatic subtitling of audiovisual contents in Basque and Spanish are presented. They are the result of the research work carried out within many European and Local R&D projects and have already been transferred to several companies which have integrated them in their internal workflows. In addition, a web platform that enables potential interested users to directly obtain automatic transcriptions and/or subtitles of their own contents during a trial period is also presented. The reported results correspond to the parliamentary domain.

In the following Section 2, a description of the state-of-the-art of the common technological components involved is given. Besides, other existing solutions focused on related tasks are enumerated. The corpora used for training and testing purposes is described in Section 3. Section 4 presents the developed systems, toghether with the performance results of each component. Finally, the web platform aiming to make the technology accessible through the Internet for trial purposes is described in Section 5 and conclusions and future lines are given in Section 6.

2 Background

Rich transcription and automatic subtitling systems share several components which work like a pipeline of technological modules, including a speech-non speech front-end, a LVCSR engine, capitalization and punctuation modules, speaker diarization and/or identification, text normalization and, in the case of subtitling, a module for proper segmentation and presentation of subtitles.

The core technology corresponds to the LVCSR engine, being the field in which more research work has been carried out during the last decades. Such research has triggered significant improvements in LVCSR technology over the last few years, mainly through the use of Deep Learning. Thereby, different research groups have shown that DNNs (Deep Neural Networks) can outperform traditional GMMs (Gaussian Mixture Models) at acoustic modeling for speech recognition on a variety of data sets [12]. Moreover, RNN (Recurrent Neural Networks) based language models have proven to outperform traditional N-grams in several challenges [16]. All these technological advances and the availability of toolkits such as Kaldi [21], which includes recipes for acoustic modeling following the latest paradigms, plus other tools such as RNNLM [17] that allow the estimation of RNN language models, have fostered the development of sophisticated LVCSR engines, enabling their integration into rich transcription and automatic subtitling systems with low error rates in bounded domains.

The raw text output of the LVCSR engines is further enriched with capitalization and punctuation marks. Automatic capitalization is commonly context-dependent and has been studied in many works through several approaches based on language models [10], rule-based taggers [5], maximum entropy Markov models [6] and Condition Random Fields (CRF) [26]. However, the ambiguity of the context and unseen words during training usually produce more errors than desired, especially in open domains. On the other hand, autopunctuation is even

more domain-dependent than capitalization, especially with different types of speech (expressive, planned, dictation, etc.) and if acoustics and prosody are employed as features to train models. Despite different punctuation marks can be used, most studies have focused on recovering the most frequent full stop and comma, although comma is quite problematic due to its multi-functionality [4].

Concerning speaker segmentation and identification, recent work in the field is mainly focused on two main open challenges: (1) how to speed up the diarization process and (2) the way to perform cross-show speaker diarization [8] correctly. Instead of the traditional GMMs, factor-analysis based techniques, such as i-vectors, which are popular in the speaker verification domain, have been adapted recently to the speaker diarization task with the aim of discriminating the variability posed by channel characteristics, ambient noises and spoken phonemes [27].

The Text Normalization module aims at converting numbers, numerals, dates and amounts (e.g. money and percentage) to their digit representation, and it is commonly performed using rule-based functions.

Regarding the automatic segmentation of subtitles, it can be considered a novel research field which aims at providing syntactically coherent breaks so that viewers can read subtitles as quick as possible. In this sense, the works presented in [1] and [2] present automatic alternatives to the Counting Character technique, which is the main technique employed in most automatic subtitling systems currently.

Finally, advances related to LVCSR technology have driven the development of commercial solutions for rich transcription and automatic subtitling. Recently, Google has started supporting the automatic generation of time-aligned draft transcriptions and subtitles of the videos uploaded to Youtube [9, 14]. Nevertheless, for the moment Youtube's automatic transcriptions do not include punctuation and capitalization marks nor follow standard professional subtitling practices [3]. Other companies such as Koemei[1], SailLabs[2], Vecsys[3] and Verbio[4] commercialize automated transcription solutions for varying pools of languages and application scenarios, but do not produce subtitles. In the subtitling field, Audimus [18] can be considered a pioneer and reference system, as it provides a complete framework for automatic subtitling in both batch and live modes for several languages [3].

3 Data resources

Two corpora have been used to train and evaluate the technological components of the systems in the parliament domain: the SAVAS corpus and a Basque Parliament corpus.

[1] https://koemei.com/
[2] https://www.sail-labs.com/
[3] https://www.bertin-it.com/vecsys/
[4] http://www.verbio.com

The SAVAS corpus [22] was compiled during the European SME-DCL SAVAS[5] project, whose aim was to collect and annotate a huge amount of audio and text corpora in the news domain for several European languages to develop LVCSR engines for automatic transcription and subtitling. For Basque and Spanish, 200 hours of broadcast news audios were collected and annotated per language. Regarding texts, 329 million words and more than one billion words of text were gathered from digital newspapers in Basque and Spanish respectively.

The second corpus was composed by audios and texts from the Basque Parliament. In terms of acoustic data, a total amount of 10 hours and 19 minutes for Basque, and 13 hours and 44 minutes for Spanish were collected and manually annotated. Annotations were done at transcription, name entity, background and speaker levels. With regard to the text corpus, texts containing 7.2 and 12.3 million words were gathered from sessions transcribed manually over the last 6 years.

The following Table 1 summarizes the audio and text data of the SAVAS and Basque Parliament corpora for each language.

Table 1. The SAVAS and Basque Parliament corpora for Basque and Spanish

Language	***Variant***	***Corpus***	***Domain***	***Audio***	***Text***
Basque	Standard Basque	SAVAS	News	200 H	329 M
Spanish	European	SAVAS	News	200 H	1009 M
Basque	Standard Basque	Basque Parliament	Politics	10 H + 19 mins	7.2 M
Spanish	European	Basque Parliament	Politics	13 H + 44 mins	12.3 M

As further detailed in Section 4, the SAVAS corpus was mainly used to construct the acoustic models of the LVCSR engines, while the remaining components, including the language models and lexicons, were built and adapted exploiting the in-domain corpus of the Basque Parliament.

4 Description of the systems

The development of the rich transcription and automatic subtitling systems has been performed employing the methods integrated in Vicomtech-IK4's proprietary Transkit-SDK[6] tool, which includes functions to train, build and evaluate all the technological components involved in these type of systems.

There are several components that both the rich transcription and automatic subtitling systems share. Such components correspond to the Speech/Non-Speech front-end, the LVCSR engine, the Capitalization and Punctuation module, the Speaker Segmentation and Identification module, and the Text Normalization module, which are described in Subsection 4.1. In addition, given the

[5] http://www.fp7-savas.eu/

[6] http://www.vicomtech.org/resources/archivosbd/sdks_documentos/Transkit.pdf

bilingual nature of the Basque Parliament's sessions, where Basque and Spanish languages are mixed interchangeably in the same audio track, a module for Language Tracking was also implemented. The component in charge of generating proper segmentation of subtitles is explained in Subsection 4.2.

4.1 Rich Transcription systems

In Figure 1, the bottom-up pipeline of our rich transcription system is shown. As it can be seen, its architecture has been designed to process bilingual audio tracks. In a first step, input audio is split into homogeneous portions containing speech and non-speech segments. The Language Tracker module is then in charge of segmenting and classifying each speech segments into Basque and Spanish. At this point, each segment is processed by a different language-dependent LVCSR engine to obtain draft transcriptions. These raw outputs are then enriched with capitalization and punctuation marks through separate technology trained for each language. Afterwards, the different speakers of the Basque Parliament are identified applying language-independent Speaker Identification technology and finally, output text is normalized converting numerals, dates and amounts into their digit representation.

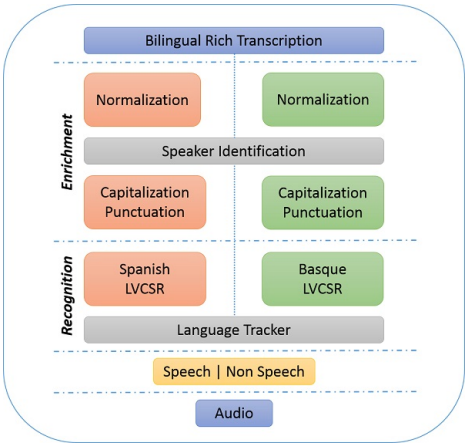


Fig. 1. Bottom-up pipeline of the bilingual Rich Transcription system

– Regarding *Speech-Non Speech* discrimination, our system has been built using the UBM/GMM (Universal Background Model/Gaussian Mixture Model) approximation described in [24]. This module is capable of discriminating between speech, silence and noise and it was trained using 2 hours per class taken from the Basque Parliament corpus. An accuracy of 94.1% at segment level in an in-domain test set with a total duration of 105 minutes has been reached.

- *Language Tracker.* Our phonotactic-based language tracker aims to identify phoneme boundaries which could be candidates of language turns, following the work presented in [15]. The followed technical approach consists in constructing an unique phone decoder combining the languages involved. To this end, an hybrid DNN-HMM acoustic model has been trained using 86 hours of audios (39 hours of Basque and 47 hours of Spanish) from the SAVAS corpus, while the LM is a trigram model at phone level estimated using bilingual texts composed of 4.3 million phones (1.9 M for Basque and 2.4 M for Spanish). The phone set is composed of all the phones in both languages, and a distinctive language tag is added to each phone to avoid mixing those that are shared in both languages.
  We have employed Language Error Rate (LER) as evaluation metric. This measure is computed in the same way as the well-know Diarization Error Rate (DER) commonly used in speaker diarization systems but using languages instead of speakers. An average LER of 10.29% has been obtained on a test set of 255 minutes, of which 92 minutes were spoken in Basque and 163 minutes in Spanish.
- *Automatic Speech Recognition.* LVCSR engines for Basque and Spanish have been built using the open-source Kaldi toolkit[21]. Acoustic models were trained using the SAVAS corpus, and following the implementation given in [25], which corresponds to a hybrid Deep Neural Network (DNN)-Hidden Markov Models (HMM) implementation where DNNs are trained to provide posterior probability estimates for the HMM states. Two types of language models (LM) were integrated per language: trigram Arpa-format LM for decoding and 9-gram constant Arpa-format LM for rescoring of the final lattices, both trained using the in-domain text gathered from the Basque Parliament. The decoding LM were estimated with Kneser-Ney modified smoothing using the KenLM [11] toolkit. Word Error Rate (WER) values of 16.73% and 9.47% were achieved for Basque and Spanish over an in-domain test set.
- *Capitalization and Punctuation.* The Capitalization module aims to re-case the lower-cased text output of the LVCSR engine. Our automatic capitalization models have been estimated using the *recasing* tool provided by the Moses open-source toolkit [13]. The Basque and Spanish models were trained using the texts collected from the SAVAS and Basque Parliament corpora. F1-score values of 82.80% and 89.94% were achieved on a test set of 10K and 28K words for Basque and Spanish, respectively.
  We have turned the Punctuation problem into a text sequence labelling task. To this end, our automatic punctuation module is constructed on top of a CRF model, in which each token is tagged with one category (NP: No punctuation; CO: Comma; FS: Full Stop) depending on whether the next token corresponds to a punctuation mark or not. The following acoustic and linguistic features are exploited: (1) the current and the surrounding 2 words on the left and right (5 words), (2) the current and the surrounding 2 words' POS information on the left and right (5 categories), (3) time between the current and the next word (1 feature), (4) Speaker Change (1 Boolean), and

(5) Language Change (1 feature). The CRF models were estimated using the corpora of the Basque Parliament, obtaining F1-score values of 65.89% and 64.34% for Basque and Spanish respectively over an in-domain test set of 2 hours per language. CRF models were constructed using the CRFSuite tool [19].

- *Speaker Identification.* This modules takes the speech segments resulting from the Speech-Non Speech module as input and applies the generalized likelihood ratio (GLR) distance measure to detect close speaker changes within each segment. Each individual portion is then modeled using an i-vector representation, for which an UBM and a TV (Total Variability) matrix, compensated with the Linear Discriminant Analysis (LDA) method, have been previously estimated using the corpus composed of 105 speakers from the Basque Parliament. A probabilistic linear discriminant analysis (PLDA) back-end computed the i-vector similarity. With this approach, an accuracy of 85.34% was achieved on an in-domain test set of 160 minutes.
- *Normalization.* Comprises all the tasks related to converting numbers into their digit representation, removing filled pauses and generating abbreviations. It is implemented using rule-based functions defined for each language.

4.2 Automatic Subtitling systems

The automatic subtitling systems developed for Basque and Spanish include all the technological components described above plus an additional module responsible of generating well-segmented subtitles. The importance of quality segmentation is supported by several works [23, 2] and by psycholinguistic studies on the readability and cognitive effort associated to a poor segmentation [20].

The *Subtitle Segmentation* component has been developed using CRF models, for which several categories have been defined to describe the function of each word within each subtitle and connected through a graphical dependence model. The categories defined represent the function of the first word of a subtitle (B-SU), the end word of a line (E-LI), the first word of the second line (B-LI), the final word of a subtitle (E-SU), inline words (I-LI), a single word in a subtitle (BE-SU), single word in the first line (BS-EL), and single word in the second line (BL-ES). The feature vectors used to describe the information extracted from each word are composed of 15 characteristics related to the words, Part-of-Speech information, Speaker Change information, time differences between surrounding words and two parameters to control the amount of characters per line and subtitle. The CRF models were trained over a corpus containing 109,006 subtitles for Basque and 81,802 for Spanish. In both corpora, subtitles were manually created by professionals following specific segmentation rules to keep linguistic and syntactic coherence. The corpora were split into train and test sets (80% and 20%), and F1-scores of 83% and 80.7% were obtained for Basque and Spanish respectively.

5 Web platform for Rich Transcription and Automatic Subtitling

The goal of this web platform is to provide interested users an online service to test by themselves if the rich transcription and automatic subtitling technology may help improve their internal services. As it can be seen in Figure 2, the technology is hosted in an server within Vicomtech-IK4's internal network, while the web platform is installed in a server allocated on the Internet, so that it can be accessed by external users. A monitoring daemon continuously checks a shared folder where the uploaded contents are saved. When a new content is detected, it is copied to the internal server to be processed automatically. The result is then sent back to the user via email.

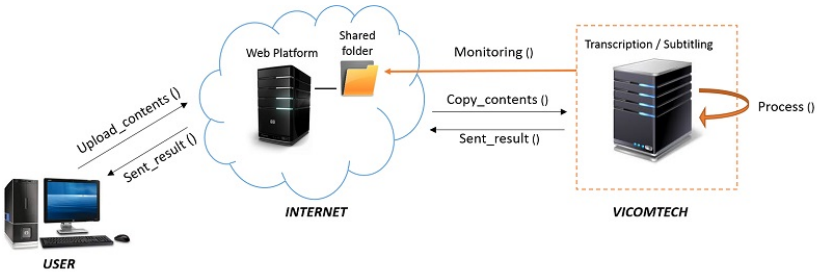


Fig. 2. Architecture of the Web Platform for Rich Transcription and Automatic Subtitling

Users can accesses the platform through a web application[7] that stays operational for them to directly test the performance of the Rich Transcription and Automatic Subtitling systems on their contents.

6 Conclusions and future work

Complete systems for rich transcription and automatic subtitling in Basque and Spanish have been presented and their performance reported for the parliamentary domain. Several local companies are already using the systems described to speed up the manual transcription of the parliament sessions and automatically subtitle political videos that are published on the Internet.

Future work will involve improving the components with lowest performance. New methods to improve automatic punctuation will be investigated following recent advances in Natural Language Processing with Neural Networks for sentence boundary detection [7]. Besides, the enhancement of subtitle segmentation will be explored using more parameters such as stop words, syntactic functions

[7] http://212.81.220.68:8086/SDK_web/

or grammatical relations and experimenting with RNNs for this task. Finally, ongoing work will continue focusing on optimizing the technology for real-time scenarios.

7 Acknowledgements

The authors would like to thank Serikat Consultoría Informática and Irekia for their provision of corpora and useful feedback.

References

1. Álvarez, A., Arzelus, H., Etchegoyhen, T.: Towards customized automatic segmentation of subtitles. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 229–238. Lecture Notes in Computer Science, Springer International Publishing (2014)
2. Álvarez, A., Matamala, A., Pozo, A.d., Balenciaga, M., Martínez-Hinarejos, C.D., Arzelus Irazusta, H.: Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 3049–3053 (2016)
3. Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., del Pozo, A.: Automating live and batch subtitling of multimedia contents for several european languages. Multimedia Tools and Applications pp. 1–31 (2015)
4. Batista, F., Moniz, H., Trancoso, I., Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 474–485 (2012)
5. Brill, E.: Some advances in transformation-based part of speech tagging. arXiv preprint cmp-lg/9406010 (1994)
6. Chelba, C., Acero, A.: Adaptation of maximum entropy capitalizer: Little data can help a lot. Computer Speech & Language 20(4), 382–399 (2006)
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12, 2493–2537 (2011)
8. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEEACM Transactions on Audio, Speech and Language Processing (TASLP) 23(12), 2286–2297 (2015)
9. Google: Automatic captions in youtube. `https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html` (2009)
10. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. pp. 4741–4744. IEEE (2009)
11. Heafield, K.: Kenlm: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197. Association for Computational Linguistics (2011)

12. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE (2012)
13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
14. Liao, H., McDermott, E., Senior, A.: Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. pp. 368–373. IEEE (2013)
15. Lyu, D.C., Chng, E.S., Li, H.: Language diarization for code-switch conversational speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 7314–7318. IEEE (2013)
16. Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH. vol. 2, p. 3 (2010)
17. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Cernocky, J.: Rnnlm-recurrent neural network language modeling toolkit. In: Proc. of the 2011 ASRU Workshop. pp. 196–201 (2011)
18. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast news subtitling system in portuguese. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. pp. 1561–1564. IEEE (2008)
19. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs). URL http://www.chokkan.orgsoftwarecrfsuite (2007)
20. Perego, E., Del Missier, F., Porta, M., Mosconi, M.: The Cognitive Effectiveness of Subtitle Processing. Media Psychology 13(3), 243–272 (2010)
21. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society (2011)
22. del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J.P., Paulo, S., Piccinini, N., Raffaelli, M.: Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling. In: LREC. pp. 432–436 (2014)
23. Rajendran, D.J., Duchowski, A.T., Orero, P., Martínez, J., Romero-Fresco, P.: Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination. Perspectives 21(1), 5–21 (2013)
24. Snyder, D., Chen, G., Povey, D.: Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484 (2015)
25. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: INTERSPEECH. pp. 2345–2349 (2013)
26. Wang, W., Knight, K., Marcu, D.: Capitalizing machine translation. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 1–8. Association for Computational Linguistics (2006)
27. Yella, S.H., Stolcke, A., Slaney, M.: Artificial neural network features for speaker diarization. In: Spoken Language Technology Workshop (SLT), 2014 IEEE. pp. 402–406. IEEE (2014)

7.12 Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction

- **Authors:** Andoni Arruti, Idoia Cearreta, Aitor Álvarez, Elena Lazkano, and Basilio Sierra
- **Journal:** PlosONE
- **Year:** 2014
- **Publisher:** Public Library of Science
- **DOI:** http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0108975

OPEN ACCESS Freely available online

PLOS ONE

Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction

Andoni Arruti[1]*, Idoia Cearreta[1], Aitor Álvarez[2], Elena Lazkano[1], Basilio Sierra[1]

1 Computer Science Faculty (University of the Basque Country), San Sebastián, Spain, **2** Vicomtech-IK4 Research Alliance, San Sebastián, Spain

Abstract

Study of emotions in human–computer interaction is a growing research area. This paper shows an attempt to select the most significant features for emotion recognition in spoken Basque and Spanish Languages using different methods for feature selection. RekEmozio database was used as the experimental data set. Several Machine Learning paradigms were used for the emotion classification task. Experiments were executed in three phases, using different sets of features as classification variables in each phase. Moreover, feature subset selection was applied at each phase in order to seek for the most relevant feature subset. The three phases approach was selected to check the validity of the proposed approach. Achieved results show that an instance-based learning algorithm using feature subset selection techniques based on evolutionary algorithms is the best Machine Learning paradigm in automatic emotion recognition, with all different feature sets, obtaining a mean of 80,05% emotion recognition rate in Basque and a 74,82% in Spanish. In order to check the goodness of the proposed process, a greedy searching approach (FSS-Forward) has been applied and a comparison between them is provided. Based on achieved results, a set of most relevant non-speaker dependent features is proposed for both languages and new perspectives are suggested.

Citation: Arruti A, Cearreta I, Álvarez A, Lazkano E, Sierra B (2014) Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction. PLoS ONE 9(10): e108975. doi:10.1371/journal.pone.0108975

Editor: Oriol Pujol, University of Barcelona, Spain

Received March 25, 2014; **Accepted** September 5, 2014; **Published** October 3, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All datasets are in supporting information files.

Funding: This work has been done within the Basque Government Research Team grant under project TIN2010-15549 of the Spanish Ministry and the University of the Basque Country UPV/EHU, under grant UFI11/45 (BAILab). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: andoni.arruti@ehu.es

Introduction

Affective computing, a discipline that develops devices for detecting and responding to user's emotions [1], is a growing research area [2] in Human Computer Interaction (HCI). The main objective of affective computing is to capture and process affective information with the aim of enhancing and naturalizing the communication between the human and the computer. Within affective computing, affective mediation uses a computer-based system as intermediary in the communication of people, reflecting the emotion the interlocutors may have [1]. Affective mediation tries to minimize the filtering of affective information carried out by communication devices, because they are usually devoted to the transmission of verbal information and therefore, miss nonverbal information [3]. There are other applications in this type of mediated communication, for example, textual telecommunication (affective electronic mail, affective chats, etc.). Speech Emotion Recognition (SER) is also a very active research field in HCI [4]. Concerning to this topic, Ramakrishnan and El Emary [5] propose several types of applications to show the importance of techniques used in SER.

Affective databases are a good chance for developing affective applications, either for affective recognizers or either for affective synthesis. This paper presents a study aimed for giving a new step towards searching relevant speech features in automatic SER area for Spanish and Basque languages, using an affective database. This study is based on two previous works and its main objective is to analyse the results using the whole set of features which come from both of them. Moreover, it tries to extract the most relevant features related with the emotions in speech. Although all studies have started being speaker-dependent, in the extraction of relevant features the aim is to achieve a speaker-independent recognizer.

The three phases are the following: (a) using a group of 32 speech features [6]; (b) using a different group containing a total of 91 features [7]; and (c) finally, merging both groups, adding up a total of 123 different features.

Several Machine Learning (ML) techniques have been applied to evaluate their usefulness for SER. In this particular case, techniques based on evolutionary algorithms (EDA) have been used in all phases to select feature subsets that noticeably optimize the automatic emotion recognition success rate.

Related work

Theories of emotions proposed by cognitive psychologists are a useful starting point for modelling human emotions. Although several theoretical emotional models exist, the most commonly used models of emotions are dimensional [8] and categorical [9,10] ones. For practical reasons, categorical models of emotions have been more frequently used in affective computing. For

example, in [11] several algorithms that recognize eight categories of emotions based on facial expressions are implemented. Oudeyer [12] has developed such algorithms for production and recognition of five emotions based on speech features. Authors such as Ekman and Friesen [13] suggest the *universality* of six basic categorical emotions and think that facial expressions for these six emotions are expressed and recognized in all cultures.

In [14], a study about the words that Basque-speaking people understand as emotions-related ones is presented and the hierarchical and family resemblance structure of the most prototypical 124 concepts that are represented as emotions are mapped. The hierarchical cluster analysis of collected data reveals two large superordinate categories (positive and negative) and five large basic level categories (love, happiness, anger, fear and sadness), which contain several subordinate level categories. They notice that those basic categories can also be found in similar studies made in Indonesia and United States of America.

Apart from models, there are also some studies related to expression and detection of emotions. In this way, Lang [8] proposed that three different systems would be implied in the expression of the emotions and that could serve like indicators to detect the emotion of the user:

- Verbal information: reports about perceived emotions described by users.
- Behavioural information: facial and postural expressions and speech paralinguistic features.
- Psychophysiological answers: such as heart rate, galvanic skin response -GSR-, and electroencephalographic response.

Verbal, behavioural and psychophysiological correlates of emotions should be taken into account when possible. Correlations among these three systems can help computers interpreting ambiguous emotions. For instance, a person with apraxia could have problems in the articulation of facial gestures, but subjective information written down with assistive technology can be used by a computer to interpret her/his emotional state. In that sense, more specific models or theories which describe the components of each system of expression can be found in the literature and selected according to the particular case, such as a dictionary of emotional speech [15], acoustic correlates of speech [10], subsyllabic and pitch spectral features [16] or facial expressions [9].

On the other hand, affective resources, such as affective stimuli databases, provide a good opportunity for training affective applications, either for affective synthesis or for affective recognizers based on classification via Artificial Neural Networks, Hidden Markov Models, Genetic Algorithms (GAs), or similar techniques (see for example, [17] and [18]). These type of databases usually record information such as images, sounds, psychophysiological values, etc. There are some references in the literature that present affective databases and their characteristics. Cowie et al. [19] listed the major contemporary databases, emphasising those which are naturalistic or induced, multimodal, and influential. Other interesting reviews are the ones provided in [20] and [21].

Most of these references of affective databases are related to English, while other languages have less resources developed, especially the ones with relatively low number of speakers; this is the case of Basque Language. To our knowledge, the first affective database in Basque is the one presented by Navas et al. [22]. Concerning to Spanish, the work of Iriondo et al. [23] stands out; and relating to Mexican Spanish, the work of Caballero-Morales [24] can be highlighted.

RekEmozio database is a multimodal bilingual database for Spanish and Basque [25], which also stores information that came from processes of some global speech features extraction for each audio recording. Some of these features are prosodic features while others are quality features.

As in the case of affective databases, most emotional speech recognition systems are related to English. For languages such as Basque and Spanish much less emotional speech recognition systems have been developed. For Basque, the work of Luengo et al. [26] is noticeable. For Spanish, works such as [27] can be found in the literature. Another example is the work of Hozjan and Kačič [28], which studies multilingual emotion recognition and includes Spanish language. In this work, 26 high-level (AHL) features and 14 database-specific emotional (DSE) features were used. AHL are statistical presentations of low-level features (low-level features are composed from pitch, derivative of pitch, energy, derivative of energy, and duration of speech segments). DSE features are a set of speaker specific emotional features. Emotion recognition was performed using artificial neural networks and results were obtained using the *max-correct* evaluation method. Taking speaker-dependent emotion recognition into account, the average of *max-correct* with AHL features was 55.21% and for recognition with DSE features 45.76%. An aspect to consider is whether cultural and linguistic variations can modify emotional speech features. This aspect has been analysed in studies such as [29], [12] and [30]. In [29], an experimental study is performed comparing Spanish and Swedish cultures. However, it must be highlighted that no reference has been found in literature about Basque language being analysed in the context of cross-cultural studies related to speech. It must also be stated that few common speech features are provided in studies where Spanish language is present and that most cross-cultural studies found in literature are based on facial expression analysis.

ML paradigms take a principal role in some works related to SER found in the literature [31]. Some papers describe works performed using several classification methods. Support Vector Machines (SVM) and Decision Trees (DT) are compared to identify relevant emotional states from prosodic, disfluency and lexical cues extracted from the real-life spoken human-human interactions in [32]. Authors such as Pan et al. [33] also apply the SVM method to classify emotions in speech, using two emotional speech databases: Berlin German and Chinese. In [34], authors developed a hybrid system capable of using information from faces and voices to recognize people's emotions. Three ML approaches are considered by Shami and Verhelst [35], K-nearest neighbours (KNN), SVM and Ada-boosted decision trees, applied to four emotional speech databases: Kismet, BabyEars, Danish, and Berlin. Rani et al. [36] presents a comparative study of four ML methods (KNN algorithm, Regression Trees (RT), Bayesian Networks and SVM) applied to the affect recognition domain using physiological signals. In [37] a system that recognizes human speech emotional states using a neural network classifier is proposed.

Different types of features (spectral, prosodic) for laughter detection were investigated by Truong and van Leeuwen [38] using different classification techniques (Gaussian Mixture Models, SVM, Multi Layer Perceptron). In [12] a large-scale data mining experiment about the automatic recognition of basic emotions in informal everyday short utterances is presented. A large set of ML algorithms is compared, ranging from Neural Networks, SVM or DT, together with 200 features, using a large database of several thousand examples, showing that the difference of performance among learning schemes can be substantial, and that some features which were previously unexplored are of crucial importance;

several schemes are emerging as candidates for describing pervasive emotion.

It has to be pointed out the work by Schröder [39], which provides a wide list of references concerning emotional speech features. Most of these references are related to English and the features used by referenced authors are the most commonly found in the literature. In terms of emotional speech features for Basque, to authors' knowledge, the work of Navas et al. [40] is the unique work and it also uses some of the most common features found. This situation is similar for Spanish, there are few references and some of most common features tend to be used [23,41,42]. On the other hand, in [43] and [44], a different approach of how to treat the signal that adds new and interesting features for the study of the emotions in the voice is presented.

Some works about feature selection for emotion recognition have been found in literature: in [45] Fast Correlation Based Filter is applied to select the attributes that take part in a Neural Network classifier; in [46], selection is performed by an expert; in [47] a non-linear dimensionality reduction is used to carry out the recognition process; Picard et al. [48] present and compare multiple algorithms for feature-based recognition of emotional state from this data; the work by Cowie et al. [19] is related with this paper in the sense that a Feature Selection method is used in order to apply a Neural Network to emotion recognition in spoken English, although both, the method chosen to perform the Feature Subset Selection (FSS) and the learning paradigms are different.

Materials and Methods

As it is mentioned before, several ML techniques have been applied to evaluate their usefulness for SER and to obtain relevant emotional speech features. To fulfil this objective, a corpus has been used to extract several features. Next subsections describe this corpus and the ML paradigms used for classification purposes during the experimental phase.

Corpus

There are few affective corpuses developed for Spanish language, and even less for Basque. The database used in this work has been Rekemozio, that contains instances of both languages and is the only alternative found for Basque. The creation and validation of this multimedia database, that includes video and audio recordings, is described in [21]. In our work, we only use the spoken material. Rekemozio uses a categorical model based on Ekman's six basic emotions [13] (Sadness, Fear, Joy, Anger, Surprise and Disgust), and also considers a Neutral emotion category. In their work Ekman and Friesen suggested that they are universal for all cultures. Table 1 summarizes the scope of RekEmozio database, presenting its relevant features.

RekEmozio database recordings were carried out by skilled actors and actresses, and contextualized by means of audiovisual stimuli (154 audio stimuli and 6 video stimuli per actor). They were asked to read a set of words and sentences (both semantically and non-semantically relevant) trying to express emotional categories by means of voice intonation and facial expression. Regarding to spoken material, in Table 2, the amount of text used is pointed out, while Table 3 shows the length of the recordings (see [25] for more details).

It should be noted that the database is validated [21]. It is considered that training affective recognizers with subject validated databases will enhance the effectiveness of recognition applications. Fifty-seven volunteers participated in the validation, and results of the categorical test allowed to conclude that the 78% of audio stimuli were valid to express the intended emotion as the recognition accuracy percentage was over 50%.

Emotional feature extraction

One of the most important questions for automatic SER is which features should be extracted from the voice signal. Previous studies show that it is difficult to find specific voice features valid as reliable indicators of the emotion present in the speech [49].

Therefore, as a first step, an in-depth literature review of emotional speech features was carried out. After reviewing the state-of-the-art, in the first phase, a number of features which had been frequently used in other similar studies [40,23,41], where selected and checked. Using a 20 ms frame-based analysis, with an overlapping of 10 ms, information related to prosody, such as the fundamental frequency, energy, intensity and speaking rate, was extracted obtaining a total of 32 features. In this phase, encouraging results were obtained applying ML classification techniques.

In a second phase it was decided to study additional features that could provide information about the emotion expressed in the speech. Tato et al. [43] proposed new interesting formulas to extract information regarding emotions from speech, and also defined a novel technique for signal treatment, not only extracting information by frames, but by regions consisting of more than three consecutive frames, either for the analysis of voice and unvoiced parts. Before adding this information consisting of 91 new features to those used in the first phase, the effectiveness of these new features was tested using the same ML paradigms, to compare the results obtained in both phases.

After verifying the effectiveness of the classification procedures and the features selected in the first two phases, it was decided to compile all the features concerning emotional information in a third and final phase, obtaining a final set of 123 speech features as input for the previous ML paradigms.

All these features are divided as follows:

- Prosodic Features: model the F0, energy, voiced and unvoiced regions, pitch derivative curve and the relations between the features as is proposed in [50] and [44] (see Table 4).
- Spectral Features: formants and energy band distribution (see Table 5).

Table 1. Summary of RekEmozio database scope for recordings.

Language	#Actors	#male/#female	Mean Age (std dev)
Basque	7	4/3	31.3 (5.2)
Spanish	10	5/5	30.7 (4.1)
Overall	17	9/8	30.9 (4.4)

doi:10.1371/journal.pone.0108975.t001

Table 2. Amount of text used for both Spanish and Basque languages.

Text unit	Specific for each emotion	Used in all emotions	Total	Per actor
Words	35	5	40	70
Sentences	21	3	24	42
Paragraphs	21	3	24	42
Total	77	11	88	154

doi:10.1371/journal.pone.0108975.t002

- Quality Features: related with the voice quality, such as harmonicity to noise ratio and active level in speech (see Table 6).

Machine Learning standard paradigms used

In the supervised learning task, the main goal is to construct a model or a classifier able to manage a classification task with an acceptable accuracy. With this aim, some variables are to be used in order to identify different elements, the so called predictor variables. In the present problem, each sample is composed by a set of speech related values, while the label value is one of the seven emotions identified.

We brieflyintroduce the single paradigms used in our experiments. These paradigms come from the ML family and are 4 well-known supervised classification algorithms. As seen before, the number of choices when selecting a classifier is very large, and in this work, being the main goal the feature selection for Speech Emotion Recognition, we have chosen to use simple paradigms, with long tradition in different classification tasks and with different approaches to learning.

Decision Trees. A *Decision Tree* consists of nodes and branches to partition a set of samples into a set of covering decision rules. In each node, a single test or decision is made to obtain a partition. The starting node is usually referred as the root node. In each node, the goal is selecting an attribute that makes the best partition between the classes of the samples in the training set [51] and [52]. In our experiments, two well-known decision tree induction algorithms are used, ID3 [53] and C4.5 [54].

Instance-Based Learning. Instance-Based Learning (IBL) has its root in the study of Nearest Neighbour algorithm [31] in the field of ML. The simplest form of Nearest Neighbour (NN) or KNN algorithms simply store the training instances and classify a new instance by predicting the same class its nearest stored instance has or the majority class of its k nearest stored instances have, respectively, according to some distance measure as described in [55]. The core of this non-parametric paradigm is the form of the similarity function that computes the distances from the new instance to the training instances, to find the nearest or k-nearest training instances to the new case. In our experiments the IB paradigm is used, an inducer developed in the MLC++ project [56] and based on the works of Aha et al. [57] and Wettschereck [58].

Naive Bayes classifiers. The Naive-Bayes (NB) rule [59] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $\mathbf{X} = (X1,X2,\ldots,Xd)$, the NB classifier applies the following rule:

$$c_{NB} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{d} p(x_i|c_j) \quad (1)$$

where c_{NB} denotes the class label predicted by the NB classifier and the possible classes of the problem are grouped in $C = \{c_1,\ldots,c_n\}$. A normal distribution is assumed to estimate the class conditional densities for predictive genes. Despite its simplicity, the NB rule has obtained better results than more complex algorithms in many domains.

Increasing the Accuracy by Feature Subset Selection

The goal of a supervised learning algorithm is to induce a classifier that allows us to classify new examples $E^* = e_{n+1},\ldots,e_{n+m}$ that are only characterized by their d descriptive features. To generate this classifier we have a set of n samples $E = e_1,\ldots,e_n$, characterized by d descriptive features $X = X_1,\ldots,X_d$ and the class label $C = w_1,\ldots,w_n$ to which they belong. ML can be seen as a data-driven process where, putting little emphasis on prior hypotheses a general rule is induced for classifying new examples using a learning algorithm. Many representations with different biases have been used to develop this classification rule. Here, the ML community has formulated the following question: "*Are all of these d descriptive features useful for learning the classification rule?*" Trying to respond to this question the FSS approach appears, which can be reformulated as follows: *given a set of candidate features, select the best subset under some learning algorithm*.

This dimensionality reduction made by a FSS process can carry out several advantages for a classification system in a specific task:

Table 3. Lengths of RekEmozio database's audio recordings.

Language	Recording's lengths
Basque	130'41''
Spanish	166'17''
Total	296'58''

doi:10.1371/journal.pone.0108975.t003

Table 4. Prosodic Features extracted for each validated recording.

Feature class	Description	Computed values
Fundamental Frequency	F0 curve in the voiced parts. Estimation based on Sun algorithm.	Maximum and its position, minimum and its position, mean, variance, standard deviation, maximum positive slope in contour, regression coefficient and its mean square error.
		Pitch derivative based features: maximum, minimum, mean, variance, regression coefficient and its mean square error.
Energy	Energy, RMS energy and Loudness.	Maximum and its position, minimum and its position, mean, variance, regression coefficient and its mean square error.
		RMS: maximum, minimum, mean, range, variance and standard deviation.
		Loudness: absolute loudness based on Zwicker's model.
Voiced/Unvoiced	Features based on Voiced and Unvoiced frames and regions.	F0 value of the first and last voiced frames, number of voiced and unvoiced frames and regions, length of the longest voiced and unvoiced regions, ratio of number of voiced and unvoiced frames and regions.
Relations	Relations among several features.	Mean, variance, mean of the maximum, variance of the maximum, mean of the pitch ranges and mean of the flatness of the pitch based on every voiced region pitch values.
		Pitch increasing and decreasing in voiced parts as well as the mean of the voiced regions duration.
		Many features related with the energy among the voiced regions, such as global energy mean, vehemence, mean of the flatness and tremor in addition to others.
Rhythm	Alternation between speech and silence.	Duration of voice, silence, maximum voice, minimum voice, maximum silence and minimum silence in the whole utterance are computed.

doi:10.1371/journal.pone.0108975.t004

- Reduction in the cost of data acquisition
- Improvement of the comprehensibility of the final classification model
- Faster induction of the final classification model
- Improvement in classification accuracy

The attainment of higher classification accuracies is the usual objective of ML processes. It has been long proved that the classification accuracy of ML algorithms is not monotonic with respect to the addition of features. Irrelevant or redundant features, depending on the specific characteristics of the learning algorithm, may degrade the predictive accuracy of the classification model. In this work, FSS objective will be the maximization of the performance of the classification algorithm. In addition, with the reduction in the number of features, it is more likely that the final classifier is less complex and more understandable by humans.

Once the objective is fixed, FSS can be viewed as a search problem, with each state in the search space specifying a subset of the possible features of the task. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice because of the large amount of computational effort required. Many search techniques have been proposed to solve FSS problem when there is no knowledge about the nature of the task, carrying out an intelligent search in the space of possible solutions. As randomized, evolutionary and population-based search algorithm, Genetic Algorithms (GAs) have long been used as the search engine in the FSS process. GAs need crossover and mutation operators to make the evolution possible.

Feature Subset Selection. As reported by Aha and Bankert [60], the objective of feature subset selection in ML is to *"reduce the number of features used to characterize a dataset so as to improve a learning algorithm's performance on a given task"*. The objective will be the maximization of the classification accuracy in a specific task for a certain learning algorithm; as a collateral effect the number of features to induce the final classification model will be reduced. The feature selection task can be exposed as a search problem, each state in the search space identifying a subset of

Table 5. Spectral Features extracted for each validated recording.

Feature class	Description	Computed values
Formants	Resonance characteristics of the vocal tract.	Mean of the first, second and third formant frequencies and their bandwidths among all voiced region as well as the mean, maximum and range of the second formant ratio.
Critical Bands	Energy in several frequency bands, using two different spectral distributions.	Energy in three frequency bands: low band (0–1300 Hz), medium band (1300–2600 Hz) and high band (2600–4000 Hz).
		Energy in four frequency bands: (0 - F0 Hz), (0–1000 Hz), (2500–3500 Hz) and (4000–5000 Hz).
		Relative energy in each band for voiced parts of utterance.

doi:10.1371/journal.pone.0108975.t005

Table 6. Quality Features extracted for each validated recording.

Feature class	Description	Computed values
Harmonicity to noise ratio	Ratio of the energy of harmonic frames to the energy of remaining part of the signal.	Maximum harmonicity, minimum, mean, range and standard deviation.
Jitter	Pitch perturbation in vocal chords vibration.	Cycle-to-cycle variation of pitch.
Shimmer	Energy perturbation in vocal chords vibration.	Cycle-to-cycle variation of energy.
Active level	Signal active level features.	Maximum, minimum, mean and variance of the speech active level among the voiced regions.

doi:10.1371/journal.pone.0108975.t006

possible features. A partial ordering on this space, with each child having exactly one more feature than its parents, can be stated.

In order to state the FSS as a search problem, the following aspects must be identified:

- The starting point in the space. It determines the direction of the search. One might start with no features and successively add them, or one might start with all the features and successively remove them. One might also select an initial state somewhere in the middle of the search space.
- The organization of the search. It determines the strategy of the search in a space of size 2d, where d is the number of features in the problem. Roughly speaking, the search strategies can be optimal or heuristic. Two classic optimal search algorithms which exhaustively evaluate all possible subsets are depth-first and breadth-first [61]. Otherwise, Branch & Bound search [62] guarantees the detection of the optimal subset for monotonic evaluation functions without the systematic examination of all subsets.
- The evaluation function. It measures the effectiveness of a particular subset of features after the search algorithm has chosen it for examination. Being the objective of the search its maximization, the search algorithm utilizes the value returned by the evaluation function to help guide the search. Many measures carry out this objective regarding only the characteristics of the data, capturing the relevance of each feature or set of features to define the target concept. As reported by John et al. [63], when the goal of FSS is the maximization of the accuracy, the features selected should depend not only on the features and the target concept to be learned, but also on the learning algorithm.

Two factors can make difficult the implementation of FSS [64]: the number of features and the number of instances. One must bear in mind that the learning algorithm used in the searching scheme requires a training phase for every possible solution visited by the FSS search engine and this can be very time consuming.

One of the first approximations to FSS mentioned in the literature consists of performing a greedy (or Hill Climbing) search. Taking an empty as the initial variable set, the method attempts to include the variable that, at each step, maximizes the accuracy. The process stops when the inclusion of any variable does not show an improvement in the accuracy. This method is known as FSS-Forward.

More complex approximations for feature selection use genetic based operators as main searching engines.

Estimation of Distribution Algorithms as searching paradigm. Genetic Algorithms [65] are one of the best known techniques for solving optimization problems. Their use has reported promising results in many areas but there are still some problems where GAs fail. These problems, known as deceptive problems, have attracted the attention of many researchers and as a consequence there has been growing interest in adapting the GAs in order to overcome their weaknesses.

The GA is a population based search method. First, a set of individuals (or candidate solutions to our optimization problem) is generated (a population), then promising individuals are selected, and finally new individuals which will form the new. population are generated using crossover and mutation operators.

An interesting adaptation of this is the Estimation of Distribution Algorithm (EDA) [66] (see Figure 1). In EDA, there are neither crossover nor mutation operators, the new population is sampled from a probability distribution which is estimated from the selected individuals.

In this way, a randomized, evolutionary, population-based search can be performed using probabilistic information to guide the search. It is shown that although EDA approach process solutions in a different way to GAs, it has been empirically proven that the results of both approaches can be very similar [67]. In this way, both approaches do the same except that EDA replaces genetic crossover and mutation operators by means of the following two steps:

- A probabilistic model of selected promising solutions is induced,
- New solutions are generated according to the induced model.

The main problem of EDA resides on how the probability distribution $p_l(x)$ is estimated. Obviously, the computation of $2n$ probabilities (for a domain with n binary variables) is impractical. This has led to several approximations where the probability distribution is assumed to factorize according to a probability model (see [67] or [68] for a review).

The simplest way to estimate the distribution of good solutions assumes the independence between the features of the domain. New candidate solutions are sampled by only regarding the proportions of the values of all features independently to the remaining solutions. Population Based Incremental Learning (PBIL) [69], Compact Genetic Algorithm (cGA) [70] and Univariate Marginal Distribution Algorithm (UMDA) [71] are three algorithms of this type. They have worked well under artificial tasks with no significant interactions among features and

$D_0 \leftarrow$ Generate N individuals (the initial population) randomly

Repeat for l = 1, 2, . . . until a stop criterion is met

$V_i \leftarrow$ Evaluate the Value V_i for each of the N individuals

$D_{l-1}^{S} \leftarrow$ Select $S \leq N$ individuals from D_{l-1} according to a selection method

$p_l(x) = p(x|D_{l-1}^{S}) \leftarrow$ Estimate the joint probability distribution of an individual being among the selected inviduals

$D_l \leftarrow$ Sample N individuals (the new population) from $p_l(x)$

Figure 1. Main scheme of the Estimation of Distribution Algorithms (EDA) approach.
doi:10.1371/journal.pone.0108975.g001

so, the need for covering higher order interactions among the variables is seen for more complex or real tasks.

Results and Discussion

The abovementioned methods have been applied over the crossvalidated datasets using the MLC++ library [56]. Each dataset corresponds to a single actor. As previously mentioned, experiments were carried out within three different phases. At first the initial 32 features have been employed; then, the second set of 91 new features has been used; finally, both sets have been joined completing a global set of 123 features. The datasets corresponding to the 17 actors can be found in Files S1–S17, each of them containing a feature matrix with 123 columns. Tables 7 to 18 show the results obtained for the three phases, applying the ML classifiers mentioned in previous section with and without FSS. Each column in these Tables represents a female (Fi) or male (Mi) actor, and mean values corresponding to each classifier/gender are also included. Last column presents the total average for each classifier in each language. Confusion Matrices corresponding to the best results obtained for each gender and language are also shown in Tables 19 to 22. In order to check the validity of proposed process, a greedy searching approach (FSS-Forward) has been applied. Tables 23 and 24 show the results obtained applying this method. A comparison among different phases and ML paradigms used is also provided (Figures 2 to 5). Finally, some statistical tests have been applied to check the significance of the results obtained in the third phase (Tables 25 and 26).

First phase

Tables 7 and 8 show the results obtained for the first phase, without FSS for Basque and Spanish languages respectively, while Tables 9 and 10 show the improvement obtained by selecting relevant features. Here, IB paradigm with FSS outperforms both Basque and Spanish results, improving previous ones in 16.75% and 21.95% respectively.

Second phase

Results obtained using the second set of 91 features are reflected in Tables 11 and 12 (without FSS) and in Tables 13 and 14 (with FSS). ID3 is the best classifier for both languages when no FSS is applied. The results are slightly better than those obtained without FSS for the first phase, although the difference is not very significant. On the contrary, when FSS is applied to these second set of features the emotion classification performance is highly increased. Again, IB classifier stands out with an accuracy of 75.5% and 70.73% for Basque and Spanish, respectively. Compared to previous phase, accuracy is increased in a 10.62% for Basque and a 7.01% for Spanish.

Third phase

In this experiment, a set of 123 predictor features is used. Here, ID3 results show a small increase of performance without FSS (1.84% Basque and 3.01% Spanish), but improvement obtained after applying FSS to this whole set is more impressive. The classification accuracy is 4.55% higher for Basque and 4.09% higher for Spanish compared to previous phase, rising the overall performance up to 80.05% (Basque) and 74.82% (Spanish) (see Tables 15 to 18).

Tables 19 to 22 show the Confusion Matrices corresponding to the best results obtained for each gender and language. As it could be seen, very few errors are found in the classification process after FSS is performed.

FSS-Forward

To show the EDA searching process goodness, a greedy FSS searching approach (FSS-Forward) has also been applied. This method has only been tested for the third phase feature set, as it is only presented for comparison purposes. Obtained results are shown in Tables 23 and 24. The best results seem to be obtained with NB classifier for both languages, but classification performances are disappointing, as far as they are similar to those obtained using the initial set of 32 features without FSS.

Results comparison among different phases

The bar diagram in Figure 2 compares the performance of the four ML paradigms used (IB, ID3, C4.5, NB) without any kind of FSS, for the Basque language. Same comparison is shown for the Spanish language in Figure 3. It can be seen how ID3 outstands for both languages; results obtained using the full set are 50.53% for Basque and 45.47% for Spanish.

Figures 4 and 5 make the same comparison (Basque and Spanish, respectively) but this time, the improvements obtained after applying FSS to the different feature subsets are shown. The first three bars in each classifier column correspond to EDA-FSS, while the fourth one represents the FSS-Forward approach. Here, IB outperforms the rest of the classifiers for both languages and best results are obtained when EDA-FSS is applied to the whole set of features.

It is worth emphasizing that the difference between the classification accuracies obtained with the initial set of 32 features without FSS and those obtained with the whole set of 123 features after applying FSS sum up a notable increase in average of

Table 7. 10-fold crossvalidation accuracy of first phase for actors in Basque.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	35.38	48.79	35.23	39.80	44.17	49.32	36.89	40.91	42.82	41.52
ID3	38.71	45.45	44.70	**42.95**	46.67	46.97	43.26	51.14	47.01	45.27
C4.5	41.52	52.20	35.00	42.90	60.38	53.26	45.08	49.47	**52.04**	**48.13**
NB	42.95	45.76	37.65	42.12	52.20	44.09	36.21	41.44	43.48	42.90

doi:10.1371/journal.pone.0108975.t007

Table 8. 10-fold crossvalidation accuracy of first phase for actors in Spanish.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	34.55	43.64	54.55	54.55	38.18	45.09	25.45	33.64	51.82	47.65	33.64	**38.44**	**41.77**
ID3	36.36	52.73	49.09	47.27	42.73	**45.63**	20.91	30.91	40.91	47.27	40.00	36.00	40.82
C4.5	30.91	50.00	46.36	43.64	42.73	42.72	29.09	31.82	46.36	42.73	35.45	37.09	39.91
NB	38.18	42.73	49.09	40.00	42.73	42.54	24.55	30.91	49.09	45.45	34.55	36.91	39.73

doi:10.1371/journal.pone.0108975.t008

Table 9. 10-fold crossvalidation accuracy of first phase for actors in Basque applying EDA-FSS.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	63.03	68.03	59.32	**63.46**	72.65	67.35	60.98	62.80	**65.94**	**64.88**
ID3	62.73	60.48	65.45	62.88	72.65	61.97	56.52	62.65	63.44	63.20
C4.5	60.23	65.98	60.00	62.07	71.82	62.80	60.08	63.56	64.56	63.49
NB	64.47	64.55	48.94	59.32	74.55	62.50	62.73	60.00	64.94	62.53

doi:10.1371/journal.pone.0108975.t009

Table 10. 10-fold crossvalidation accuracy of first phase for actors in Spanish applying EDA-FSS.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	61.82	66.36	75.45	71.82	68.18	**68.72**	42.73	57.27	69.09	63.64	60.91	**58.72**	**63.72**
ID3	59.09	66.36	66.36	60.00	61.81	62.72	42.73	51.82	66.36	61.82	60.00	56.54	59.63
C4.5	57.27	62.73	64.55	65.45	63.64	62.72	43.64	56.36	65.45	64.55	56.36	57.27	60.00
NB	54.55	59.09	68.18	65.45	60.00	61.45	40.91	48.18	64.55	59.09	51.82	52.91	57.18

doi:10.1371/journal.pone.0108975.t010

Table 11. 10-fold crossvalidation accuracy of second phase for actors in Basque.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	34.00	42.91	33.91	36.94	56.18	41.00	36.91	36.82	42.73	40.25
ID3	49.45	45.91	46.78	**47.38**	54.27	44.00	51.45	49.45	**49.79**	**48.75**
C4.5	42.73	40.09	42.73	41.85	60.36	39.55	48.45	37.82	46.55	44.54
NB	39.82	31.00	46.45	39.09	60.36	29.91	36.91	41.44	42.16	40.84

doi:10.1371/journal.pone.0108975.t011

Table 12. 10-fold crossvalidation accuracy of second phase for actors in Spanish.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	36.46	41.92	41.92	43.64	33.64	39.52	30.00	36.46	44.55	36.46	30.00	35.49	37.51
ID3	38.18	47.27	55.45	43.64	44.55	45.82	24.55	40.00	50.00	46.36	34.55	**39.09**	**42.46**
C4.5	42.73	48.18	50.91	50.91	45.45	**47.64**	21.82	39.09	46.36	48.18	27.27	36.54	42.00
NB	34.55	34.45	40.91	32.73	31.82	34.89	20.91	39.09	40.00	35.45	21.82	31.45	33.17

doi:10.1371/journal.pone.0108975.t012

Table 13. 10-fold crossvalidation accuracy of second phase for actors in Basque applying EDA-FSS.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	72.55	79.73	62.27	**71.52**	91.36	73.00	77.82	71.82	**78.50**	**75.50**
ID3	71.00	71.73	66.64	69.79	78.73	65.82	72.64	66.91	71.03	70.50
C4.5	67.73	75.91	68.09	70.58	76.73	65.82	69.91	68.91	70.34	70.44
NB	73.00	77.73	63.36	71.36	89.45	67.27	66.18	65.36	72.07	71.76

doi:10.1371/journal.pone.0108975.t013

Table 14. 10-fold crossvalidation accuracy of second phase for actors in Spanish applying EDA-FSS.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	72.73	72.73	80.91	76.36	64.55	**73.46**	58.18	72.73	76.36	70.00	62.73	**68.00**	**70.73**
ID3	67.27	75.45	73.64	72.73	68.18	71.45	51.82	63.64	76.36	69.09	59.09	64.00	67.72
C4.5	70.91	75.45	74.55	64.55	66.36	70.36	54.55	63.64	80.91	66.36	56.36	64.36	67.35
NB	75.45	73.64	68.18	67.27	64.55	69.82	50.00	60.00	76.36	68.18	58.18	62.54	66.18

doi:10.1371/journal.pone.0108975.t014

Table 15. 10-fold crossvalidation accuracy of third phase for actors in Basque.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	36.00	46.82	33.82	38.88	59.45	44.36	40.45	36.55	45.20	42.49
ID3	49.55	47.64	39.91	**45.70**	61.00	49.27	53.36	50.36	**54.25**	**50.59**
C4.5	50.73	47.36	35.82	44.64	63.91	35.09	48.18	38.64	46.46	45.68
NB	43.73	40.91	40.91	41.85	58.36	37.09	46.64	40.82	45.73	44.07

doi:10.1371/journal.pone.0108975.t015

Table 16. 10-fold crossvalidation accuracy of third phase for actors in Spanish.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	32.73	36.36	48.18	45.45	40.00	40.54	28.18	40.91	47.27	37.27	31.82	37.09	38.82
ID3	35.45	50.00	55.45	41.92	50.91	46.75	30.00	49.09	55.45	47.27	39.09	**44.18**	**45.47**
C4.5	44.55	51.82	57.27	49.09	45.45	**49.64**	25.45	44.55	46.36	45.45	34.55	39.27	44.46
NB	30.91	38.18	44.55	32.73	40.91	37.46	20.91	37.27	46.36	40.91	26.36	34.36	35.91

doi:10.1371/journal.pone.0108975.t016

Table 17. 10-fold crossvalidation accuracy of third phase for actors in Basque applying EDA-FSS.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	75.36	82.55	73.73	**77.21**	90.45	84.27	76.27	77.73	**82.18**	**80.05**
ID3	68.09	75.64	71.64	71.79	78.82	69.55	73.73	69.73	72.96	72.46
C4.5	69.82	77.73	68.09	71.88	78.64	64.91	66.91	71.45	70.48	71.04
NB	74.82	82.55	67.27	74.88	91.27	78.73	67.91	74.73	78.16	76.75

doi:10.1371/journal.pone.0108975.t017

Table 18. 10-fold crossvalidation accuracy of third phase for actors in Spanish applying EDA-FSS.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	71.82	77.27	80.91	80.91	78.18	**77.82**	59.09	73.64	80.91	74.55	69.09	**71.42**	**74.82**
ID3	68.18	75.45	80.00	70.00	75.45	73.82	50.00	70.00	80.00	72.73	67.27	68.00	70.91
C4.5	67.27	73.64	80.00	71.82	70.91	72.73	52.73	70.00	76.36	75.45	66.36	68.18	70.46
NB	70.00	77.27	78.18	77.27	62.73	73.09	51.82	63.64	74.55	69.09	60.00	63.82	68.46

doi:10.1371/journal.pone.0108975.t018

Table 19. Confusion Matrix of the *F2* Basque actor.

	Sadness	Fear	Joy	Anger	Surprise	Disgust	Neutral
Sadness	**20**	0	0	0	0	0	0
Fear	0	**6**	1	0	0	0	1
Joy	0	0	**14**	1	0	0	0
Anger	0	0	2	**14**	0	0	1
Surprise	1	1	1	0	**5**	1	1
Disgust	2	0	2	0	0	**5**	0
Neutral	1	0	0	0	0	0	**21**

doi:10.1371/journal.pone.0108975.t019

Table 20. Confusion Matrix of the *M1* Basque actor.

	Sadness	Fear	Joy	Anger	Surprise	Disgust	Neutral
Sadness	**18**	0	0	0	0	0	2
Fear	1	**7**	0	0	0	0	0
Joy	0	0	**16**	1	0	0	0
Anger	0	0	2	**13**	0	1	1
Surprise	0	0	0	2	**8**	0	0
Disgust	0	0	0	0	0	**9**	0
Neutral	0	0	0	0	0	0	**22**

doi:10.1371/journal.pone.0108975.t020

Table 21. Confusion Matrix of the *F3* Spanish actor.

	Sadness	Fear	Joy	Anger	Surprise	Disgust	Neutral
Sadness	**17**	0	0	0	0	0	1
Fear	0	**7**	2	0	1	1	0
Joy	0	0	**14**	2	2	0	2
Anger	0	0	1	**16**	0	0	0
Surprise	0	0	2	1	**11**	1	0
Disgust	1	0	1	0	0	**4**	3
Neutral	0	0	0	0	0	0	**20**

doi:10.1371/journal.pone.0108975.t021

Table 22. Confusion Matrix of the *M3* Spanish actor.

	Sadness	Fear	Joy	Anger	Surprise	Disgust	Neutral
Sadness	**17**	0	0	0	0	0	1
Fear	0	**7**	2	1	0	1	0
Joy	0	1	**18**	0	0	0	1
Anger	0	0	1	**14**	0	2	0
Surprise	0	1	0	1	**10**	2	1
Disgust	0	0	3	1	2	**3**	0
Neutral	0	0	0	0	0	0	**20**

doi:10.1371/journal.pone.0108975.t022

Table 23. 10-fold crossvalidation accuracy for Basque applying FSS-FORWARD to the whole set.

	Female				Male					Total
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	38.91	46.55	44.00	43.15	66.18	51.45	47.45	48.55	53.41	49.01
ID3	42.73	43.55	52.45	46.24	59.27	42.82	49.36	45.45	49.23	47.95
C4.5	47.18	49.45	36.00	44.21	63.00	33.73	39.64	43.64	45.00	44.66
NB	47.45	62.09	31.09	**46.88**	69.73	56.18	46.45	50.36	**55.68**	**51.91**

doi:10.1371/journal.pone.0108975.t023

Table 24. 10-fold crossvalidation accuracy for Spanish using FSS-FORWARD to the whole set.

	Female						Male						Total
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	45.45	46.36	56.36	52.73	32.73	46.73	23.64	26.36	47.27	44.55	35.45	35.45	41.09
ID3	38.18	45.45	60.00	48.18	45.45	47.45	26.36	40.91	44.55	42.73	40.00	38.91	43.18
C4.5	35.45	46.36	57.27	55.45	39.09	46.72	29.09	28.18	44.55	35.45	37.27	34.91	40.82
NB	45.45	54.55	53.64	61.72	40.91	**51.25**	28.18	38.18	53.64	49.09	34.55	**40.73**	**45.99**

doi:10.1371/journal.pone.0108975.t024

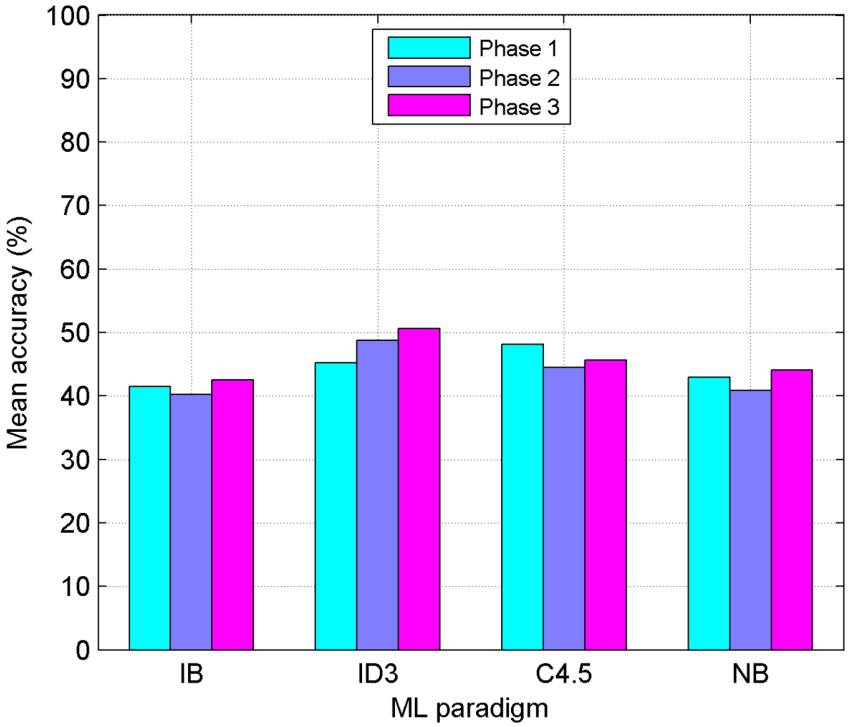


Figure 2. Results for the Basque Language without Feature Subset Selection. Performance comparison between four Machine Learning paradigms (IB: Instance Based, ID3: Decision Tree, C4.5: Decision Tree, NB: Naive-Bayes) without any kind of FSS. Mean accuracy obtained in the three phases, for the Basque language, is shown.
doi:10.1371/journal.pone.0108975.g002

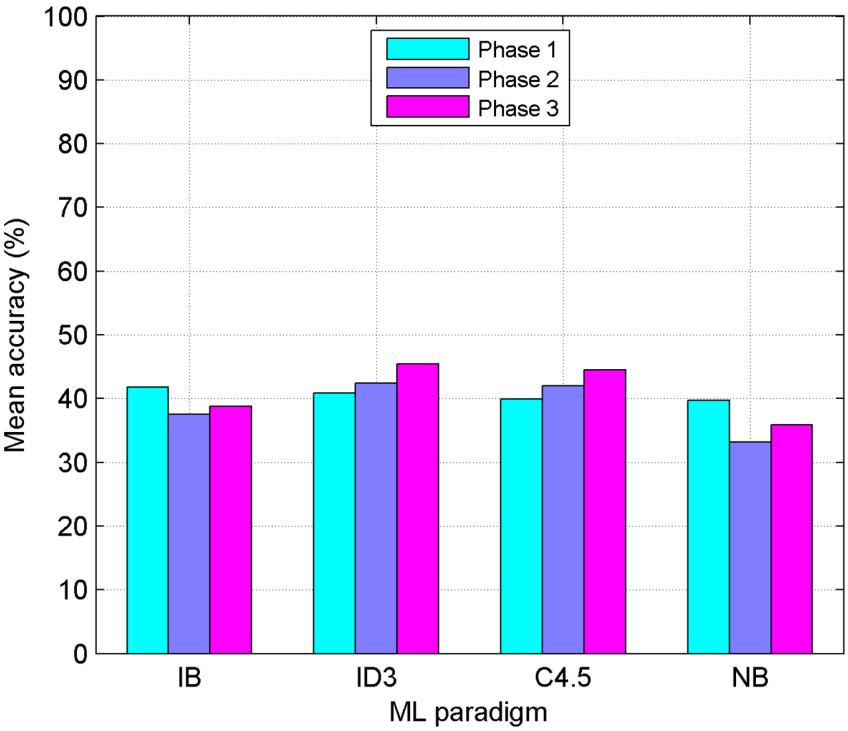


Figure 3. Results for the Spanish Language without Feature Subset Selection. Performance comparison between four Machine Learning paradigms (IB: Instance Based, ID3: Decision Tree, C4.5: Decision Tree, NB: Naive-Bayes) without any kind of FSS. Mean accuracy obtained in the three phases, for the Spanish language, is shown.
doi:10.1371/journal.pone.0108975.g003

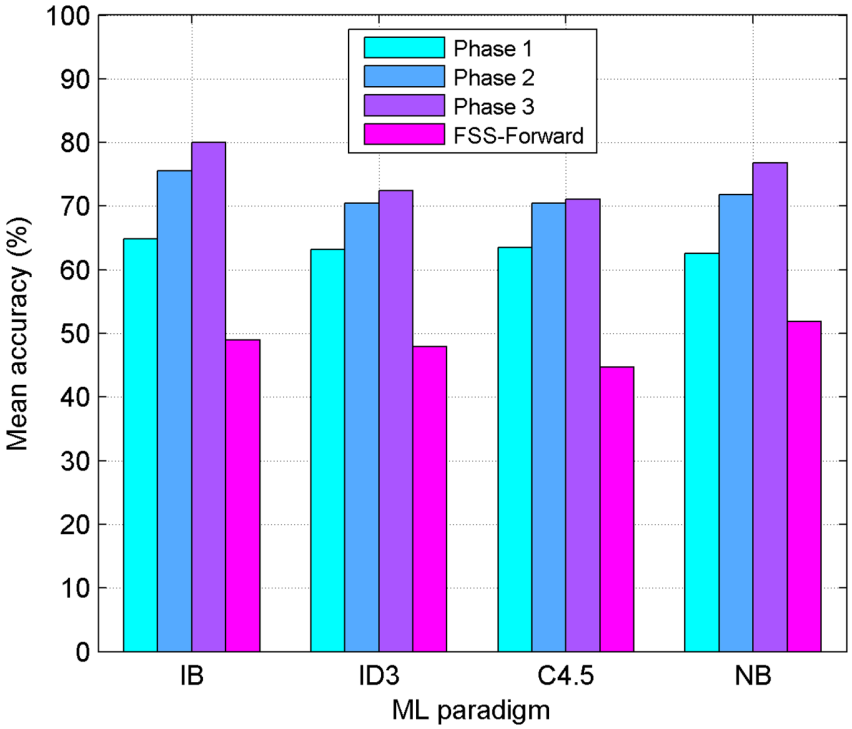


Figure 4. Results for the Basque Language using EDA Feature Subset Selection. Performance comparison between four Machine Learning paradigms (IB: Instance Based, ID3: Decision Tree, C4.5: Decision Tree, NB: Naive-Bayes) using EDA-FSS. Mean accuracy obtained in the three phases, for the Basque language, is shown. Results obtained with a standard FSS-Forward approach are also shown.
doi:10.1371/journal.pone.0108975.g004

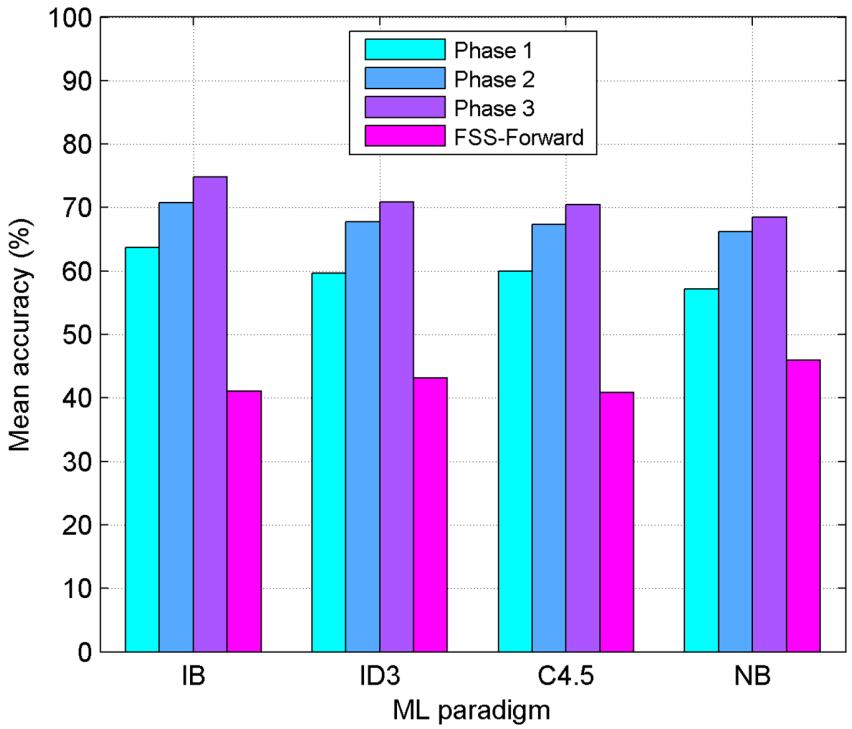


Figure 5. Results for the Spanish Language using EDA Feature Subset Selection. Performance comparison between four Machine Learning paradigms (IB: Instance Based, ID3: Decision Tree, C4.5: Decision Tree, NB: Naive-Bayes) using EDA-FSS. Mean accuracy obtained in the three phases, for the Basque language, is shown. Results obtained with a standard FSS-Forward approach are also shown.
doi:10.1371/journal.pone.0108975.g005

Table 25. p-values obtained with Wilcoxon test comparing FSS methods.

Classifier	FSS-FWD > without FSS ?	EDA > FSS-FWD ?
All	**0,03667**	**3,89e-13**
IB	**0,02323**	**0,00016**
ID3	0,95359	**0,00016**
C4.5	0,92620	**0,00016**
NB	**0,00174**	**0,00016**

doi:10.1371/journal.pone.0108975.t025

30.62% for the Basque language and 30.61% for the Spanish language.

Statistical tests

As seen in previous subsections, EDA based FSS clearly improves classification accuracies for all subjects, in both languages, and with all the classifiers, but to extract other interesting conclusions about the goodness of classifiers and FSS-Forward procedure, the mean values for all subjects are not sufficiently significant, and some type of statistical test should be made.

We have used Wilcoxon signed-rank test [72], that is a non-parametric paired difference test, used to assess whether two population mean ranks differ. Specifically, we have used the right-sided version, which tests a hypothesis of the form X>Y?

Tables 25 and 26 show the p-values obtained by applying the test to various hypotheses. Only third phase feature set has been used for tests, and in all cases the sample to test is constructed using the classification accuracies obtained for all subjects (17 without distinguishing languages), for a given classifier and FSS strategy. In some cases we have put together the four types of classifiers, working with samples of 61 values.

A p-value is a nonnegative scalar from 0 to 1 that represents the probability of observing, under the null hypothesis, data as or more extreme than the obtained values. If the p-value is less than a certain significance level we say that the hypothesis is significantly valid. In tables 25 and 26, significant values (<5%) are in bold.

In Table 25 the improvement obtained with the different FSS strategies are compared. The second column shows that if we do not distinguish between classifiers, FSS-forward is significantly better than not using FSS, but the p-value is just down 5%. In fact, its behaviour depends strongly on the classifier, obtaining the best results for NB, but not improving significantly with ID3 and C4.5. The third column shows, as we already knew, that EDA-FSS significantly improves the results of FSS-forward in all cases.

In Table 26, the classifier with best results for each FSS methods is compared with the others. Without FSS, ID3 is significantly better than IB and NB. When features are selected with greedy FSS-forward method, NB is significantly better than IB and C4.5. Finally, when EDA-FSS is applied, IB clearly outperforms all the other classifiers.

Most relevant features

The procedure employed to extract the most relevant features is based on the results and the features used in the third phase, where the best classification rates have been obtained and the whole set of features have been employed.

EDA based FSS has been applied for each of previous described ML paradigms, so each classifier has found its own relevant features for each actor. In order to identify the most relevant speech features for SER this estimation has been based on the paradigm which obtains the higher classification rate after applying FSS. As mentioned before, the classifier with the best results in most of the cases is the IB paradigm, except in a case of a male actor (M1) for Basque language (see Table 17). As overall IB can be considered the most adequate option for the defined task, IB paradigm resulting features have been taken into account to select the most relevant features, which have been extracted separately for Spanish and Basque languages on one hand and for gender on the other (see Tables 27 and 28).

This information concerns to the features that EDA evolutionary algorithm selects more frequently for each actor. Given that the classification is speaker dependent, each actor may have different relevant features for each ML paradigm. These relevant features have been analyzed grouping actors by language and gender aiming at a partial independence of the actor. The purpose of this grouping is to shed more light on the impact that gender and language can have in the final features of each subgroup. The criterion to consider relevant a feature in a subgroup is that more than the 50% of the actors have that feature selected by the algorithm.

It must be highlighted that several features are common for all the categories, both for Spanish and Basque languages and for male and female gender, principally the prosodic features related with the Fundamental Frequency - the mean, variance, the mean square error of the regression coefficient and mean of the pitch means in every voiced region; Energy - maximum, mean and variance; RMS energy - maximum and mean - and Loudness. The features related with the voice quality and shared by all the

Table 26. p-values obtained with Wilcoxon test comparing the best classifier for each FSS method with the others classifiers.

FSS method	Classifier	> IB ?	> ID3 ?	> C4.5 ?	> NB ?
None	ID3	**0,00038**	1	0,06477	**0,00019**
Forward	NB	**0,00930**	0,05119	**0,00260**	1
EDA	IB	1	**0,00016**	**0,00019**	**9,155e-05**

doi:10.1371/journal.pone.0108975.t026

Table 27. The most relevant features using the IB paradigm with EDA for Basque.

Feature class	Female	Male
Fundamental Frequency	Position of the maximum, minimum and its position, mean, variance and mean square error of the regression coefficient.	Mean, variance, maximum positive slope in contour, mean square error of the regression coefficient.
		Mean of the derivative and mean square error of the regression coefficient of the derivative.
Energy	Maximum, mean, variance and regression coefficient.	Maximum, minimum, mean, variance, mean square error of the regression coefficient.
	RMS maximum and mean.	RMS maximum and mean.
	Loudness.	Loudness
Voiced/Unvoiced	F0 value of the first and last voiced frames and length of the longest unvoiced region.	Ratio of number of voiced and unvoiced frames and number of frames.
Relations	Mean of the pitch means in every regions and duration from beginning to pitch maximum.	Mean of the pitch means in every regions.
	Ratio of the energy maximum.	
Formants	Mean of the second and third formant frequency, the bandwidths of the first and second formants and mean of the second formant ratio.	Mean of the first, second and third formant frequency and the bandwidths of the first and second formants
Critical Bands	Energy in bands (0–1300 Hz), (0 - F0 Hz) and (2500–3500 Hz).	Energy in band (1300–2600 Hz). Energy in band (2500–3500 Hz) of whole the utterance divided by the energy over all frequencies
	Rate of the energy of the longest region and energy over all the utterance.	Rate of energy in longest region and energy over all the utterance.
Harmonicity to noise ratio	Range.	Range.
Jitter	Cycle-to-cycle variation of pitch.	
Shimmer	Cycle-to-cycle variation of energy.	
Active level	Maximum and mean.	Maximum and mean.

doi:10.1371/journal.pone.0108975.t027

categories are less than the prosodic and they specially refer to the third formant mean, the first and second formants bandwidth and the level of the activation of the speech signal; in this case, the maximum and mean stand out among all the voiced regions. These common features in all groups could be considered as the more relevant in order to design a system that intends to achieve full speaker independence. This system should be able to classify automatically emotions no matter who the speaker is.

The non-shared features in each subgroup should be analyzed in order to establish the relationships between these features and language and gender dependent characteristics.

Conclusions and Future Work

This paper shows an attempt to select the most significant features for emotion recognition in spoken Basque and Spanish Languages. RekEmozio database was used as experimental data set. Several ML paradigms were used for the emotion classification task. Experiments were executed in three different phases, using different sets of features as classification variables in each phase. Moreover, feature subset selection was applied at each phase in order to seek for the most relevant feature subset. The three phases approach has proven to be useful in order to check which ML paradigms provide the best results in emotion automatic recognition and provide initial results with different sets of features.

Results show an encouraging improvement in the accuracies obtained. From an initial emotion classification performance of about 48% for the initial set of 32 features, performance has increased up to 80% when EDA-FSS is applied to the whole set of features for the case of Basque language. For the Spanish language, although a bit smaller, the performance has also shown a noticeable increase from 41% up to almost 75%. It is worth noting that achieved results are approaching the emotion recognition rate obtained by humans when validating RekEmozio database.

Therefore, emotion recognition rates have been improved using the features defined in this paper, but it must also be taken into account that such improvement has been achieved after applying EDA for FSS. Concerning the classifiers used, accuracies have

Table 28. The most relevant features using the IB paradigm with EDA for Spanish.

Feature class	Female	Male
Fundamental Frequency	Minimum, mean, variance and regression coefficient and its mean square error.	Maximum, minimum, mean, variance and mean square error of the regression coefficient.
	Maximum, mean and mean square error of the regression coefficient of the derivative	Mean of the derivative.
Energy	Maximum, minimum, mean, variance and regression coefficient and its mean square error.	Maximum and mean.
	RMS maximum, minimum and mean.	RMS value, maximum, mean.
	Loudness.	Loudness.
Voiced/Unvoiced	F0 value of the first and last voiced frames and length of the longest unvoiced region, ratio of number of voiced frames and number of frames.	F0 value of the first voiced frame, number of unvoiced frames, length of the longest unvoiced region, ratio of unvoiced regions.
Relations	Mean and variance of the pitch means in every regions.	Mean, variance, variance of the maximum, mean of the pitch ranges and mean of the flatness of the pitch based on every voiced region pitch values.
		Global energy mean among voiced regions
Rhythm	Duration of silence and maximum voiced parts.	Duration of silence parts.
Formants	Mean of the first, second and third formant frequency and the bandwidths of the second and third formants.	Mean of the first formant frequency and the bandwidths of the first, second and third formants.
Critical Bands	Energy in bands (0–1300 Hz) and (2600–4000 Hz). Energy in bands (0–1000 Hz), (2500–3500 Hz) and of whole the utterance divided by the energy over all frequencies.	Energy in bands (0–1300 Hz) and (2600–4000 Hz). Energy in band (4000–5000 Hz) of whole the utterance divided by the energy over all frequencies.
	Rate of the energy of the longest region and energy over all the utterance.	Rate of the energy of the longest region and energy over all the utterance.
Harmonicity to noise ratio		Minimum
Shimmer		Perturbation cycle to cycle of the energy.
Active level	Maximum, minimum, mean and variance.	Maximum, mean and variance.

doi:10.1371/journal.pone.0108975.t028

clearly improved over the results obtained using the full set of features. IB appears as the best classifier in most experiments, if EDA-FSS is applied, and ID3 when no FSS is applied. In order to check the validity of achieved results, a greedy FSS searching approach (FSS-Forward) has been applied, but providing disappointing classification performances, and showing the best results when NB classifier is used. As future work, the authors will extend the study to other classifiers (SVM,…) and other methods of feature selection.

Authors have developed affective recognizers for speech using the categorical theory of emotions. However, currently they are studying emotions according to dimensional and appraisal models, information from other modalities (such as verbal and psycho physiological information) and also, other models such as user context models. In the future, the authors will perform studies related with the meaning of the utterances, comparing the results with semantically meaningful content and with non-semantically meaningful content. Moreover, more languages will be taken into account (such as the Catalan language).

Supporting Information

File S1 Feature matrix corresponding to Basque male actor M1.
(CSV)

File S2 Feature matrix corresponding to Basque male actor M2.
(CSV)

File S3 Feature matrix corresponding to Basque male actor M3.
(CSV)

File S4 Feature matrix corresponding to Basque male actor M4.
(CSV)

File S5 Feature matrix corresponding to Basque female actress F1.
(CSV)

File S6 Feature matrix corresponding to Basque female actress F2.
(CSV)

File S7 Feature matrix corresponding to Basque female actress F3.
(CSV)

File S8 Feature matrix corresponding to Spanish male actor M1.
(CSV)

File S9 Feature matrix corresponding to Spanish male actor M2.
(CSV)

File S10 Feature matrix corresponding to Spanish male actor M3.
(CSV)

File S11 Feature matrix corresponding to Spanish male actor M4.
(CSV)

File S12 Feature matrix corresponding to Spanish male actor M5.
(CSV)

File S13 Feature matrix corresponding to Spanish female actress F1.
(CSV)

File 14 Feature matrix corresponding to Spanish female actress F2.
(CSV)

File S15 Feature matrix corresponding to Spanish female actress F3.
(CSV)

File S16 Feature matrix corresponding to Spanish female actress F4.
(CSV)

File S17 Feature matrix corresponding to Spanish female actress F5.
(CSV)

Author Contributions

Conceived and designed the experiments: A. Arruti IC A. Álvarez EL BS. Performed the experiments: A. Arruti A. Álvarez BS. Analyzed the data: A. Arruti IC A. Álvarez EL BS. Contributed reagents/materials/analysis tools: A. Arruti IC A. Álvarez EL BS. Contributed to the writing of the manuscript: A. Arruti IC A. Álvarez EL BS.

References

1. Picard RW (1997) Affective Computing. Cambridge, MA: MIT Press.
2. Tao J, Tan T (2005) Affective computing: A review. In: Proceedings of The First International Conference on Affective Computing & Intelligent Interaction (ACII'05), pp. 981–995.
3. Garay N, Cearreta I, López JM, Fajardo I (2006) Assistive technology and affective mediation. Human technology, 2(1): 55–83.
4. Koolagudi S, Rao KS (2012) Emotion recognition from speech: a review. International Journal of Speech Technology, 15: 99–117.
5. Ramakrishnan S, El Emary IMM (2013) Speech emotion recognition approaches in human computer interaction. Telecommunication Systems, 52(3): 1467–1478.
6. Álvarez A, Cearreta I, López JM, Arruti A, Lazkano E, et al. (2006) Feature Subset Selection based on Evolutionary Algorithms for automatic emotion recognition in spoken Spanish and Standard Basque languages. In: Proceedings of Ninth International Conference on Text, Speech and Dialog (TSD'06), pp. 565–572.
7. Álvarez A, Cearreta I, López JM, Arruti A, Lazkano E, et al. (2007) A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms. In: Proceedings of Tenth International Conference on Text, Speech and Dialog (TSD'07), pp. 423–430.
8. Lang PJ (1979) A bio-informational theory of emotional imagery. Psychophysiology, 16: 495–512.
9. Ekman P (1984) Expression and nature of emotion. In: Scherer K, Ekman P, editors. Approaches to emotion. Hillsdale, New Jersey: Erlbaum.
10. Scherer KR (1986) Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99: 143–165.
11. Picard RW (1998) Towards Agents that Recognize Emotion. In: Proceedings IMAGINA, pp. 153–165.
12. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 59(1–2): 157–183.
13. Ekman P, Friesen W (1976) Pictures of facial affect. Palo Alto, CA, Consulting Psychologist Press.
14. Alonso-Arbiol I, Shaver PR, Fraley RC, Oronoz B, Unzurrunzaga E, et al. (2006) Structure of the Basque emotion lexicon. Cognition and Emotion, 20(6): 836–865.
15. Bradley MM, Lang PJ, Cuthbert NB (1997) Affective Norms for English Words (ANEW). University of Florida, NIMH Center for the Study of Emotion and Attention.
16. Koolagudi S, Krothapalli S (2012) Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. International Journal of Speech Technology 15: 495–511.
17. Athanaselis T, Bakamidis S, Dologlou I, Cowie R, Douglas-Cowie E, et al. (2005) ASR for emotional speech: clarifying the issues and enhancing performance. Neural Networks, 18: 437–444.
18. Fragopanagos NF, Taylor JG (2005) Emotion recognition in human-computer interaction, Neural Networks, 18: 389–405.
19. Cowie R, Douglas-Cowie E, Cox C (2005) Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural Networks, 18: 371–388.
20. Humaine (2007) Available: http://emotion-research.net/. Accessed 11 March 2007.
21. López JM, Cearreta I, Fajardo I, Garay N (2007) Validating a multimodal and multilingual affective database. In: Proceedings of the 2nd international conference on Usability and internationalization (UI-HCII'07), pp. 422–431.
22. Navas E, Hernáez I, Castelruiz A, Luengo I (2004) Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque. in Proceedings of Seventh International Conference on Text, Speech and Dialog (TSD'04), pp. 393–400.
23. Iriondo I, Guaus R, Rodríguez A, Lázaro P, Montoya N, et al. (2000) Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 161–166.
24. Caballero-Morales SO (2013) Recognition of emotions in Mexican Spanish speech: an approach based on acoustic modelling of emotion-specific vowels. In: Scientific World Journal, vol. 13 pages.
25. López JM, Cearreta I, Garay N, López de Ipiña K, Beristain A (2006) Creación de una base de datos emocional bilingüe y multimodal. In: Proceedings of the 7th Spanish Human Computer Interaction Conference, pp. 55–66.
26. Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic Emotion Recognition using Prosodic Parameters. In: Proceedings of the ninth European Conference on Speech Communication and Technology (Eurospeech'05), pp. 493–496.
27. Nogueiras A, Moreno A, Bonafonte A, Mariño JB (2001) Speech emotion recognition using hidden Markov models. In: Proceedings of the seventh European Conference on Speech Communication and Technology (Eurospeech'01), pp. 2679–2682.
28. Hozjan V, Kačič Z (2003) Context-independent multilingual emotion recognition. International Journal of Speech Technology, 6(3): 311–320.

29. Abelin A, Allwood J (2000) Cross-linguistic interpretation of emotional prosody. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 110–113.
30. Tickle A (2000) English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 104–109.
31. Dellaert F, Polzin T, Waibel A (1996) Recognizing Emotion in Speech. In: Proceedings of the fourth International Conference on Spoken Language (ICSLP'96).
32. Taylor JG, Scherer K, Cowie R (2005) Introduction to Emotion and Brain: Understanding Emotions and Modelling their recognition. Neural Networks, 18(4): 313–316.
33. Pan Y, Shen P, Shen L (2012) Speech Emotion Recognition Using Support Vector Machine. In: International Journal of Smart Home, 6(2).
34. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, et al. (2001) Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1): 32–80.
35. Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication, 49(3): 201–212.
36. Rani P, Liu C, Sarkar N, Vanman E (2006) An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Analysis and Applications, 9(1): 58–69.
37. Partila P, Voznak M (2013) Speech Emotions Recognition Using 2-D Neural Classifier. In: Advances in Intelligent Systems and Computing, 210: 221–231.
38. Truong KP, van Leeuwen DA (2007) Automatic discrimination between laughter and speech. Speech Communication, 49(2): 144–158.
39. Schröder M (2004) Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (Ph.D thesis). Saarland University, Institute of Phonetics.
40. Navas E, Hernáez I, Castelruiz A, Sánchez A, Luengo I, et al. (2004) Acoustical Analysis of Emotional Speech in Standard Basque for Emotions Recognition. In: Proceedings of the ninth Iberoamerican Congress on Pattern Recognition (CIARP'04), pp. 386–393.
41. Montero JM, Gutiérrez-Arriola J, Colás J, Enríquez E, Pardo JM (1999) Analysis and Modelling of Emotional Speech in Spanish. In: Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS'99), pp. 957–960.
42. Cordoba R, Montero JM, Gutierrez JM, Vallejo JA, Enriquez E, et al. (2002) Selection of the most significant parameters for duration modelling in a Spanish text-to-speech system using neural networks. Computer Speech and Language, 16: 183–203.
43. Tato R, Santos R, Kompe R, Pardo JM (2002) Emotional space improves emotion recognition. In: Proceedings of 7th International Conference on Spoken Language Processing (ICSLP'02), pp. 2029–2032.
44. Batliner A, Fisher K, Huber R, Spilker J, Nöth E (2000) Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In Proceedings of the ISCA Workshop on Speech and Emotion, pp. 195–200.
45. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S (2011) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computing and Applications, 21: 2115–2126.
46. Petrushin V (1999) Emotion in Speech: Recognition and Application to Call Centers. In: Proceedings of Conference on Artificial Neural Networks in Engineering (ANNIE'99), pp. 7–10.
47. Zhang S, Zhao X (2013) Dimensionality reduction-based spoken emotion recognition. Multimedia Tools and Applications, 63: 615–646.
48. Picard RE, Vyzas E, Healy J (2001) Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions Pattern Analysis and Machine Intelligence, 23(10): 1175–1191.
49. Laukka P (2004) Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts (Ph.D thesis). Uppsala University.
50. Huber R, Batliner A, Buckow J, Nöth E, Warnke V, et al. (2000) Recognition of emotion in a realistic dialogue scenario. In: Proceedings of the fourth International Conference on Spoken Language (ICSLP'00), pp. 665–668.
51. Martin JK (1997) An exact probability metric for Decision Tree splitting and stopping. Machine Learning, 28(2/3): 257–291.
52. Mingers J (1988) A comparison of methods of pruning induced Rule Trees (Technical Report). University of Warwick, School of Industrial and Business Studies.
53. Quinlan JR (1986) Induction of Decision Trees. Machine Learning, 1: 81–106.
54. Quinlan JR (1993) C4.5: Programs for Machine Learning. California, Morgan Kaufmann Publishers.
55. Ting KM (1995) Common issues in Instance-Based and Naive-Bayesian classifiers (Ph.D. Thesis). The Univesity of Sidney Basser, Department of Computer Science.
56. Kohavi R, Sommerfield D, Dougherty J (1997) Data mining using MLC++, a Machine Learning Library in C++. International Journal of Artificial Intelligence Tools, 6 (4): 537–566.
57. Aha D, Kibler D, Albert MK (1991) Instance-Based learning algorithms. Machine Learning, 6: 37–66.
58. Wettschereck D (1994) A study of distance-based Machine Learning Algorithms (Ph.D. Thesis), Oregon State University.
59. Minsky M (1961) Steps towards artificial intelligence. Proceedings of the IRE, 49: 8–30.
60. Aha DW, Bankert RL (1994) Feature selection for case-based classification of cloud types: An empirical comparison. In: Proceedings of the AAAI'94 Workshop on Case-Based Reasoning, pp. 106–112.
61. Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery and Data Mining. Norwell, MA, Kluwer Academic Publishers.
62. Narendra P, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. IEEE Transactions on Computer, C-26 (9): 917–922.
63. John G, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: Machine Learning: Proceedings of the Eleventh International Conference, pp. 121–129.
64. Liu H, Setiono R (1998) Incremental Feature Selection. Applied Intelligence, 9(3): 217–230.
65. Holland JH (1975) Adaptation in Natural and Artificial Systems. Ann Arbor, MI University of Michigan Press.
66. Mühlenbein H, Paass G (1996) From recombination of genes to the estimation of distributions. Binary parameters. In: Lecture Notes in Computer Science: Parallel Problem Solving from Nature (PPSN IV), 1411: 188–197.
67. Pelikan M, Goldberg DE, Lobo F (1999) A Survey of Optimization by Building and Using Probabilistic Model (IlliGAL Report 99018), University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory.
68. Larrañaga P, Etxeberria R, Lozano JA, Sierra B, Inza I, et al. (1999) A review of the cooperation between evolutionary computation and probabilistic graphical models. In: Proceedings of the II Symposium on Artificial Intelligence (CIMAF99), pp. 314–324.
69. Baluja S (1994) Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning (Technical Report CMU-CS-94-163), Pittsburgh, PA, Carnegie Mellon University.
70. Harik GR, Lobo FG, Goldberg DE (1997) The compact genetic algorithm (IlliGAL Report 97006). University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory.
71. Mühlenbein H (1997) The equation for response to selection and its use for prediction. Evolutionary Computation, 5(3): 303–346.
72. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7: 1–30.

7.13 A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms

- **Authors:** Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay
- **Booktitle:** Proceedings of the 10th International Conference on Text, Speech and Dialogue (TSD)
- **Year:** 2007
- **Publisher:** Springer

A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms

Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay

Computer Science Faculty (University of the Basque Country)
Manuel Lardizabal 1, E-20018 Donostia (Gipuzkoa), Spain
aalvarez031@ikasle.ehu.es

Abstract. Study of emotions in human-computer interaction is a growing research area. Focusing on automatic emotion recognition, work is being performed in order to achieve good results particularly in speech and facial gesture recognition. This paper presents a study where, using a wide range of speech parameters, improvement in emotion recognition rates is analyzed. Using an emotional multimodal bilingual database for Spanish and Basque, emotion recognition rates in speech have significantly improved for both languages comparing with previous studies. In this particular case, as in previous studies, machine learning techniques based on evolutive algorithms (EDA) have proven to be the best emotion recognition rate optimizers.

1 Introduction

Affective computing, a discipline that develops devices for detecting and responding to users' emotions [20], is a growing research area [22]. The main objective of affective computing is to capture and process affective information with the aim of enhancing and naturalizing the communication between the human and the computer.

Development of affective systems is a challenge that involves analysing different multimodal data sources. A large amount of data is needed in order to include a wide range of emotionally significant material. Affective databases are a good chance for developing such applications, either for affective recognizers or either for affective synthesis.

This papers presents a study aimed at giving a new step towards researching relevant speech parameters in automatic speech recognition area. Based on the same classification techniques used in [1], where basic speech parameters were used, a wide group of new parameters have been used in order to analyse if emotion recognition rates have improved. This work has also served to check the efficiency of these new parameters in the study of emotions.

After a brief review on related work, corpus, speech parameters and machine learning techniques used for this study are detailed. Achieved experimental results are shown next. Finally, some conclusions and future work are highlighted.

V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 423–430, 2007.

2 Related Work

Affective resources, such as affective databases, provide a good opportunity for training affective applications, either for affective synthesis or for affective recognizers based on classification via artificial neural networks, Hidden Markov Models, genetic algorithms, or similar techniques (e.g., [2, 8]). There are some references in the literature that present affective databases and their characteristics. [4] carried out a wide review of affective databases. Other interesting reviews are the ones provided in [11] and [16].

Most of these references are related to English, while other languages have less resources developed, especially the ones with relatively low number of speakers. This is the case of Standard Basque. To our knowledge, the first affective database in Standard Basque is the one presented by [18]. Concerning to Spanish, the work of [12] stands out.

This type of databases usually record information such as images, sounds, psychophysiological values, etc. RekEmozio database is a multimodal bilingual database for Spanish and Basque [17], which also restores information came from processes of some global speech parameters extraction for each audio recording. Some of these parameters are prosodic features while others are quality features.

Machine Learning (ML) paradigms take a principal role in some works related to emotion recognition that can be found in literature. [5] presented a good reference paper. The Neural Networks Journal devoted special issue to emotion treatment from a Neural Networks perspective [23]. The work by [4] is related with this paper in the sense of using a Feature Selection method in order to apply a Neural Network to emotion recognition in spoken English, although both the methods to perform the FSS and the paradigms are different. In this line it has to be pointed out the work by [10] which uses a reduced number of emotions and a greedy approach to select the features.

3 Study of Automatic Emotion Recognition Relevant Parameters Using Machine Learning Paradigms

3.1 Corpus

The emotions used were chosen based on Ekman's six basic emotions [6], and neutral emotion was added. RekEmozio database has been used for this study. Its characteristics are described in [17].

3.2 Emotional Feature Extraction

One of the most important questions for automatic recognition of emotions in speech is which features should be extracted from the voice signal. Previous studies show that it is difficult to find specific voice features that could be used as reliable indicators of the emotion present in the speech [14]. In order to improve the results of previous works, new parameters have been calculated using the same recordings.

These parameters have been collected from the work carried out of [7] and consist of a total of 91. Parameters are divided in this way:

3.2.1 Prosodic Features

Many parameters have been computed that model the F0, energy, voiced and unvoiced regions, pitch derivative curve and the relations between the features as proposed in [3, 21].

- F0 based features: Values of the F0 curve in the voiced parts. Maximum, position of the maximum, minimum, position of the minimum, mean, variance, regression coefficient and its mean square error are computed.
- Energy: maximum and its position curve in the whole utterance, minimum and its relative time position, mean, variance, regression coefficient and its mean square error values are computed.
- Voiced/unvoiced regions based features: F0 value of the first and last voiced frames, number of regions with more than three successive voiced and unvoiced frames, amount of voiced and unvoiced frames in the utterance, length of the longest voiced and unvoiced regions, ratio of number of voiced and voiced regions and the ratio of number of voiced and unvoiced frames in the whole utterance are computed.
- Pitch contour derivative based features: Based on the derivative of the F0, maximum, minimum, mean, variance, regression coefficient and its mean square error values are computed.
- Relations among features: Mean, variance, mean of the maximum, variance of the maximum, mean of the pitch ranges and mean of the flatness of the pitch based on every voiced region pitch values are computed. The pitch increasing and decreasing in voiced parts and in the whole utterance is also measured as well as the mean of the voiced regions duration. Many features related with the energy among the voiced regions have been taken account, like global energy mean, vehemence, mean of the flatness and tremor in addition to others.

3.2.2 Quality Features

Features have been computed related with the voice quality parameters, such as formants, energy band distribution, harmonicity to noise ratio and active level in speech.

- Formant frequency based features: Mean of the first, second and third formant frequencies and their bandwidths among all voiced region are computed as well as the maximum and the range of the second formant ratio.
- Energy band distribution: The four frequency bands used are the following: 0 Hz to F0 Hz, 0Hz to 1000 Hz, 2500 Hz to 3500 Hz and 4000 Hz to 5000 Hz. The energy contained in the corresponding band is calculated for all voiced parts and divided by the energy over all frequencies of the voiced parts of utterance. The longest region is calculated and the energy values are also computed in that region as well as rate and relative energy contained in voiced regions and energy over all utterance.

- Harmonicity to noise ratio: The ratio of the energy of the harmonic frames to the energy of the remaining part of the signal is computed. In this sense, maximum harmonicity, minimum, mean, range and standard deviation have been analysed.
- Active level features: Maximum, minimum, mean and variance of the speech active level among the voiced regions are computed.

3.3 Machine Learning Standard Paradigms Used

The models that have been constructed to solve this problem of classification made up of 91 speech related values for each sample are constructed for the previous study, while the label value is one of the seven emotions identified. Therefore, *Decision Trees, Instance-Based Learning* and *Naive Bayes* compose the sort of classifiers.

Most of the supervised learning algorithms perform rather poorly when faced with many irrelevant or redundant features (depending on the specific characteristics of the classifier). Bearing it in mind, Feature Subset Selection (FSS) [15] is applicated, as it can be reformulated as follows: *given a set of candidate features, select the 'best' subset in a classification problem*. In this case, the 'best' subset will be the one with the best predictive accuracy. The FSS proposes additional methods to reduce the number of features so as to improve the performance of the supervised classification algorithm.

For applicating FSS, an Estimation of Distribution Algorithm (EDA) [19] has been used having model accuracy as fitness function. It is necessary to clarify that given the number of dimensions of the classification problem (search space for EDA is 2^91=2,4758E27), it is not possible to consider all the possibilities.

4 Experimental Results

The above mentioned methods have been applied over the crossvalidated data sets using the MLC++ library [13]. Each dataset corresponds to a single actor. Experiments were carried out with and without FSS in order to extract the accuracy improvement introduced by the feature selection process and then comparated with the results obtained in the previous study. The first two tables show classification results obtained using the whole set of variables, for Basque and Spanish languages respectively. First column presents used Machine Learning paradigms (as Decision Trees classifiers, ID3 and C4.5 paradigms; as Instance-Based Learning classifiers, IB paradigm; and finally, Naive Bayes classifier (NB) have been used). Last column presents the total average for each classifier. Each remaining column represents a female (Fi) or male (Mi) actor, and mean values corresponding to each classifier/gender are also included.

Results do not seem very impressive. In fact, as it can be see in Figure 1, results from previous studies are not improved, except for one case. For both Basque and Spanish languages, ID3 best classifies emotions for female actresses, as well as for male actors for Basque. IB appears as the best classifier for Spanish for male actors.

Table 1. 10-fold crossvalidation accuracy for Basque using the whole variable set

	Female				*Male*					*Total*
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	34.00	42.91	33.91	36.94	56.18	41.00	36.91	36.82	42.73	40.25
ID3	49.45	45.91	46.78	**47.38**	54.27	44.00	51.45	49.45	**49.79**	**48.75**
C4.5	42.73	40.09	42.73	41.85	60.36	39.55	48.45	37.82	46.55	44.54
NB	39.82	31.00	46.45	39.09	60.36	29.91	36.91	41.44	42.16	40.84

Table 2. 10-fold crossvalidation accuracy for Spanish using the whole variable set

	Female						*Male*						*Total*
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	36.46	41.92	41.92	43.64	33.64	39.52	30.00	36.46	44.55	36.46	30.00	**35.49**	37.51
ID3	38.18	47.27	55.45	43.64	44.55	**45.82**	24.55	40.00	50.00	46.36	34.55	39.09	**42.46**
C4.5	42.73	48.18	50.91	50.91	45.45	47.64	21.82	39.09	46.36	48.18	27.27	36.54	42.00
NB	34.55	34.45	40.91	32.73	31.82	34.89	20.91	39.09	40.00	35.45	21.82	31.45	33.17

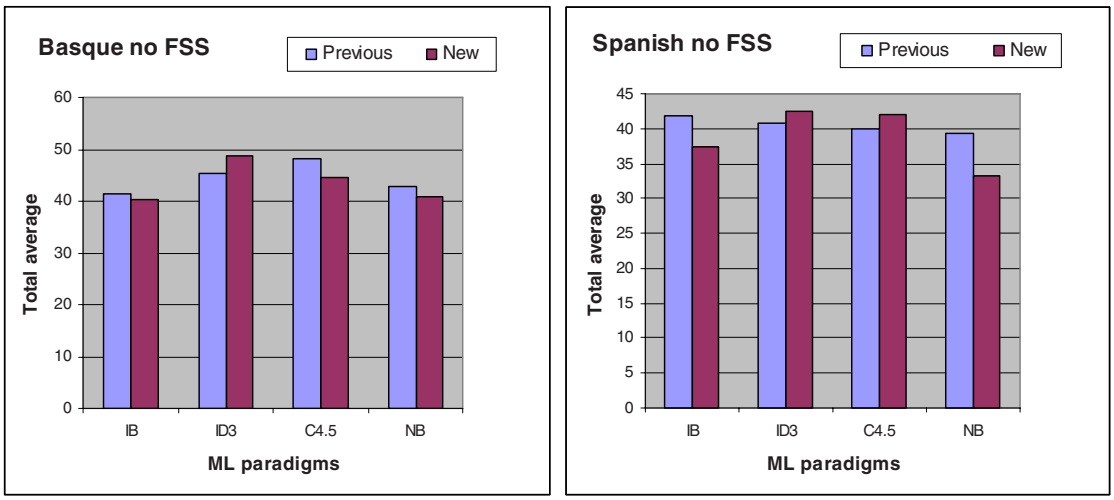


Fig. 1. Comparison between previous and new results in all classifiers using the whole variable set

Results obtained after applying FSS are more appealing, as it can be seen in Tables 3 and 4. In fact, there is a substantial improvement in all cases with regard to previous results. IB classifier appears once again as the best paradigm for all categories, both female and male, and Basque and Spanish languages. Moreover, as it can be seen in Figure 2, accuracies outperform previous ones between 6 and 10%. It must also be highlighted once more that FSS improves the well classified rate for all ML paradigms.

Table 3. 10-fold crossvalidation accuracy for Basque using FSS

	Female				*Male*					*Total*
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	72.55	79.73	62.27	**71.52**	91.36	73.00	77.82	71.82	**78.50**	**75.50**
ID3	71.00	71.73	66.64	69.79	78.73	65.82	72.64	66.91	71.03	70.50
C4.5	67.73	75.91	68.09	70.58	76.73	65.82	69.91	68.91	70.34	70.44
NB	73.00	77.73	63.36	71.36	89.45	67.27	66.18	65.36	72.07	71.76

Table 4. 10-fold crossvalidation accuracy for Spanish using FSS

	Female						*Male*						*Total*
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	72.73	72.73	80.91	76.36	64.55	**73.46**	58.18	72.73	76.36	70.00	62.73	**68.00**	**70.73**
ID3	67.27	75.45	73.64	72.73	68.18	71.45	51.82	63.64	76.36	69.09	59.09	64.00	67.72
C4.5	70.91	75.45	74.55	64.55	66.36	70.36	54.55	63.64	80.91	66.36	56.36	64.36	67.35
NB	75.45	73.64	68.18	67.27	64.55	69.82	50.00	60.00	76.36	68.18	58.18	62.54	66.18

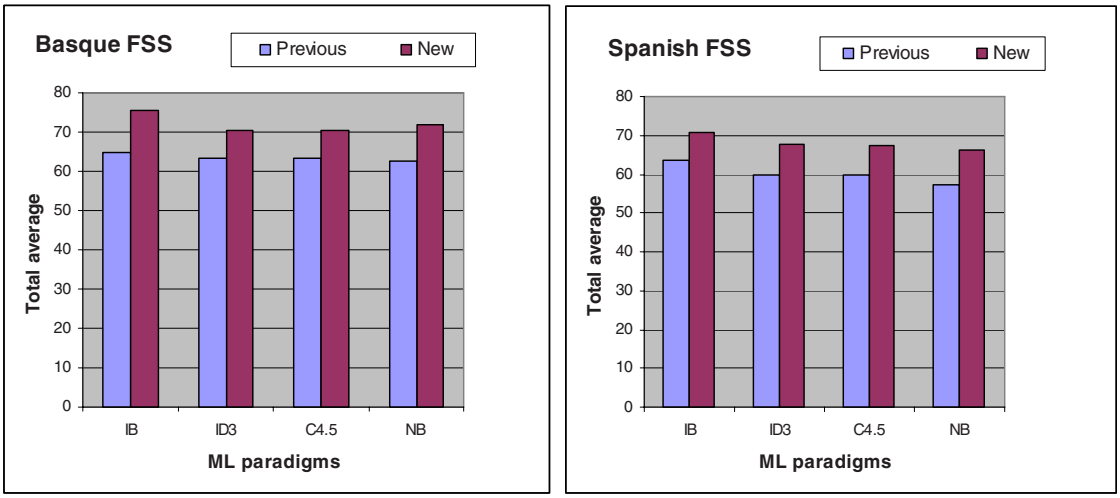


Fig. 2. Improvement with new parameters in Basque and Spanish languages using FSS

4.1 Most Relevant Features

In case of new parameters, the ones that appear most times in best subsets for each actor after using EDA are selected. These parameters consider both languages and are different for men and women.

- Most used parameters for men (Basque and Spanish languages): *regression coefficient for F0 and its mean square error, variance of the pitch values over the voiced regions, mean of the pitch means in every voiced regions, variance of the pitch means in every voiced regions, mean of the flatness of the pitch for every voiced regions, relation between the maximum of the energy in all voiced regions and the maximum of the utterance, maximum of the energy curve, slope coefficient of the regression line for the energy curve and its mean square error, variance of the energy curve, F0 value for the first voiced frame, number of voiced frames in the utterance, maximum, mean and variance of the active level in the speech signal over the voiced regions.*
- Most used parameters for women (Basque and Spanish languages): *mean F0 value calculated over the voiced regions of the utterance, relation between the maximum of the energy in all voiced regions and the maximum of the utterance, tremor or number of zero-crossings over a window of the energy curve derivative, maximum and minimum of the energy curve in the whole utterance, regression coefficient for the energy curve values and its mean square error, mean and variance of the energy values over the whole utterance, ratio of number of voiced frames vs.*

number of all frames, mean of the harmonicity to noise ratio, mean and variance of the active level in speech signal in every voiced regions.

5 Conclusions and Future Work

Affective databases have been very useful for developing affective computing systems, being primarily used for training affective recognition systems. RekEmozio database is being used to training some automatic recognition systems applied to the localization where authors make their research.

Emotion recognition rates have improved using the parameters defined in this paper, but it must also be taken into account that such improvement has been achieved after applying EDA for FSS. In the future, new voice features related to emotions are expected be taken into account, with the aim of improving current results. Another option in order to improve emotion recognition rate would be to merge all parameters in one single classification.

This paper also describes how results obtained by Machine Learning techniques applied to emotion classification can be improved automatically by selecting an appropriate subset of classifying variables by FSS. Classification accuracies, although not very impressive yet, are clearly improved over the results obtained using the full set of variables. Still, an analysis of the features selected by FSS is required as an effort to extract meaningful information from selected set. Merging or combining information from multiple sources by means of a multiclassifier model [9] can help obtaining better classification accuracies.

Acknowledgements

The involved work has received financial support from the Department of Economy of the local government “Gipuzkoako Foru Aldundia” and from the University of the Basque Country.

References

1. Álvarez, A., Cearreta, I., López, J.M., Arruti, A., Lazkano, E., Sierra, B., Garay, N.: Feature Subset Selection based on Evolutionary Algorithms for automatic emotion recognition in spoken Spanish and Standard Basque languages. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 565–572. Springer, Heidelberg (2006)
2. Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C.: Asr for emotional speech: clarifying the issues and enhancing performance. Neural Networks 18, 437–444 (2005)
3. Batliner, A., Fisher, K., Huber, R., Spilker, J., Nöth, E.: Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In: Cowie, R., Douglas-Cowie, E., Schröder, Manuela (Hrsg.) Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research Newcastle, Northern Ireland, pp. 195–200 (September 2000)

4. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural Networks 18, 371–388 (2005)
5. Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. In: Proc. of ICSLP'96 (1996)
6. Ekman, P., Friesen, W.: Pictures of facial affect. Consulting Psychologist Press, Palo Alto, CA (1976)
7. Emotion Recognition in Speech Signal: Retrieved (March 30, 2007), from http://lorien.die.upm.es/partners/sony/main.html
8. Fragopanagos, N.F., Taylor, J.G.: Emotion recognition in human-computer interaction. Neural Networks 18, 389–405 (2005)
9. Gunes, V., Menard, M., Loonis, P., Petit-Renaud, S.: Combination, cooperation and selection of classiers: A state of the art. International Journal of Pattern Recognition 17, 1303–1324 (2003)
10. Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H.: Recognition of emotion in a realistic dialogue scenario. In: Proc. ICSLP'00, pp. 665–668 (2000)
11. Humaine: Retrieved (March 26, 2007) (n.d.), from http://emotion-research.net/
12. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J.M., Bernadas, D., Oliver, J.M., Tena, D., Longhi, L.: Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In: SpeechEmotion'00, pp. 161–166 (2000)
13. Kohavi, R., Sommerfield, D., Dougherty, J.: Data mining using MLC++, a Machine Learning Library in C++. International Journal of Artificial Intelligence Tools 6(4), 537–566 (1997), http://www.sgi.com/Technology/mlc/
14. Laukka, P.: Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, Uppsala 141, 80 (2004) ISBN 91-554-6091-7
15. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Dordrecht (1998)
16. López, J.M., Cearreta, I., Fajardo, I., Garay, N.: Validating a multimodal and multilingual affective database. In: To be published in Proceedings of HCI International, Springer, Heidelberg (2007)
17. López, J.M., Cearreta, I., Garay, N., de Ipiña, K.L., Beristain, A.: Creación de una base de datos emocional bilingüe y multimodal. In: Redondo, M.A., Bravo, C., Ortega, M. (eds) Proceedings of the 7th Spanish Human Computer Interaction Conference, Interaccion'06, Puertollano, pp. 55–66 (2006)
18. Navas, E., Hernáez, I., Castelruiz, A., Luengo, I.: Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 393–400. Springer, Heidelberg (2004)
19. Pelikan, M., Goldberg, D.E., Lobo, F.: A Survey of Optimization by Building and Using Probabilistic Models. Technical Report 99018, IlliGAL (1999)
20. Picard, R.W.: Affective Computing. MIT Press, Cambridge, MA (1997)
21. Tato, R., Santos, R., Kompe, R., Pardo, J.M.: Emotional space improves emotion recognition. In: Hansen, J.H.L., Pellom, B. (eds.) Proceedings of 7th International Conference on Spoken Language Processing (ICSLP'02 – INTERSPEECH'02). Denver, Colorado, USA, pp. 2029–2032 (2002)
22. Tao, J., Tan, T.: Affective computing: A review. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 981–995. Springer, Heidelberg (2005)
23. Taylor, J.G., Scherer, K., Cowie, R.: Neural Networks. special issue on Emotion and Brain 18(4), 313–455 (2005)

7.14 Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech

- **Authors:** Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay
- **Booktitle:** Advances in Nonlinear Speech Processing
- **Year:** 2007
- **Publisher:** Springer

Application of Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in Speech

Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay

Computer Science Faculty (University of the Basque Country)
Manuel Lardizabal 1, E-20018 Donostia (Gipuzkoa), Spain
aitor.alvarez@ehu.es

Abstract. The study of emotions in human-computer interaction is a growing research area. Focusing on automatic emotion recognition, work is being performed in order to achieve good results particularly in speech and facial gesture recognition. In this paper we present a study performed to analyze different machine learning techniques validity in automatic speech emotion recognition area. Using a bilingual affective database, different speech parameters have been calculated for each audio recording. Then, several machine learning techniques have been applied to evaluate their usefulness in speech emotion recognition, including techniques based on evolutive algorithms (EDA) to select speech feature subsets that optimize automatic emotion recognition success rate. Achieved experimental results show a representative increase in the success rate.

Keywords: Affective computing, Machine Learning, speech features extraction, emotion recognition in speech.

1 Introduction

Human beings are eminently emotional, as their social interaction is based on the ability to communicate their emotions and perceive the emotional states of others [1]. Affective computing, a discipline that develops devices for detecting and responding to users' emotions [2], is a growing research area [3]. The main objective of affective computation is to capture and process affective information with the aim of enhancing the communication between the human and the computer. Within the scope of affective computing, the development of affective applications is a challenge that involves analyzing different multimodal data sources. In order to develop such applications, a large amount of data is needed in order to include a wide range of emotionally significant material. Affective databases are a good chance for developing affective recognizers or affective synthesizers. In this paper different speech paralinguistic parameters have been calculated for the analysis of the human emotional voice, using several audio recordings. These recordings are stored in a bilingual and multimodal affective database. Several works have already been done in which the use of Machine Learning paradigms takes a principal role.

M. Chetouani et al. (Eds.): NOLISP 2007, LNAI 4885, pp. 273–281, 2007.

2 Related Work

As previously mentioned affective databases provide a good opportunity for training affective applications. This type of databases usually record information such as images, sounds, psychophysiological values, etc. There are some references in the literature that present affective databases and their characteristics [4],[5],[6]. Many studies have been focused on the different features used in human emotional speech analysis [7],[8]. The number of voice features analysed varies among the studies, but basically most of these are based in fundamental frequency, energy and timing parameters, such as speech rate or mean phone duration. Works where the use of Machine Learning paradigms take a principal role can also be found in the literature [9],[10]. The work by [4] is related with this paper in the sense of using a Feature Selection method in order to apply a Neural Network to emotion recognition in speech, although both, the methods to perform the FSS and the paradigms used, are different. In this line it has to be pointed out the work by [11] which uses a reduced number of emotions and a greedy approach to select the features.

3 Study of Automatic Emotion Recognition Relevant Parameters Using Machine Learning Paradigms

3.1 RekEmozio Database

The RekEmozio bilingual database was created with the aim of serving as an information repository for performing research on user emotion, adding descriptive information about the performed recordings, so that processes such as extracting speech parameters and video features could be carried out on them. Members of different work groups involved in research projects related to RekEmozio have performed several processes for extracting speech and video features; this information was subsequently added to the database. The emotions used were chosen based on [12], and the neutral emotion was added. The characteristics of the RekEmozio database are described in [13]. The languages that are considered in RekEmozio database are Spanish and Basque.

3.2 Emotional Feature Extraction

For emotion recognition in speech, one of the most important questions is which features should be extracted from the voice signal. Previous studies show that it is difficult to find specific voice features that could be used as reliable indicators of the emotion present in the speech [14]. In this work, RekEmozio database audio recordings (stereo wave files, sampled at 44100 Hz) have been processed using standard signal processing techniques (windowing, Fast Fourier Transform, auto-correlation...) to extract a wide group of 32 features which are described below. Supposing that each recording in the database corresponds to one single emotion, only one global vector of features has been obtained for each recording

by using some statistical operations. Parameters used are calculated over entire recordings. Selected features are detailed next (in italics):

- **Fundamental Frequency (F0):** The most common feature analyzed in several studies [7],[8]. For F0 estimation we used Sun algorithm [15] and statistics are computed: *Maximum, Minimum, Mean, Range, Variance, Standard deviation and Maximum positive slope in F0 contour.*
- **RMS Energy:** The mean energy of speech quantified by calculating root mean square (RMS) value and 6 statistics: *Maximum, Minimum, Mean, Range, Variance and Standard Deviation.*
- **Loudness:** *Absolute loudness* based on Zwicker's model [16].
- **Spectral distribution of energy:** Each emotion requires a different effort in the speech and it is known that the spectral distribution of energy varies with speech effort [7]. We have computed energy in *Low band*, between 0 and 1300 Hz, *Medium band*, between 1300 and 2600 Hz and *High band* from 2600 to 4000 Hz [17].
- **Mean Formants and Bandwidth:** Energy from the sound source (vocal folds) is modified by the resonance characteristics of the vocal tract (formants). Acoustic variations dues to emotion are reflected in formants [18]. The *first three mean Formants*, and their corresponding *mean Bandwidths.*
- **Jitter:** *Perturbation in vibration of vocal chords.* It is estimated based on the model presented by [19].
- **Shimmer:** *Perturbation cycle to cycle of the energy.* Its estimation is based on the previously calculated absolute loudness.
- **Speaking Rate:** Rhythm is known to be an important aspect in recognition of emotion in speech. Progress has been made on a simple aspect of rhythm, the alternation between speech and silence [7]. The speaking rate estimation has been divided in 6 values based on their duration with respect to the whole elocution: *Duration of voice part, Silence part, Maximum voice part, Minimum voice part, Maximum silence part and Minimum silence part.*

3.3 Machine Learning Standard Paradigms Used

In the supervised learning task, a classification problem has been defined where the main goal is to construct a model or a classifier able to manage the classification itself with acceptable accuracy. With this aim, some variables are to be used in order to identify different elements, the so called predictor variables. For the current problem, each sample is composed by the set of 32 speech related values, while the label value is one of the seven emotions identified. The single paradigms used in our experiments that come from the family of Machine Learning (ML) are briefly introduced:

- **Decision trees:** A decision tree consists of nodes and branches to partition a set of samples into a set of covering decision rules. In each node, a single test or decision is made to obtain a partition. The starting node is usually referred as the root node. In each node, the goal is to select an attribute that

makes the best partition between the classes of the samples in the training set [20],[21]. In our experiments, two well-known decision tree induction algorithms are used, ID3 [22] and C4.5 [23].

- **Instance-Based Learning:** Instance-Based Learning (IBL) has its root in the study of nearest neighbor algorithm [24] in the field of machine learning. The simplest form of nearest neighbor (NN) or k-nearest neighbor (k-NN) algorithms simply stores the training instances and classifies a new instance by predicting the same class its nearest stored instance has or the majority class of its k nearest stored instances have, respectively, according to some distance measure as described in [25]. The core of this non-parametric paradigm is the form of the similarity function that computes the distances from the new instance to the training instances, to find the nearest or k-nearest training instances to the new case. In our experiments the IB paradigm is used, an inducer developed in the *MLC++* project [26] and based on the works of [27] and [28].
- **Naive Bayes classifiers:** The Naive-Bayes (NB) rule [29] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes X = (X_1,X_2, ... ,X_d), the NB classifier applies the following rule:

$$c_{NB} = arg\ max\ p(c_j) \prod_{i=1}^{d} p(x_i|c_j) \tag{1}$$

  where c_{NB} denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in C = c_1, ... , c_l. A normal distribution is assumed to estimate the class conditional densities for predictive genes. Despite its simplicity, the NB rule obtains better results than more complex algorithms in many domains.
- **Naive Bayesian Tree learner:** The naive Bayesian tree learner, NBTree [30], combines naive Bayesian classification and decision tree learning. It uses a tree structure to split the instance space into sub-spaces defined by the paths of the tree, and generates one naive Bayesian classifier in each sub-space.

Feature Subset Selection by Estimation of Distribution Algorithms. The basic problem of ML is concerned with the induction of a model that classifies a given object into one of several known classes. In order to induce the classification model, each object is described by a pattern of d features. Here, the ML community has formulated the following question: are all of these d descriptive features useful for learning the 'classification rule'? On trying to respond to this question, we come up with the Feature Subset Selection (FSS) [31] approach which can be reformulated as follows: given a set of candidate features, select the 'best' subset in a classification problem. In our case, the 'best' subset will be the one with the best predictive accuracy. Most of the supervised learning algorithms perform rather poorly when faced with many irrelevant or redundant (depending on the specific characteristics of the classifier) features. In this way, the FSS proposes additional methods to reduce the number of features so as to improve the performance of the supervised classification algorithm. FSS

can be viewed as a search problem [32], with each state in the search space specifying a subset of the possible features of the task. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice due to the large amount of computational effort required. In this way, any feature selection method must determine the nature of the search process. In the experiments performed, an Estimation of Distribution Algorithm (EDA) has been used which has the model accuracy as fitness function. To assess the goodness of each proposed gene subset for a specific classifier, a wrapper approach is applied. In the same way as supervised classifiers when no gene selection is applied, this wrapper approach estimates, by using the 10-fold crossvalidation [33] procedure, the goodness of the classifier using only the variable subset found by the search algorithm.

Other Feature Subset Selection Approaches. Several approaches have been developed to search a good attribute selection. On Filter Based approach the attribute selection takes a statistical measure of each variable and sorts them according to the obtained value. Among them, Principal Component analysis takes a measure of the correlation and covariance among the predictor variables and the class. Transformation based Feature Selection is a second approach which transforms the representation space to obtain a reduced one. For instance, Singular Value Decomposition transforms the classification problem in another one after projecting the variables through a Matrix, while the number of singular selected values determines the dimension (size) of the vectors.

Another Feature Subset Selection technique, the one used in this paper, is the so called Wrapper approach in which a subset of variables is selected based on the accuracy of the classifier itself. Two simple ways to perform this selection are the following:

- **Forward:** starts with an empty set of variables and adds to the set, at each step, the variable which most increases the obtained accuracy, until no increase is obtained.
- **Backward:** starts with all the variables and deselects variables while no decrease is produced.

These approaches are outperformed by a more powerfull method based on Evolutionary Algorithms; such as EDA. As it can be seen, the rest of the approaches are local guided, while the used one is a more sophisticated search engine, which is supposed to outperform the rest of the approaches. In fact, this happens with any greedy search when compared to an evolutionary one when applied to the same search problem.

4 Experimental Results

The above mentioned methods have been applied over the crossvalidated data sets using the *MLC++* library [26]. Each dataset corresponds to a single actor. Experiments were carried out with and without FSS in order to extract the accuracy improvement introduced by the feature selection process. Tables 1 and 2

show the classification results obtained using the whole set of variables, for Basque and Spanish languages respectively. Each column represents a female (Fi) of male (Mi) actor, and mean values corresponding to each classifier/gender are also included. Last column presents the total average for each classifier.

Results don't seem very impressive. ID3 best classifies the emotions for female actresses, for both Basque and Spanish languages, while C4.5 outstands for Basque male actors and IB for Spanish male actors. Results obtained after applying FSS are more appealing, as can be seen in Tables 3 and 4. There, classifier IB appears as the best paradigm for all the categories, female and male, and Basque and Spanish languages. Moreover, the accuracies outperform the previous ones in more than 15%. It must also be highlighted that FSS improves the well classified rate for all the ML paradigms, as it can be seen in Figure 1.

Table 1. 10-fold crossvalidation accuracy for Basque language using the whole variable set

	Female				***Male***					***Total***
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	35.4	48.8	35.2	39.8	44.2	49.3	36.9	40.9	42.8	41.5
ID3	38.7	45.5	44.7	**43.0**	46.7	46.9	43.3	51.1	47.0	45.3
C4.5	41.5	52.2	35.0	42.9	60.4	53.3	45.1	49.5	**52.0**	**48.1**
NB	42.9	45.8	37.7	42.1	52.2	44.1	36.2	41.4	43.5	42.9
NBT	42.3	39.8	35.2	39.1	53.1	46.2	45.2	43.3	46.9	43.6

Table 2. 10-fold crossvalidation accuracy for Spanish language using the whole variable set

	Female						***Male***						***Total***
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	34.6	43.6	54.6	54.6	38.2	45.1	25.5	33.6	51.8	47.7	33.6	**38.4**	**41.8**
ID3	36.4	52.7	49.1	47.3	42.7	**45.6**	20.9	30.9	40.9	47.3	40.0	36.0	40.8
C4.5	30.9	50.0	46.4	43.6	42.7	42.7	29.1	31.8	46.4	42.7	35.5	37.1	39.9
NB	38.2	42.7	49.1	40.0	42.7	42.5	24.6	30.9	49.1	45.5	34.6	36.9	39.7
NBT	42.7	43.6	49.1	50.0	39.1	44.9	18.2	27.3	40.9	48.2	42.7	35.5	40.2

Table 3. 10-fold crossvalidation accuracy for Basque language using FSS

	Female				***Male***					***Total***
	F1	F2	F3	mean	M1	M2	M3	M4	mean	
IB	63.0	68.0	59.3	**63.5**	72.7	67.4	61.0	62.8	**65.9**	**64.9**
ID3	62.7	60.5	65.5	62.9	72.7	62.0	56.5	62.7	63.4	63.2
C4.5	60.2	66.0	60.0	62.1	71.8	62.8	60.1	63.6	64.6	63.5
NB	64.5	64.6	48.9	59.3	74.6	62.5	62.7	60.0	64.9	62.5
NBT	58.6	61.1	54.8	58.1	74.4	59.9	62.7	59.4	64.1	61.6

Table 4. 10-fold crossvalidation accuracy for Spanish language using FSS

	Female						*Male*						*Total*
	F1	F2	F3	F4	F5	mean	M1	M2	M3	M4	M5	mean	
IB	61.8	66.4	75.5	71.8	68.2	**68.7**	42.7	57.3	69.1	63.6	60.9	**58.7**	**63.7**
ID3	59.1	66.4	66.4	60.0	61.8	62.7	42.7	51.8	66.4	61.8	60.0	56.5	59.6
C4.5	57.3	62.7	64.6	65.5	63.6	62.7	43.6	56.4	65.5	64.6	56.4	57.3	60.0
NB	54.6	59.1	68.2	65.5	60.0	61.5	40.9	48.2	64.6	59.1	51.8	52.9	57.2
NBT	53.6	66.4	63.6	58.2	60.0	60.4	38.2	47.3	60.0	63.6	59.1	53.6	57.0

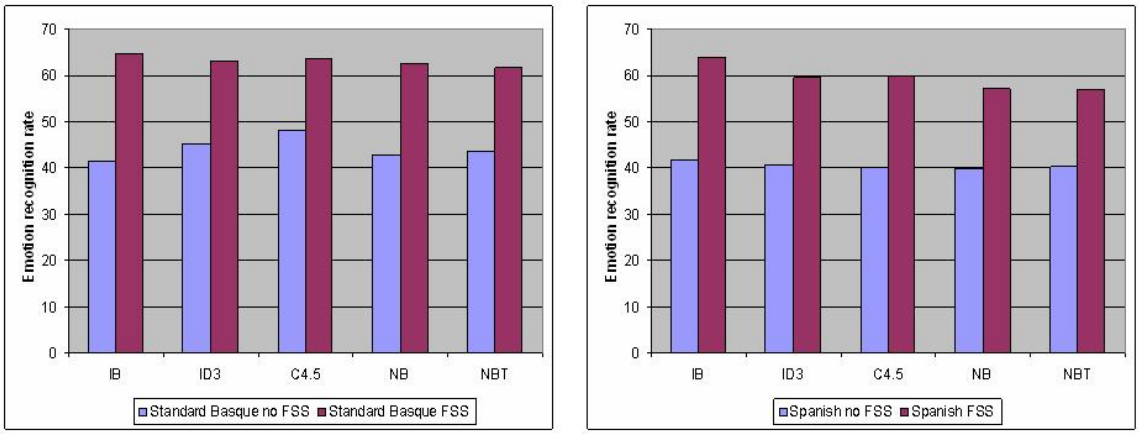


Fig. 1. Improvement in Basque and Spanish languages using FSS in all classifiers

5 Conclusions and Future Work

RekEmozio database has been used to training some automatic recognition systems. In this paper we have shown that applying FSS enhances classification rates for the ML paradigms that we have used (IB, ID3, C4.5, NB and NBTree). An analysis of the selected features by FSS is required. Moreover, the speech data should be combined with visual information. This combination could be performed by means of a multiclassifier model [34].

References

1. Casacuberta, D.: La mente humana: Diez Enigmas y 100 preguntas, Océano, Barcelona, Spain (2001)
2. Picard, R.W.: Affective Computing. The MIT Press, Cambridge, Massachusetts (1997)
3. Tao, J., Tan, T.: Affective computing: A review. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 981–995. Springer, Heidelberg (2005)
4. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Network 18(4), 371–388 (2005)
5. Humaine, Retrieved (January 10, 2007), `http://emotion-research.net/wiki/databases`

6. López, J.M., Cearreta, I., Fajardo, I., Garay, N.: Validating a multilingual and multimodal affective database. In: Proc. HCII, Beijing, China. LNCS, vol. 4560, pp. 422–431. Springer, Heidelberg (2007)
7. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S.D., Fellenz, W.A., Taylor, J.G.: Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE 18(1), 32–80 (2001)
8. Schröder, M.: Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, Institute of Phonetics, Saarland University (2004)
9. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotions in speech. In: Proc. ICSLP 1996, Philadelphia, PA, vol. 3, pp. 1970–1973 (1996)
10. Taylor, J.G., Scherer, K.R., Cowie, R.: Neural network. Special issue: Emotion and brain 18(4), 313–455 (2005)
11. Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H.: Recognition of emotion in a realistic dialogue scenario. In: Proc. Int. Conf. on Spoken Language Processing, Beijing, China, vol. 1, pp. 665–668 (October 2000)
12. Ekman, P., Friesen, W.V.: Pictures of facial affect. Consulting Psychologist Press, Palo Alto, CA (1976)
13. López, J.M., Cearreta, I., Garay, N., López de Ipiña, K., Beristain, A.: Creación de una base de datos emocional bilingüe y multimodal. In: Redondo, M.A., Bravo, C., Ortega, M. (eds.) Proceeding of the 7th Spanish Human Computer Interaction Conference, Interacción 2006, Puertollano, pp. 55–66 (2006)
14. Laukka, P.: Vocal Expression of Emotion: Discrete-emotions and Dimensional Accounts. PhD thesis, Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences (2004)
15. Sun, X.: Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, Florida (2002)
16. Fernandez, R.: A Computational Model for the Automatic Recognition of Affect in Speech. PhD thesis, Massachusetts Institute of Technology (2004)
17. Kazemzadeh, A., Lee, S., Narayanan, S.: Acoustic correlates of user response to errors in human-computer dialogues. In: Proc. IEEE ASRU, St. Thomas, U.S. Virgin Islands (December 2003)
18. Bachorowski, J.-A., Owren, M.J.: Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. Psychological Science 6(4), 219–224 (1995)
19. Rothkrantz, L.J.M., Wiggers, P., van Wees, J.W.A., van Vark, R.J.: Voice stress analysis. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 449–456. Springer, Heidelberg (2004)
20. Martin, K.: An exact probability metric for decision tree splitting and stopping. Mach. Learn. 28(2-3), 257–291 (1997)
21. Mingers, J.: A comparison of methods of pruning induced rule trees, Technical Report, Coventry, England: University of Warwick, School of Industrial and Business Studies (1988)
22. Quinlan, J.R.: Induction of decision trees. Mach. Learn. 1(1), 81–106 (2003)
23. Quinlan, R.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
24. Dasarathy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Recognition Classification Techniques. IEEE Computer Society Press, Los Alamitos (1991)

25. Ting, K.M.: Common issues in Instance-Based and Naive-Bayesian classifiers. PhD thesis, Baser Department of Computer Science, The University of Sidney, Australia (1995)
26. Kohavi, R., Sommerfield, D., Dougherty, J.: Data mining using MLC++: A machine learning library in C++. In: Tools with Artificial Intelligence, pp. 234–245. IEEE Computer Society Press, Los Alamitos (1996)
27. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6(1), 37–66 (1991)
28. Wettschereck, D.: A study of distance-based machine learning algorithms. PhD thesis, Adviser-Thomas G. Dietterich (1994)
29. Minsky, M.: Steps towards artificial intelligence. In: Feigenbaum, E.A., Feldman, J. (eds.) Computers and Thought, pp. 406–450. McGraw-Hill, New York (1963)
30. Kohavi, R.: Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 202–207 (1996)
31. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Dordrecht (1998)
32. Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B.: Feature subset selection by bayesian network-based optimization. Artificial Intelligence 123(1-2), 157–184 (2000)
33. Stone, M.: Cross-validatory choice and assessment of statistical procedures. Journal of the Royal Statistical Society 36, 111–157 (1974)
34. Gunes, V., Menard, M., Loonis, P., Petit-Renaud, S.: Combination, cooperation and selection of classifiers: A state of the art. International Journal of Pattern Recognition 17, 1303–1324 (2003)

7.15 Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in Spoken Spanish and Standard Basque Language

- **Authors:** Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay
- **Booktitle:** Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD)
- **Year:** 2006
- **Publisher:** Springer

Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in Spoken Spanish and Standard Basque Language

Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay

Dept. of Computer Science and Artificial Intelligence,
Computer Science Faculty (University of the Basque Country)
Manuel Lardizabal 1, E-20018 Donostia (Gipuzkoa), Spain
e-mail: aalvarez031@ikasle.ehu.es

Abstract. The study of emotions in human-computer interaction is a growing research area. Focusing on automatic emotion recognition, work is being performed in order to achieve good results particularly in speech and facial gesture recognition. In this paper we present a study performed to analyze different Machine Learning techniques validity in automatic speech emotion recognition area. Using a bilingual affective database, different speech parameters have been calculated for each audio recording. Then, several Machine Learning techniques have been applied to evaluate their usefulness in speech emotion recognition. In this particular case, techniques based on evolutive algorithms (EDA) have been used to select speech feature subsets that optimize automatic emotion recognition success rate. Achieved experimental results show a representative increase in the abovementioned success rate.

1 Introduction

Human beings are eminently emotional, as their social interaction is based on the ability to communicate their emotions and perceive the emotional states of others [3]. Affective computing, a discipline that develops devices for detecting and responding to users emotions [27] is a growing research area [35]. The main objective of affective computation is to capture and process affective information with the aim of enhancing the communication between the human and the computer. Within the scope of affective computing, the development of affective applications is a challenge that involves analyzing different multimodal data sources. In order to develop such applications, a large amount of data is needed in order to include a wide range of emotionally significant material. Affective databases are a good chance for developing such applications, either for affective recognizers or either for affective synthesis. In this paper different speech paralinguistic parameters have been calculated for the analysis of the human emotional voice, using several audio recordings. This recordings are stored in a bilingual and multimodal affective database. Several works have already been done in which the use of Machine Learning paradigms take a principal role.

2 Related Work

As previously mentioned, affective databases provide a good opportunity for training affective applications. This type of databases usually record information such as images,

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 565–572, 2006.

sounds, psychophysiological values, etc. There are some references in the literature that present affective databases and their characteristics. [4] carried out a wide review of affective databases. Other interesting reviews are the ones provided in [12] and [19]. Most references found in literature are related to English, while other languages have less resources developed, especially the ones with relatively low number of speakers. This is the case of Standard Basque. To our knowledge, the first affective database in Standard Basque is the one presented by [25]. Concerning to Spanish, the work of [30] stands out. Several studies have been realized about the different features used in human emotional speech analysis [5,32]. However, the studies related to languages such as Standard Basque and Spanish are not numerous. In one hand, the works by [14] and [24] can be emphasized on Spanish language; on the other hand, the study presented by [25] related to Standard Basque is remarkable. The number of voice features analysed varies among the studies, but basically most of these are based in fundamental frequency, energy and timing parameters, like speech rate or mean phone duration. The use of Maching Learning paradigms takes a principal role in some works that can be found in literature. [7] presented a good reference paper. The Neural Networks Journal recently devoted special issue to emotion treatment from a Neural Networks perspective [36]. The work by [4] is related with this paper in the sense of using a Feature Selection method in order to apply a Neural Network to emotion recognition in spoken English, although both the methods to perform the FSS and the paradigms are different. In this line it has to be pointed out the work by [11] which uses a reduced number of emotions and a greedy approach to select the features.

3 Study of Automatic Emotion Recognition Relevant Parameters Using Machine Learning Paradigms

3.1 RekEmozio Database

The RekEmozio database was created with the aim of serving as an information repository for performing research on user emotion. The aim when building the RekEmozio resource was to add descriptive information about the performed recordings, so that processes such as extracting speech parameters and video features could be carried out on them later. Members of different work groups involved in research projects related to RekEmozio performed several processes for extracting speech and video features; this information was subsequently added to the database. The emotions used were chosen based on Ekman's six basic emotions [8], and neutral emotion was added. The characteristics of the RekEmozio database are described in [20]. A normative study of affective values in the RekEmozio database has been performed with the aim of finding out which recordings obtained better emotion recognition rates with experimental subjects [19]. Validation of the RekEmozio database attempts to extract recordings with relevant affective information in order to assist in the development of affective applications applied to local culture and languages. In this particular case, only recordings in which emotion detection success rate was over 50% were used in later work.

3.2 Emotional Feature Extraction

For recognition of emotions in speech, the most important question is which features should be extracted from the voice signal. Previous studies show us that it is difficult

to find specific voice features that could be used as reliable indicators of the emotion present in the speech [17]. In this work, RekEmozio database audio recordings (stereo wave files, sampled at 44100 Hz) have been processed using standard signal processing techniques (windowing, Fast Fourier Transform, auto-correlation,...) to extract a wide group of 32 features which are described below. Supposing that each recording in database corresponds to one single emotion, one global vector of features has been obtained for each recording, using some statistical operations. Parameters used are global parameters calculated over entire recordings. Selected features are described next (in italics):

- **Fundamental Frequency F0**: Is the most common feature analyzed in several studies [5,32]. For F0 estimation we have used Sun algorithm [34] and statistics are computed: *maximum, minimum, mean, range, variance, standard deviation* and *maximum positive slope in F0 contour.*
- **RMS Energy**: The mean energy of speech quantified by calculating root mean square value (RMS) and 6 statistics *maximum, minimum, mean, range, variance* and *standard deviation.*
- **Loudness**: *absolute loudness* based on Zwicker's model [9].
- **Spectral distribution of energy**: Each emotion requires a different effort in the speech and it is known that the spectral distribution of energy varies with speech effort [5]. Effortful speech, like anger or surprise tends to contain relatively greater energy in low and mid spectral bands than the speech that does not need as much effort, like sadness or neutral. We have computed energy in *low band*, between 0 and 1300 Hz, *medium band*, between 1300 and 2600 Hz and *high band* from 2600 to 4000 Hz [15].
- **Mean Formants and Bandwidth**: Energy from the sound source (vocal folds) is modified by the resonance characteristics of the vocal tract (formants). Acoustic variations dues to emotion are reflected in formants [2]. We have computed the *first three mean formants*, and their corresponding *mean bandwidths*.
- **Jitter**: Defined as the *perturbation in vocal chords vibration.* Its estimation is based on the model presented by [31].
- **Shimmer**: *Perturbation cycle to cycle of the energy.* We based its estimation on the previously calculated absolute loudness.
- **Speaking Rate**: Progress has been made on a simple aspect of rhythm, the alternation between speech and silence [5]. We divided the speaking rate estimation in 6 values based on their duration with respect to the whole elocution: *duration of voice, silence, maximum voice, minimum voice, maximum silence* and *minimum silence.*

3.3 Machine Learning Standard Paradigms Used

In the supervised learning task, we have defined a classification problem where the main goal is constructing a model or a classifier able to manage the classification itself with acceptable accuracy. With this aim, some variables are to be used in order to identify different elements, the so called predictor variables. In the present problem, each sample is composed by the set of 32 speech related values, while the label value is one of the seven emotions identified. We briefly introduce the single paradigms used in our experiments. These paradigms come from the family of Machine Learning (ML). A state of the art description and deep explanation about FSS methods can be found in [13] and [18].

Decision Trees. A decision tree consists of nodes and branches to partition a set of samples into a set of covering decision rules. In each node, a single test or decision is made to obtain a partition. The starting node is usually referred as the root node. In each node, the goal is selecting an attribute that makes the best partition between the classes of the samples in the training set [21] and [22]. In our experiments, two well-known decision tree induction algorithms are used, ID3 [28] and C4.5 [29].

Instance-Based Learning. Instance-Based Learning (IBL) has its root in the study of nearest neighbor algorithm [6] in the field of Machine Learning. The simplest form of nearest neighbor (NN) or k-nearest neighbor (k-NN) algorithms simply stores the training instances and classifies a new instance by predicting the same class its nearest stored instance has or the majority class of its k nearest stored instances have, respectively, according to some distance measure as described in [37]. The core of this non-parametric paradigm is the form of the similarity function that computes the distances from the new instance to the training instances, to find the nearest or k-nearest training instances to the new case. In our experiments the IB paradigm is used, a inducer developed in the $\mathcal{MLC}^{++}$ project [16] and based on the works of [1] and [38].

Naive Bayes Classifiers. The Naive-Bayes (NB) rule [23] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $X = (X1, X2, \ldots, Xd)$, the NB classifier applies the following rule:

$$C_N - B = \arg\max_{c_j \in C} \; p(c_j) \prod_{i=1}^{d} p(x_i|c_j)$$

where $c_N - B$ denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \cdots, c_l\}$. A normal distribution is assumed to estimate the class conditional densities for predictive genes. Despite its simplicity, the NB rule has obtained better results than more complex algorithms in many domains.

Feature Subset Selection by Estimation of Distribution Algorithms. The basic problem of ML is concerned with the induction of a model that classifies a given object into one of several known classes. In order to induce the classification model, each object is described by a pattern of d features. Here, the ML community has formulated the following question: are all of these d descriptive features useful for learning the "classification rule"? On trying to respond to this question, we come up with the Feature Subset Selection (FSS) [18] approach which can be reformulated as follows: given a set of candidate features, select the "best" subset in a classification problem. In our case, the "best" subset will be the one with the best predictive accuracy. Most of the supervised learning algorithms perform rather poorly when faced with many irrelevant or redundant (depending on the specific characteristics of the classifier) features. In this way, the FSS proposes additional methods to reduce the number of features so as to improve the performance of the supervised classification algorithm. FSS can be viewed as a search problem [13], with each state in the search space specifying a subset of the possible features of the task. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice due to the large amount of computational effort required. In

this way, any feature selection method must determine the nature of the search process. In the experiments performed, an Estimation of Distribution Algorithm (EDA) [26] has been used which has the model accuracy as fitness function. To assess the goodness of each proposed gene subset for a specific classifier, a wrapper approach is applied. In the same way as supervised classifiers when no gene selection is applied, this wrapper approach estimates, by the 10-fold crossvalidation [33] procedure, the goodness of the classifier using only the variable subset found by the search algorithm.

4 Experimental Results

The above mentioned methods have been applied over the crossvalidated data sets using the $\mathcal{MLC}$++library [16]. Each dataset corresponds to a single actor. Experiments were carried out with and without FSS in order to extract the accuracy improvement introduced by the feature selection process. Tables 1 and 2 show the classification results obtained using the whole set of variables, for Standard Basque and Spanish languages respectively. Each column represents a female (Fi) of male (Mi) actor, and mean values corresponding to each classifier/gender is also included. Last column presents the total average for each classifier. Results don't seem very impressing; ID3 best classifies the emotions for female actresses, both Standard Basque and Spanish, while C4.5 outstands for Standard Basque male actors and IB for Spanish male actors.

Table 1. 10-fold crossvalidation accuracy for Standard Basque Language using the whole variable set

	Female				**Male**					**Total**
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	35,38	48,79	35,23	39,80	44,17	49,32	36,89	40,91	42,82	41,52
ID3	38,71	45,45	44,70	**42,95**	46,67	46,97	43,26	51,14	47,01	45,27
C4.5	41,52	52,20	35,00	42,90	60,38	53,26	45,08	49,47	**52,04**	**48,13**
NB	42,95	45,76	37,65	42,12	52,20	44,09	36,21	41,44	43,48	42,90

Table 2. 10-fold crossvalidation accuracy for Spanish Language using the whole variable set

	Female						**Male**						**Total**
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	34,55	43,64	54,55	54,55	38,18	45,09	25,45	33,64	51,82	47,65	33,64	**38,44**	**38,40**
ID3	36,36	52,73	49,09	47,27	42,73	**45,63**	20,91	30,91	40,91	47,27	40,00	36,00	36,81
C4.5	30,91	50,00	46,36	43,64	42,73	42,72	29,09	31,82	46,36	42,73	35,45	37,09	36,36
NB	38,18	42,73	49,09	40,00	42,73	42,54	24,55	30,91	49,09	45,45	34,55	36,91	36,27

Results obtained after applying FSS are more appealing, as can be seen in Tables 3 and 4. There, classifier IB appears as the best paradigm for all the categories, female and male, and Standard Basque and Spanish languages. Moreover, the accuracies outperform the previous ones in more than 15%. It must also be highlighted that FSS improves the well classified rate for all the ML paradigms.

Table 3. 10-fold crossvalidation accuracy for Standard Basque Language using FSS

	Female				**Male**					**Total**
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	63,03	68,03	59,32	**63,46**	72,65	67,35	60,98	62,80	**65,94**	**64,88**
ID3	62,73	60,48	65,45	62,88	72,65	61,97	56,52	62,65	63,44	63,20
C4.5	60,23	65,98	60,00	62,07	71,82	62,8	60,08	63,56	64,56	63,49
NB	64,47	64,55	48,94	59,32	74,55	62,5	62,73	60,00	64,94	62,53

Table 4. 10-fold crossvalidation accuracy for Spanish Language using FSS

	Female						**Male**						**Total**
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	61,82	66,36	75,45	71,82	68,18	**68,72**	42,73	57,27	69,09	63,64	60,91	**58,72**	57,63
ID3	59,09	66,36	66,36	60,00	61,81	62,72	42,73	51,82	66,36	61,82	60,00	56,54	53,63
C4.5	57,27	62,73	64,55	65,45	63,64	62,72	43,64	56,36	65,45	64,55	56,36	57,27	54,36
NB	54,55	59,09	68,18	65,45	60,00	61,45	40,91	48,18	64,55	59,09	51,82	52,91	52,00

5 Conclusions and Future Work

Affective databases have been very useful for developing affective computing systems, being primarily used for training affective recognition systems. RekEmoziodatabase, either validated or not, is being used to training some automatic recognition systems applied to the localization where authors make their research. In the future, new voice features related to emotions will be taken into account, with the aim of to improve the current results. This paper describes how results obtained by Machine Learning techniques applied to emotion classification can be improved automatically selecting the appropriate subset of classifying variables by FSS. The classification accuracies, although not very impressive yet, are clearly improved over the results obtained using the full set of variables. Still, an analysis of the features selected by FSS is required as an effort to extract meaningful information from that set. Merging or combining information from multiple sources by means of a multiclassifier model [10] could help to obtain better classification accuracies.

Acknowledgements

The involved work has received financial support from the Department of Economy of the local government “Gipuzkoako Foru Aldundia” and from the University of the Basque Country (in the University-Industry projects modality).

References

1. Aha, D., Kibler, D. & Albert, M.K. (1991). *Instance-Based learning algorithms*, Machine Learning **6**, 37–66.
2. Bachorowski, J.A., Owren, M. J. (1995) *Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context*, Psychological Science 6 219–224.

3. Casacuberta, D. *La mente humana: Diez Enigmas y 100 preguntas (The human mind: Ten Enigmas and 100 questions)*. Océano (Ed), Barcelona, Spain (2001) ISBN: 84-7556-122-5.
4. Cowie, R., Douglas-Cowie, E., Cox, C. *Beyond emotion archetypes: Databases for emotion modelling using neural networks*. Neural Networks 18 (2005) 371–388.
5. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction (2001).
6. Dasarathy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Recognition Classification Techniques. IEEE Computer Society Press (1991).
7. Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. In Proc. of ICSLP (1996).
8. Ekman, P., Friesen, W.: Pictures of facial affect. Consulting Psychologist Press, Palo Alto, CA (1976).
9. Fernández, R.: A Computational Model for the Automatic Recognition of Affect in Speech. Massachusetts Institute of Technology (2004).
10. Gunes, V., Menard, M., Loonis, P., Petit-Renaud, S.: Combination, cooperation and selection of classiers: A state of the art. International Journal of Pattern Recognition, 17 (2003) 1303–1324.
11. Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H.: Recognition of emotion in a realistic dialogue scenario. In Proc. ICSLP (2000) 665–668.
12. Humaine: Retrieved March 10, 2006, from `http://emotion-research.net/` (n.d.).
13. Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B.: Feature subsetselection by Bayesian network-based optimization. Artificial Intelligence 123 (2000) 157–184.
14. Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J.M., Tena, D., Longhi, L.: Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In: SpeechEmotion (2000) 161–166.
15. Kazemzadeh, A., Lee, S., Narayanan, S.: Acoustic correlates of user response to errors in human-computer dialogues. Proc. IEEE ASRU, (St. Thomas, U.S. Virgin Islands), December (2003).
16. Kohavi, R., Sommerfield,D., Dougherty,J.: Data mining using MLC++, a Machine Learning Library in C++, International Journal of Artificial Intelligence Tools 6 (4) (1997) 537–566 `http://www.sgi.com/Technology/mlc/`.
17. Laukka, P.: Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, 141, 80 pp. Uppsala (2004) ISBN 91-554-6091-7.
18. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers (1998).
19. López, J.M., Cearreta, I., Fajardo, I., Garay, N.: Evaluating the validity of RekEmozio affective multimodal database with experimental subjects. Technical Report EHU-KAT-IK-04-06. Computer Architecture and Technology department, University of the Basque Country (2006).
20. López, J.M., Cearreta, I., Garay, N., López de Ipiña, K., Beristain, A.: RekEmozio project: bilingual and multimodal affective database. Technical Report EHU-KAT-IK-03-06. Computer Architecture and Technology department, University of the Basque Country (2006).
21. Martin, J.K.: An exact probability metric for Decision Tree splitting and stopping, Machine Learning 28(2/3) (1997).
22. Mingers, J.: A comparison of methods of pruning induced Rule Trees, Technical Report. Coventry, England: University of Warwick, School of Indutrial and Business Studies, (1988).
23. Minsky, M.: Steps towards artificial intelligence. Proceedings of the IRE, 49 (1961) 8–30.
24. Montero, J.M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., Pardo, J.M.: Emotional speech synthesis: from speech database to tts. Proceedings of the 5th International Conference of Spoken Language Processing. Sydney, Australia (1998) 923–926.
25. Navas, E., Hernáez, I., Castelruiz, A., Luengo, I.: Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque. Lecture Notes on Artificial Intelligence, Vol 3206. Springer-Verlag, Berlin (2004) 393–400.

26. Pelikan, M., Goldberg, D.E., Lobo, F.: A Survey of Optimization by Building and Using Probabilistic Models. Technical Report 99018, IlliGAL (1999).
27. Picard, R.W.: Affective Computing. MIT Press, Cambridge, MA (1997).
28. Quinlan, J.R.: Induction of Decision Trees, Machine Learning 1 (1986) 81–106.
29. Quinlan, J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann. Publishers, Inc. Los Altos, California (1993).
30. Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J.M., Bernadas, D., Oliver, J.M., Longhi, L.: Modelización acústica de la expresión emocional en el español. Procesamiento del Lenguaje Natural, No. 25, Lérida, España (1999) 159–166. issn: 1135-5948.
31. Rothkrantz, L.J.M., Wiggers, P., van Wees, J.W.A., van Vark, R.J.: Voice stress analysis. Proceedings of Text, Speech and Dialogues 2004 (2004).
32. Schröder, M.: Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. Ph.D. thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University (2004).
33. Stone, M.: Cross-validation choice and assessment of statistical procedures. Journal Royal of Statistical Society 36 (1974) 111–147.
34. Sun, X.: Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio `http://mel.speech.nwu.edu/sunxj/pda.htm` (2002).
35. Tao, J., Tan, T.: Affective computing: A review. In: J. Tao, T. Tan, R. W. Picard (eds.): Lecture Notes in Computer Science, Vol. 3784 – Proceedings of The First International Conference on Affective Computing & Intelligent Interaction (ACII '05). Beijing, China (2005) 981–995.
36. Taylor, J.G., Scherer, K., Cowie, R.: Neural Networks, special issue on Emotion and Brain. Vol. 18, Issue 4 (2005) 313–455.
37. Ting, K.M.: Common issues in Instance-Based and Naive-Bayesian classifiers, Ph.D. Thesis, Basser Department of Computer Science. The Univesity of Sydney, Australia (1995).
38. Wettschereck, D.: A study of distance-based Machine Learning Algorithms, Ph.D. Thesis, Oregon State University (1994).

7.16 Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech

- **Authors:** Aitor Álvarez, Basilio Sierra, Andoni Arruti, Juan-Miguel López-Gil and Nestor Garay-Vitoria
- **Journal:** Sensors
- **Year:** 2016
- **URL:** http://www.mdpi.com/1424-8220/16/1/21
- **DOI:** 10.3390/s16010021

sensors

MDPI

Article

Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech

Aitor Álvarez [1,*], Basilio Sierra [2], Andoni Arruti [2], Juan-Miguel López-Gil [2] and Nestor Garay-Vitoria [2]

Received: 22 September 2015; Accepted: 17 December 2015; Published: 25 December 2015
Academic Editor: Vittorio M. N. Passaro

1 Vicomtech-IK4. Human Speech and Language Technologies Department, Paseo Mikeletegi 57, Parque Científico y Tecnológico de Gipuzkoa, 20009 Donostia-San Sebastián, Spain
2 University of the Basque Country (UPV/EHU), Paseo de Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain; b.sierra@ehu.eus (B.S.); andoni.arruti@ehu.eus (A.A.); juanmiguel.lopez@ehu.eus (J.-M.L.-G.); nestor.garay@ehu.eus (N.G.-V.)
* Correspondence: aalvarez@vicomtech.org; Tel.: +34-943-309-230

Abstract: In this paper, a new supervised classification paradigm, called classifier subset selection for stacked generalization (CSS stacking), is presented to deal with speech emotion recognition. The new approach consists of an improvement of a bi-level multi-classifier system known as stacking generalization by means of an integration of an estimation of distribution algorithm (EDA) in the first layer to select the optimal subset from the standard base classifiers. The good performance of the proposed new paradigm was demonstrated over different configurations and datasets. First, several CSS stacking classifiers were constructed on the RekEmozio dataset, using some specific standard base classifiers and a total of 123 spectral, quality and prosodic features computed using in-house feature extraction algorithms. These initial CSS stacking classifiers were compared to other multi-classifier systems and the employed standard classifiers built on the same set of speech features. Then, new CSS stacking classifiers were built on RekEmozio using a different set of both acoustic parameters (extended version of the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)) and standard classifiers and employing the best meta-classifier of the initial experiments. The performance of these two CSS stacking classifiers was evaluated and compared. Finally, the new paradigm was tested on the well-known Berlin Emotional Speech database. We compared the performance of single, standard stacking and CSS stacking systems using the same parametrization of the second phase. All of the classifications were performed at the categorical level, including the six primary emotions plus the neutral one.

Keywords: affective computing; machine learning; speech emotion recognition

1. Introduction

Affective computing is an emerging area that tries to make human-computer interaction (HCI) more natural to humans. This area covers topics, such as affect or emotion recognition, understanding and synthesis. Computing systems can better adapt to human behavior taking non-verbal information into account. As Mehrabian suggested [1], verbal information comprises around 10% of the information transmitted between humans, while around 90% is non-verbal. This is why the inclusion of emotion-related knowledge in HCI applications improves the interaction by increasing the level of understanding and decreasing the ambiguity of the messages.

The expression of emotions by humans is multimodal [2]. Apart from verbal information (written or spoken text), emotions are expressed through speech [3–5], facial expressions [6], gestures [7] and other nonverbal clues (mainly psycho-physiological). With regard to the speech communication modality, the literature shows that several parameters (e.g., volume, pitch and speed) are appropriate to generate or recognize emotions [8]. This knowledge is important either to emulate diverse moods reflecting the user's affective states or, in the case of a recognizer, to create patterns for classifying the emotions transmitted by the user.

Affective speech analysis refers to the analysis of spoken behavior as a marker of emotion, with a focus on the nonverbal aspects of speech [9]. Speech emotion recognition is particularly useful for applications that require natural human-machine interaction, in which the response to the user may depend on the detected emotion. Furthermore, it has been demonstrated that emotion recognition through speech can also be helpful in a wide range of other several scenarios, such as e-learning, in-card board safety systems, medical diagnostic tools, call centers for frustration detection, robotics, mobile communication or psychotherapy, among others.

Nevertheless, recognizing emotions from a human's voice is a challenging task due to multiple issues. First, it must be considered that emotions' expression is highly speaker, culture and language dependent. In addition, one spoken utterance can include more than one emotion, either as a combination of different underlying emotions in the same portion or as an individual expression of each emotion in different speech segments. Another interesting aspect is that there is no definitive consensus among the research community regarding which are the most useful speech features for emotion recognition. One possible cause may be the high impact of the variability introduced by the different speakers in commonly-used prosodic features. Finally, selecting the set of emotions to classify is an important decision, which can affect the performance of the speech emotion recognizer. Many works on the topic agree that any emotion is a combination of primary emotions. The primary six emotions include anger, disgust, fear, joy, sadness and surprise [10].

In this paper, we present a study on emotion recognition based on two different sets of speech features extracted from emotional audio signals recorded by professional actors. The analysis was performed on two datasets called RekEmozioand the Berlin Emotional Speech (Emo-DB) database. RekEmozio contains bilingual utterances in Basque and Spanish languages [11], whilst Emo-DB [12] includes sentences recorded in German. Both databases were designed to cover the six primary emotions plus the neutral one, and each recording contained one acted emotion. The classification approach was focused on the categorical recognition of the seven emotions included in the open Emo-DB and the RekEmozio dataset, which is currently in the process of becoming publicly available to the community.

The experiments were divided into three main phases. The first phase corresponded to the construction and evaluation of 10 base supervised classifiers, multi-classifier systems (bagging, boosting and standard stacking generalization) and bi-level multi-classifiers based on the classifier subset selection for stacked generalization (CSS stacking) method on the RekEmozio dataset. For this end, local and global speech parameters containing prosodic, quality and spectral information were computed from each recording through in-house feature extraction algorithms. The selected supervised classifiers for this phase were the following: Bayesian Network (BN), C4.5, k-Nearest Neighbors (kNN), KStar, Naive Bayes Tree (NBT), Naive Bayes (NB), One Rule (OneR), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Random Forest (RandomF) and Support Vector Machines (SVM). These classifiers were also used to build the CSS stacking classifiers in this first phase.

The aim of the second phase was to verify the efficiency of the CSS stacking classification paradigm on the RekEmozio dataset using: (1) a well-known set of acoustic parameters (extended version of the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)); and (2) different base classifiers in the first layer. For this purpose, CSS stacking classifiers were built using the best meta-classifier of the first phase. In the second phase, we applied the following base classifiers in the

first layer: MultiLayer Perceptron (MLP), Radial Basis Function network (RBF), Logistic Regression (LR), C4.5, kNN, NB, OneR, RIPPER, RandomF and SVM. Hence, the MLP, RBF and LR classifiers were added with respect to the first phase, and the BN, KStar and NBT were discarded.

The third phase consisted of testing the CSS stacking paradigm over the well-known and open Emo-DB. To this end, the same standard classifiers and acoustic features (eGeMAPS) of the second phase were used to build single, standard stacking and CSS stacking classifiers. We decided to leave out the bagging and boosting classifiers because of their poor performance in the first phase.

The paper presents the results from three phases. Regarding the first phase, the results obtained when applying each classification method to each actor were presented, providing a comparison and discussion between each of the several classification paradigms proposed. Concerning the second phase, the results obtained with the CSS stacking classifier are given for each actor, and a comparison with the CSS stacking classifiers from the first phase is also provided. Finally, in the third phase, only the three classifiers with a better score have been presented for each of the constructed systems (single, standard stacking and CSS stacking). The performances of these classifier systems have been compared to each other and to other results obtained in related works in the literature over the same Emo-DB.

In addition, this paper aims to serve as a forum to announce that the RekEmozio dataset will be publicly available soon for research purposes. The aim is to provide the scientific community a new resource to make experiments in the speech emotion recognition field over audio and video acted recordings, several made by actors, others by amateurs, in the Spanish and Basque languages.

The rest of the paper is structured as follows. Section 2 introduces related work. Section 3 details the RekEmozio and Emo-DB datasets, in addition to the two sets of speech features used in this work. In Section 4, how EDA was applied for the stacking classification method is explained. Section 5 describes how the experiments that have been carried out were performed, specifying which techniques have been used in each step of the process. Section 6 explains the obtained results and provides a discussion. Section 7 concludes the paper and presents future work.

2. Related Work

Many studies in psychology have examined vocal expressions of emotions. Eyben *et al.* [8], Schuller *et al.* [3,9], Scherer [4] and Scherer *et al.* [5] provide reviews of these works. Besides, during recent years, the field of emotional content analysis of speech signals has been gaining growing attention. Scherer [4] described the state of research on emotion effects on voice and speech and discussed issues for future research efforts. The analyses performed by Sundberg *et al.* [13] suggested that the emotional samples could be better described by three physiological mechanisms, namely the parameters that quantified subglottal pressure, glottal adduction and vocal fold length and tension. Ntalampiras and Fakotakis [14] presented a framework for speech emotion recognition based on feature sets from diverse domains, as well as on modeling their evolution in time. Wu *et al.* [15] proposed modulation spectral features (MSFs) for the automatic recognition of human affective information from speech. More recently, [16] proposed a novel feature extraction based on multi-resolution texture image information (MRTII), including a BS-entropy-based acoustic activity detection (AAD) module and using an SVM classifier. They improve the performance of other systems based on Mel-frequency cepstral coefficients (MFCC), prosodic and low-level descriptor (LLD) features for three artificial corpora (Emo-DB, eNTERFACE, KHUSC-EmoDB) and a mixed database. There have been several challenges on emotion and paralinguistics in INTERSPEECH, as shown in [3,9].

An important issue to be considered in the evaluation of an emotional speech recognizer is the quality of the data used to assess its performance. The proper design of emotional speech databases is critical to the classification task. Work in this area has made use of material that was recorded during naturally-occurring emotional states of various sorts, that recorded speech samples of experimentally-induced specific emotional states in groups of speakers and that recorded

professional or lay actors asked to produce vocal expressions of emotion as based on emotion labels and/or typical scenarios [4].

Several reviews on emotional speech databases have been published. Douglas-Cowie *et al.* [17] provided a list of 19 data collections, while El Ayadi *et al.* [18] and Ververidis and Kotropoulos [19] provided a record of an overview of 17 and 64 emotional speech data collections, respectively. Most of these references of affective databases are related to English, while fewer resources have been developed for other languages. This is particularly true to languages with a relatively low number of speakers, such as the Basque language. To the authors' knowledge, the first affective database in Basque is the one presented by Navas *et al.* [20]. Concerning Spanish, the work of Iriondo *et al.* [21] stands out; and relating to Mexican Spanish, the work of Caballero-Morales [22] can be highlighted. On the other hand, the RekEmozio dataset is a multimodal bilingual database for Spanish and Basque [11], which also stores information that came from processes of some global speech feature extractions for each audio recording.

Popular classification models used for emotional speech classification include, among others, different decision trees [23], SVM [8,24–26], neural networks [27] and hidden Markov models (HMM) [28,29]. Which one is the best classifier often depends on the application and corpus [30]. El Ayadi *et al.*[18] and Ververidis and Kotropoulos [19] provide a review of appropriate techniques in order to classify speech into emotional states.

In order to combine the benefits of different classifiers, classifier fusion is starting to become common, and several different examples can be found in the literature [31]. Pfister and Robinson [30] proposed an emotion classification framework that consists of n(n-1)/2 pairwise SVMs for n labels, each with a differing set of features selected by the correlation-based feature selection algorithm. Arruti *et al.* [32] used four machine learning paradigms (IB, ID3, C4.5, NB) and evolutionary algorithms to select feature subsets that noticeably optimize the automatic emotion recognition success rate. Schuller *et al.* [24] combined SVMs, decision trees and Bayesian classifiers to yield higher classification accuracy. Scherer *et al.* [33] combined three different KNN classifiers to improve the results. Chen *et al.* [34] proposed a three-level speech emotion recognition model combining Fisher rate, SVM and artificial NN in comparative experiments. Attabi and Dumouchel [35] proved that, in the context of highly unbalanced data classes, back-end systems, such as SVMs or a multilayer perceptron (MLP), can improve the emotion recognition performance achieved by using generative models, such as Gaussian mixture models (GMMs), as front-end systems, provided that an appropriate sampling or importance weighting technique is applied. Morrison *et al.* [36] explored two classification methods that had not previously been applied in affective recognition in speech: stacked generalization and unweighted vote. They showed how these techniques can yield an improvement over traditional classification methods. Huang *et al.* [37] developed an emotion recognition system for a robot pet using stacked generalization ensemble neural networks as the classifier for determining human affective state in the speech signal. Wu and Liang [38] presented an approach to emotion recognition of affective speech based on multi-classifiers using acoustic-prosodic information (AP) and semantic labels. Three types of models, GMMs, SVMs and MLPs, are adopted as the base-level classifiers. A meta decision tree (MDT) is then employed for classifier fusion to obtain the AP-based emotion recognition confidence. Several methods have been used for decision fusion in speech emotion recognition. Kuang and Li [39] proposed the Dempster–Shafer evidence theory to execute decision fusion among the three kinds of emotion classifiers to improve the accuracy of the speech emotion recognition. Huang *et al.* [40] used FoCalfusion, AdaBoost fusion and simple fusion on their studies of the effects of acoustic features, speaker normalization methods and statistical modeling techniques on speaker state classification.

3. Case Study

In this section, the main characteristics of the RekEmozio and Emo-DB datasets used for the experiments are presented first. In addition, the speech features used to train and test classifiers are described.

3.1. RekEmozio Dataset

The RekEmozio dataset was created with the aim of serving as an information repository to perform research on user emotions. The RekEmozio dataset is based on data acquired through user interaction and metadata used to describe and label each interaction and provides access to the data stored and the faculty of performing transactions over them, so new information can be added to the dataset by analyzing the data included in it. When building the RekEmozio dataset, the aim was adding descriptive information about the performed recordings, so processes, such as extracting speech parameters and video features, may be done currently on them.

The RekEmozio dataset is composed of audio and video acted recordings, several made by professional actors, while others are by amateurs. In this study, we use the audio recordings made by professional actors. Those recordings are either in the Basque or Spanish languages.

The classification of emotions was performed at the categorical level. For this purpose, seven emotions were used: the six basic emotions described by [6], that is sadness, fear, joy, anger, surprise and disgust, and a neutral emotion. The selection of these specific emotions was based on the work by Ekman and Friesen [6], which suggested that these emotions are universal for all cultures. This is interesting considering the bilingualism of the RekEmozio dataset.

There are 88 different sentences with 154 recordings over them for each actor. Seven actors recorded sentences for Basque, while 10 recorded for Spanish. The total length of the audio recordings was $130'41''$ for Basque and $166'17''$ for Spanish.

A validation for normative study was performed by experimental subjects in order to obtain affective values for each recording and to see what the validity of the recorded material and the affective values for each recording are [41]. Achieved results show that the material recorded in the RekEmozio database was correctly identified by 57 experimental subjects, with a mean accuracy of 66.5% for audio recordings. In Table 1, audio recognition accuracy percentages for the different types of utterances (depending on the language) are presented. It has also to be noted that several automatic emotion recognition systems have used the RekEmozio dataset in previous works, such as [32,42].

Table 1. Human recognition accuracy percentages for utterances as a function of language and emotions (taken from [41]).

	Sadness	Fear	Joy	Anger	Surprise	Disgust	Neutral
Spanish	75%	51%	78%	71%	66%	52%	80%
Basque	77%	52%	68%	74%	59%	51%	77%

The RekEmozio dataset is currently in the process of being made publicly available (until the process is completed and as the RekEmozio dataset remains unavailable from a public repository, anyone interested can contact Karmele López de Ipiña or the co-author Nestor Garay-Vitoria with the aim of the community having access to the dataset for research purposes.

A complete description of the RekEmozio dataset characteristics can be seen in [11].

3.2. Emo-DB

The widely extended German Emo-DB [12] is composed of recordings of 10 actors (five female and five male), which simulated the six primary emotions defined by [6] plus the neutral one. The complete database was evaluated through a perception test with 20 subjects, achieving a human performance of 84% accuracy [43]. The Emo-DB is publicy available via the Internet.

3.3. Speech Features

The selection of suitable features to extract from the voice signal is one of the most difficult and important decisions to be made in the speech emotion recognition task. It is even more critical when pattern recognition techniques are involved, since they are highly dependent on the domain and training material. The voice characteristics most commonly employed in the literature involve the computation of prosodic and continuous features, qualitative features, spectral features and Teager energy operator (TEO)-based features. A deep description of these categories is given in the survey on speech emotion recognition presented in [18]. With the aim of creating a common baseline and agreed set of speech features to use by the speech emotion recognition community, a minimalistic set of voice parameters were recently compiled and presented in [8].

The feature extraction method is also a regular topic of discussion within the speech emotion recognition field. Because of the non-stationary nature of speech signals, the features are usually extracted from overlapped small frames, which consist of a few milliseconds portions of signal. The features extracted at the frame level are known as local features. Using these local features and computing statistics among them, global features are also usually calculated at the utterance level. Even if the best results were obtained in many works [44–46] using global features instead of local features, it is not clear whether global features performed better for any emotion classification. In fact, in the work presented in [28], they proved that global features do not perform correctly when recognizing emotions with similar arousal, e.g., happiness and anger.

In this work, two sets of speech features were computed along the three phases. In the first phase, local and global features containing prosodic, spectral and quality information were extracted using in-house algorithms, considering a total set of 123 features for each spoken utterance. The extraction of local features was done at both the frame and region levels. In the first case, a 20-millisecond frame-based analysis window was used, with an overlapping of 10 milliseconds. Concerning the feature extraction at the region level, the work presented in Tato *et al.* [47] was followed. They defined a technique for signal treatment and information extraction from emotional speech, not only extracting information by frames, but also by regions consisting of more than three consecutive speech frames. With regard to global features, statistics containing measures, such as the mean, variance, standard deviation and the maximum and minimum values and their positions, were computed, among others. The full set of the 123 features we used in the first phase of this work, including local characteristics and their correlated global statistics, were described in more detail in [32].

With regard to the second and third phases, the extended version (eGeMAPS) of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) was used to extract a different set of speech features. The complete description of the parameters involved in the eGeMAPS set is given in [8]. The extraction of this set of features was done through the OpenSMILE toolkit presented in [48].

4. Classifier Subset Selection to Improve the Stacked Generalization Method

One of the main goals of this work was the construction of a multi-classifier system with optimal selection of the base classifiers in the speech emotion recognition domain. For this purpose, a method proposed in [49] was applied to select an optimal classifier subset by means of the estimation of distribution algorithms (EDAs).

In order to combine the results of the base classifiers, we employed stacked generalization (SG) as a multi-classifier system. Stacked generalization is a well-known ensemble approach, and it is also called stacking [50,51]. While ensemble strategies, such as bagging or boosting, obtain the final decision after a vote among the predictions of the individual classifiers, SG applies another individual classifier to the predictions in order to detect patterns and improve the performance of the vote.

As can be seen in Figure 1, SG is divided into two levels: for Level 0, each individual classifier makes a prediction independently, and for Level 1, these predictions are treated as the input values of another classifier, known as the meta-classifier, which returns the final decision.

The data for training the meta-classifier is obtained after a validation process, where the outputs of the Level 0 classifiers are taken as attributes, and the class is the real class of the example. This implies that a new dataset is created in which the number of predictor variables corresponds to the number of classifiers of the bottom layer, and all of the variables have the same value range as the class variable.

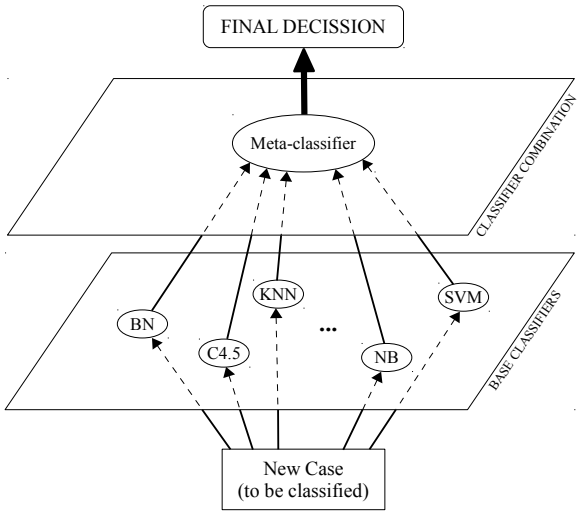


Figure 1. Stacked generalization schemata.

Within this approach, using many classifiers can be very effective, but selecting a subset of them can reduce the computational cost and improve the accuracy, assuming that the selected classifiers are diverse and independent. It is worth mentioning that a set of accurate and diverse classifiers is needed in order to be able to improve the classification results obtained by each of the individual classifiers that are to be combined. This fact has been taken into account to select the classifiers that take part in the first layer of the stacked generalization multi-classifier used.

In [49], an extension of the staking generalization approach is proposed, reducing the number of classifiers to be used in the final model. This new approach is called classifier subset selection (CSS), and a graphical example is illustrated in Figure 2. As can be seen, an intermediate phase is added to the multi-classifier to select a subset of Level 0 classifiers. The classification accuracy is the main criterion to make this selection. As can be seen in Figure 2, discarded classifiers, those with an X, are not used in the multi-classifier.

The method used to select the classifiers could be any, but in this type of scenario, evolutionary approaches are often used. Currently, some of the best known evolutionary algorithms for feature subset selection (FSS) are based on EDAs [52]. EDA combines statistical learning with population-based search in order to automatically identify and exploit certain structural properties of optimization problems. Inza *et al.* [53] proposed an approach that used an EDA called the estimation of Bayesian network algorithm (EBNA) [54] for an FSS problem. Seeing that in [55], EBNA shows better behavior than genetic and sequential search algorithms for FSS problems (and hence, for CSS in this approach), we decided to use EBNA. Moreover, EBNA has been selected as the model in the recent work that analyses the behavior of the EDAs [56].

In our approach, an individual in the EDA algorithm is defined as an n-tuple with $0, 1$ binary values, so-called binary encoding. Each position in the tuple refers to a concrete base classifier, and the value indicates whether this classifier is used (1 value) or not (0 value). An example with 10 classifiers (the value used in this paper) can be seen in Figure 3. In this example, Classifiers 1, 4 and 7 (Cl1, Cl4 and Cl7) are the selected classifiers, and the remaining seven are not used.

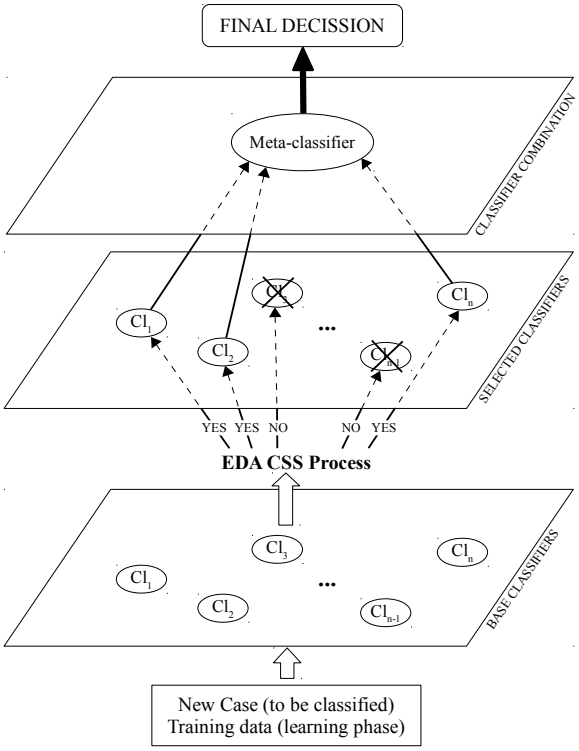


Figure 2. Classifier subset selection stacked generalization.

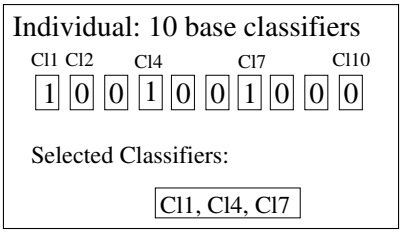


Figure 3. The combinations of base classifiers as the estimation of distribution algorithm (EDA) individuals.

Once an individual has been sampled, it has to be evaluated. The aim is to consider the predictive power of each subset of base classifiers. To this end, a multi-classifier is built for each individual using the corresponding subset of classifiers, and the obtained validated accuracy is used as the fitness function. Thus, when looking for the individual that maximizes the fitness function, the EDA algorithm is also searching the optimal subset of base classifiers.

5. Experiments

In this section, the whole experimental design is described. Firstly, the single classifiers employed in all phases are presented, followed by the definition of the experiment steps. In the end, the experimental setup and the main measure used for the analysis of the obtained results are detailed.

5.1. Base Classifiers

5.1.1. First Phase

The experiments of the first phase were carried out over 10 well-known machine-learning (ML) supervised classification algorithms through the Weka software package [57], which includes a

collection of machine learning algorithms for data mining tasks. A brief description of the classifiers of the first phase is presented below.

- Bayesian Networks (BN): A Bayesian network [58], belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG).
- C4.5: C4.5 [59] represents a classification model by a decision tree. The tree is constructed in a top-down way, dividing the training set and beginning with the selection of the best variable in the root of the tree.
- k-Nearest Neighbors (KNN): This algorithm is a case-based, nearest-neighbor classifier [60]. To classify a new test sample, a simple distance measure is used to find the training instance closest to the given test instance, and then, it predicts the same class as this nearest training instance.
- KStar: This classifier is an instance-based algorithm that uses an entropy-based distance function [61].
- Naive Bayes Tree (NBT): This classification method uses a decision tree with naive Bayes classifiers at the leaves [62].
- Naive Bayes (NB): The naive Bayes rule [63] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $\mathbf{X} = (X_1, X_2, ..., X_d)$, the NB classifier applies the following rule:

$$c_{NB} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{d} p(x_i|c_j)$$

  where c_{NB} denotes the class label predicted by the naive Bayes classifier, and the possible classes of the problem are grouped in $C = \{c_1, \ldots, c_l\}$.
- One Rule (OneR): This simple classification algorithm is a one-level decision tree, which tests just one attribute [64]. The chosen attribute is the one that produces the minimum error.
- Repeated Incremental Pruning to Produce Error Reduction (RIPPER): The rule-based learner presented in [65] forms rules through a process of repeated growing (to fit training data) and pruning (to avoid overfitting). RIPPER handles multiple classes by ordering them from least to most prevalent and then treating each in order as a distinct two-class problem.
- Random Forest (RandomF): This constructs a combination of many unpruned decision trees [66]. The output class is the mode of the classes output by individual trees.
- Support Vector Machines (SVM): These are a set of related supervised learning methods used for classification and regression [67]. Viewing input data as two sets of vectors in a n-dimensional space, an SVM will construct a separating hyperplane in that space, one that maximizes the margin between the two datasets.

5.1.2. Second and Third Phases

For the second and third phases, the BN, K-Star and NBT classifiers of the first phase were discarded, the rest of the base classifiers were kept, and three new classifiers were included for experimentation, including multilayer perceptron, radial basis function networks and logistic regression, as they are described below.

- Multilayer Perceptron (MLP): A multilayer perceptron is a feedforward artificial neural network model to map sets of input data onto a set of appropriate outputs [68]. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a processing element with a nonlinear activation function.

- Radial Basis Function (RBF) network: A radial basis function network is an artificial neural network using radial basis functions as activation functions [69]. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters.
- Logistic Regression: A logistic regression (also known as logit regression or logit model) [70] is considered in statistics a regression model where the dependent variable is categorical.

As it can be seen from the three phases, classifiers with different approaches for learning and widely used in different classification tasks were selected. The goal was to combine them in a multi-classifier to maximize the benefits of each modality by intelligently fusing their information and by overcoming the limitations of each modality alone.

5.2. Experimental Steps

As described above, the experiments were organized in three phases. In the first phase, single classifiers, standard multi-classifier systems and CSS stacking classifiers were built over the RekEmozio dataset and compared. During the second phase, new CSS stacking classifiers were built for each of the 17 actors in the same dataset, using new parametrization and configuration of the base classifiers in the first layer. These CSS stacking systems were compared to the CSS stacking classifiers of the first phase. Finally, new single, standard stacking and CSS stacking classifiers were built on the Emo-DB, employing the same acoustic features and standard classifiers of the second phase.

5.2.1. First Phase

1. Single classifiers: build 10 classifiers on the RekEmozio dataset, applying the 10 base machine learning algorithms of the first phase to the training dataset and get validated classification accuracies.
2. Standard multi-classifiers: build one classifier on the RekEmozio dataset applying bagging, another one applying boosting and ten more applying stacking generalization, one for each base classifier at Level 1, and get validated classification accuracies.
3. Classifier subset selection for stacked generalization: build 10 stacking generalization classifiers on the RekEmozio dataset, one for each base classifier acting as a meta-classifier at Level 1, and select, by means of an evolutionary algorithm, a subset of the ten classifiers to participate in the Level 0 layer.

It is worth mentioning that in all of the experiments, a 10-fold cross-validation technique was used. In the case of the classifier subset selection method, this validation was also employed to select the classifier configuration that performed better on average.

5.2.2. Second Phase

1. Classifier subset selection for stacked generalization: build one stacking generalization classifier on the RekEmozio dataset, using the best meta-classifiers from the first phase, and select, by means of an evolutionary algorithm, a subset of the ten classifiers to participate in the Level 0 layer.

5.2.3. Third Phase

1. Single classifiers: build 10 classifiers on the Emo-DB, applying the 10 base machine learning algorithms of the second phase to the training dataset, and get validated classification accuracies.
2. Standard multi-classifiers: build ten classifiers applying stacking generalization, one for each base classifier at Level 1, and get validated classification accuracies.
3. Classifier subset selection for stacked generalization: build 10 stacking generalization classifiers on the Emo-DB, one for each base classifier acting as the meta-classifier at Level 1, and select, by means of an evolutionary algorithm, a subset of the classifiers to participate in the Level 0 layer.

5.3. Experimental Setup

In all of the experiments, 10-fold cross-validation [71] was applied to get a validated classification accuracy (well-classified rate), and this accuracy has been the criterion to define the fitness of an individual, inside the evolutionary algorithm.

For classifier subset selection, the selected EDA algorithm was EBNA, with Algorithm B [72] for structural learning of the Bayesian network. Population size N was set to 50 individuals, representing 50 combinations of classifiers; the number S of selected individuals at each generation was 20 (40% of the population size); and the maximum number of generations of new individuals was set to 10.

5.4. Obtained Results Analysis

The main measure that has been used in this study to evaluate classification methods was the accuracy. The accuracy reflects how many times the emotions are recognized, comparing this to the metadata stored in the RekEmozio and Emo-DB datasets. Accuracy is expressed as a percentage with respect to the total of the recordings.

6. Results and Discussion

6.1. First Phase

Table 2 presents the results obtained for each of the 17 actors in the first phase when a single classification is applied for categorical emotion recognition, in addition to the mean values and standard deviation (SD) of each classifier in the last two rows. The best accuracies obtained per actor are highlighted in bold. The results suggest that SVM is the classifier that performs better when the single classifier method is applied, as for 13 of the 17 actors, the SVM classifier obtains the best results compared to the rest of the single classifiers, and its mean value is 6.43 percentage points higher than the second mean value. Only BN and RandomF get the better accuracies than SVM in the single classification, in the case of two actors for each one. The best accuracy (73.79%) is achieved for the actor P1. The rest of best accuracies for each actor range from 41.82% (P8) to 68.93% (P6).

Table 2. First phase. Accuracy percentages for each person using single classifiers. Mean and SD rows denote the average and standard deviation for each classifier considering all of the actors. BN, Bayesian network; NBT, naive Bayes tree; OneR, one rule; RIPPER, repeated incremental pruning to produce error reduction; RandomF, random forest.

	BN	C4.5	KNN	KStar	NBT	NB	OneR	RIPPER	RandomF	SVM
P1	69.90%	64.08%	65.05%	54.37%	55.34%	61.17%	53.40%	48.54%	64.08%	**73.79%**
P2	58.25%	45.63%	49.51%	36.89%	44.66%	39.81%	34.95%	46.60%	53.40%	**66.02%**
P3	46.60%	45.63%	49.51%	36.89%	44.66%	39.81%	34.95%	46.60%	34.95%	**53.40%**
P4	**59.22%**	43.69%	39.81%	35.92%	43.69%	34.95%	45.63%	33.01%	54.37%	**59.22%**
P5	52.43%	42.72%	41.75%	36.89%	54.37%	49.51%	42.72%	48.54%	52.43%	**68.93%**
P6	53.40%	48.54%	46.60%	38.83%	56.31%	38.83%	43.69%	46.60%	63.11%	**66.99%**
P7	42.72%	33.98%	32.04%	25.24%	41.75%	37.86%	38.83%	38.83%	**50.49%**	45.63%
P8	18.18%	29.09%	29.09%	20.91%	29.09%	19.09%	12.73%	12.73%	26.36%	**41.82%**
P9	44.55%	43.64%	37.27%	33.64%	40.00%	38.18%	30.00%	36.36%	43.64%	**52.73%**
P10	54.55%	43.64%	50.91%	26.36%	52.73%	55.45%	30.00%	41.82%	58.18%	**64.55%**
P11	56.36%	42.73%	37.27%	28.18%	50.00%	43.64%	38.18%	42.73%	55.45%	54.55%
P12	44.55%	32.73%	37.27%	27.27%	31.82%	27.27%	38.18%	39.09%	37.27%	45.45%
P13	44.55%	40.91%	33.64%	35.45%	44.55%	27.27%	45.45%	42.73%	50.91%	**61.82%**
P14	**64.55%**	51.82%	37.27%	31.82%	54.55%	35.45%	40.00%	45.45%	60.00%	56.36%
P15	51.82%	53.64%	53.64%	37.27%	52.73%	40.91%	40.00%	57.27%	**62.73%**	**62.73%**
P16	58.18%	48.18%	47.27%	36.36%	50.00%	40.91%	43.64%	54.55%	53.64%	**59.09%**
P17	50.91%	46.36%	40.91%	23.64%	45.45%	37.27%	40.00%	40.91%	**53.64%**	50.91%
Mean	51.22%	44.53%	42.87%	33.29%	46.57%	39.26%	38.37%	42.49%	51.45%	**57.88%**
SD	10.98	7.91	8.83	7.57	7.69	9.72	8.52	9.51	10.13	8.69

Table 3. First phase. Accuracy percentages for each person using stacking and bagging and boosting multi-classifiers. Mean and SD rows denote the average and standard deviation for each standard multi-classifier considering all of the actors.

	BN	C4.5	KNN	KStar	NBT	NB	OneR	RIPPER	RandomF	SVM	Bagging	Boosting
P1	64.08%	59.22%	72.82%	57.28%	67.96%	61.17%	43.69%	62.14%	66.99%	**73.79%**	65.05%	34.95%
P2	47.57%	54.37%	42.72%	36.89%	**60.19%**	47.57%	36.89%	49.51%	49.51%	52.43%	57.28%	30.10%
P3	49.51%	44.66%	48.54%	41.75%	46.60%	**58.25%**	45.63%	46.60%	52.43%	49.51%	49.51%	35.92%
P4	49.51%	48.54%	41.75%	33.98%	53.40%	45.63%	39.81%	42.72%	50.49%	44.66%	**55.34%**	33.01%
P5	51.46%	45.63%	48.54%	32.04%	**55.34%**	50.49%	41.75%	52.43%	**55.34%**	**55.34%**	54.37%	32.04%
P6	**59.22%**	56.31%	53.40%	42.72%	**59.22%**	54.37%	55.34%	54.37%	**59.22%**	55.34%	52.43%	32.04%
P7	46.60%	33.98%	41.75%	39.81%	46.60%	44.66%	37.86%	41.75%	46.60%	40.78%	**48.54%**	37.86%
P8	**31.82%**	23.64%	30.91%	24.55%	29.09%	25.45%	19.09%	20.91%	23.64%	24.55%	30.91%	17.27%
P9	45.45%	41.82%	47.27%	36.36%	49.09%	45.45%	41.82%	36.36%	**50.91%**	46.36%	**50.91%**	30.91%
P10	55.45%	55.45%	53.64%	40.91%	56.36%	48.18%	37.27%	52.73%	**60.91%**	55.45%	54.55%	35.45%
P11	49.09%	48.18%	43.64%	39.09%	49.09%	40.91%	39.09%	50.00%	55.45%	52.73%	**58.18%**	35.45%
P12	39.09%	35.45%	36.36%	20.91%	38.18%	48.18%	30.91%	35.45%	35.45%	43.64%	**47.27%**	35.45%
P13	39.09%	41.82%	34.55%	30.91%	42.73%	43.64%	40.91%	**46.36%**	40.91%	40.91%	44.55%	34.55%
P14	50.00%	56.36%	57.27%	40.91%	63.64%	62.73%	37.27%	56.36%	**60.00%**	**60.00%**	44.55%	34.55%
P15	44.55%	57.27%	50.91%	40.91%	57.27%	52.73%	37.27%	50.91%	56.36%	59.09%	**60.91%**	36.36%
P16	55.45%	45.45%	49.09%	40.91%	50.00%	49.09%	40.00%	48.18%	54.55%	48.18%	**56.36%**	32.73%
P17	41.82%	44.55%	38.18%	38.18%	46.36%	44.55%	36.36%	43.64%	43.64%	**50.00%**	43.64%	30.91%
Mean	48.22%	46.63%	46.55%	37.54%	51.24%	48.41%	38.88%	46.50%	50.73%	50.16%	**51.43%**	32.91%
SD	7.44	9.05	9.33	7.53	9.04	8.08	6.82	8.96	9.86	9.86	7.53	4.32

The performance of the standard multi-classifiers systems for all of the actors in the first phase is presented in Table 3, with mean and SD values in the last two rows. In the first 10 columns, the results obtained by the stacked generalization method with the single classifiers as meta classifiers are presented. In addition, the accuracy achieved by the bagging and boosting multi-classifiers are shown in the last two columns. The best results per actor are marked in bold. In contrast to single classifiers, there is no meta classifier that performs much better than the others. This is evident looking at their mean values, with four classifiers in the range from 50.16% to 51.43%, showing low differences between them. For seven actors, the best accuracies are reached using the bagging multi-classifier; RandomF gets the best accuracies for five actors, SVM for four actors, NBT for three actors, BN for two actors, and NB and RIPPER get the best accuracy for one actor each. This happens because in some cases, there are several meta-classifiers that get the best accuracies for a given actor (for example, P5). On the other hand, for 14 actors, the worst results are obtained with the boosting multi-classifier. Compared to the results from the single classifiers in Table 2, only for three of the 17 actors (P3, P11 and P12) are improvements achieved on their best classification results using multi-classifiers. For the rest of the actors, the accuracies are lower when compared to single classifiers.

The results reached by CSS stacking classifiers in the first phase are shown in Table 4, including their mean and SD values. If we focus on the highlighted values, which correspond to the best accuracies for each of the actors, the SVM classifier achieves the best scores, on average, when it is used as a meta classifier (an increase of 2.22 percentage points over the second one) and for 13 actors. The other actors obtained best accuracies with C4.5, NBT, NB and RIPPER meta-classifiers. In general, the best accuracies are improved using the CSS stacking classification method against the standard multi-classifiers. Besides, if we compared the results from CSS stacking with the best accuracies achieved by the single classifiers, 13 of the 17 actors obtain higher classification results. This point is clearly demonstrated in Table 5, where the best accuracies obtained per actor are presented for each of the classification methods, including single classifiers, multi-classifiers (boosting, bagging and stacking) and CSS stacking classifiers. In addition, two columns are presented that show the differences obtained when comparing the best accuracies achieved by multi-classifiers against the single classifiers (Differences_1), and the ones obtained by the CSS stacking classifiers against the best between the single and multi-classifiers (Differences_2).

Table 4. First phase. Accuracy percentages for each person applying CSS stacking with the EDA classification method. Mean and SD rows denote the average and standard deviation for each classifier working as meta classifiers and considering all of the actors.

	BN	C4.5	KNN	KStar	NBT	NB	OneR	RIPPER	RandomF	SVM
P1	71.84%	70.87%	**73.79%**	72.82%	68.93%	70.87%	44.66%	68.93%	71.84%	**73.79%**
P2	66.02%	70.87%	**73.79%**	72.82%	60.19%	70.87%	36.89%	68.93%	71.84%	**73.79%**
P3	57.28%	70.87%	**73.79%**	72.82%	65.05%	70.87%	45.63%	68.93%	71.84%	**73.79%**
P4	55.34%	60.19%	51.46%	53.40%	51.46%	**62.14%**	39.81%	51.46%	59.22%	61.17%
P5	55.34%	63.11%	57.28%	49.51%	51.46%	62.14%	42.72%	58.25%	62.14%	**65.05%**
P6	64.08%	**66.99%**	59.22%	58.25%	50.49%	54.37%	57.28%	60.19%	66.02%	59.22%
P7	51.46%	47.57%	51.46%	41.75%	39.81%	51.46%	37.86%	47.57%	50.49%	**52.43%**
P8	30.91%	34.55%	35.45%	30.00%	**43.64%**	34.55%	26.36%	33.64%	33.64%	32.73%
P9	48.18%	50.91%	48.18%	47.27%	46.36%	45.45%	42.73%	47.27%	50.00%	**54.55%**
P10	58.18%	60.91%	58.18%	56.36%	45.45%	60.91%	37.27%	**61.82%**	60.91%	60.91%
P11	53.64%	53.64%	54.55%	54.55%	46.36%	55.45%	41.82%	50.91%	56.36%	**60.00%**
P12	34.55%	49.09%	45.45%	42.73%	32.73%	49.09%	40.91%	41.82%	49.09%	**51.82%**
P13	47.27%	**59.09%**	51.82%	54.55%	52.73%	52.73%	40.91%	52.73%	50.91%	**59.09%**
P14	50.91%	62.73%	64.55%	59.09%	55.45%	64.55%	38.18%	64.55%	63.64%	**66.36%**
P15	58.18%	62.73%	59.09%	59.09%	50.00%	60.91%	37.27%	59.09%	59.09%	**63.64%**
P16	55.45%	50.91%	59.09%	53.64%	44.55%	60.00%	40.91%	52.73%	56.36%	**60.91%**
P17	48.18%	50.91%	52.73%	45.45%	45.45%	49.09%	37.27%	44.55%	48.18%	**54.55%**
Mean	53.34%	58.00%	57.05%	54.36%	50.01%	57.38%	40.50%	54.90%	57.74%	**60.22%**
SD	9.57	9.34	9.73	10.87	8.40	9.26	5.75	9.63	9.55	9.37

The results from Table 5 show that using multi-classifiers does not outperform the classification accuracies in this classification problem. Nevertheless, when applying the CSS stacking classification method, the improvements are noticeable for many of the actors. As is detailed in the last column Differences_2, 11 actors outperform the best accuracies when compared to the ones obtained with the single and multi-classifiers, giving a mean increase of 1.48 percentage points. The highest improvement is achieved by the actor P3, which increases the accuracy by 15.54 percentage points. The rest of the improvements are in the range from 0.91 to 7.77 points. In addition, two of the actors (P1 and P6) reached the same best accuracy with no significant improvements, and there are four cases (P5, P8, P10 and P13) where the single classifiers reach the best accuracies. A comparison of the best accuracies obtained per actor for each of the classification methods is presented in Figure 4.

Table 5. First phase. Best accuracy per person by using each classification method. Improvements comparing the best accuracy from multi-classifiers (bagging, boosting and stacking) against single classifiers are presented in the Differences_1 column. In addition, the improvements between the CSS stacking with EDA and the best accuracy from both single and standard multi-classifiers are shown in the Differences_2 column. Mean and SD rows denote the average and standard deviation for each classification method and the type of differences considering all of the actors. Differences are expressed in percentage points.

	Single	Bagging	Boosting	Stacking	Differences_1	CSS Stacking	Differences_2
P1	**73.79%**	65.05%	34.95%	**73.79%**	0.00	**73.79%**	0.00
P2	66.02%	57.28%	30.10%	60.19%	−5.83	**73.79%**	+7.77
P3	53.40%	49.51%	35.92%	58.25%	+4.85	**73.79%**	+15.54
P4	59.22%	55.34%	33.01%	53.40%	−3.88	**62.14%**	+2.92
P5	**68.93%**	54.37%	32.04%	55.34%	−13.59	65.05%	−3.88
P6	**66.99%**	52.43%	32.04%	59.22%	−7.77	**66.99%**	0.00
P7	50.49%	48.54%	37.86%	46.60%	−1.95	**52.43%**	+1.94
P8	**41.82%**	30.91%	17.27%	31.82%	−10.00	35.45%	−6.37
P9	52.73%	50.91%	30.91%	50.91%	−1.82	**54.55%**	+1.82
P10	**64.55%**	54.55%	35.45%	60.91%	−3.64	61.82%	−2.73
P11	56.36%	58.18%	35.45%	55.45%	+1.82	**60.00%**	+1.82
P12	45.45%	47.27%	35.45%	48.18%	+2.73	**51.82%**	+3.64
P13	**61.82%**	44.55%	34.55%	46.36%	−15.45	59.09%	−2.73
P14	64.55%	44.55%	34.55%	63.64%	−0.91	**66.36%**	+1.82
P15	62.73%	60.91%	36.36%	59.09%	−1.82	**63.64%**	+0.91
P16	59.09%	56.36%	32.73%	55.45%	−2.73	**60.91%**	+1.82
P17	53.64%	43.64%	30.91%	50.00%	−3.64	**54.55%**	+0.91
Mean	58.92%	51.43%	32.91%	54.62%	−3.74	**60.95%**	+1.48
SD	8.30%	7.75%	4.45%	8.77%	5.28	9.33%	4.70

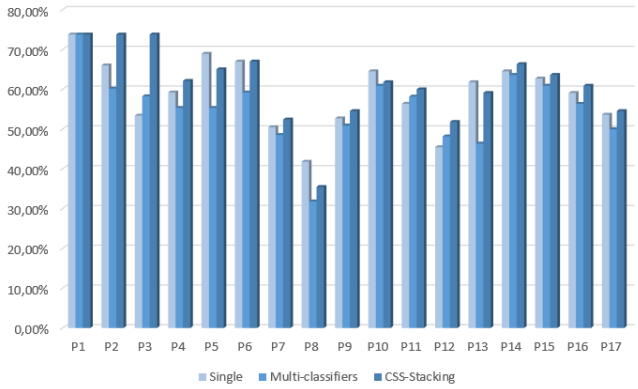


Figure 4. First phase. Best accuracies per person considering single, multi-classifiers and CSS stacking with EDA classification methods.

Finally, we selected one of the classifiers as the meta classifier (SVM) for both stacking and CSS stacking classification methods and presented the results obtained per actor in Table 6 and the mean and SD values at the end. The results prove that using the CSS stacking classification method, the recognition accuracy is outperformed for all of the actors, except for actor P1, in which no improvements are appreciated. The improvements using the CSS stacking classification method range from 3.88 to 24.27 percentage points, with an average improvement of 10.06 points.

Table 6. First phase. Accuracies and improvements per person in percentage points comparing stacking and CSS stacking with EDA classification methods using SVM as the meta classifier. Mean and SD rows denote the average and standard deviation for each classification method and improvements considering all of the actors.

	Stacking	CSS Stacking	Improvements
P1	73.79%	73.79%	0.00
P2	52.43%	73.79%	+21.36
P3	49.51%	73.79%	+24.27
P4	44.66%	61.17%	+16.50
P5	55.34%	65.05%	+9.71
P6	55.34%	59.22%	+3.88
P7	40.78%	52.43%	+11.65
P8	24.55%	32.73%	+8.18
P9	46.36%	54.55%	+8.18
P10	55.45%	60.91%	+5.45
P11	52.73%	60.00%	+7.27
P12	43.64%	51.82%	+8.18
P13	40.91%	59.09%	+18.18
P14	60.00%	66.36%	+6.36
P15	59.09%	63.64%	+4.55
P16	48.18%	60.91%	+12.73
P17	50.00%	54.55%	+4,55
Mean	50.16%	60.22%	+10.06
SD	10.14	9.64	6.42

Statistical Tests

According to [73], we employed the Iman and Davenport test to detect statistical differences among the different classification paradigms. This test rejects the null hypothesis of equivalence between algorithms, since the p-value (0.000216) is lower than the α-value (0.1). Thus, Shaffer *post hoc* test is applied in order to find out which algorithms are distinctive among them. Table 7 shows the statistical differences obtained. As can be seen, the new approach statistically outperforms the results obtained with the standard multi-classifier systems (p-value <0.01). It is worth mentioning that there were no significant differences between CSS stacking and the best single paradigm. This

is indeed due to the selection phase of the best approach among all of the single approaches used, before applying meta-classification, as explained in Section 4 of this paper.

Table 7. First phase. *p*-values of the pair-wise comparison between CSS stacking and the other multi-classifiers.

Hypothesis	Adjusted *p*
CSS Stacking vs. Boosting	**1.2094622076166072E-10**
CSS Stacking vs. Bagging	**2.2567292727265824E-4**
CSS Stacking vs. Stacking	**0.004635715398394891**

If the comparison is done pair-wise, the new approach shows better accuracy than each of the single classifiers used. For instance, comparing the SVM single classifier (the best one) with the new approach obtained using SVM as the meta classifier, the new paradigm outperforms the single one in 11 up to 17 actors.

6.2. Second Phase

Table 8 presents the results obtained by the CSS stacking classification method during the second phase, in which eGeMAPS parameters and a new combination of base classifiers in the first layer were employed for classification. Besides, a comparison with the CSS stacking built in the first phase and the corresponding improvements achieved are also presented. Both CSS stacking classifiers were constructed using the SVM as the meta-classifier, as it was the best meta-classifier in the first phase. As can be seen, the integration in the first layer of new base classifiers that performed well as single classifiers (especially the MLP classifier) and the employment of the eGeMAPS acoustic parameters, which also demonstrated their efficiency when comparing the results of single classifiers in both phases, helped improve the results for most actors. The most appreciable improvements are given by the actors P13, P12 and P8, which outperformed the previous results in the first phase by 20.70, 20.26 and 16.62 percentage points, respectively. In global terms, the average accuracy of the CSS stacking classifiers of the second phase outperformed the mean accuracy of the first phase by 4.56 percentage points, which demonstrated the effectiveness of the eGeMAPS parameters and the new classifiers included in the first layer of the CSS stacking classifiers of the second phase.

Table 8. Second phase. Accuracy percentages per actor for the CSS stacking classifier systems of the second phase (CSS stacking 2nd_Phase) and the comparison with the CSS stacking classifiers of the first phase (CSS stacking 1st_Phase). Mean and SD rows denote the average and standard deviation for each classifier for all of the actors.

	CSS Stacking 2nd_Phase	CSS Stacking 1st_Phase	Differences
P1	85.06%	73.79%	+11.27
P2	69.48%	73.79%	−4.31
P3	75.32%	73.79%	+1.53
P4	70.78%	61.17%	+9.61
P5	77.27%	65.05%	+12.22
P6	64.29%	59.22%	+5.07
P7	47.4%	52.43%	−5.03
P8	49.35%	32.73%	+16.62
P9	46.01%	54.55%	−8.54
P10	73.38%	60.91%	+12.47
P11	66.88%	60.00%	+6.88
P12	72.08%	51.82%	+20.26
P13	79.87%	59.09%	+20.78
P14	61.69%	66.36%	−4.67
P15	59.09%	63.64%	−4.55
P16	46.1%	60.91%	−14.81
P17	57.14%	54.55%	+2.59
Mean	64.78%	60.22%	+4.56
SD	12.34	9.94	10.46

In Appendix A, the confusion matrices scored by the CSS stacking classifiers in the second phase are presented for all of the actors.

6.3. Third Phase

In the third phase, ten classifiers were built for each of the classification systems (single, standard stacking and CSS stacking) employed on the Emo-DB. In Table 9, the results of the three best classifiers of each system are shown. The best result of the three classification systems is highlighted in bold per actor. Interestingly, MLP, RandomF and SVM are the best three classifiers for each of the classification systems.

Looking at the results, only for the A5 and A9 actors, the single classifier (RandomF) system scored the best accuracies; 80.00% and 82.14%, respectively, whilst the standard stacking classifiers achieved the worst results. However, the CSS stacking systems outperformed the results of single and standard stacking classifiers for the rest of the actors. The best result is achieved by the A2 actor, which scored an accuracy of 96.55% when the SVM acted as the meta-classifier. On average, the CSS stacking classifier with the SVM acting as the meta-classifier reached higher results, obtaining a mean of 82.45% accuracy for all of the actors. Considering that the human perception rate for the Emo-DB was set to 84% [43], this mean value of 82.45% can be seen as a promising result. Moreover, this score outperforms the results of other works in the literature over the Emo-DB, like the scores obtained in [43,74], which reached accuracies of 79% and 77%, respectively, although these works analyzed the whole database and used different machine learning algorithms and audio features. The overall results demonstrate the good performance of the CSS stacking classification paradigm and confirms the robustness of this classification system to deal with the emotion recognition in speech over several conditions and datasets.

Table 9. Third phase. Accuracy percentages per actor for the best three classifiers of each system built on the Berlin Emotional Speech database (Emo-DB). Mean and SD rows represent the average and standard deviation considering all of the actors.

	Single			**Standard Stacking**			**CSS Stacking**		
	MLP	**RandomF**	**SVM**	**MLP**	**RandomF**	**SVM**	**MLP**	**RandomF**	**SVM**
A1	79.59%	73.46%	77.55%	63.26%	71.42%	61.22%	79.59%	**81.63%**	79.59%
A2	94.82%	87.93%	86.20%	79.31%	89.65%	72.41%	93.10%	94.83%	**96.55%**
A3	74.41%	62.79%	67.44%	62.79%	67.44%	62.79%	74.42%	74.42%	**76.74%**
A4	84.21%	84.21%	81.57%	68.42%	71.05%	68.42%	**89.47%**	84.21%	86.84%
A5	63.63%	**80.00%**	72.72%	56.36%	65.45%	54.54%	67.27%	72.73%	78.18%
A6	77.14%	74.28%	80.00%	71.42%	68.57%	68.57%	**82.86%**	**82.86%**	**82.86%**
A7	78.68%	75.40%	72.13%	67.21%	70.49%	65.57%	77.05%	**80.33%**	78.69%
A8	78.26%	75.36%	78.26%	73.91%	76.81%	78.26%	82.61%	**86.96%**	85.51%
A9	67.85%	**82.14%**	66.07%	69.64%	71.42%	64.28%	76.79%	75.00%	75.00%
A10	74.64%	83.09%	76.05%	73.23%	71.83%	76.05%	83.10%	80.28%	**84.51%**
Mean	77.32%	77.87%	75.80%	68.55%	72.41%	67.21%	80.63%	81.32%	**82.45%**
SD	8.52	7.17	6.28	6.56	6.76	7.13	7.41	6.57	6.33

In Appendix A, the confusion matrices scored by the CSS stacking system with the SVM classifier acting as the meta-classifier are presented for all of the actors.

7. Conclusions and Future Work

Enabling computers the ability to recognize human emotions is an emergent research area. Continuing the authors' previous work on the topic, in this article, different classification approaches have been presented and compared for the speech emotion recognition task. The experimentation was divided into three main phases, which differ from each other in: (1) the speech parametrization; (2) the base classifiers used to construct the classification systems; and (3) the dataset employed.

The experiments were performed over the RekEmozio and Emo-DB datasets, which contain audio recordings in Basque, Spanish and German from several actors. As the emotional annotation in both datasets was performed using categories, the statistical approach was also turned into a categorical classification problem.

In the first phase, 10 single classifiers, 12 multi-classifiers (bagging, boosting and standard stacking generalization) and 10 final CSS stacking classifiers with the EDA classification method were built, evaluated and compared to each other. For single classifiers, the SVM became the best classifier among the ten algorithms employed, as it obtained the best accuracy for 13 of the 17 actors. If we focus on the performance of multi-classifiers, in most cases, they did not achieve better results compared to single classifiers. In addition, it is noticeable that although bagging was the classifier that reached the best results in most cases, it performed better only for seven of the 17 actors. The best accuracies for multi-classifiers ranged between 31.82% and 73.79%.

In comparison, the CSS stacking multi-classifier with EDA achieved higher accuracies than the single and multi-classifiers in most cases. Table 5 shows that, except for four out of 17 actors, CSS stacking with EDA outperformed the results of all of the other single and multi-classifiers tested in the first phase of this work. Furthermore, these results were statistically significant when comparing pair-wise with the other multi-classifiers. Therefore, it can be concluded from this first phase that multi-classifiers based on the CSS stacking method with EDA are a promising approach for emotion recognition in speech.

With regard to the second phase, a new parametrization based on the eGeMAPS acoustic parameters in addition to new base classifiers was employed to construct new CSS stacking classifiers using the best meta-classifier of the first phase. These new CSS stacking classifiers were compared to the CSS stacking classifiers from the first phase, in order to evaluate the impact of the new parameters and base classifiers included. The results from Table 8 concluded that the new configuration of the CSS stacking classifiers of the second phase outperformed the results obtained in the first phase in most cases. This demonstrated the good performance of the acoustic parameters and the new base classifiers employed in the second phase.

Finally, the third phase was focused on constructing single, standard stacking and CSS stacking classifiers for each of the actors in the well-known and freely-available Emo-DB. The results confirmed the good performance of the CSS stacking classifier system, which improved the accuracies obtained by the other classification systems for all actors, except two.

A future work for this research will be to perform new experiments on different databases, such as the Belfast naturalistic emotion database [10], the Vera am Mittag German audio-visual emotional speech database [75] and the FAUAibo Emotion Corpus [76], which include spontaneous speech, and the Berlin Database of Emotional Speech [12] and EMOVO[77] databases, in order to test out the efficiency of the presented new classification paradigm in other dataset conditions and domains. Besides, new standard classifiers will be explored, and a combination of data from several databases will be used with the aim of building speaker- and language-independent classification systems.

Acknowledgments: This research work was partially funded by the Spanish Ministry of Economy and Competitiveness (Project TIN2014-52665-C2-1-R) and by the Department of Education, Universities and Research of the Basque Government (Grants IT395-10 and IT313-10). Egokituz Laboratory of HCI for Special Needs, Galan research group and Robotika eta Sistema Autonomoen Ikerketa Taldea (RSAIT) are part of the Basque Advanced Informatics Laboratory (BAILab) unit for research and teaching supported by the University of the Basque Country (UFI11/45). The authors would like to thank Karmele López de Ipiña and Innovae Vision S.L. for giving permission to use RekEmozio database for this research.

Author Contributions: The current research was completed through the collaboration of all of the authors. Aitor Álvarez was the team leader and responsible for the speech processing part, selecting and extracting the features to be classified from the speech utterances. Basilio Sierra managed the machine learning part, training and evaluating the classifiers used for the project. Andoni Arruti helped with the audio analysis and with designing the new classification paradigm. Juan-Miguel López-Gil worked preparing the RekEmozio and Emo-DB and provided the state of the art to the team. Nestor Garay-Vitoria completed the state of the art, helped in the data and results interpretation and guided the focus of the article writing.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

Confusion Matrices for the CSS Stacking Classification Method of the Second and Third Phases

In this Appendix, one table per actor is presented, in which the confusion matrices obtained by the CSS stacking classifiers of the second and third phases are detailed. First, the confusion matrices from the RekEmozio database are shown from Tables A1 to A17. Results of the Emo-DB are then presented from Tables A18 to A27.

Table A1. P1 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	19	2	0	0	0	1	0
Fear	0	20	0	0	0	1	1
Joy	0	0	18	3	1	0	0
Anger	1	0	2	17	1	1	0
Surprise	0	0	3	2	17	0	0
Disgust	0	0	0	2	0	20	0
Neutral	1	0	0	0	0	1	20

Table A2. P2 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	16	2	0	0	0	1	3
Fear	1	20	0	0	0	1	0
Joy	0	0	17	2	3	0	0
Anger	0	0	3	19	0	0	0
Surprise	0	2	7	0	13	0	0
Disgust	2	1	2	0	3	5	9
Neutral	0	1	0	0	1	8	12

Table A3. P3 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	18	2	0	0	0	1	1
Fear	1	18	0	0	2	1	0
Joy	0	0	10	8	4	0	0
Anger	0	0	8	14	0	0	0
Surprise	0	1	2	2	16	1	0
Disgust	1	2	0	0	0	16	3
Neutral	3	0	0	0	0	1	18

Table A4. P4 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	21	0	0	0	0	0	1
Fear	0	16	5	0	0	1	0
Joy	0	4	12	4	2	0	0
Anger	0	1	4	13	2	0	2
Surprise	0	2	2	1	14	2	1
Disgust	0	2	1	2	0	15	2
Neutral	2	0	0	1	0	1	18

Table A5. P5 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	21	0	0	0	0	0	1
Fear	0	15	2	2	1	1	1
Joy	0	1	13	5	1	2	0
Anger	0	2	2	17	1	0	0
Surprise	0	0	1	5	16	0	0
Disgust	1	2	2	0	0	17	0
Neutral	1	0	0	0	0	1	20

Table A6. P6 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	20	0	1	0	1	0	0
Fear	0	16	0	1	3	2	0
Joy	2	0	16	1	0	1	2
Anger	2	5	3	8	3	1	0
Surprise	0	6	0	0	14	2	0
Disgust	3	2	2	3	0	9	3
Neutral	5	0	3	0	1	0	13

Table A7. P7 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	10	3	0	0	3	4	2
Fear	4	8	0	3	3	4	0
Joy	1	5	10	1	2	1	2
Anger	1	2	7	8	1	2	1
Surprise	2	4	1	0	12	2	1
Disgust	6	0	0	2	3	11	0
Neutral	4	1	1	2	0	0	14

Table A8. P8 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	12	5	0	0	0	2	3
Fear	5	8	1	0	0	6	2
Joy	1	1	10	5	4	0	1
Anger	0	1	5	11	5	0	0
Surprise	0	1	8	4	9	0	0
Disgust	3	7	0	0	0	11	1
Neutral	2	3	0	0	0	2	15

Table A9. P9 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	12	5	1	0	0	4	0
Fear	5	5	1	0	0	10	1
Joy	0	3	6	7	4	1	1
Anger	0	1	4	7	8	0	2
Surprise	1	1	4	5	11	0	0
Disgust	2	8	1	0	0	11	0
Neutral	0	2	0	2	0	2	16

Table A10. P10 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	19	0	0	0	0	1	2
Fear	0	16	1	0	1	4	0
Joy	0	1	12	4	1	4	0
Anger	0	1	3	16	0	2	0
Surprise	1	1	1	0	19	0	0
Disgust	0	4	3	4	1	10	0
Neutral	0	0	0	0	0	1	21

Table A11. P11 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	20	0	0	1	0	0	1
Fear	0	14	3	1	2	2	0
Joy	0	2	14	4	2	0	0
Anger	2	0	4	13	1	2	0
Surprise	0	3	3	0	11	4	1
Disgust	0	1	2	5	2	12	0
Neutral	2	0	0	0	0	1	19

Table A12. P12 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	20	0	0	0	0	1	1
Fear	1	13	1	0	0	6	1
Joy	0	1	18	2	1	0	0
Anger	0	1	2	16	2	1	0
Surprise	0	0	2	3	17	0	0
Disgust	2	3	0	1	1	11	4
Neutral	1	0	0	0	1	4	16

Table A13. P13 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	20	1	0	0	0	1	0
Fear	2	17	1	1	1	0	0
Joy	0	1	18	1	0	0	2
Anger	0	0	1	20	0	0	1
Surprise	0	3	1	1	17	0	0
Disgust	1	0	1	0	1	15	4
Neutral	0	1	1	0	1	3	16

Table A14. P14 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	7	1	0	0	0	7	7
Fear	2	14	1	0	2	3	0
Joy	0	1	14	2	5	0	0
Anger	0	0	2	17	2	0	1
Surprise	0	2	10	3	7	0	0
Disgust	5	2	0	0	0	14	1
Neutral	2	0	0	0	0	1	19

Table A15. P15 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	10	0	0	0	0	5	7
Fear	3	14	1	1	1	1	1
Joy	0	0	16	3	3	0	0
Anger	0	1	4	8	6	2	1
Surprise	0	0	5	3	13	1	0
Disgust	4	4	0	0	0	13	1
Neutral	4	0	0	0	0	1	17

Table A16. P16 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	16	0	0	0	0	4	2
Fear	0	9	5	4	2	0	2
Joy	1	3	7	3	4	3	1
Anger	2	4	3	8	4	1	0
Surprise	0	4	2	3	13	0	0
Disgust	4	0	0	0	2	12	4
Neutral	1	3	2	0	0	3	13

Table A17. P17 actor confusion matrix from the RekEmozio dataset in the second phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	7	1	0	0	0	8	6
Fear	1	9	2	3	4	3	0
Joy	0	2	7	9	3	1	0
Anger	0	6	7	4	5	0	0
Surprise	0	2	3	2	10	5	0
Disgust	3	3	0	1	5	9	1
Neutral	8	0	1	0	0	2	11

Table A18. A1 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	13	0	0	1	0	0	0
Fear	0	2	0	0	0	0	3
Joy	0	0	0	0	1	0	0
Anger	3	0	0	1	0	0	0
Surprise	0	0	0	0	7	0	0
Disgust	0	0	0	0	0	7	0
Neutral	0	2	0	0	0	0	9

Table A19. A2 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	11	0	0	0	1	0	0
Fear	0	10	0	0	0	0	0
Joy	0	0	0	0	0	0	0
Anger	0	0	0	6	0	0	0
Surprise	0	0	0	0	11	0	0
Disgust	0	0	0	0	0	9	0
Neutral	0	1	0	0	0	0	9

Table A20. A3 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	12	0	1	0	0	0	0
Fear	0	0	0	0	0	0	4
Joy	1	0	7	0	0	0	0
Anger	1	0	0	0	0	0	0
Surprise	2	0	0	0	2	0	0
Disgust	0	0	0	0	0	4	0
Neutral	0	1	0	0	0	0	8

Table A21. A4 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	10	0	0	0	0	0	0
Fear	0	7	0	0	0	1	0
Joy	0	0	0	1	0	0	0
Anger	0	0	0	7	1	0	0
Surprise	0	0	0	0	4	0	0
Disgust	0	0	0	0	0	3	0
Neutral	0	2	0	0	0	0	2

Table A22. A5 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	8	0	0	1	2	0	0
Fear	0	4	0	0	0	3	1
Joy	0	0	0	2	0	0	0
Anger	0	0	0	10	0	0	0
Surprise	2	0	0	1	5	0	0
Disgust	0	0	0	0	0	7	0
Neutral	0	0	0	0	0	0	9

Table A23. A6 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	12	0	0	0	0	0	0
Fear	0	3	0	0	2	0	0
Joy	0	0	1	1	0	0	0
Anger	0	1	0	5	0	0	0
Surprise	1	1	0	0	0	0	0
Disgust	0	0	0	0	0	4	0
Neutral	0	0	0	0	0	0	4

Table A24. A7 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	11	0	0	0	1	0	0
Fear	0	9	0	0	0	0	1
Joy	0	1	6	0	0	0	1
Anger	1	0	1	5	0	0	0
Surprise	1	0	0	0	9	0	0
Disgust	0	0	0	0	0	5	0
Neutral	0	6	0	0	0	0	3

Table A25. A8 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	16	0	0	0	0	0	0
Fear	0	7	0	0	0	0	1
Joy	0	0	7	1	0	0	0
Anger	0	0	1	10	1	0	0
Surprise	6	0	0	0	2	0	0
Disgust	0	0	0	0	0	10	0
Neutral	0	0	0	0	0	0	7

Table A26. A9 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	11	0	0	1	1	0	0
Fear	0	7	0	0	0	1	1
Joy	0	0	4	0	0	0	1
Anger	2	0	0	6	0	0	0
Surprise	2	0	0	2	2	0	0
Disgust	0	0	0	0	0	4	0
Neutral	0	2	0	0	0	1	8

Table A27. A10 actor confusion matrix from the Emo-DB in the third phase.

	Sadness	**Fear**	**Joy**	**Anger**	**Surprise**	**Disgust**	**Neutral**
Sadness	12	0	0	1	1	0	0
Fear	0	12	1	0	0	0	1
Joy	0	1	10	0	0	0	0
Anger	1	0	1	4	1	0	0
Surprise	1	0	0	0	10	0	0
Disgust	0	0	0	0	0	9	0
Neutral	0	2	0	0	0	0	3

References

1. Albert, M. *Silent Messages*; Wadsworth: Belmont, CA, USA, 1971.
2. Lang, P.J. The emotion probe: Studies of motivation and attention. *Am. Psychol.* **1995**, *50*, 372.
3. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087.
4. Scherer, K.R. Vocal communication of emotion: A review of research paradigms. *Speech Commun.* **2003**, *40*, 227–256.
5. Scherer, K.R.; Johnstone, T.; Klasmeyer, G. Vocal expression of emotion. In *Handbook of Affective Sciences*; Oxford University Press: London, UK, 2003; pp. 433–456.
6. Ekman, P.; Friesen, W.V.; Press, C.P. *Pictures of Facial Affect*; Consulting Psychologists Press: Palo Alto, CA, USA, 1975.
7. Lefter, I.; Burghouts, G.B.; Rothkrantz, L.J. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Trans. Affect. Comput.* **2015**, in press.
8. Eyben, F.; Scherer, K.; Schuller, B.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.; Epps, J.; Laukka, P.; Narayanan, S.; *et al*. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, in press.
9. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. Paralinguistics in speech and language—State-of-the-art and the challenge. *Comput. Speech Lang.* **2013**, *27*, 4–39.
10. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80.

11. López, J.M.; Cearreta, I.; Garay-Vitoria, N.; de Ipiña, K.L.; Beristain, A. A methodological approach for building multimodal acted affective databases. In *Engineering the User Interface*; Springer: London, UK, 2009; pp. 1–17.
12. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech 2005, Lissabon, Portugal, 4–8 September 2005; pp. 1517–1520.
13. Sundberg, J.; Patel, S.; Bjorkner, E.; Scherer, K.R. Interdependencies among voice source parameters in emotional speech. *IEEE Trans. Affect. Comput.* **2011**, *2*, 162–174.
14. Ntalampiras, S.; Fakotakis, N. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Trans. Affect. Comput.* **2012**, *3*, 116–125.
15. Wu, S.; Falk, T.H.; Chan, W.Y. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* **2011**, *53*, 768–785.
16. Wang, K.C. Time-Frequency Feature Representation Using Multi-Resolution Texture Analysis and Acoustic Activity Detector for Real-Life Speech Emotion Recognition. *Sensors* **2015**, *15*, 1458–1478.
17. Douglas-Cowie, E.; Campbell, N.; Cowie, R.; Roach, P. Emotional speech: Towards a new generation of databases. *Speech Commun.* **2003**, *40*, 33–60.
18. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587.
19. Ververidis, D.; Kotropoulos, C. Emotional speech recognition: Resources, features, and methods. *Speech Commun.* **2006**, *48*, 1162–1181.
20. Navas, E.; Hernáez, I.; Castelruiz, A.; Luengo, I. Obtaining and evaluating an emotional database for prosody modelling in standard Basque. In *Text, Speech and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 393–400.
21. Iriondo, I.; Guaus, R.; Rodríguez, A.; Lázaro, P.; Montoya, N.; Blanco, J.M.; Bernadas, D.; Oliver, J.M.; Tena, D.; Longhi, L. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, 5–7 September 2000.
22. Caballero-Morales, S.O. Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *Sci. World J.* **2013**, *2013*, 162093.
23. Sobol-Shikler, T.; Robinson, P. Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1284–1297.
24. Schuller, B.; Reiter, S.; Muller, R.; Al-Hames, M.; Lang, M.; Rigoll, G. Speaker independent speech emotion recognition by ensemble classification. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, The Netherland, 6 July 2005; pp. 864–867.
25. Lee, C.C.; Mower, E.; Busso, C.; Lee, S.; Narayanan, S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* **2011**, *53*, 1162–1171.
26. Pan, Y.; Shen, P.; Shen, L. Speech emotion recognition using support vector machine. *Int. J. Smart Home* **2012**, *6*, 101–107.
27. Batliner, A.; Fischer, K.; Huber, R.; Spilker, J.; Nöth, E. Desperately seeking emotions or: Actors, wizards, and human beings. In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, 5–7 September 2000.
28. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623.
29. Shahin, I. Speaker identification in emotional talking environments based on CSPHMM2s. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1652–1659.
30. Pfister, T.; Robinson, P. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Trans. Affect. Comput.* **2011**, *2*, 66–78.
31. Alhamdoosh, M.; Wang, D. Fast decorrelated neural network ensembles with random weights. *Inf. Sci.* **2014**, *264*, 104–117.
32. Arruti, A.; Cearreta, I.; Álvarez, A.; Lazkano, E.; Sierra, B. Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction. *PLoS ONE* **2014**, *9*, e108975.
33. Scherer, S.; Schwenker, F.; Palm, G. Classifier fusion for emotion recognition from speech. In *Advanced Intelligent Environments*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 95–117.

34. Chen, L.; Mao, X.; Xue, Y.; Cheng, L.L. Speech emotion recognition: Features and classification models. *Digit. Signal Process.* **2012**, *22*, 1154–1160.
35. Attabi, Y.; Dumouchel, P. Anchor models for emotion recognition from speech. *IEEE Trans. Affect. Comput.* **2013**, *4*, 280–290.
36. Morrison, D.; Wang, R.; de Silva, L.C. Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.* **2007**, *49*, 98–112.
37. Huang, Y.; Zhang, G.; Xu, X. Speech Emotion Recognition Research Based on the Stacked Generalization Ensemble Neural Network for Robot Pet. In Proceedings of the Chinese Conference on Pattern Recognition, 2009, CCPR 2009, Nanjing, China, 4–6 November 2009; pp. 1–5.
38. Wu, C.H.; Liang, W.B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans. Affect. Comput.* **2011**, *2*, 10–21.
39. Kuang, Y.; Li, L. Speech emotion recognition of decision fusion based on DS evidence theory. In Proceedings of the 2013 4th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 23–25 May 2013; pp. 795–798.
40. Huang, D.Y.; Zhang, Z.; Ge, S.S. Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines. *Comput. Speech Lang.* **2014**, *28*, 392–419.
41. López, J.M.; Cearreta, I.; Fajardo, I.; Garay, N. Validating a multilingual and multimodal affective database. In *Usability and Internationalization. Global and Local User Interfaces*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 422–431.
42. Álvarez, A.; Cearreta, I.; López, J.M.; Arruti, A.; Lazkano, E.; Sierra, B.; Garay, N. A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms. In *Text, Speech and Dialogue*; Springer: Berlin/ Heidelberg, Germany, 2007; pp. 423–430.
43. Esparza, J.; Scherer, S.; Brechmann, A.; Schwenker, F. Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; pp. 1253–1258.
44. Ververidis, D.; Kotropoulos, C. Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm. In Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, ICME 2005, Amsterdam, The Netherland, 6 July 2005; pp. 1500–1503.
45. Hu, H.; Xu, M.X.; Wu, W. Fusion of global statistical and segmental spectral features for speech emotion recognition. In Proceedings of the INTERSPEECH, Antwerp, Belgium, 27–31 August 2007; pp. 2269–2272.
46. Shami, M.T.; Kamel, M.S. Segment-based approach to the recognition of emotions in speech. In Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, ICME 2005, Amsterdam, The Netherlands, 6–8 July 2005; pp. 366-369.
47. Tato, R.; Santos, R.; Kompe, R.; Pardo, J.M. Emotional space improves emotion recognition. In Proceedings of the INTERSPEECH, Denver, CO, USA, 16–20 September 2002; pp. 2029–2032.
48. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Catalunya, Spain, 21–25 October 2013; pp. 835–838.
49. Mendialdua, I.; Arruti, A.; Jauregi, E.; Lazkano, E.; Sierra, B. Classifier Subset Selection to construct multi-classifiers by means of estimation of distribution algorithms. *Neurocomputing* **2015**, *157*, 46–60.
50. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259.
51. Sierra, B.; Serrano, N.; LarrañAga, P.; Plasencia, E.J.; Inza, I.; JiméNez, J.J.; Revuelta, P.; Mora, M.L. Using Bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data. *Artif. Intell. Med.* **2001**, *22*, 233–248.
52. Larrañaga, P.; Lozano, J.A. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*; Springer Science & Business Media: New York, NY, USA, 2002; Volume 2.
53. Inza, I.; Larrañaga, P.; Etxeberria, R.; Sierra, B. Feature subset selection by Bayesian network-based optimization. *Artif. Intell.* **2000**, *123*, 157–184.
54. Etxeberria, R.; Larranaga, P. Global optimization using Bayesian networks. In Proceedings of the Second Symposium on Artificial Intelligence (CIMAF-99), Habana, Cuba, March 1999; pp. 332–339.

55. Inza, I.; Larrañaga, P.; Sierra, B. Feature subset selection by Bayesian networks: A comparison with genetic and sequential algorithms. *Int. J. Approx. Reason.* **2001**, *27*, 143–164.
56. Echegoyen, C.; Mendiburu, A.; Santana, R.; Lozano, J.A. Toward understanding EDAs based on Bayesian networks through a quantitative analysis. *IEEE Trans. Evolut. Comput.* **2012**, *16*, 173–189.
57. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.
58. Sierra, B.; Lazkano, E.; Jauregi, E.; Irigoien, I. Histogram distance-based Bayesian Network structure learning: A supervised classification specific approach. *Decis. Support Syst.* **2009**, *48*, 180–190.
59. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Elsevier: San Francisco, CA, USA, 1993.
60. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66.
61. Cleary, J.G.; Trigg, L.E. K*: An instance-based learner using an entropic distance measure. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; Volume 5, pp. 108–114.
62. Kohavi, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996; pp. 202–207.
63. Cestnik, B. Estimating probabilities: A crucial task in machine learning. In Proceedings of the 9th European Conference on Artificial Intelligence (ECAI-90), Stockholm, Sweden, 6 August 1990, Volume 90, pp. 147–149.
64. Holte, R.C. Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.* **1993**, *11*, 63–90.
65. Cohen, W.W. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9-12 July 1995; pp. 115–123.
66. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
67. Meyer, D.; Leisch, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55*, 169–186.
68. Rosenblatt, F. *Principles oF Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*; Spartan Books: Washington, DC, USA, 1961.
69. Broomhead, D.; Lowe, D. Multivariable functional interpolation and adaptive networks. *Complex Syst.* **1988**, *2*, 321–355.
70. Freedman, D.A. *Statistical Models: Theory and Practice*; Cambridge University Press: New York, NY, USA, 2009.
71. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* **1974**, *36*, 111–147.
72. Buntine, W. Theory refinement on Bayesian networks. In Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence, Los Angeles, CA, USA, 13–15 July 1991; pp. 52–60.
73. García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Inf. Sci.* **2010**, *180*, 2044–2064.
74. Schwenker, F.; Scherer, S.; Magdi, Y.M.; Palm, G. The GMM-SVM supervector approach for the recognition of the emotional status from speech. In *Artificial Neural Networks–ICANN 2009*; Springer: Berlin/Heidelberg, Germany, 14–17 September 2009; pp. 894–903.
75. Grimm, M.; Kroschel, K.; Narayanan, S. The Vera am Mittag German audio-visual emotional speech database. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23 June 2008.
76. Batliner, A.; Steidl, S.; Nöth, E. Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus. In Proceedings of the Satellite Workshop of LREC, Marrakesh, Morocco, 26 May 2008; pp. 28–31.
77. Costantini, G.; Iaderola, I.; Paoloni, A.; Todisco, M. EMOVO Corpus: An Italian Emotional Speech Database. In Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 3501–3504.

7.17 High-Realistic and Flexible Virtual Presenters

- **Authors:** David Oyarzun, Andoni Mujika, Aitor Álvarez, Aritz Legarretaetxeberria, Aitor Arrieta, and María del Puy Carretero
- **Booktitle:** Articulated Motion and Deformable Objects
- **Year:** 2010
- **Publisher:** Springer

High-Realistic and Flexible Virtual Presenters

David Oyarzun, Andoni Mujika, Aitor Álvarez, Aritz Legarretaetxeberria, Aitor Arrieta, and María del Puy Carretero

Vicomtech Research Centre
P. Mikeletegi, 57
20009 San Sebastián, Spain
doyarzun@vicomtech.org

Abstract. This paper presents the research steps that have been necessaries for creating a mixed reality prototype called PUPPET. The prototype provides a 3D virtual presenter that is embedded in a real TV scenario and is driven by an actor in real time. In this way it can interact with real presenters and/or public. The key modules of this prototype improve the state-of-the-art in such systems in four different aspects: real time management of high-realistic 3D characters, animations generated automatically from actor's speech, less equipment needs, and flexibility in the real/virtual integration. The paper describes the architecture and main modules of the prototype.

Keywords: 3D virtual presenters, mixed reality, real time animation.

1 Introduction

Television is a world where technologies with some level of maturity are sooner or later applied. And 3D computer graphics are not an exception. In fact, 3D virtual images have been appearing together with real ones during the last years. For example, they are very common in some weather reports.

In 2006, an Australian TV channel (Channel Ten) went a step forward and broadcasted a talk-show called 'David Tench Tonight show', which was conducted by a 3D virtual character. This character performed interviews to real people, in a mixed reality system shown on live [1].

Its conceptual way of working was very simple. A real actor drove the virtual presenter in real time. His voice caused a synchronized animation of the virtual presenter's lips and he drove the corporal animations by means of a motion capture system. The system was developed by Animal Logic.

Although the show was cancelled some months later, it was initially successful, being one of the 10 most watched programs in Australia[1].

Nowadays, the interest on this kind of mixed reality applications for TV is still alive. For example, companies like Nazooka have created 3D characters that have

[1] Statistics from eBroadcast:
http://www.ebroadcast.com.au/enews/Third_Time_Lucky_for_Seven_180806.html

F.J. Perales and R.B. Fisher (Eds.): AMDO 2010, LNCS 6169, pp. 108–117, 2010.

been broadcasted in different TV programs *via* mixed reality and focus its business model in this kind of technology [2].

However, the applicability of 3D real-time virtual presenters to the TV environment presents some lacks yet. Although most of these lacks are not appreciated by the audience, they imply costs that could be reduced, and interfaces that are not very comfortable for the actors. Concretely, some of the main lacks are:

- Character flexibility. Companies provide the whole system, including the character modeling. Costs would be considerably reduced if TV producers could (re)use characters not created exclusively for the mixed reality system.
- Actor's comfort. Some of current applications require the actor wear a motion capture system or s/he needs to memorize and launch in real time a lot of facial and corporal animations *via* joysticks or keyboards.
- Equipment needs. Apart from the possible motion capture system need, some applications require a chroma system for creating the mixed reality. It implies space, high cost equipment and time for setting up the TV program.
- Mixed reality flexibility. Real cameras are usually fixed when the 3D virtual presenter appears. Cameramen cannot make zoom or change cameras while virtual character is on-screen. Being able to *play* with camera parameters would increase the mixed reality illusion.

This work presents a research project, called PUPPET. Its main requisites are on the whole to improve current state-of-the-art on these applications. The initial prototype developed solves the lacks explained above and so, it provides a low-cost and very flexible solution.

Sections below explain the TV virtual presenter prototype in detail, going into development related research lines in depth. Section 2 present the state-of-the-art about systems related with this prototype and section 3 explains briefly our system architecture. Sections 4, 5 and 6 explain the main modules that solve the lacks mentioned above: section 4, the animation engine that provides character flexibility; section 5, the speech analyzer that improves the actor's comfort and section 6 the mixing module that reduces the equipment needs and improves the real/virtual flexibility. Finally, section 7 explains the resulting prototype tests and section 8 presents the conclusions and future work.

2 State of the Art

The prototype that is presented in this article involves several research fields like mixed reality, real time animation, speech technologies, etc.

Probably the applications that mix these fields in the most related way to this prototype are the works of Nazooka [2] and the David Tench Tonight show, developed by Animal Logic [1].

Nazooka presents some nice developments, however they present limitations regarding the change of camera parameters. That is, the camera remains fixed while the virtual presenter is visible.

On the other hand, David Tench Tonight was initially a successful TV program both from technological and audience level points of view. However, the actor had to use a motion capture system for reproducing all his/her movements in real time. It implied a complex setup and high hardware costs.

There is not many companies and prototypes for creating the mixed reality on TV, however, mixed reality applications are used in several fields such as marketing [3], leisure [4], medicine [5], education [6], etc.

In this way, techniques for obtaining realistic mixing between virtual and real world has been widely studied:

- Lighting, for achieving the shadows of virtual objects over the real world and *vice versa*. Methods like *shadow mapping* [7] or *shadow volumes* [8] are used frequently.
- Occlusions, for calculating virtual world elements occluding real ones and *vice versa*. Different models like a 3D representation of the real world [8], stereo vision-based depth maps [9] or multi-camera 3D reconstruction [10] are used. Each of them has their advantages and disadvantages.

Regarding speech driven animations, most of the previous works are related to the lip synchronization and coarticulation. Phoneme analysis has to transform the speech into phonetic sounds [11] and map them to visemes (the visual representation of each phoneme). However, most of them concerns English [12, 13]. On the other hand, some approaches have been presented for the generation of non-verbal facial expressions from speech. For example, works in [14, 15] generate head movements from fundamental frequency and real time speech driven facial animation is addressed in [16]. However, obtaining a coherent and realistic animation is a state-of-the-art field of research yet.

3 System Overview

The PUPPET prototype system architecture is designed for achieving independence among concrete input devices and the animation and mixed reality modules. In Fig. 1, the conceptual schema of the architecture is presented.

Basically, the input devices are on the one hand the microphone, command devices like keyboards, joysticks, data gloves… and, on the other hand, the cameras of the TV studio.

Microphone input, that is, the voice signal, is managed by the *Speech Analyzer* module. Command devices inputs are retrieved by the *Command Manager*. It is an abstraction layer that avoids device-related dependences in the Animation Engine.

The *Animation Engine* creates the 3D virtual scene that is sent to the *Positioning and Mixing Module*. This module creates the visually correct mixing between the virtual scene and the real image, taking into account real camera changes. Sections below detail the technical aspects about the modules that improve the current systems' lacks. They are:

- The *Speech Analyzer*, which provides not only the analysis for synchronizing the real speech with the virtual presenter lips, but facial animations and expressions too.
- The *Animation Engine*, which is able to load characters created by means of commercial tools like Maya or Poser and animate them through standard BVH files.
- The *Positioning and Mixing Module*, which receives the virtual scene and the real camera parameters in real time and creates a coherent real/virtual mixing.

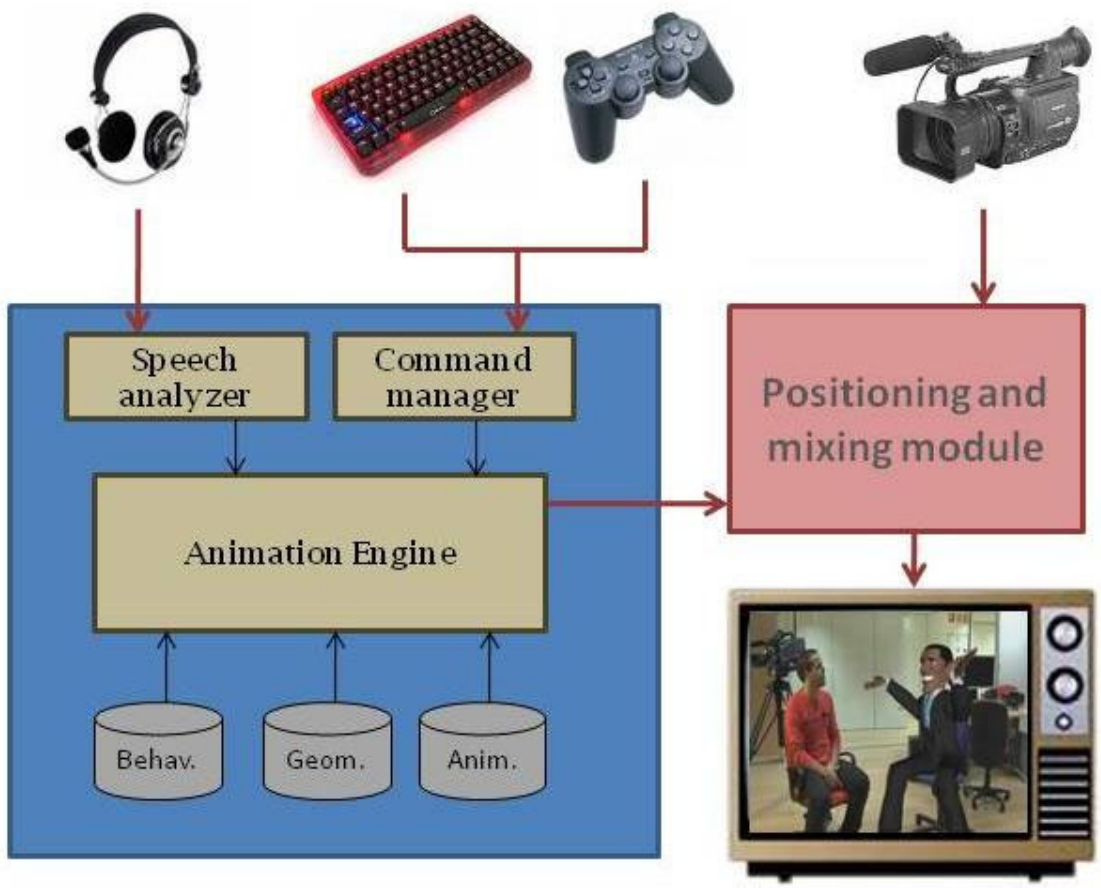


Fig. 1. Simplified architecture schema

4 Speech Analyzer

The Speech Analyzer provides the synchronization between the actor's voice and the avatar lips as well as some facial expressions and animations.

The analyzer captures the speech signal from the input using a microphone and identifies the appropriate phonemes. As phonemes are recognized, they are mapped to their corresponding visemes. In a parallel process, the speech signal is processed by a pitch and energy tracking algorithm, in order to analyze its behavior and decide non-verbal facial movements. The virtual character is then animated in real time and synchronized with the speaker's voice. Therefore, the speech analyzer developed in this paper is composed of four main sub-modules:

- The phoneme recognition system (described in the 4.1 subsection).
- The non-verbal facial animations sub-module (described in the 4.2 subsection).
- The sub-module that sends the input audio to the recognition system and to the pitch/energy tracking algorithm in real-time. To develop this interface we used the ATK API [17].
- The communication interface between the speech application and the animation platform that was developed with sockets, based on the TCP/IP communication protocol. Through this module we fed the animation module with the recognized unit and facial movements for realistic animation.

Using this module the actor has not to get worried about the facial animation of the virtual presenter. All the aspects including lips synchronization and facial expressions will be automatically and coherently launched by the prototype when s/he speaks.

4.1 Real-Time Phoneme Recognition System

The main goal of this sub-module is to obtain the suitable data to animate the lips of the virtual character in real time. To obtain these data, we trained a triphoneme model using HTK Toolkit [18]. The corpus used for training and testing was Albayzin [19], a phonetic database for the development and evaluation of speech recognition and processing systems. It consists of 6800 sentences and 204 speakers. We divided this corpus in two data sets, training (4800 recordings) and test (2000 recordings). All of them are in WAV format (16 kHz/ 16bits/ mono). The feature extraction was performed over 25 ms segments every 10 ms. The parametrization of the speech signal was based on MFCCs, delta and delta-delta coefficients. The Spanish version of SAMPA was used as phoneme set for the recognizer. This set contains 29 phonemes plus the silence and short pause ones. Thiphoneme models were created, which consisted of non-emitting start and end states and three emitting states (except from the short pause model) using Gaussian density functions. Their number of components of these functions was increased until no further recognition improvements were observed. The states are connected left-to-right with no skips. The models were trained iteratively using the embedded Baum-Welch re-estimation and the Viterbi alignment, while the resulting was tested using a Viterbi decoder. Algorithm results are resumed in Table 1.

Table 1. Experimental results (phoneme recognition rate)

Training		**Testing**	
Correctly Words	*Word Accuracy*	*Correctly Words*	*Word Accuracy*
90.41 %	**84.20 %**	**81.18 %**	**71.23 %**

4.2 Non-verbal Facial Animations Sub-module

The recognized phonemes are mapped in real-time to their corresponding visemes in order to make the lip-synchronization process. This is the first step for the facial animation, which has been enriched using prosodic information of speech. A statistical model adapted to current speaker is created during the first steps of the recognition, based on the fundamental frequency (pitch) and energy of the speech signal, in addition to some related statistics. According to the values given in real-time by both pitch and energy trackers, some facial animations are shot, mainly related to the head and eyebrows up and down movement, and eyes and mouth more or less expressively movements.

Fig. 2. Several facial expressions automatically generated from the actor's voice

5 Animation Engine

The animation engine has been designed in order to obtain high-quality real time animations and at the same time be able to load and animate characters not exclusively created for the PUPPET system.

The animation engine is divided in two main modules, the facial animation engine and the body animation engine.

- The facial animation engine uses advanced morphing techniques [20] for generating a high quality animation in real time. This technique is quite extended and it creates the resulting animation by means of the linear interpolation among a set of predefined key faces. The animation engine includes a technique to avoid tests among vertices that are equal to get a lower computational cost.
- The body animation engine implements a set of techniques that aim a realistic movements with a low computational cost. For obtaining realistic movements the engine supports the loading of animations created by professional animators. They are loaded in the system by means of BVH files [21], a semi-standard format to which almost all commercial modeling applications are able to export. Moreover, it includes a set of optimizations that achieve their execution in real time in a standard desktop computer.

  The animation is based on smooth skinning techniques. That is, the vertices of the geometry (or geometries) that conforms the virtual presenter are affected by a virtual skeleton. Transformations over this skeleton influence each vertex taking into account weights assigned to this vertex. These weights provide a way for avoiding *cracks* in the geometry and achieving smooth deformations.

  The conceptual equation for the animation is:

$$v_r = v + w_i * M_r * v_i$$

  Where v_r is the resulting vertex, v is the vertex with the previous transformations in the hierarchy, w_i is the assigned weight, M_r is the rotation matrix corresponding to current node and v_i is the vertex in its initial position.

So as not to depend on specific modeling formats a separation has been established between the geometrical and the smooth skinning information.

- Geometrical information. The 3D character can be loaded in any common geometrical format (3ds, obj, vrml, etc.)
- Smooth skinning information. A new file format, called SHF (Simple Hierarchical Format), has been designed for storing the skeleton and weighting information (Fig. 3). A plug-in to *connect* Maya [22] to SHF has been developed. It allows designers to obtain this information from any Maya modeled character.

The animation engine relates both files in execution time and applies the BVH and morphing animations to them. This way, smooth deformations and high-realistic animations are obtained in real time over any virtual character designed with a standard modeling tool.

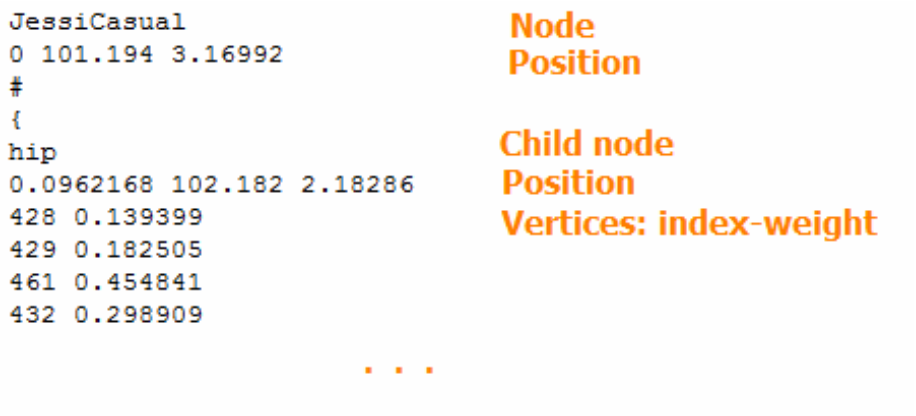


Fig. 3. SHF format file description

6 Positioning and Mixing Module

The Positioning and Mixing Module is designed for creating the mixed reality in a coherent way, without need of physical chroma systems or similar. It works in the opposite way than chroma systems. The virtual presenter background is one uniform color and the real scene replaces directly that color.

Moreover, since our application will be used in television, it would be useful to allow the cameras to translate and zoom. Then, the cameras will be able to follow either the real presenters or the virtual characters and get a more detailed view of them, without losing synchronization between real world and virtual worlds.

The camera is motorized and can be handled remotely. With a remote control three parameters of the camera can be changed: pan, rotation with respect to the vertical axis; tilt, rotation that makes the camera look up and down and zoom.

Fig. 4. Playing with the real camera parameters: changes in translation and zoom (chroma system is not necessary; it is just for having a clean background. Virtual character's chair is real, non virtual).

The robot that moves the camera is connected with the computer through a serial port and transmits the values of the parameters to the computer in real-time. The animation engine receives those values and with simple linear transformations parameters' values in degrees are calculated and transferred to the virtual camera.

In conclusion, the real camera is controlled remotely, but the virtual objects change their position in the screen coherently because of the information traffic between the real and the virtual camera. Fig. 4 shows some screenshots changing the camera parameters.

7 PUPPET Prototype Tests

Modules described before conforms the PUPPET prototype. It has been tested by professional actors and staff from a Basque TV production company called Pausoka[23]. They all agreed that the system is easy to use and avoids limitations and lacks found in the state-of-the-art.

The system has been tested in a standard desktop PC and using virtual characters from different sources. Concretely, along this paper, Fig. 2 shows some screenshots detailing the speech-based facial animation. The virtual character has been obtained from the Poser commercial tool [24].

Fig 4. showed changes in the parameters of the real camera, concretely translations and zooms, and the coherence between the virtual and real images. In this case, the virtual presenter, that is a caricature of Barack Obama, had been designed by a professional modeling company.

8 Conclusions and Future Work

This paper presents a prototype that provides a 3D virtual presenter that is immersed in a real TV scenario. It can be driven by an actor in real time and interact with real presenters and/or public.

The prototype solves some lacks existing in state-of-the-art similar developments. Concretely:

- Character flexibility. There is no need to model animations or virtual characters specifically for their use in the mixed reality platform. The platform supports standard file formats for animating the character and a new file format that supports the smooth skinning data store has been designed.
- Actor's comfort. The platform does not need motion capture systems. It can be handled just with a microphone and usual devices like keyboards or joysticks. Speech signal automates not only the lip animation but also some facial animations.
- Equipment needs. There is no need to use chroma systems or similar. The computer creates the real/virtual mix directly.

– Mixed reality flexibility. Almost all current platforms that do not use chroma systems need to fix the camera, without moving. The platform of this work allows the cameraman to change the camera parameters (zoom, movements…) in real time.

Next steps are to include lighting and occlusion techniques that improve the realism and possibilities of the virtual presenter.

References

1. Animal Logic web page, http://www.animallogic.com
2. Nazooka web page, http://www.nazooka.com/site/
3. Metaio Augmented Solutions, http://www.metaio.com
4. Oda, O., Lister, L.J., White, S., Feiner, S.: Developing an augmented reality racing game. In: Proceedings of the 2nd international conference on INtelligent TEchnologies for interactive enterTAINment (2008)
5. Carlin, A.S., Hoffman, H.G., Weghorst, S.: Virtual reality and tactile augmentation in the treatment of spider phobia: a case report. Behaviour research and therapy (1997)
6. Tan, K.T.W., Lewis, E.M., Avis, N.J., Withers, P.J.: Using augmented reality to promote an understanding of materials science to school children. In: International Conference on Computer Graphics and Interactive Techniques (2008)
7. McCool, M.D.: Shadow volume reconstruction from depth maps. ACM Transactions on Graphics (TOG) 19, 1–26 (2000)
8. Fuhrmann, A., Hesina, G., Faure, F., Gervautz, M.: Occlusion in collaborative augmented environments. Computers and Graphics 23 (1999)
9. Fortin, P., Herbert, P.: Handling occlusions in realtime augmented reality: Dealing with movable real and virtual objects. In: Proceedings of the Canadian Conf. on Computer and Robot Vision, Vol. 54 (2006)
10. Matusik, W., Buehler, C., McMillan, L.: Polyhedral visual hulls for real-time rendering. In: Proc. 12th Eurographics Workshop on Rendering EGWR '01, London (2001)
11. Lehr, M., Arruti, A., Ortiz, A., Oyarzun, D., Obach, M.: Speech Driven Facial Animation using HMMs in Basque. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 415–422. Springer, Heidelberg (2006)
12. Goldenthal, W., Waters, K., Van Thong, J.M., Glickman, O.: Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe. In: Eurospeech, Rhodes, Greece (1997)
13. Massaro, D., Beskow, S., Cohen, M., Fry, C., Rodriguez, T.: Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. In: AVSP, Santa Cruz, California (1999)
14. Deng, Z., Busso, C., Narayanan, S., Neumann, U. : Audio-based Head Motion Synthesis for Avatar-based Telepresence Systems. In: ACM SIGMM Workshop on Effective Telepresence (ETP) (2004)
15. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. ACM Transactions on Graphics (TOG) (2005)
16. Malcangi, M., de Tintis, R.: Audio Based Real-Time Speech Animation of Embodied Conversational Agents. LNCS. Springer, Heidelberg (2004)
17. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book
18. Young, S.: The ATK Real-Time API for HTK

19. Casacuberta, F., Garcia, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A.: Development of Spanish Corpora for Speech Research (ALBAYZIN). In: Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy (1991)
20. Alexa, M., Behr, J., Müller, W.: The morph node. In: Proceedings of the fifth symposium on Virtual reality modeling language (Web3D-VRML), Monterey, California, United States, pp. 29–34 (2000)
21. Meredith, M., Maddock, S.: Motion Capture File Formats Explained. Department of Computer Science, University of Sheffield (2001)
22. Maya Home Page, `http://usa.autodesk.com/adsk/servlet/pc/index?siteID=123112&id=13577897`
23. Poser Home Page, `http://my.smithmicro.com/win/poser/`

7.18 Realistic Visual Speech Synthesis in WebGL

- **Authors:** Andoni Mujika, Helen Diez, Aitor Álvarez, Miren Urteaga, and David Oyarzun
- **Booktitle:** Proceedings of the 18th International Conference on 3D Web Technology
- **Year:** 2013
- **Organization:** ACM

Realistic Visual Speech Synthesis in WebGL

Andoni Mujika⋆, Helen Diez⋆, Aitor Alvarez⋆, Miren Urteaga†, David Oyarzun⋆ *
⋆ Vicomtech-ik4
† University of Basque Country

Abstract

This paper presents the work that has been done to develop a web application that shows the face of a virtual character pronouncing the sentences the user sets. The level of realism was high and the performance was fast enough. The application makes use of WebGL, speech processing, text to speech and co-articulation technologies to obtain the virtual pronunciation.

1 Introduction

Since the pioneer works [Parke 1972] in facial animation, hundreds of methods have been presented to make a virtual character pronounce a sentence, but very few [Benin et al. 2012] are based on the emerging technology WebGL. This paper presents a project that works in this direction, SPEEP, partly funded by the Basque Government. The project creates a system for foreign language pronunciation learning where the key part is the visualization of a virtual character pronouncing the corresponding sentences in a web. Figure 1 shows the interface designed for the project integrated in the language learning system. Since the sentences that have to be synthesized in the virtual character are not predefined i.e. the user writes the desired sentence, the system cannot work with animations that have been generated in a previous phase of the project. Thus, we can define our project as a work in real-time Visual Speech Synthesis.

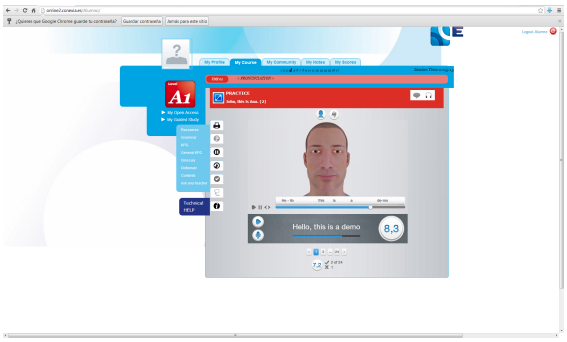


Figure 1: *Interface of the system.*

2 Implementation

In a web browser, the user of the application will find a text field to write the desired sentence and once it is written and confirmed, it is sent to the server. In the server, the text is converted to speech with the desired voice by the module for text-to-speech conversion and the module for voice transformation. On the other hand, the text is used to get the phonemes and the syllables of the sentence and align them with the audio file generated by the text-to-speech module. Once the three files needed (audio, phonemes and syllables) are generated, they are sent to the client.

*e-mail:{amujika, hdiez, aalvarez, doyarzun}@vicomtech.org, miren.urteaga@ehu.es

The client receives the information about the phonemes that must be rendered and their exact times. Then, during the rendering of the virtual character, in each frame the co-articulation engine takes the phonemes that surround the actual time and interpolates the values of the Facial Animation Parameters following the method presented by Cohen and Massaro [Cohen et al. 1993]. Besides, several rules have been defined that change the value of some parameters and the position of the maximum point of the phoneme in its interval, depending on the type of syllable and the phonemes involved. For example, the consonant of a syllable CV (consonant-vowel) is placed in a different position of its interval comparing to a consonant of a syllable VC.

A study of the computational cost of the different methods that are called in each frame during the facial animation showed that the most time consuming function was the skinning i.e. the computation of the new position of the vertices according to the transformations of their neighbor bones. So, we decided to make use of GPU's power and we implement the skinning using shaders, obtaining a considerable increase in the number of frames per second.

For the speech generation, the server takes the text written by the user and converts it into an audio file. Then, speech and text are automatically aligned using an acoustic model obtained by training. Finally, in order to identify each syllable for a natural mouth articulation of the avatar, a syllabification module was built.

3 Conclusion

The results obtained so far in the project SPEEP are satisfactory. A big number of vertices are needed to obtain a realistic virtual face and although all the vertices are not transformed in all frames, the amount of vertices that are moved is big enough to become a problem. Specifically, the virtual face that is used in the project SPEEP is compound of 22802 vertices that form 43449 polygons. Nevertheless, the problems that slower performance of web applications can cause have been overcome. Several strategies have been implemented to make the application work faster; fast enough even in computers with commodity graphic card. Moreover, a high level of realism has been achieved in Visual Speech Synthesis.

References

BENIN, A., LEONE, G. R., AND COSI, P. 2012. A 3d talking head for mobile devices based on unofficial ios webgl support. In *Proceedings of the 17th International Conference on 3D Web Technology*, ACM, New York, NY, USA, Web3D '12, 117–120.

COHEN, M. M., MASSARO, D. W., ET AL. 1993. Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation 92*.

PARKE, F. I. 1972. Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, ACM, 451–457.

7.19 Other Publications

7.19.1 BerbaTek: euskararako hizkuntza teknologien garapena itzulpengintza, edukien kudeaketa eta irakaskuntza arloetan

- **Authors:** Igor Leturia, Eva Navas, Iñaki Sainz, David Baranda, Urtza Iturraspe, Kepa Sarasola, Xabier Arregi, Arantza Diaz de Ilarraza, Arantza del Pozo, and Aitor Álvarez
- **Journal:** Euskalingua
- **Year:** 2013
- **Publisher:** Mendebalde Kultur Alkartea
- **Abstract:** Basque is both a minority language (only a small proportion of the population of the Basque Country speaks it) and also a less-resourced language. Fortunately, the Basque regional government is committed to its recovery, and has adopted policies for funding, among other things, language technologies, a field which a language aiming to survive cannot dispense with. BerbaTek was a 3-year (2009-2011) strategic research project on language, speech and multimedia technologies for Basque carried out by a consortium of five members, all prominent local organizations dedicated to research in the above-mentioned areas, and partially funded by the Departments for Industry and Culture of the Basque Government. Collaboration in BerbaTek allowed to carry out a great amount of both basic and applied research. In addition, various prototypes were developed to show the potential of integrating the developed technologies to the language industry sector.

7.19.2 Assisted subtitling: a new opportunity for access services

- **Authors:** Carlo Aliprandi, Isabella Gallucci, Nicola Piccinini, Matteo Raffaelli, Arantza del Pozo, Aitor Álvarez, Renato Cassaca, Joao Neto, Carlos Mendes, and Marcos Viveiros
- **Year:** 2014
- **Publisher:** IET

- **Abstract:** The demand for Access Services has quickly grown over the years, mainly due to National and International laws. This trend is expected to consolidate for subtitling in particular, as almost every broadcaster is nowadays working with digital content: large amounts of existing assets are going to be digitized in the near future. In terms of accessibility, digitalization is a very challenging task that can be turned into a profitable process if addressed with adequate technology. In this paper we will focus on an emerging technique: Assisted Subtitling. Assisted Subtitling consists in the application of Automatic Speech Recognition (ASR) to generate transcripts of programs and to use the transcripts as the basis for subtitles. This paper will report on recent advances in ASR, presenting SAVAS, a novel Speaker Independent ASR technology specifically designed for Live Subtitling. We will describe the technology and, evaluating its performances, we will present the promising results we have so far achieved.

7.19.3 Automatic Live Subtitling: state of the art, expectations and current trends

- **Authors:** Carlo Aliprandi, Cristina Scudellari, Isabella Gallucci, Nicola Piccinini, Matteo Raffaelli, Arantza del Pozo, Aitor Álvarez, Haritz Arzelus, Renato Cassaca, Tiago Luis, Joao Neto, Carlos Mendes, Sérgio Paulo, and Marcio Viveiros

- **Booktitle:** Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies (Las Vegas)

- **Year:** 2014

- **Abstract:** The subtitling demand has grown quickly over the years. The path of manual subtitling is no longer feasible, due to increased costs and reduced production times. Assisted Subtitling is an emerging technique, consisting in the application of Automatic Speech Recognition (ASR) to automatically generate program transcripts. This paper will report on recent advances in ASR, presenting SAVAS, a novel Speaker Independent ASR technology specifically designed for Live Subtitling. We will describe the technology, presenting its features and detailing language and domain-specific tunings that we have carried out. We will also introduce the S.Scribe!, S.Live! and S.Respeak! systems, which are based on SAVAS. S.Scribe! is a batch Speaker Independent Transcription system for subtitling. S.Live! is a first-of-a-kind Speaker Independent Transcription System, with real-time performances for

online subtitling. S.Respeak! is a collaborative Respeaking System, for live and batch production of multilingual subtitles. S.Respeak! has proven to be sufficiently robust for programs where the acoustic conditions are challenging and for spontaneous speech. Similar results are expected to be achieved also for S.Live! and S.Scribe!, which are currently being tested under real conditions at different broadcasters premises, to subtitle live programs, in both assisted and unassisted tasks. We will finally detail performances of the systems for 7 languages (English, Spanish, Italian, French, German, Portuguese and Basque).

7.19.4 The reception of Intralingual and Interlingual Automatic Subtitling: An Exploratory Study within The HBB4ALL Project

- **Authors:** Anna Matamala, Andreu Oliver, Aitor Álvarez, and Andoni Azpeitia
- **Booktitle:** Translating and the Computer Conference
- **Year:** 2015
- **Abstract:** HBB4ALL is a EC funded project (CIP-ICT-PSP-2013-5.1.) that builds on HbbTV, the European standard for broadcast and broadband multimedia converged services, and looks at how HbbTV technologies may be used to enhance access services such as subtitling. This work presents the results of user testing on automatic subtitling carried out within the project. A first preliminary test with 56 students allowed to: (a) compare the comprehension of clips with automatic intralingual subtitling versus automatic interlingual subtitles in students with low English proficiency, (b) compare the comprehension reached by students with a low level of English using subtitles to that of students with a higher level using no subtitles, and most importantly (c) test the methodology. A second experiment aimed to determine if intralingual or interlingual automatic subtitling help to better understand news content. Three breaking news clips in English automatically subtitled in English and Spanish were used. Data from 30 volunteers exposed to the three conditions (no subtitles/intralingual/interlingual) were analysed. Comprehension was measured through a summarisation task and a questionnaire. An English proficiency control test as well as a demographics questionnaire were administered. The paper presents the results of both tests and discusses methodological issues.

7.19.5 Interactive Multimodal Platform for Digital Signage

- **Authors:** Helen V. Diez, Javier Barbadillo, Sara García, María del Puy Carretero, Aitor Álvarez, Jairo R. Sánchez, and David Oyarzun
- **Booktitle:** Articulated Motion and Deformable Objects
- **Year:** 2014
- **Publisher:** Springer
- **Abstract:** The main objective of the platform presented in this paper is the integration of various modules into Web3D technology for Digital Signage systems. The innovation of the platform consists on the development and integration of the following technologies; 1) autonomous virtual character with natural behaviour, 2) text-to-speech synthesizer and voice recognition 3) gesture recognition. The integration of these technologies will enhance the user interface interaction and will improve the existing Digital Signage solutions offering a new way of marketing to engage the audience. The goal of this work is also to prove whether this new way of e-commerce may improve sales and customer fidelity.

Bibliography

[ÁAE14] Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. „Towards customized automatic segmentation of subtitles“. In: *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014, pp. 229–238 (cit. on pp. 7, 8, 29, 65, 68–70).

[ÁAR14] Aitor Álvarez, Haritz Arzelus, and Pablo Ruiz. „Long audio alignment for automatic subtitling using different phone-relatedness measures“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE. 2014, pp. 6280–6284 (cit. on pp. 6, 7, 28, 59, 64).

[AD13] Yazid Attabi and Pierre Dumouchel. „Anchor models for emotion recognition from speech“. In: *IEEE Transactions on Affective Computing* 4.3 (2013), pp. 280–290 (cit. on p. 82).

[Aen12] Aenor. *Subtitulado para personas sordas y personas con discapacidad auditiva*. Tech. rep. 2012 (cit. on p. 49).

[Ahm+13] Imran Ahmed, Sunil Kumar Kopparapu, TCS Innovation, et al. „Technique for automatic sentence level alignment of long speech and transcripts.“ In: *INTERSPEECH*. 2013, pp. 1516–1519 (cit. on p. 49).

[AKA91] David W Aha, Dennis Kibler, and Marc K Albert. „Instance-based learning algorithms“. In: *Machine learning* 6.1 (1991), pp. 37–66 (cit. on p. 86).

[ALG14] Xavier Anguera, Jordi Luque, and Ciro Gracia. „Audio-to-text alignment for speech recognition with very limited resources.“ In: *INTERSPEECH*. 2014, pp. 1405–1409 (cit. on p. 49).

[Ali+14a] C Aliprandi, I Gallucci, N Piccinini, et al. „Assisted subtitling: a new opportunity for access services“. In: (2014) (cit. on p. 23).

[Ali+14b] Carlo Aliprandi, Cristina Scudellari, Isabella Gallucci, et al. „Automatic Live Subtitling: state of the art, expectations and current trends“. In: *Proceedings of NAB Broadcast Engineering Conference: Papers on Advanced Media Technologies, Las Vegas*. 2014 (cit. on p. 23).

[Álv+06] Aitor Álvarez, Idoia Cearreta, Juan Miguel López, et al. „Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken spanish and standard basque language“. In: *Text, Speech and Dialogue*. Springer. 2006, pp. 565–572 (cit. on pp. 10, 35, 86, 91).

[Álv+07a] Aitor Álvarez, Idoia Cearreta, Juan Miguel López, et al. „A comparison using different speech parameters in the automatic emotion recognition using Feature Subset Selection based on Evolutionary Algorithms“. In: *Text, Speech and Dialogue*. Springer. 2007, pp. 423–430 (cit. on pp. 9, 35, 86, 90).

[Álv+07b] Aitor Álvarez, Idoia Cearreta, Juan Miguel López, et al. „Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech“. In: *Advances in Nonlinear Speech Processing*. Springer, 2007, pp. 273–281 (cit. on pp. 10, 35, 90).

[Álv+15] Aitor Álvarez, Carlos Mendes, Matteo Raffaelli, et al. „Automating live and batch subtitling of multimedia contents for several European languages“. In: *Multimedia Tools and Applications* (2015), pp. 1–31 (cit. on pp. 5, 22, 23, 46, 47, 52, 55, 56, 58).

[Álv+16a] Aitor Álvarez, Haritz Arzelus, Santiago Prieto, and Arantza del Pozo. „Rich Transcription and Automatic Subtitling for Basque and Spanish“. In: *Advances in Speech and Language Technologies for Iberian Languages*. Status: Submitted. Springer, 2016 (cit. on pp. 8, 33, 77).

[Álv+16b] Aitor Álvarez, Marina Balenciaga, Arantza del Pozo, et al. „Impact of Automatic Segmentation on the Quality, Productivity and Self-reported Post-editing Effort of Intralingual Subtitles“. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portoroz, Slovenia: European Language Resources Association (ELRA), May 2016 (cit. on pp. 8, 29, 54, 69).

[Álv+16c] Aitor Álvarez, Carlos-D. Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. „Improving the Automatic Segmentation of Subtitles through Conditional Random Field“. In: *Speech Communication* (2016). Status: In 2nd revision (cit. on pp. 7, 28, 65, 69, 75).

[Álv+16d] Aitor Álvarez, Basilio Sierra, Andoni Arruti, Juan-Miguel López-Gil, and Nestor Garay-Vitoria. „Classifier Subset Selection for the Stacked Generalization Method Applied to Emotion Recognition in Speech“. In: *Sensors* 16.1 (2016), p. 21 (cit. on pp. 10, 35, 94, 95).

[ÁPA10] Aitor Álvarez, Arantza del Pozo, and Andonin Arruti. „APyCA: Towards the automatic subtitling of television content in Spanish“. In: *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*. IEEE. 2010, pp. 567–574 (cit. on pp. 5, 17).

[ÁRA14] Aitor Álvarez, Pablo Ruiz, and Haritz Arzelus. „Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque“. In: *Text, Speech and Dialogue*. Springer. 2014, pp. 473–480 (cit. on pp. 6, 28, 59, 61, 63, 64).

[Arr+14] Andoni Arruti, Idoia Cearreta, Aitor Álvarez, Elena Lazkano, and Basilio Sierra. „Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction“. In: *PloS one* 9.10 (2014), e108975 (cit. on pp. 9, 35, 87, 90, 93).

[AW14] Monther Alhamdoosh and Dianhui Wang. „Fast decorrelated neural network ensembles with random weights“. In: *Information Sciences* 264 (2014), pp. 104–117 (cit. on p. 82).

[Azk+13] Igor Leturia Azkarate, Eva Navas Cordón, Iñaki Sainz Moncalvillo, et al. „BerbaTek: euskararako hizkuntza teknologien garapena itzulpengintza, edukien kudeaketa eta irakaskuntza arloetan“. In: *Euskalingua* 23 (2013), pp. 66–76 (cit. on p. 20).

[Bar+01] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. „Transcriber: development and use of a tool for assisting speech corpora production“. In: *Speech Communication* 33.1 (2001), pp. 5–22 (cit. on p. 57).

[Bat+12] Fernando Batista, Helena Moniz, Isabel Trancoso, and Nuno Mamede. „Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts“. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 474–485 (cit. on p. 52).

[Bor+12] Germán Bordel, Mikel Peñagarikano, Luis Javier Rodríguez-Fuentes, and Amparo Varona. „A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions.“ In: *INTERSPEECH*. 2012, pp. 1840–1843 (cit. on pp. 7, 47, 49, 59).

[Bor+16] German Bordel, Mikel Penagarikano, Luis Javier Rodríguez-Fuentes, Aitor Álvarez, and Amparo Varona. „Probabilistic Kernels for improved text-to-speech alignment in long audio tracks“. In: *IEEE Signal Processing Letters* 23.1 (2016), pp. 126–129 (cit. on pp. 6, 28, 61, 64).

[Bri94] Eric Brill. „Some advances in transformation-based part of speech tagging“. In: *arXiv preprint cmp-lg/9406010* (1994) (cit. on p. 52).

[BSN08] A Batliner, S Steidl, and E Nöth. „Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus“. In: *Proceedings of a satellite workshop of LREC*. 2008, pp. 28–31 (cit. on p. 105).

[Bur+05] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. „A database of German emotional speech.“ In: *INTERSPEECH*. Vol. 5. 2005, pp. 1517–1520 (cit. on pp. 85, 94).

[Bur+08] Juan José Burred, Martin Haller, Shan Jin, Amjad Samour, and Thomas Sikora. *Audio content analysis*. Springer, 2008 (cit. on p. 45).

[Bus+05] Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. „Natural head motion synthesis driven by acoustic prosodic features“. In: *Computer Animation and Virtual Worlds* 16.3-4 (2005), pp. 283–290 (cit. on p. 83).

[BWM05] Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier. „ALIZE, a free toolkit for speaker recognition.“ In: *ICASSP (1)*. 2005, pp. 737–740 (cit. on p. 41).

[CA06] Ciprian Chelba and Alex Acero. „Adaptation of maximum entropy capitalizer: Little data can help a lot“. In: *Computer Speech & Language* 20.4 (2006), pp. 382–399 (cit. on p. 52).

[CB05] Erika Chuang and Christoph Bregler. „Mood swings: expressive speech animation“. In: *ACM Transactions on Graphics (TOG)* 24.2 (2005), pp. 331–347 (cit. on p. 83).

[Che+12] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng. „Speech emotion recognition: Features and classification models“. In: *Digital signal processing* 22.6 (2012), pp. 1154–1160 (cit. on p. 82).

[CM93] Michael M Cohen and Dominic W Massaro. „Modeling coarticulation in synthetic visual speech“. In: *Models and techniques in computer animation*. Springer, 1993, pp. 139–156 (cit. on pp. 40, 82, 96).

[Col+11] Ronan Collobert, Jason Weston, Léon Bottou, et al. „Natural language processing (almost) from scratch“. In: *The Journal of Machine Learning Research* 12 (2011), pp. 2493–2537 (cit. on p. 78).

[Cos+14] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. „EMOVO Corpus: an Italian Emotional Speech Database.“ In: *LREC*. 2014, pp. 3501–3504 (cit. on p. 94).

[Cow+01] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, et al. „Emotion recognition in human-computer interaction“. In: *IEEE Signal Processing Magazine* 18.1 (2001), pp. 32–80 (cit. on p. 80).

[Cut+04] Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. „Patterns of English phoneme confusions by native and non-native listeners“. In: *The Journal of the Acoustical Society of America* 116.6 (2004), pp. 3668–3678 (cit. on p. 61).

[Dav+64] Joel R Davitz, M Beldoch, S Blau, et al. „Personality, perceptual, and cognitive correlates of emotional sensitivity“. In: *The communication of emotional meaning* (1964), pp. 57–68 (cit. on p. 105).

[Del+15] Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano. „Fast single-and cross-show speaker diarization using binary key speaker modeling“. In: *IEEE ACM Transactions on Audio, Speech and Language Processing (TASLP)* 23.12 (2015), pp. 2286–2297 (cit. on p. 53).

[Dix09] Alan Dix. *Human-computer interaction*. Springer, 2009 (cit. on p. 79).

[Dou+03] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. „Emotional speech: Towards a new generation of databases“. In: *Speech Communication* 40.1 (2003), pp. 33–60 (cit. on p. 105).

[DR07] Jorge Díaz-Cintas and Aline Remael. *Audiovisual Translation, Subtitling*. St. Jerome Publishing, 2007 (cit. on p. 49).

[DR89] Géry D'Ydewalle and Johan Van Rensbergen. „13 Developmental Studies of Text-Picture Interactions in the Perception of Animated Cartoons with Text“. In: *Advances in Psychology* 58 (1989), pp. 233–248 (cit. on p. 46).

[DRW39] Homer Dudley, RR Riesz, and SSA Watkins. „A synthetic speaker“. In: *Journal of the Franklin Institute* 227.6 (1939), pp. 739–764 (cit. on p. 1).

[EBS15] Florian Eyben, Anton Batliner, and Bjoern Schuller. „Towards a standard set of acoustic features for the processing of emotion in speech“. In: *Proceedings of Meetings on Acoustics*. Vol. 9. 1. Acoustical Society of America. 2015, p. 060006 (cit. on p. 81).

[EFP75] Paul Ekman, Wallace V Friesen, and Consulting Psychologists Press. *Pictures of facial affect*. consulting psychologists press, 1975 (cit. on p. 86).

[EGP02] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. *Trainable video realistic speech animation*. Vol. 21. 3. ACM, 2002 (cit. on p. 82).

[EL99] Ramon Etxeberria and Pedro Larranaga. „Global optimization using Bayesian networks“. In: *Second Symposium on Artificial Intelligence (CIMAF-99)*. Habana, Cuba. 1999, pp. 332–339 (cit. on p. 92).

[Esp+12] Jose Esparza, Stefan Scherer, Andre Brechmann, and Friedheim Schwenker. „Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark“. In: *11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012*. IEEE. 2012, pp. 1253–1258 (cit. on p. 94).

[EWS10] Florian Eyben, Martin Wollmer, and Bjorn Schuller. „Opensmile: the munich versatile and fast open-source audio feature extractor“. In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 1459–1462 (cit. on p. 81).

[EWS13] Florian Eyben, Felix Weninger, and Bjorn Schuller. „Affect recognition in real-life acoustic conditions-a new perspective on feature selection.“ In: *INTERSPEECH*. Citeseer. 2013, pp. 2044–2048 (cit. on p. 81).

[Eyb+16] Florian Eyben, Klaus R Scherer, Björn W Schuller, et al. „The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing“. In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202 (cit. on pp. 81, 82, 85, 93).

[For09] G. Ford Williams. *Online Subtitling Editorial Guidelines v1.1*. Tech. rep. BBC, 2009 (cit. on p. 49).

[Gal98] Mark JF Gales. „Maximum likelihood linear transformations for HMM-based speech recognition“. In: *Computer speech & language* 12.2 (1998), pp. 75–98 (cit. on p. 48).

[Gar+93] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett. „DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1“. In: *NASA STI/Recon technical report n* 93 (1993) (cit. on p. 40).

[GB06] Erdan Gu and Norman I Badler. „Visual attention and eye gaze during multiparty conversations with distractions“. In: *Intelligent Virtual Agents*. Springer. 2006, pp. 193–204 (cit. on p. 83).

[GFG02] D Graff, J Fiscus, and J Garofolo. „1997 HUB4 English evaluation speech and transcripts“. In: *Linguistic Data Consortium, Philadelphia* 133 (2002) (cit. on p. 63).

[GJB09] Agustin Gravano, Martin Jansche, and Michiel Bacchiani. „Restoring punctuation and capitalization in transcribed speech“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.* IEEE. 2009, pp. 4741–4744 (cit. on p. 52).

[GKN] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. „The Vera am Mittag German audio-visual emotional speech database“. In: *IEEE International Conference on Multimedia and Expo* (cit. on p. 105).

[Gol95] Daniel Goleman. *Emotional intelligence*. Bantam, 1995 (cit. on p. 105).

[Goo09] Google. *Automatic Captions in Youtube*. `https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html`. 2009 (cit. on p. 47).

[Gra03] Guillaume Gravier. „SPro:Speech Signal Processing Toolkit“. In: *Software available at http://gforge.inria.fr/projects/spro* (2003) (cit. on p. 41).

[Gré+07] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocky. „Probabilistic and bottle-neck features for LVCSR of meetings“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.* Vol. 4. IEEE. 2007, pp. IV–757 (cit. on p. 48).

[GS10] Martijn Goudbeek and Klaus Scherer. „Beyond arousal: Valence and potency control cues in the vocal expression of emotion“. In: *The Journal of the Acoustical Society of America* 128.3 (2010), pp. 1322–1336 (cit. on p. 81).

[GSR91] Herbert Gish, M-H Siu, and Robin Rohlicek. „Segregation of speakers for speech recognition and speaker identification“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.* IEEE. 1991, pp. 873–876 (cit. on p. 74).

[Gut+05] Ricardo Gutierrez-Osuna, Praveen K Kakumanu, Anna Esposito, et al. „Speech-driven facial animation with realistic dynamics“. In: *IEEE Transactions on Multimedia* 7.1 (2005), pp. 33–42 (cit. on p. 82).

[Han+12] Wei Han, Lijuan Wang, Frank Soong, and Bo Yuan. „Improved minimum converted trajectory error training for real-time speech-to-lips conversion“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2012, pp. 4513–4516 (cit. on p. 82).

[Hea11] Kenneth Heafield. „KenLM: Faster and smaller language model queries“. In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics. 2011, pp. 187–197 (cit. on pp. 26, 70, 71).

[HES00] Hynek Hermansky, Daniel W Ellis, and Shantanu Sharma. „Tandem connectionist feature extraction for conventional HMM systems“. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings.* Vol. 3. IEEE. 2000, pp. 1635–1638 (cit. on p. 48).

[Hin+12] Geoffrey Hinton, Li Deng, Dong Yu, et al. „Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups“. In: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 82–97 (cit. on p. 48).

[Hir75] Daniel S. Hirschberg. „A linear space algorithm for computing maximal common subsequences“. In: *Communications of the ACM* 18.6 (1975), pp. 341–343 (cit. on pp. 49, 59).

[HK07] Alexander Haubold and John R Kender. „Alignment of speech to highly imperfect text transcriptions“. In: *2007 IEEE International Conference on Multimedia and Expo*. IEEE. 2007, pp. 224–227 (cit. on p. 49).

[HN06] Florian Hilger and Hermann Ney. „Quantile based histogram equalization for noise robust large vocabulary speech recognition“. In: *IEEE Transactions on Audio, Speech, and Language Processing.* 14.3 (2006), pp. 845–854 (cit. on p. 48).

[Hol+04] Constance Holden et al. „The origin of speech“. In: *Science* 303.5662 (2004), pp. 1316–1319 (cit. on p. 1).

[Hol75] John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press, 1975 (cit. on p. 87).

[HP13] Sarah Hoffmann and Beat Pfister. „Text-to-speech alignment of long recordings using universal phone models.“ In: *INTERSPEECH*. Citeseer. 2013, pp. 1520–1524 (cit. on p. 49).

[HW07] Marijn Huijbregts and Chuck Wooters. „The Blame Game: Performance Analysis of Speaker Diarization System Components“. In: *Eighth Annual Conference of the International Speech Communication Association*. 2007 (cit. on p. 52).

[HWH02] Pengyu Hong, Zhen Wen, and Thomas S Huang. „Real-time speech-driven face animation with expressions using neural networks“. In: *IEEE Transactions on Neural Networks* 13.4 (2002), pp. 916–927 (cit. on p. 82).

[HYS08] Gregor Hofer, Junichi Yamagishi, and Hiroshi Shimodaira. „Speech-driven lip motion generation with a trajectory HMM“. In: *INTERSPEECH*. Citeseer. 2008 (cit. on p. 82).

[HYT14] Kun Han, Dong Yu, and Ivan Tashev. „Speech emotion recognition using deep neural network and extreme learning machine“. In: *INTERSPEECH*. 2014, pp. 223–227 (cit. on p. 80).

[HZG14] Dong-Yan Huang, Zhengchen Zhang, and Shuzhi Sam Ge. „Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines“. In: *Computer Speech & Language* 28.2 (2014), pp. 392–419 (cit. on p. 82).

[IC98] J. Ivarsson and M. Carroll. *Code of Good Subtitling Practice*. Tech. rep. Berlin: ESIST (European Association for Studies in Screen Translation), 1998 (cit. on p. 49).

[Inz+00] Iñaki Inza, Pedro Larrañaga, Ramón Etxeberria, and Basilio Sierra. „Feature subset selection by Bayesian network-based optimization“. In: *Artificial intelligence* 123.1 (2000), pp. 157–184 (cit. on pp. 87, 92).

[Iri+00] Ignasi Iriondo, Roger Guaus, Angel Rodríguez, et al. „Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques“. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. 2000 (cit. on p. 86).

[Joz+16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. „Exploring the limits of language modeling“. In: *arXiv preprint arXiv:1602.02410* (2016) (cit. on p. 104).

[Kar98] F. Karamitroglu. „A Proposed Set of Subtitling Standards in Europe“. In: *Translation Journal* 2 (1998) (cit. on p. 49).

[KN95] Reinhard Kneser and Hermann Ney. „Improved backing-off for m-gram language modeling“. In: *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP.* Vol. 1. IEEE. 1995, pp. 181–184 (cit. on p. 104).

[Koe+07] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. „Moses: Open source toolkit for statistical machine translation“. In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics. 2007, pp. 177–180 (cit. on p. 74).

[Kon02] Grzegorz Kondrak. *Algorithms for Language Reconstruction (PDF). University of Toronto, Ontario*. Tech. rep. Retrieved 2007-01-21, 2002 (cit. on p. 60).

[KR12] Shashidhar G Koolagudi and K Sreenivasa Rao. „Emotion recognition from speech: a review“. In: *International journal of speech technology* 15.2 (2012), pp. 99–117 (cit. on pp. 80, 81).

[Kra82] Christian Gottlieb Kratzenstein. „Sur la naissance de la formation des voyelles“. In: *Journal de Physique* 21 (1782), pp. 358–380 (cit. on p. 1).

[Lam+03] Paul Lamere, Philip Kwok, Evandro Gouvea, et al. „The CMU SPHINX-4 speech recognition system“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*. Vol. 1. Citeseer. 2003, pp. 2–5 (cit. on p. 39).

[Lan95] Peter J Lang. „The emotion probe: studies of motivation and attention“. In: *American psychologist* 50.5 (1995), p. 372 (cit. on p. 80).

[Lar+12] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li. „RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases.“ In: *INTERSPEECH*. 2012, pp. 1580–1583 (cit. on p. 41).

[LCL13] Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. „Language diarization for code-switch conversational speech“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE. 2013, pp. 7314–7318 (cit. on p. 72).

[Lec+06] Benjamin Lecouteux, Georges Linares, Pascal Nocera, and Jean-François Bonastre. „Imperfect transcript driven speech recognition.“ In: *INTERSPEECH*. 2006 (cit. on p. 49).

[Lec+13] Maria Luisa Garcia Lecumberri, Máté Attila Tóth, Yan Tang, and Martin Cooke. „Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish.“ In: *INTERSPEECH*. Citeseer. 2013, pp. 2807–2811 (cit. on p. 61).

[Lev+10] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. „Gesture controllers“. In: *ACM Transactions on Graphics (TOG)*. Vol. 29. 4. ACM. 2010, p. 124 (cit. on p. 83).

[LG08] Hank Liao and MJF Gales. „Issues with uncertainty decoding for noise robust automatic speech recognition“. In: *Speech Communication* 50.4 (2008), pp. 265–277 (cit. on p. 48).

[LJ14] Peter Ladefoged and Keith Johnson. *A course in phonetics*. Nelson Education, 2014 (cit. on p. 60).

[LMD12] Binh H Le, Xiaohan Ma, and Zhigang Deng. „Live speech driven head-and-eye motion generators“. In: *IEEE Transactions on Visualization and Computer Graphics* 18.11 (2012), pp. 1902–1914 (cit. on p. 83).

[LMS13] Haitao Liao, Erik McDermott, and Alan Senior. „Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription“. In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013*. IEEE. 2013, pp. 368–373 (cit. on p. 47).

[Lóp+06] Juan M López, Idoia Cearreta, Nestor Garay, K López de Ipiña, and Andoni Beristain. „Creación de una base de datos emocional bilingue y multimodal“. In: *Proceedings of the 7th Spanish Human Computer Interaction Conference, Interaccion*. Vol. 6. 2006, pp. 55–66 (cit. on p. 86).

[Lop+07] Juan Miguel L'opez, Idoia Cearreta, Inmaculada Fajardo, and Nestor Garay. „Validating a multilingual and multimodal affective database". In: *Usability and Internationalization. Global and Local User Interfaces*. Springer, 2007, pp. 422–431 (cit. on pp. 85, 86).

[Lóp+09] Juan Miguel López, Idoia Cearreta, Nestor Garay-Vitoria, Karmele López de Ipiña, and Andoni Beristain. „A methodological approach for building multimodal acted affective databases". In: *Engineering the user interface*. Springer, 2009, pp. 1–17 (cit. on p. 34).

[LR98] Li Lee and Richard Rose. „A frequency warping approach to speaker normalization". In: *IEEE Transactions on Speech and Audio Processing*. 6.1 (1998), pp. 49–60 (cit. on p. 48).

[LTK09] Sergey Levine, Christian Theobalt, and Vladlen Koltun. „Real-time prosody-driven synthesis of body language". In: *ACM Transactions on Graphics (TOG)* 28.5 (2009), p. 172 (cit. on p. 83).

[LYW14] Changwei Luo, Jun Yu, and Zengfu Wang. „Synthesizing real-time speech-driven facial animation". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 4568–4572 (cit. on p. 80).

[Mat+15] Anna Matamala, Aitor Álvarez Muniain, Andoni Azpeitia Zaldua, and Andreu Oliver Moreno. „The reception of intralingual and interlingual subtitling". In: *Translating and the Computer Conference*. 37. 2015 (cit. on p. 26).

[MD12] Xiaohan Ma and Zhigang Deng. „A statistical quality model for data-driven speech animation". In: *IEEE Transactions on Visualization and Computer Graphics* 18.11 (2012), pp. 1915–1927 (cit. on p. 82).

[MH07] Soh Masuko and Junichi Hoshino. „Head-eye Animation Corresponding to a Conversation for CG Characters". In: *Computer Graphics Forum*. Vol. 26. 3. Wiley Online Library. 2007, pp. 303–312 (cit. on p. 83).

[Mik+11] Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. „RNNLM-Recurrent neural network language modeling toolkit". In: *Proc. of the 2011 ASRU Workshop*. 2011, pp. 196–201 (cit. on pp. 48, 49, 72).

[Min61] Marvin Minsky. „Steps toward artificial intelligence". In: *Proceedings of the IRE* 49.1 (1961), pp. 8–30 (cit. on p. 86).

[Mir+12] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, et al. „Speaker diarization: A review of recent research". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 356–370 (cit. on p. 53).

[MM10] Sylvain Meignier and Teva Merlin. „LIUM SpkDiarization: an open source toolkit for diarization". In: *CMU SPUD Workshop*. Vol. 2010. 2010 (cit. on p. 16).

[Moo+13] Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. „Appraisal theories of emotion: State of the art and future development". In: *Emotion Review* 5.2 (2013), pp. 119–124 (cit. on p. 83).

[Mor+98] Pedro J Moreno, Christopher F Joerg, Jean-Manuel Van Thong, and Oren Glickman. „A recursive algorithm for the forced alignment of very long audio segments." In: *ICSLP*. Vol. 98. 1998, pp. 2711–2714 (cit. on p. 49).

[Muj+13] Andoni Mujika, Helen Diez, Aitor Alvarez, Miren Urteaga, and David Oyarzun. „Realistic visual speech synthesis in WebGL“. In: *Proceedings of the 18th International Conference on 3D Web Technology*. ACM. 2013, pp. 207–207 (cit. on pp. 11, 40, 96).

[Nav+04] Eva Navas, Inmaculada Hernáez, Amaia Castelruiz, Jon Sánchez, and Iker Luengo. „Acoustical analysis of emotional speech in standard Basque for emotions recognition“. In: *Progress in Pattern Recognition, Image Analysis and Applications*. Springer, 2004, pp. 386–393 (cit. on p. 86).

[Net+08] J Neto, Hugo Meinedo, Márcio Viveiros, et al. „Broadcast news subtitling system in Portuguese“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* IEEE. 2008, pp. 1561–1564 (cit. on pp. 30, 47, 56).

[Ofc15] Ofcom. *Code on Television Access Services*. Tech. rep. 2015 (cit. on p. 49).

[Oka07] Naoaki Okazaki. „CRFsuite: a fast implementation of conditional random fields (CRFs)“. In: *URL http://www.chokkan.orgsoftwarecrfsuite* (2007) (cit. on p. 73).

[OW04] Joern Ostermann and Axel Weissenfeld. „Talking faces-technologies and applications“. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. IEEE. 2004, pp. 826–833 (cit. on p. 80).

[Oya+10] David Oyarzun, Andoni Mujika, Aitor Álvarez, et al. „High-realistic and flexible virtual presenters“. In: *Articulated Motion and Deformable Objects*. Springer, 2010, pp. 108–117 (cit. on pp. 11, 37, 96).

[PB92] Douglas B Paul and Janet M Baker. „The design for the Wall Street Journal-based CSR corpus“. In: *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics. 1992, pp. 357–362 (cit. on p. 63).

[Per+10] Elisa Perego, Fabio Del Missier, Marco Porta, and Mauro Mosconi. „The Cognitive Effectiveness of Subtitle Processing“. In: *Media Psychology* 13.3 (2010), pp. 243–272 (cit. on pp. 46, 50).

[Per08] Elisa Perego. „Subtitles and line-breaks: Towards improved readability“. In: vol. 78. John Benjamins Publishing, 2008, pp. 211–223 (cit. on p. 50).

[Pov+05] Daniel Povey, Brian Kingsbury, Lidia Mangu, et al. „fMPE: Discriminatively Trained Features for Speech Recognition.“ In: *ICASSP (1)*. 2005, pp. 961–964 (cit. on p. 48).

[Pov+08] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, et al. „Boosted MMI for model and feature-space discriminative training“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* IEEE. 2008, pp. 4057–4060 (cit. on p. 48).

[Pov+11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al. „The Kaldi speech recognition toolkit“. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. EPFL-CONF-192584. IEEE Signal Processing Society. 2011 (cit. on pp. 19, 26, 48, 70).

[Poz+14] Arantza Del Pozo, Carlo Aliprandi, Aitor Álvarez, et al. „SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling“. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC1́4)*. Reykjavik, Iceland, May 2014 (cit. on pp. 6, 22, 23, 32, 55, 57, 58).

[PP97] Rosalind W. Picard and Rosalind Picard. *Affective computing*. Vol. 252. MIT press Cambridge, 1997 (cit. on pp. 34, 79).

[PR11] Tomas Pfister and Peter Robinson. „Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis“. In: *IEEE Transactions on Affective Computing* 2.2 (2011), pp. 66–78 (cit. on p. 82).

[Qui14] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014 (cit. on p. 86).

[Qui86] J. Ross Quinlan. „Induction of decision trees“. In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 86).

[RÁA14] Pablo Ruiz, Aitor Álvarez, and Haritz Arzelus. „Phoneme similarity matrices to improve long audio alignment for automatic subtitling“. In: *LREC, Ninth International Conference on Language Resources and Evaluation*. 2014 (cit. on pp. 7, 28, 59, 64).

[Raj+13] Dhevi J Rajendran, Andrew T Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. „Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination“. In: *Perspectives* 21.1 (2013), pp. 5–21 (cit. on pp. 46, 50).

[RE13] S Ramakrishnan and Ibrahiem MM El Emary. „Speech emotion recognition approaches in human computer interaction“. In: *Telecommunication Systems* 52.3 (2013), pp. 1467–1478 (cit. on pp. 80, 81).

[RM11] Pablo Romero-Fresco and Juan Martínez. „Accuracy rate in live subtitling. The NER model“. In: *Audiovisual Translation: Taking Stocki* (2011) (cit. on p. 50).

[Rom11] Pablo Romero-Fresco. *Subtitling through speech recognition: Respeaking*. St. Jerome Publishing, 2011 (cit. on p. 50).

[Ros96] Roni Rosenfeld. „A maximum entropy approach to adaptive statistical language modeling“. In: (1996) (cit. on p. 48).

[RTS06] A Rathinavelu, H Thiagarajan, and SR Savithri. „Evaluation of a computer aided 3D lip sync instructional model using virtual reality objects“. In: *Proc. 6th Intl Conf. Disability, Virtual Reality & Assoc. Tech., Esbjerg, Denmark*. 2006, pp. 67–73 (cit. on p. 82).

[Sao+00] George Saon, Mukund Padmanabhan, Ramesh Gopinath, and Scott Chen. „Maximum likelihood discriminant feature spaces“. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings.* Vol. 2. IEEE. 2000, pp. II1129–II1132 (cit. on p. 48).

[SC12] George Saon and Jen-Tzung Chien. „Large-vocabulary continuous speech recognition systems: A look at some recent advances“. In: *IEEE Signal Processing Magazine.* 29.6 (2012), pp. 18–33 (cit. on p. 51).

[Sch+10] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, et al. „Cross-corpus acoustic emotion recognition: variances and strategies“. In: *IEEE Transactions on Affective Computing* 1.2 (2010), pp. 119–131 (cit. on p. 81).

[Sch+11] Bjorn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. „Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge“. In: *Speech Communication* 53.9 (2011), pp. 1062–1087 (cit. on p. 81).

[Sch+13] Bjorn Schuller, Stefan Steidl, Anton Batliner, et al. „Paralinguistics in speech and language. State-of-the-art and the challenge“. In: *Computer Speech & Language* 27.1 (2013), pp. 4–39 (cit. on p. 81).

[Sch03] Klaus R Scherer. „Vocal communication of emotion: A review of research paradigms“. In: *Speech Communication* 40.1 (2003), pp. 227–256 (cit. on p. 81).

[SCP15] David Snyder, Guoguo Chen, and Daniel Povey. „MUSAN: A Music, Speech, and Noise Corpus“. In: *arXiv preprint arXiv:1510.08484* (2015) (cit. on p. 72).

[SEA13] Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. „Speaker variability in speech based emotion models-Analysis and normalisation“. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2013, pp. 7522–7526 (cit. on p. 82).

[SEA15] Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. „Speech based emotion recognition“. In: *Speech and Audio Processing for Coding, Enhancement and Recognition*. Springer, 2015, pp. 197–228 (cit. on pp. 82, 83).

[Set09] Vidhyasaharan Sethu. „Automatic Emotion Recognition: An Investigation of Acoustic and Prosodic Parameters“. PhD thesis. The University of New South Wales, 2009 (cit. on p. 82).

[Sha13] Ismail Shahin. „Speaker identification in emotional talking environments based on CSPHMM2s“. In: *Engineering Applications of Artificial Intelligence* 26.7 (2013), pp. 1652–1659 (cit. on p. 82).

[Sin+13] Sabato Marco Siniscalchi, Dong Yu, Li Deng, and Chin-Hui Lee. „Exploiting deep neural networks for detection-based speech recognition“. In: *Neurocomputing* 106 (2013), pp. 148–157 (cit. on p. 51).

[SJK03] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. „Vocal expression of emotion“. In: *Handbook of affective sciences* (2003), pp. 433–456 (cit. on p. 81).

[SK13] Mark Sinclair and Simon King. „Where are the challenges in speaker diarization?“ In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2013, pp. 7741–7745 (cit. on p. 53).

[SM11] Charles Sutton and Andrew McCallum. „An introduction to conditional random fields“. In: *Machine Learning* 4.4 (2011), pp. 267–373 (cit. on p. 66).

[SR10] Tal Sobol-Shikler and Peter Robinson. „Classification of complex information: Inference of co-occurring affective states from their expressions in speech“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.7 (2010), pp. 1284–1297 (cit. on p. 82).

[SS09] Hagen Soltau and George Saon. „Dynamic network decoding revisited“. In: *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*. IEEE. 2009, pp. 276–281 (cit. on p. 48).

[SS12a] George Saon and Hagen Soltau. „Boosting systems for large vocabulary continuous speech recognition“. In: *Speech Communication* 54.2 (2012), pp. 212–218 (cit. on p. 48).

[SS12b] Ronanki Srikanth and Li Bo2 James Salsman. „Automatic Pronunciation Evaluation And Mispronunciation Detection Using CMUSphinx“. In: *24th International Conference on Computational Linguistics*. Citeseer. 2012, p. 61 (cit. on p. 39).

[SYW13] Michael L Seltzer, Dong Yu, and Yongqiang Wang. „An investigation of deep neural networks for noise robust speech recognition“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2013, pp. 7398–7402 (cit. on p. 51).

[Tac+13] Yuuki Tachioka, Shinji Watanabe, Jonathan Le Roux, and John R Hershey. „Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark“. In: *Proc. CHiME* (2013), pp. 19–24 (cit. on p. 51).

[Tak09] György Takács. „Direct, modular and hybrid audio to visual speech conversion methods-a comparative study.“ In: *INTERSPEECH*. 2009, pp. 2267–2270 (cit. on p. 82).

[Tat+02] Raquel Tato, Rocio Santos, Ralf Kompe, and José María Pardo. „Emotional space improves emotion recognition.“ In: *INTERSPEECH*. 2002 (cit. on p. 86).

[Tay+12] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. „Dynamic units of visual speech“. In: *Proceedings of the 11th ACM SIGGRAPH Eurographics conference on Computer Animation*. Eurographics Association. 2012, pp. 275–284 (cit. on p. 82).

[TBT08] Tomoki Toda, Alan W Black, and Keiichi Tokuda. „Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model“. In: *Speech Communication* 50.3 (2008), pp. 215–227 (cit. on p. 82).

[Teh06] Yee Whye Teh. „A hierarchical Bayesian language model based on Pitman-Yor processes“. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2006, pp. 985–992 (cit. on p. 48).

[Ves+13] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey. „Sequence-discriminative training of deep neural networks.“ In: *INTERSPEECH*. 2013, pp. 2345–2349 (cit. on pp. 26, 70, 72).

[WEF07] Alice Wang, Michael Emmi, and Petros Faloutsos. „Assembling an expressive facial animation system“. In: *Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*. ACM. 2007, pp. 21–26 (cit. on p. 82).

[Wen+14] Chao Weng, Dong Yu, Shigetaka Watanabe, and Biing-Hwang Fred Juang. „Recurrent deep neural networks for robust speech recognition“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 5532–5536 (cit. on p. 51).

[WHS12] Lijuan Wang, Wei Han, and Frank K Soong. „High quality lip-sync animation for 3D photo-realistic talking head“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2012, pp. 4529–4532 (cit. on p. 82).

[WKM06] Wei Wang, Kevin Knight, and Daniel Marcu. „Capitalizing machine translation“. In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics. 2006, pp. 1–8 (cit. on p. 52).

[WL11] Chung-Hsien Wu and Wei-Bin Liang. „Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels“. In: *IEEE Transactions on Affective Computing* 2.1 (2011), pp. 10–21 (cit. on p. 82).

[Wöl+09] Martin Wöllmer, Florian Eyben, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. „Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks.“ In: *INTERSPEECH*. Citeseer. 2009, pp. 1595–1598 (cit. on p. 82).

[Wol92] David H Wolpert. „Stacked generalization“. In: *Neural networks* 5.2 (1992), pp. 241–259 (cit. on p. 91).

[WW04] Lan Wang and Philip C Woodland. „MPE-based discriminative linear transform for speaker adaptation“. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*. Vol. 1. IEEE. 2004, pp. I–321 (cit. on p. 48).

[Xu+13] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. „A practical and configurable lip sync method for games“. In: *Proceedings of Motion on Games*. ACM. 2013, pp. 131–140 (cit. on p. 82).

[You+97] Steve Young, Gunnar Evermann, Mark Gales, et al. *The HTK book*. Vol. 2. Entropic Cambridge Research Laboratory Cambridge, 1997 (cit. on pp. 17, 59).

[You07] Steve Young. „ATK real-time API for HTK, ver. 1.6“. In: *Machine Intelligence Laboratory, University of Cambridge, University of Cambridge, UK* (2007) (cit. on p. 37).

[Yu+08] Dong Yu, Li Deng, Jasha Droppo, et al. „A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* IEEE. 2008, pp. 4041–4044 (cit. on p. 51).

[Zha+14] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. „Improving deep neural network acoustic models using generalized maxout networks“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 215–219 (cit. on p. 48).

List of Figures

List of Tables