eman ta zabal zazu

Universidad del País Vasco / Euskal Herriko Unibertsitatea

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
FACULTY OF SCIENCE AND TECHNOLOGY

Doctoral Programme in Physics Engineering
Department of Electricity and Electronics

ANALYSIS AND AUTOMATIC IDENTIFICATION OF SPONTANEOUS EMOTIONS IN SPEECH FROM HUMAN-HUMAN AND HUMAN-MACHINE COMMUNICATION

Doctoral Thesis

Mikel de Velasco Vázquez

Advisors

Dr. Raquel Justo Blaco
Prof. María Inés Torres Barañano

Leioa, Spain, 2023

Preface

This thesis report follows the modality known as a “thesis by published papers” according to the University of the Basque Country (UPV/EHU) regulations[1]. It is a particular option for presenting and defending a PhD thesis, which requires the candidate to compose a series of scientific papers centred around a single theme, of which they must provide proof of authorship and publication.

In order to meet the requirements for a thesis by published papers, the PhD student must fulfil specific criteria related to their research contributions. Specifically, they must have three published or accepted contributions during their doctoral program. All of these contributions must be centred around a single research theme and must be published in reputable scientific journals, such as those listed in the latest Journal Citation Reports or SCOPUS databases or in databases recognized by the National Assessment Committee for Research Activities (CNEAI). Additionally, at least one of these contributions must fall within the first or second quartile of its category.

A PhD thesis by published papers must consist of a set of papers following this structure:

- **Synthesis:** This section provides a detailed overview of the thesis and emphasises the coherence of the papers’ themes. The thesis’s central theme is introduced, along with the methodology employed. The general and specific hypotheses and the research objectives are clearly defined, and publications related to each objective are indicated. Additionally, the section includes a summary of the results and a detailed discussion of their implications. Finally, the section concludes with a comprehensive list of all sources cited throughout the thesis.
- **Conclusions:** This section presents the final conclusions of the thesis.
- **Contributions:** This section includes the published works under the title “A Compilation of Published Articles”. The complete version of each contribution is included. The complete list of bibliographic references and relevant quality indicators, such as the journal’s impact factor and its relative position in its category, is also included. The report include abbreviations to refer to each of the compiled papers, namely P1, P2, and so on.

Specifically, this document is divided into three parts. **Part I** encompasses the Synthesis and Conclusions. **Part II** includes a single Appendix which contains supplementary material, yet to be published, that offers insight into the thesis. This material is relevant for a comprehensive understanding of the synthesis. Finally, **Part III** encompasses the Contributions made in this study, i.e the published articles.

[1] https://www.ehu.eus/en/web/doktoregoa/doctoral-thesis/thesis-by-published-papers

Contents

List of Figures

List of Tables

List of Abbreviations

Part I.

Synthesis: Summary of Contributions and Conclusions

Chapter 1

Introduction

Emotions arise from psychological and physiological factors and can impact decision-making, relationships, and well-being. They differ across cultures and individuals but play a fundamental role in shaping our perception of the world. Emotion recognition in machine learning involves developing algorithms to identify and classify emotions expressed in spoken language accurately. Challenges include the inherent variability and subjectivity of emotional expression, handling cultural and individual differences in speech patterns, and addressing the limitations of available training data. Successful recognition has applications in human-computer interaction, virtual assistants, and mental health support.

Emotion theories agree that an emotional episode consists of several components, such as the process of a stimulus, motivation for actions, central and peripheral physiological responses, behaviour (e.g., facial and vocal expressions, among others), and subjective experiences or feelings (Moors, 2012; Statharakos et al., 2022). In addition, the expression of emotions depends on the particular person and the specific scenario because the stimulus and the behavioural and physiological expression of emotions differ from one scenario to another and from one person to another.

Emotions can be expressed differently, including facial expressions, speech and gestures. In this work, we focus on analysing and developing systems that can detect emotions from speech. *Speech* can be defined as human vocal communication using language (Lleida and Rodriguez-Fuentes, 2018), and it is inseparably linked with emotional status during the cognitive process of human communication (Li et al., 2022, 2023). Furthermore, it seems to be a good indicator of depression (Kiss and Vicsi, 2017), valid for psychosomatic monitoring (Hayette Hadjar et al., 2021), very related to emotional status or even to Parkinson's disease (Sztahó et al., 2017; Jeancolas et al., 2021).

The expression and perception of emotions are very important issues in human interactions and one of the bases upon which human communication is established. Therefore, the automatic detection of emotions by a computer has become a very attractive topic due to its impact on more natural and empathic Human-Machine Interaction (HMI) systems. In fact, during the last few years, the Scientific Community has shown an increasing interest in affective computing and its potential capability to change how HMI is carried out by better understanding Human-Human Communication (HHC). In recent years, the significant amount of multimedia information available and new computational methodologies related to machine learning have led the scientific community to put a great effort into this area (Irastorza and Torres, 2016; Eskimez et al., 2016; Scibilia et al., 2022).

In the literature, a significant number of studies have been conducted on the identification of the emotional status of the speaker, with a focus on a limited set of acted emotions (Kim and Clements, 2015; Eskimez et al., 2016; Schuller et al., 2011; Horvat et al., 2022). Specifically, Eckman's basic set of emotions (Ekman, 1999) is usually employed. In this way, a considerable amount of labelled data can be obtained to train machine learning algorithms with a little effort. Moreover, corpora can be reused, and the results achieved with different models can be easily compared. However, the emotions found in real scenarios are different. The intensity of spontaneous emotions is generally lower than the one of acted emotions, resulting in a smaller emotional space where emotions are closer, as shown in Chakraborty et al. (2017) and also in deVelasco et al. (2022b) (in P15). Thus, the automatic recognition of spontaneous emotions from speech results in a very complex task. In fact, the surface expressions of spontaneous emotions differ from those displayed in acted emotions, as noted by Schuller et al. (2019). This discrepancy makes it challenging to directly apply research findings based on acted emotions and to use acted data for training purposes. Furthermore, the set of emotions that appear in each specific real scenario is task-dependent and, thus, the related automatic detection is.

An additional challenge of spontaneous emotions is the labelling procedure since the emotional status of the speaker cannot be unequivocally established. The emotional label assigned by a speaker to his utterance might differ from the one assigned by a listener to the same utterance, being the first one closer to the current emotion, as shown in Chakraborty et al. (2017) and deVelasco et al. (2022a) (in P16). However, speaker self-annotation is only occasionally a feasible approach. Consequently, the annotation of utterances in terms of spontaneous emotions is generally carried out through perception experiments, which are based on the particular judgment of every annotator. Therefore, the disagreement among annotators and the distance between the emotion expressed and the perceived emotion can be significant. In contrast, if emotions are expressed by professional actors or just elicited, the annotation procedure is not required (Bänziger et al., 2012). In such a case, the actor's intent always labels the generated emotion.

HMIs emerge within specific contexts where individuals have some degree of familiarity with each other. However, current artificial agents cannot imitate a real user resulting in shallow interactions (Vinciarelli et al., 2015). Users find it hard to interact with agents with rudimentary visual and speech capacities, as shown by Chiba et al. (2017) and also by Justo et al. (2020) (in P8). The literature suggests that a human's behaviour guides other human's behaviour in HHCs and is thus a reactionary behaviour (Aïsha et al., 2017). However, comparisons between these two scenarios have almost exclusively been conducted at the interaction and dialogue levels (Vinciarelli et al., 2015; Aïsha et al., 2017; Firdaus et al., 2023) without addressing emotional differences. When it comes to emotional levels, individuals seem to exhibit less intense emotional responses when interacting with a machine than with other humans due to the limited emotional capacity of the artificial agent, which results in more neutral emotional expressions.

Through multiple projects, the primary objective of this doctoral research was to identify emotional states in various real-life situations. Some corpora were acquired and then annotated using crowdsourcing[1] methods as well as with the help of expert annotators. Furthermore, the created corpora were analysed to understand better spontaneous emotions in speech and the emotional space related to each corpus. These two objectives have been crucial in laying the foundations for developing computer models that accurately identify emotional states in real-life situations.

The development of computational models capable of accurately detecting emotional states has been conducted during the doctoral period building on the objectives mentioned above. To this end, this thesis report is structured into the following chapters in **Part I**:

- **Corpora definition, annotation and analysis:** This chapter consists of several sections. The first section presents an overview of the categorical and dimensional emotional models, while the second section investigates two annotation methods, namely crowd and expert. The third, fourth, and fifth sections provide a detailed analysis of the three specific corpora we have worked on: "La Sexta Noche" (HHC in Spanish), "EMPATHIC" (HMI in Spanish, French and Norwegian), and "MENHIR" (HHC in English). This analysis covers essential aspects such as the investigation of the real emotional spaces, the annotation process, inter-annotator agreement, and the annotated data's quality.

[1]Crowdsourcing involves a large group of participants producing goods or services (including ideas, votes, micro-tasks, etc.) for payment or as volunteers.

- **Methods and Models:** In this chapter, the focus is on the methods and models that have been used to identify emotion in speech. First, speech signal representations, including acoustic features, spectrograms, and wave embeddings, are discussed. Second, the chapter presents several classification and regression models across several neural network architectures, model versions and configurations. In addition, some issues of the emotion recognition task are discussed.

- **Experimentation:** This chapter focuses on designing and evaluating speech-emotion recognition systems. The chapter includes a variety of experiments to analyse and evaluate different models of emotions in the heterogeneity of real scenarios, a variety of speech representations, network architectures, training configurations and a diversity of involved hyperparameters, which are published and included in **Part III** (from P1 to P19). However, they aimed to make local comparisons between some specific aspects and thus are hardly comparable. Thus, additional, unpublished and comparable results are presented in the Appendix of **Part II**. They offer an overview of the advantages and disadvantages of the tested methodologies [2].

- **XAI: Analysis of system behaviour:** This chapter presents some methods to analyse and understand the decision-making process in a neural architecture to identify emotions in speech. To this end, we first build and evaluate a simple convolutional neural model. Then, the chapter presents a layer-by-layer analysis of the model's behaviour and proposes a class model visualisation using deep dream techniques.

- **Conclusions:** This chapter presents the final conclusions of the thesis.

Then, **Part II** includes a single Appendix which contains supplementary material, yet to be published, that offers insight into the thesis.

The research findings presented in this report have been published in five journals indexed in JRC (1 Q1, 2 Q2 and 2 Q3). Also, an ongoing manuscript is aimed to be submitted to a Q1 JRC journal. Moreover, the work has also been published in ten conference papers. The list of publications can be found at the end of this report in **Part III**.

[2] As this is a "Thesis-by-published papers" and these experiments have not still been published they are described in the Appendix instead of in a Chapter

Chapter 2

Emotion, corpora, annotation and analysis

Identifying human emotions is a task commonly referred to as emotion recognition. Humans perform this process effortlessly by considering various cues such as facial expressions, body language, and verbal cues. However, implementing this process in an automated system presents significant challenges.

This chapter begins with an overview of categorical and dimensional models of emotions. The second section presents two corpus annotation methods: crowd and expert annotation. Finally, sections 3, 4, and 5 provide an in-depth analysis of three distinct corpora, which have been developed in collaboration with different partners: "La Sexta Noche", "EMPATHIC", and "MENHIR". The analysis includes critical aspects such as annotation procedures, inter-annotator agreement, and the quality of annotated data.

1 | Categorical and Dimensional models of emotions

The literature on Affective Computing shows that different models can represent emotional states. Most works in this area adopt the categorical model of emotions that involves discrete labels, such as boredom, frustration and anger (Calvo and D'Mello, 2010; Calvo and Mac Kim, 2013). This framework is commonly based on Ekman's proposal of the "Big Six" emotions, consisting of six basic emotions that are considered to be biologically universal and present in all cultures: surprise, disgust, sadness, anger, fear, and happiness (Ekman, 1999; Prinz, 2004). However, other works exist, such as the "Hourglass of Emotions" (Susanto et al., 2020), which further contribute to the understanding and exploration of a broader range of categorical emotions.

However, the specific emotions in a particular real-world scenario can vary significantly. For instance, in a call centre setting, the goal may be to recognise anger through a binary classification of anger/no anger (Pappas et al., 2015) or to identify levels of annoyance activation (Irastorza and Torres, 2016, 2019) in customer service calls.

Moreover, research on emotions has revealed that regular communication involves many complex feeling states that cannot adequately represent using a limited set of categories (Cambria et al., 2022). As a result, some researchers have proposed a dimensional representation of emotions. This representation involves using a two-dimensional space (Schlosberg, 1952), with Valence representing the polarity of the emotion and Arousal representing the degree of excitement. Some authors have also extended this model to three dimensions by adding Dominance, representing the degree of control over a situation (Schlosberg, 1954). This approach aims to improve our comprehension of human emotions by representing them as a few continuous signals, Valence, Arousal, and Dominance, which results in the VAD model. This methodology offers multiple advantages, such as encoding slight emotional changes over time and differentiating between subtly different emotional expressions (Valstar et al., 2014). Due to its versatility, in recent years, dimensional models have been widely used in emotion recognition (Gunes and Pantic, 2010; Schuller et al., 2011).

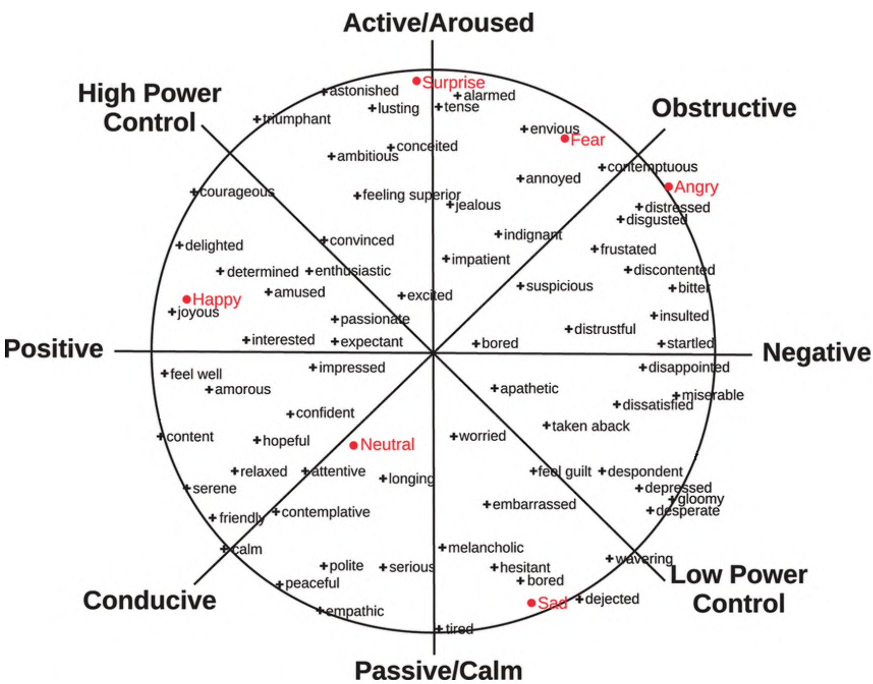


Figure 2.1.: Illustration of Scherer's circumplex (Scherer, 2005), which shows categories represented by the dimensions arousal and valence.

According to the Core Affect theory, the two models of emotions, categorical and dimensional, are closely related (Russell, 1980, 2003). This theory posits that each emotion category can be represented in the dimensional model as a point or area in the two-dimensional space. For example, Sherer's circumplex (Scherer, 2005) uses the Arousal/Valence two-dimensional model to represent categories of emotions (Figure 2.1).

In this thesis, both the categorical and the dimensional models were considered to represent the emotional state of the speakers.

2 | Annotation procedures

Supervised learning needs a reference, namely ground truth, to train computational models. These supervised methods use many labelled data to produce robust patterns. However, the labelling process requires effort, time and an appropriate budget. Moreover, labelling emotion-related data involves a perception experiment that depends on socio-cultural aspects (Mesquita and Leu, 2007; Averill, 2015; Vea, 2020; Riviello et al., 2012), adding subjectivity and complexity to these complicated tasks. Thus, the labelling procedure could lead to a low-quality annotated corpus due to the socio-cultural differences between the annotator and the speaker.

Two different annotation procedures were used to capture the perceived emotion in the audio files. The first procedure consists of segmenting the audio with automatic tools and annotating them with the crowd (Section 2.1). In the second procedure, a reduced set of expert annotators make the segmentation and provide the corresponding label to each segment (Section 2.2). In both cases, the perception of multiple annotators has been considered for each audio recording.

2.1 | Crowd Annotation

Before carrying out the crowd annotation, an audio segmentation methodology was implemented. The proposed solution for automatic segmentation was based on selecting utterances that were compatible with clauses to be annotated. The clause, defined as "a sequence of words grouped together on semantic or functional basis" (Esposito et al., 2008, 2004), was used as the unit of work under the hypothesis that the emotional status does not change within a clause.

An algorithm was developed to extract clauses from speech signals. This algorithm, described in deVelasco et al. (2022b) (in P15), considers factors such as silences, pauses, and text transcriptions. The algorithm returns audio segments that range from two to five seconds and correspond to the clauses under consideration.

The annotation of the extracted segments is challenging due to the subjectivity of the emotion labelling task (Ortony and Turner, 1990), making it challenging to obtain ground truth. This ambiguity may lead to disagreement between opinions, such as the ones arising from the listener's perception or their specific cultural environment (Gurney, 1884; Scherer, 1999). Therefore, as suggested by Aroyo and Welty (2015), the notion of a single "correct" truth is outdated in specific contexts and needs to be re-examined. They propose using "crowd truth", which is based on the idea that human interpretation is subjective and that measuring annotations on the same objects of interpretation across a crowd will provide a valuable representation of their subjectivity and the range of reasonable interpretations.

Considering this, crowdsourcing is an appropriate method for labelling since it allows us to consider as many opinions as possible. This method has emerged as an alternative to traditional labelling and is widely used in different fields, including information retrieval and natural language processing. Various crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk)[1], SamaSource[2], and CrowdZientzia[3], are available. The CrowdZientzia platform, developed by the University of the Basque Country (UPV/EHU) research group, was used in this research.

When considering the categorical model, each annotator assigned an emotion category to each audio segment. In the case of the dimensional model, emotions were discretised for the labelling process, making it similar to the one used for the categorical model. Consequently, a new numerical label was extracted in the vector space by averaging the set of labels the crowd annotators provide for each segment. This way, the annotation procedure led to a representation of the emotion as a continuous value.

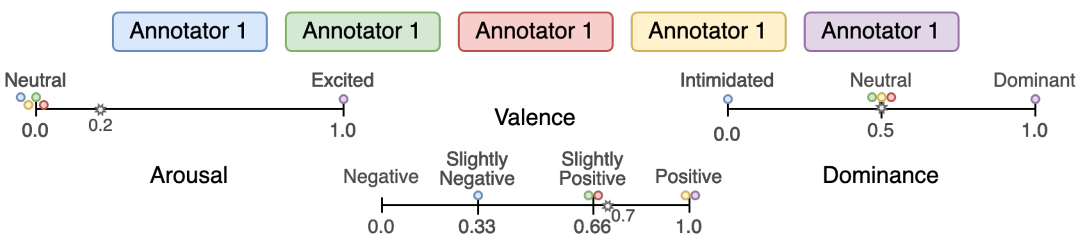


Figure 2.2.: Crowd dimensional annotation procedure.

[1] www.mturk.com
[2] www.samasource.org
[3] crowdzientzia.ehu.eus

For example, Figure 2.2 shows how five different annotators (each with a different colour) chose a label for each dimension (each coloured ball represents the annotation). The grey point represents the average of 5 annotations, resulting in a continuous value in the vector space representing the final annotation.

2.2 | Expert annotation

Expert annotation is one of the most widely used annotations in artificial intelligence. This section describes the expert annotations methodology for labelling emotions in long speech sessions.

This labelling process requires that the annotators perform the segmentation and the labelling simultaneously. Thus, annotators' disagreement might appear in the labelling and audio alignment processes. In order to have fixed-length segments that can be processed by a system working in real-time, 3-second windows were considered with a 1-second overlapping, as illustrated in Figure 2.3.

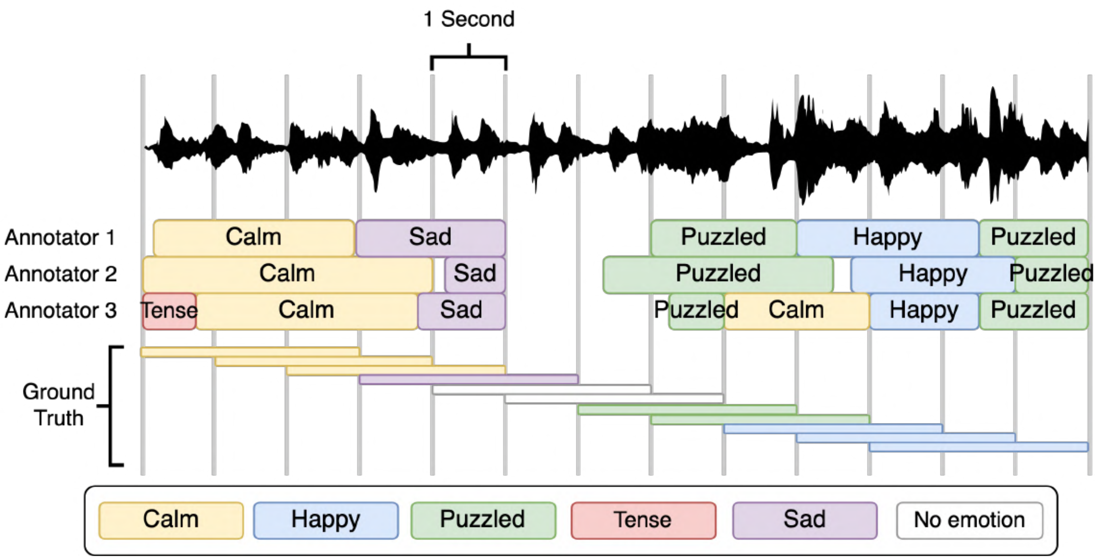


Figure 2.3.: Multiple expert annotation procedure. Each of the colours represents a different category chosen by annotators.

The labels of the three annotators were considered to assign an emotional label to each segment. Each audio segment was assigned the emotion that received the most votes as long as it covered a specific percentage (defined by a threshold) of time of that segment. Oppositely, the segment was left without annotation if none of the emotions reached the percentage threshold. Each segment was annotated with a categorical label and three-dimensional labels for Valence, Arousal, and Dominance.

3 | La Sexta Noche (L6N)

This corpus was developed by a consortium of Spanish Universities in the framework of AMIC's "Affective Multimedia Analytics with Inclusive and Natural Communication" project (Ortega Giménez et al., 2018). The partners are the University of the Basque Country (UPV/EHU), the Polytechnic University of Valencia, the Polytechnic University of Madrid and the University of Zaragoza.

This project tackles a specific scenario of Human-Human interactions from a Spanish TV show for emotion detection. In this context, a corpus was built collecting real Human-Human conversations extracted from the Spanish TV program "La 6 Noche"[4].

In this weekly broadcast show, two moderators lead social and political debate panels addressing the week's hot news. There is an extensive range of talk-show guests, such as politicians and journalists, who analyse, from their perspective, social topics. Their interventions are mixed with edited videos and research reports. People in the set can give their opinion regarding the current discussion topics, and also, people following the program at home using social networks. The topics under discussion are usually controversial, so emotionally rich interactions can be expected.

However, the participants are used to speaking in public, so they do not lose control of the situation. Even if they might overreact sometimes, it is a real scenario where emotions are subtle. The spontaneity in this situation makes a significant difference from scenarios with acted emotions, as shown in de-Velasco Vazquez et al. (2019) (in P7). The selected programs were broadcasted during the electoral campaign of the Spanish general elections in December 2015.

It is worth noting that TV shows are broadcasted to the general public, and their format framed with specific roles, such as moderator and guest, affects the emotions expressed and their intensity. Table 2.1 shows a small excerpt of a dialogue taken from the corpus.

[4] ATRESMEDIA, producer and owner of the copyright of LaSextaNoche program's contents, provided the consortium with the rights to use the audio files only for research purposes.

Table 2.1.: Excerpt from L6N corpus showing a discussion about politics between two guests, providing a rich emotional example. The excerpt is displayed in both Spanish (the original language) and English.

	Spanish
Speaker 1:	Yo entiendo que de España y de datos y de hechos no quieras hablar, pero resulta...
Speaker 2:	Claro que puedo hablar...
Speaker 1:	Que acaban de imputar también al quinto tesorero en la historia de tu partido.
Speaker 2:	Y dale.
Speaker 1:	De cinco...
	English
Speaker 1:	I understand that you do not want to talk about Spain, about neither data nor facts, but it turns out...
Speaker 2:	Of course I can talk...
Speaker 1:	That they have just imputed the fifth treasurer in the history of your party as well.
Speaker 2:	And hit it.
Speaker 1:	Five out...

3.1 | Annotation procedure

Each TV program consists of a video and audio recording of approximately six hours, which includes the program's headline, advertisements, and other irrelevant content. The algorithm mentioned in Section 2.1 extracted 5,500 segments 2-5 seconds long. According to our assumptions, these segments corresponded to clauses, containing a constant emotional state.

Additionally, another corpus (referred to as L6N interventions) was created with 2964 audio segments. The audio segments are longer in this case, they have the duration of the speakers' interventions, and the transcription has been added as additional information.

A crowd annotation was carried out, using the methodology explained in Section 2.1, to get emotional labels for the audio segments. Each audio segment was labelled by five different annotators using the questionnaire from Table 2.2. It encompassed questions related to the three-dimensional VAD and categorical models and an additional question to detect poor audio quality, such as music or overlaps.

As a result of previous experiments, the questionnaire was refined to its current version, shown in Table 2.2. In the process, some questions and possible answers were modified. The categorical model used in this questionnaire was adapted from the emotions identified by Cowen and Keltner (2017) to fit the specific task.

Table 2.2.: Questionnaire used for annotating the L6N corpus.

1	How do you perceive the speaker?	• Excited • Slightly Excited • Neutral	
2	His/Her mood is:	• Positive • Slightly Positive • Neutral • Slightly Negative • Negative	
3	How do you perceive the speaker in relation to the situation in which he/she is?	• Rather dominant / controlling the situation • Rather intimidated / defensive • Neither dominant nor intimidated	
4	Select the word that you think better describes the speaker's mood:	• Embarrassed • Bored/Tired • Disconcerted/Surprised • Angry • Interested	• Satisfied/Pleased • Worried • Enthusiastic • Annoyed/Tense • Calm/Indifferent
5	If none of the above emotional states applies, indicate the state you see as most appropriate. (Open Answer)		
6	Quality of the audio	• Correct • Overlapping of several speakers • Advertisement • Others	

3.2 | AGREEMENT AND QUALITY OF ANNOTATED DATA

The 5500 extracted segments were annotated using the crowdsourcing platform CrowdZientzia, where five annotators annotated each segment. Thus, a total of 27500 micro-task were created, one per each segment and annotator, and 136 annotators were involved in the annotation procedure. The commonly used method to identify potentially unreliable annotations involves incorporating predefined *gold-standard* micro-tasks alongside regular microtasks (Rothwell et al., 2015). However, in our case, the task is subjective, and the *gold-standard* method makes no sense because choosing unambiguous cases is impossible. Other solutions have been developed for cases where the *gold-standard* methodology is inappropriate, such as using an algorithm based on majority voting to control quality (Ipeirotis et al., 2010; Raquel Justo, 2017). We used the following measures to detect fraudulent workers according to the work described in deVelasco Vázquez et al. (2019) (in P4).

Data Quality Measures

A set of measurements was carefully selected and implemented to ensure the accuracy and reliability of the results. The work deVelasco Vázquez et al. (2019) (in P4) provides a comprehensive explanation of each measurement and method used. In summary, the following measurements were used:

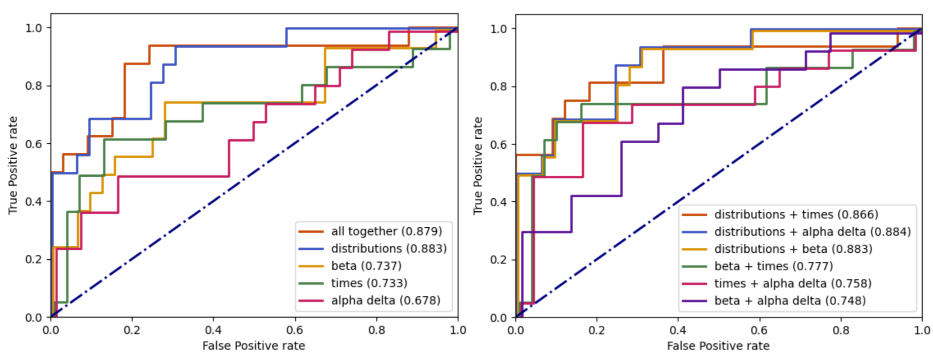
- **Measurements based on labelling time:** The average and the standard deviation of time spent to perform a micro-work.
- **Measurements based on Inter-Annotator Agreement (IAA):** The Delta of Krippendorff's α value ($\Delta\alpha$) (Krippendorff, 2004) and the β_l measure proposed in Raquel Justo (2017).
- **Measurements based on user's answers and the overall answer distribution:** The distance calculation that takes into account the ratio of the highest and lowest values for each distribution element, along with the Euclidean distance.

Each of the metrics provides a score for each annotator. These scores are then analysed to identify which annotators deviate significantly from the overall average in each evaluation metric. This process is essential to determine which annotators may not perform the task accurately. By identifying these outliers, the overall reliability and accuracy of the annotation process can be improved, and the final results will be more trustworthy.

Validating Data Quality Measures

An experiment was conducted to compare the different measures. Specifically, we selected 45 workers who ranked among the 20 worst-performing workers in some metrics we studied, many of whom performed poorly across multiple measures. We manually examined the work of each of these workers with a pair of evaluators and discussed whether to classify their work as fraudulent or suitable. As a result, we identified six fraudulent workers and ten workers whose classification was uncertain due to issues with some of the questionnaire questions.

Then we trained several linear classifiers that analysed 45 workers (6 fraudulent and 39 suitable) to determine which measures provide the most information for detecting bad annotators. We experimented with adjusting the classifiers to different measures and evaluated their performance using the ROC curve and the surface below the ROC curve.

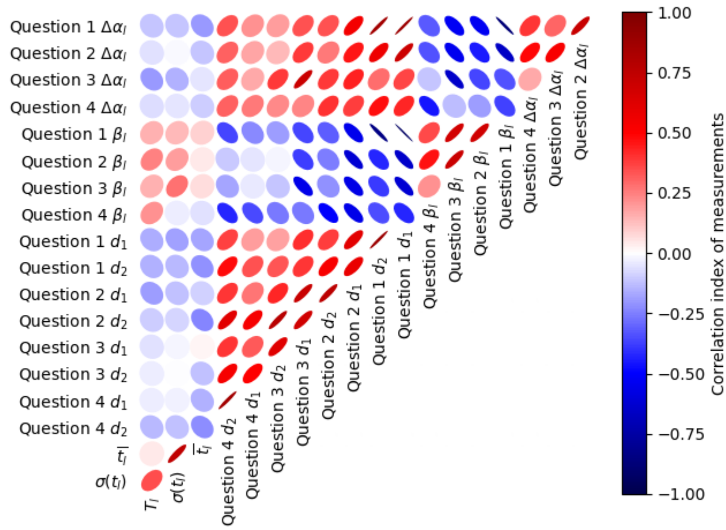


(a) Individually (b) By pairs

Figure 2.4.: The ROC curve is drawn from different sets of measures.

Figure 2.4a shows that the most prominent surface under the ROC curve was the *distributions* set with a value of 0.883. Below it is the *beta* set (0.737), the *time* set (0.733), and finally, the *alpha delta* set (0.678). However, by adding up all the sets, we can see that we improved the scores with a value of 0.879. In Figure 2.4b, we combined the clusters created in pairs to verify that the new clusters are better than individual clusters.

Figure 2.5.: Correlation between all the studied measurements



On the other hand, we represented the correlation between all the measures in Figure 2.5. The narrower and more coloured ellipses indicate an increasing correlation between the two measures.

In conclusion, the measures presented in this section are highly effective in detecting the bad attitudes of workers or annotators. These measures allow a thorough data analysis, providing a detailed understanding of the quality of the results and the potential errors. Furthermore, the measures can also be applied in similar research projects in the future, making them a valuable tool for validating data quality. In a field where data quality is essential, these measures provide a reliable and efficient way of assessing and improving the reliability of the results.

3.3 | Analysis of the annotations

The annotations will be first analysed regarding the categorical model. Five annotators provided a label for each segment, and a majority voting criterion was implemented to decide the final label. However, a minimum agreement of 60% was required to assign a specific category to a sample. This approach ensured that at least 3 of 5 annotators provided the same label to an audio segment. From 5500 audio segments, only 4118 fulfilled this requirement.

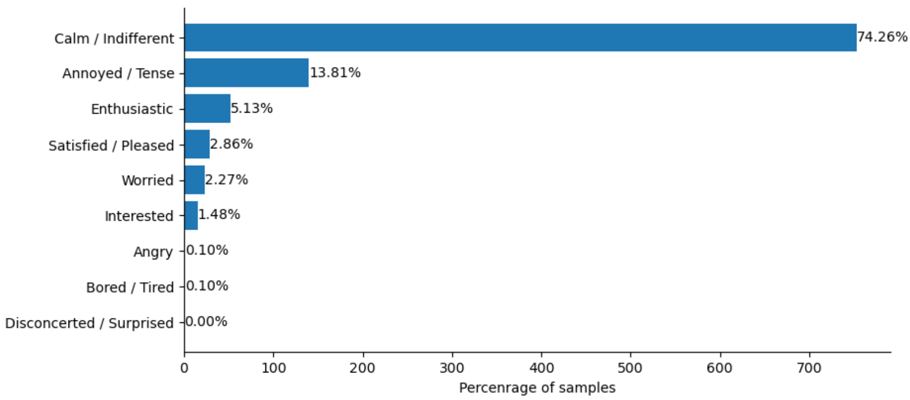


Figure 2.6.: The frequency of the annotations in each category in the categorical model of emotions with an agreement grater than 60%.

As shown in Figure 2.6, the category "Calm / Indifferent" is the predominant one, encompassing nearly 75% of the samples. The remaining categories fall far behind it, with "Annoyed / Tense" at around 14%, "Enthusiastic" at 5%, and so on. This suggests that most audio samples can be described as having a calm emotional state, which can be expected when dealing with spontaneous tasks.

On the other hand, an analysis of the annotations was conducted using the VAD (Valence, Arousal, and Dominance) dimensional model. The answers of the annotators to the first three questions from the questionnaire in Table 2.2 were used to represent each audio chunk in a 3D space, averaging the value of the five annotations. This approach provided a dimensional representation of the emotional state of the audio segments.

Figure 2.7 shows three projections of the points obtained in the 3D space, namely arousal/valence, arousal/dominance and valence/dominance. These figures show that speakers are neutral regarding arousal but tend to be slightly positive and dominant. This observation correlates well with the nature of the task, in which speakers assertively express themselves without getting excited or angry.

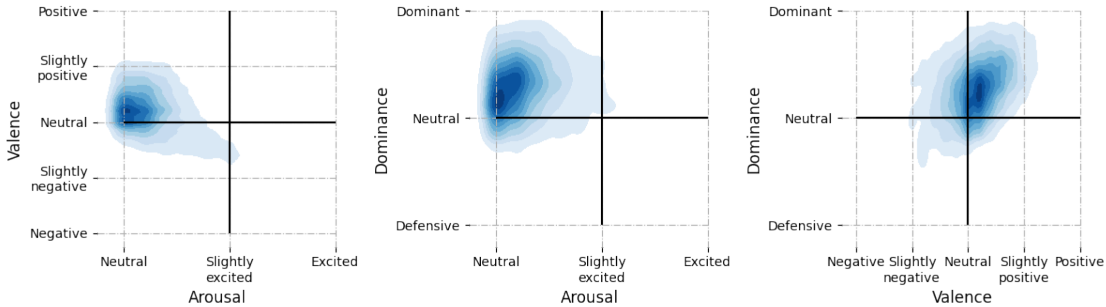


Figure 2.7.: Three projections of the data in the VAD space.

Later, the average of the valence and arousal values for all audio samples labelled with a specific category were calculated. The resulting point is depicted in Figure 2.8 as a circle, with each colour representing a different category. The triangles in the figure correspond to the theoretical position of the same category according to the map presented in references Scherer (2005) and A. Russell (1983).

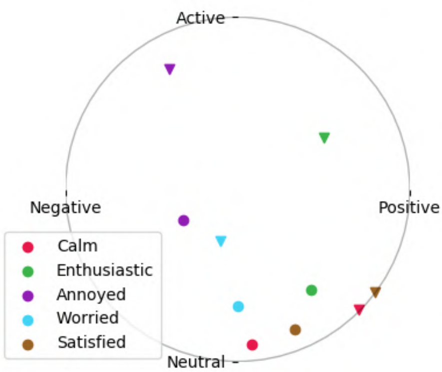


Figure 2.8.: Categorical average of the valence and arousal (dots) vs associated theoretical value Scherer (2005); A. Russell (1983) (triangles).

This map illustrates the relationship between discrete categories and their representation within a valence/arousal space. Based on this figure, the arousal level is consistently lower for realistic emotions compared to the theoretical values. Additionally, the valence tends to be more neutral overall. Moreover, it is observed that the space occupied by the circles representing spontaneous emotions is smaller than those occupied by the triangles representing theoretical emotions. This observation suggests that emotions in spontaneous scenarios are more subtle, reflected by the reduction of the space associated with them.

Considering the low number of samples in some categories and their location (very close to each other) in Figure 2.8, some emotion categories were removed or fused, resulting in a set of only three different categories: calm/indifferent, interested and worried under the label “CALM”; enthusiastic and satisfied were grouped under the label “ENT”; and Annoyed/tense with the label, “ANN”.

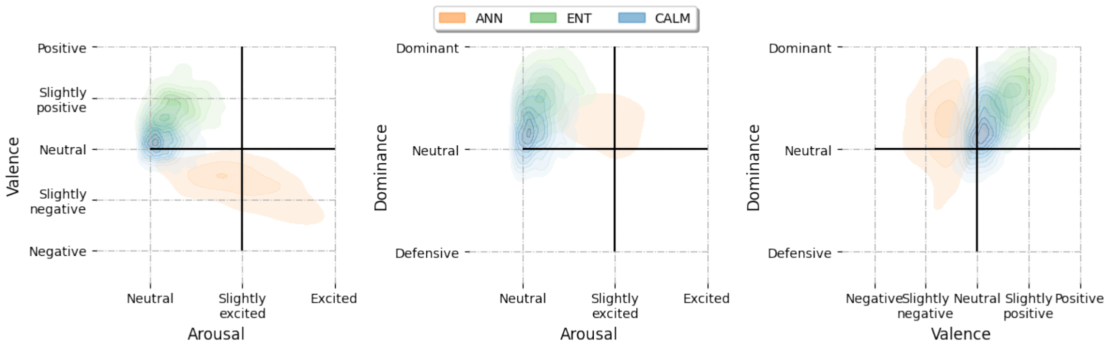


Figure 2.9.: 2D projections of 3D spaces.

The projection of the resulting categories in the VAD space can be observed in Figure 2.9. “CALM” is primarily centred at Neutral for Valence and Arousal, with a tendency towards Dominant. “ENT” is centred at Slightly Positive for Valence and located between Neutral and Slightly Excited for Arousal and mostly Dominant for Dominance. “ANN” is centred at Slightly Negative for Valence and located between Neutral and Excited for Arousal, with the most spread distribution for Dominance between Defensive and Dominant. These findings reflect the nature of the data, where speakers aim to maintain a calm mood and control over the situation.

The most natural way of tackling the VAD representation is to consider a regression problem. However, due to the limited data available in this study, it was decided to discretise each dimension, as shown in Figure 2.10, to simplify the task into a classification problem and make it more manageable. The frontiers were chosen manually to ensure a well-balanced dataset that makes sense regarding the resulting classes. For doing that, the probability density functions of the annotations associated with each dimension are given in Figure 2.10. The arousal value was divided into Neutral and Excited, with a split at point 0.25. Valence was divided into three categories: Negative, Neutral, and Positive, with splits points at 0.4 and 0.6. Dominance was divided into Neutral and Dominant, with 0.75 being the dividing point. Dominance values below 0.25 were not considered because they have different meanings than Neutral values, and there were very few samples with these values.

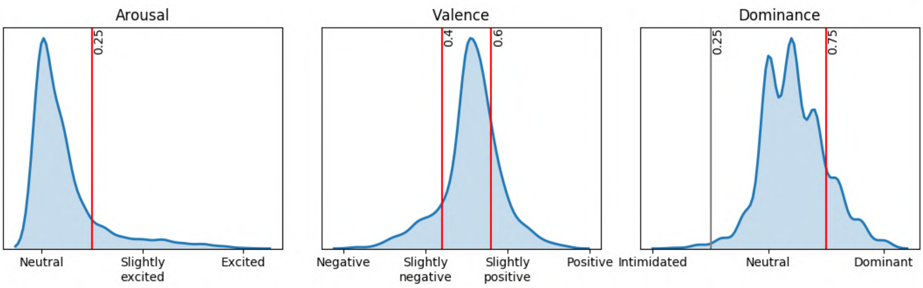


Figure 2.10.: The Probability Density Function for each VAD dimension, along with the chosen cutting points for discretization.

4 | EMPATHIC (EMP)

The EMPATHIC[5] (Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly) project (Torres et al., 2019b,a; Brinkschulte et al., 2021) in P1 was founded by the European Commission H2020-SC1-2017-RIA grant number 769872. The European Consortium consisted of 10 partners from 6 EU and associated countries, which included 5 universities or research centres, 2 healthcare centres related to institutions, and 3 companies. The University of the Basque Country (UPV/EHU), particularly Prof. M. Inés Torres, coordinated the project.

This project aims to design and validate new interaction paradigms for personalized Virtual Coaches to promote healthy and independent ageing. To this end, the Virtual Agent implemented coaching sessions to make older adults aware of unhealthful habits (Torres et al., 2019b; López Zorrilla et al., 2018; Justo et al., 2021). The conversations follow the GROW coaching model[6] to help users make positive behavioural changes (Torres et al., 2019a; Justo et al., 2020). These conversations were related to four main topics: leisure, nutrition, physical activity and social and family relationships. In this context, a module was developed to detect the user's emotional state during the coaching session. This information will enable the system to respond empathetically to the user's emotional needs.

[5] www.empathic-project.eu

[6] The GROW coaching model is a framework designed to help individuals, teams, and organizations achieve their goals through a structured approach. It stands for Goal, Reality, Options, and Will.

4.1 | Data acquisition process

A Wizard of Oz (WoZ) platform (Schlögl et al., 2014, 2010) was used to acquire the database. The WoZ constitutes a method that uses a human operator (the so-called wizard) to simulate a non-existing system. This way, senior users could interact with a simulated system and provide valuable input for future developments, such as audiovisual data used to train the machine learning models.

The project methodology involved the acquisition of a corpus consisting of video recordings collected during user interactions with the Wizard of Oz. The conversations were carried out on several topics in three different languages: Spanish, French, and Norwegian. This database is part of the Empathic database distributed by ELRA[7].

Table 2.3 presents an example of a conversation about nutrition between a user and the WoZ, where some emotional content can be found. A detailed overview of the annotated data can be found in Justo et al. (2020) (in P8) and Justo et al. (2021) (in P11).

Table 2.3.: Small excerpt taken from the *Empathic VA* corpus. This sample is a segment of a conversation between a user and the WoZ, where the WoZ asks questions trying to obtain extensive answers from the user.

	Spanish
User:	Para cenar suelo tomar o huevos con... o pescado con algo de también de verdura, y luego suelo tomar o tomo algunas frutas, suelo tener fruta. Y luego pues no tomo nada más.
System:	Sueles merendar?
User:	A veces sí, depende... A veces sí meriendo y suelo merendar pues un...un café y algún dulce o algún bocadillito de jamón también suelo tomar. Depende de la ocasión, si estoy en casa o no.
System:	Entiendo, ¿Crees que comes bien?
User:	Creo que sí.
System:	Dime Ana, ¿Hay algo que te gustaría cambiar en tu forma de comer?
User:	Pues nada porque como a gusto, no... no quiero otra cosa.
	English
User:	For dinner, I usually have either eggs with ... or fish with some vegetables too, and then I usually take or I take some fruits, I usually have fruit. And then, well, I do not take anything else.
System:	Do you usually have a snack?
User:	Sometimes, yes, it depends... Sometimes, yes, I take a snack and I use to have a snack; let us say a...a coffee and I also used to take some sweets or some small ham sandwich. It depends on the occasion, if I am at home or not.
System:	I understand. Do you think you eat well?
User:	I think so.
System:	Tell me Ana, Is there anything you would like to change in the way you eat?
User:	Well, nothing because I eat at ease; no, ... I do not want anything else.

[7] http://catalog.elra.info/en-us/repository/browse/ELRA-S0414/

4.2 | EXPERT ANNOTATION

Nine annotators were hired for this emotion annotation task: three Spanish, three French, and three Norwegian native speakers. The procedure was designed to achieve high inter-rater agreement scores. It was carried out as follows: An initial set of files were provided to each annotator for individual annotation, and then the inter-annotator agreement was computed. In cases where the agreement score fell below a predefined threshold, the annotators discussed and re-annotated the same files. When the agreement score exceeded the threshold, the annotators were provided with additional files to annotate. Finally, the annotators provided a set of emotionally annotated segments associated with the acoustic signal. The resulting corpus statistics for the three countries are detailed in Table 2.4.

Table 2.4.: Duration and number of segments of annotations in the Empathic corpus for the three countries annotated by experts.

	Spain		France		Norway	
	Duration	**No. Seg.**	**Duration**	**No. Seg.**	**Duration**	**No. Seg.**
Annotation 1	7h 17’ 55”	7654	5h 00’ 09”	4849	4h 06’ 46”	3228
Annotation 2	7h 11’ 05”	8427	4h 56’ 54”	5831	4h 05’ 14”	3813
Annotation 3	7h 09’ 02”	8301	5h 38’ 47”	3640	4h 06’ 52”	3802

In this study, the annotators segmented the audio files and assigned an emotional label to each segment. The annotators had to select from the following range of possible emotional labels, both the categorical and the three-dimensional VAD model:

- Categories: Calm, Sad, Happy, Puzzled, and Tense.
- Valence: Positive, Neither Positive nor Negative, Negative
- Arousal: Excited, Slightly Excited, Neutral
- Dominance: Dominant, Neither Dominant nor Intimidated, Defensive

Once the three annotators fully labelled the corpus, the methodology outlined in Section 2.2 was used to synchronise segments across various modalities. This synchronisation was essential to facilitate the detection of emotions from multiple modalities. After completing the final annotation, Table 2.5 presents the distribution of categories separated by country, revealing a significant imbalance with a tendency towards neutrality.

Table 2.5.: Final distributions of the expert annotated Empathic corpus for each emotional model and language.

(a) Categorical model

	Spain	France	Norway
Calm	94.84% (38359)	95.62% (19875)	96.42% (13960)
Sad	0.37% (151)	0.00% (1)	0.00% (0)
Puzzled	2.53% (1022)	2.18% (453)	0.30% (44)
Happy	2.06% (833)	2.14% (445)	3.27% (474)
Tense	0.20% (81)	0.05% (11)	0.00% (0)
Unknown	4607	2819	1775
Silence	37910	15978	15764

(b) Arousal

	Spain	France	Norway
Neutral	98.24% (40336)	77.61% (15634)	97.11% (14389)
Slightly excited	0.00% (0)	0.00% (0)	0.00% (0)
Excited	1.76% (722)	22.39% (4510)	2.89% (428)
Unknown	3995	3460	1436
Silence	37910	15978	15764

(c) Valence

	Spain	France	Norway
Negative	0.54% (216)	0.99% (203)	0.77% (114)
Neutral	29.07% (11570)	84.87% (17327)	96.73% (14280)
Positive	70.39% (28016)	14.13% (2885)	2.49% (368)
Unknown	5251	3189	1491
Silence	37910	15978	15764

(d) Dominance

	Spain	France	Norway
Defensive	11.37% (5251)	0.28% (60)	0.07% (10)
Neutral	84.84% (39197)	95.23% (20123)	98.91% (14791)
Dominant	3.80% (1754)	4.48% (947)	1.02% (153)
Unknown	3882	2474	1299
Silence	37910	15978	15764

Most data (over 90% of the segments) were annotated as calm. Additionally, 2% of the segments were labelled as happy and puzzled for all countries, and emotions such as sad and tense are almost absent. Concerning the Arousal dimension, French annotators labelled a higher percentage of “slightly excited” segments (22%) compared to Spanish and Norwegian ones (2%). Regarding the Valence dimension, Spanish annotators labelled most segments positive, while French annotators labelled most neutral (70-80%). Almost 100% of the Norwegian annotators labelled neutral the segments in this dimension. Concerning the Dominance dimension, most segments in all three datasets were labelled neither dominant nor intimidated.

4.3 | Crowd Annotation

In the crowd annotation task, 4525 new audio segments extracted using the methodology described in Section 2.1 were labelled. 56 different annotators were involved in this process. It is worth noting that this corpus only represents the Spanish part of the collected conversations in the Empathic project.

Table 2.7.: Questionnaire used for annotating the Empathic crowd corpus.

1	How do you perceive the audio voice?	• Male voice • Female voice
2	How do you perceive the speaker?	• Excited • Slightly Excited • Neutral
3	His/Her mood is:	• Rather Positive • Neither Positive nor Negative • Rather Negative
4	How do you perceive the speaker in relation to the situation in which he/she is?	• Rather dominant / controlling the situation • Rather intimidated / defensive • Neither dominant nor intimidated
5	Select the emotion that you think best describes the speaker's mood:	• Calm/Bored/Tired • Sad • Happy/Amused • Puzzled • Annoyed/Tense

The questionnaire used in this study (Table 2.7) is similar to the one used in the L6N corpus annotation and was carefully designed to ensure accurate and meaningful results. However, some modifications were made to adapt it to the current task. Firstly, a control question (question 1) was added to assess participant attention and remove non-serious responses. Specifically, the participants were asked to identify the gender of the speaker, which was expected to be a straightforward answer. Secondly, the number of categories for the valence dimension (question 2) was changed to three, unifying the number of categories for all three VAD model questions (questions 2, 3, and 4). Lastly, the categorical question was reduced to five categories. The categories were carefully selected to be the most relevant for this task and were aimed to increase agreement among annotators.

Table 2.8.: Krippendorff's Alpha value for each question in the Empathic crowd annotation.

Voice	Arousal	Valence	Dominance	Categorical
0.897	0.154	0.098	0.148	0.198

Krippendorff's alpha value was utilised to evaluate inter-annotator agreement, as shown in Table 2.8. Although the emotion annotation task got low Krippendorff's alpha values, it is important to mention that the control question received a high score of 0.897. This value validates the annotation process and suggests that the subjectivity of the emotion annotation tasks might explain the low values achieved. Other studies, such as Blanco et al. (2018), also reported lower Krippendorff's alpha values in subjective tasks.

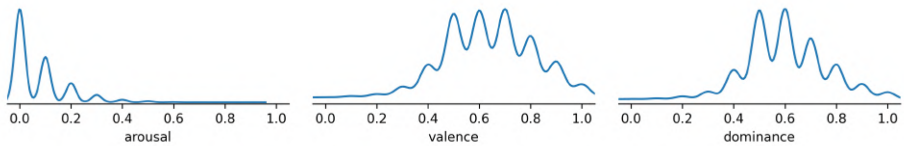


Figure 2.11.: Probability density function of the annotated VAD model in Empathic crowd annotation.

A consistent emotional state characterised by a calm mood is revealed after converting the VAD dimensional values into real numbers and analysing the annotations of Figure 2.11. The results demonstrate that the corpus is mainly characterised by neutral arousal levels, slightly positive valence, and neutral states regarding dominance. These findings align with the expected outcomes of older adults interacting with a machine.

Furthermore, for the categorical model, nearly 89% of the samples display an agreement of 60% or higher (4023 out of 4525 samples). Nevertheless, the imbalance between classes is evident, and the classes that exhibit the most significant degree of disagreement are those less represented, such as "Annoyed/Tense" and "Sad".

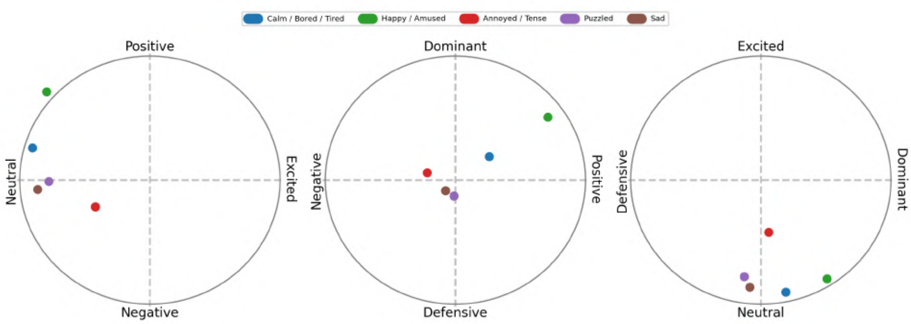


Figure 2.12.: Mean of categorical values displayed in VAD model in Empathic last pilot annotation.

Figure 2.12 shows the average values for each category in the VAD dimensional model. According to this representation, a decision was made to merge the emotions "Puzzled" and "Sad" into a single category, as they are very close in this space. Table 2.9 shows the number of categories after combining the two emotions.

Table 2.9.: Final number of samples per category in Empathic crowd corpus after merging categories.

Emotion	Name	N. samples	N. samples (>60% agree)
Calm / Bored / Tired	CALM	3395	3197
Happy / Amused	HAPPY	644	545
Puzzled + Sad	PUZZ	288	208
Annoyed / Tense	ANN	198	114

Figure 2.13 shows the density distribution of the merged dataset categories in the dimensional model. Each distinct category is represented with a different colour.

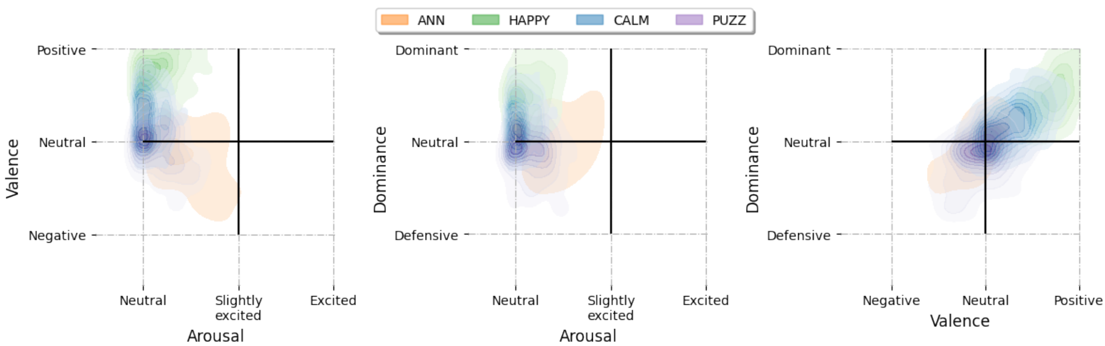


Figure 2.13.: 2D projections of 3D dimensional emotion model coloured with categorical values in Empathic's crowd annotation.

Figure 2.13 shows the density distribution of the merged categories in the dimensional model. Each category is represented with a different colour.

The joint analysis of the dimensional and categorical models shows a more significant separation of emotions. The *HAPPY* category is characterised by positive valence, slightly high dominance, and neutral to slightly high arousal levels. The *CALM* category, instead, is neutral across all dimensions but exhibits slightly positive valence and dominance. *ANN* category is distinct from the rest, with high arousal levels, negative valence, and a range of dominance values. Finally, the *PUZZ* category is positioned near the *CALM* category.

5 | MENHIR

The MENHIR project[8], funded by the European H2020 MSCA-RISE program under grant number 823907 (Callejas et al., 2019), aims to develop interactive conversations for mental health monitoring.

As a project partner, we focused on emotion recognition in speech. In this context, we were also involved in finding the most appropriate solution for storing all the project's data and metadata, including annotated emotions, in a structured format. This task was carried out in collaboration with GLOBIT, another project partner.

The following sections explain the whole process. Section 5.1, Data Acquisition Process, outlines how the conversations were recorded. In section 5.2, we discuss the representation of the data and metadata. Finally, section 5.3 analyses the emotional corpus created throughout this process.

5.1 | Data acquisition process

The data collection process recorded 41 conversations between a counsellor and a participant. Participants were divided into the Action Mental Health group (AMH_G) and the Ulster University Control Group (UU_CG). The AMH group comprised 21 individuals diagnosed with soft mental disorders like depression and anxiety. In comparison, the Control group consisted of 20 individuals who had never been diagnosed with any mental disorder.

Before starting the interaction, the AMH_G and the UU_CG participants must sign an informed consent. Additionally, participants from the UU_CG group were required to complete the 21-item Depression, Anxiety, and Stress Scale (DASS21) questionnaire, which took approximately 15 minutes. The questionnaire results were used to define inclusion criteria; thus, only individuals with low levels of anxiety or depression were included in the control group.

[8]MENHIR website: https://menhir-project.eu/

Table 2.10.: Template for emotion annotation by counsellors.

	CLIENT ANSWER					CONSELLOR PERCEPTION	
Statements	**None of the time**	**Rarely**	**Some of the time**	**Often**	**All the time**	**Arousal**	**Valence**
I'm feeling optimistic about the future						a++ a+ a0	n sn sp p
I'm feeling useful						a++ a+ a0	n sn sp p
I'm feeling relaxed						a++ a+ a0	n sn sp p
I'm feeling interested in other people						a++ a+ a0	n sn sp p
I've energy to spare						a++ a+ a0	n sn sp p
I'm dealing with problems well						a++ a+ a0	n sn sp p
I'm thinking clearly						a++ a+ a0	n sn sp p
I'm feeling good about myself						a++ a+ a0	n sn sp p
I'm feeling close to other people						a++ a+ a0	n sn sp p
I'm feeling confident						a++ a+ a0	n sn sp p
I'm able to make up my own mind about things						a++ a+ a0	n sn sp p
I'm feeling loved						a++ a+ a0	n sn sp p
I'm interested in new things						a++ a+ a0	n sn sp p
I'm feeling cheerful						a++ a+ a0	n sn sp p

The conversations were divided into Introduction, Main Interaction, and Reading.

In the Introduction, the counsellor informed the participant that the recording had started and asked open questions to make the participant feel comfortable and confident. Then, the counsellor asked if the participant was ready to begin and how they felt. The answers to these questions were noted in the final report of the counsellor.

In the Main Interaction, the counsellor administered the Warwick Edinburgh Mental Wellbeing Scale test (WEMWBS) orally to assess participants' mental well-being. The questionnaire was adapted for MENHIR purposes by changing the verb forms to the present tense, as shown in Table 2.10. The counsellor formulated the 14 WEMWBS items as questions to the participant and aimed to get extended answers rather than just a yes/no answer, asking follow-up questions like "why?" or "how?" to encourage the participant to provide more information. The participant's response was noted as one mark on the questionnaire scale. Additionally, the counsellor described their perception of the participant's current emotional status, rating their level of arousal and valence after each question provided. During this part, the counsellor will also ensure that the information provided by the participant does not result in the identification of the participant or their relatives.

In the Reading phase, the counsellor informed the participant that the structured interview had finished and would establish a clear separation from the previous phase using informal sentences. The participant was asked to read a short but phonetically rich and emotionally neutral text, such as a text passage from the popular tale "The Boy and the Wolf". This reading is intended to provide an unemotional sample of the participant's speech.

5.2 | Metadata representation of recorded audios

The data collected during the project had to be stored in a scalable high-performance repository to meet future user demands. A system that allows the classification of content, knowledge, analysis results, and datasets was critical for the success of MENHIR. For example, an emotion detection server could analyse a large volume of subject data and make the results available to be used by other applications. Sharing and exchanging research results was also essential for collaboration and co-creative networking among project participants. To this end, the MENHIR project developed a cloud-based ecosystem for comprehensive data-management support. For more detailed information, please refer to the work of Vu et al. (2021) (in P12).

In this context, the University of Ulster generated the MENHIR Content and Knowledge Management Ecosystem (KM-EP)[9], a cloud-based platform for managing scientific and educational content and knowledge resources. It comprises four essential components: Media Archive (MA), Digital Library (DL), Taxonomy Manager (TM), and Asset Manager (AM).

[9] https://menhir.projekt.onl

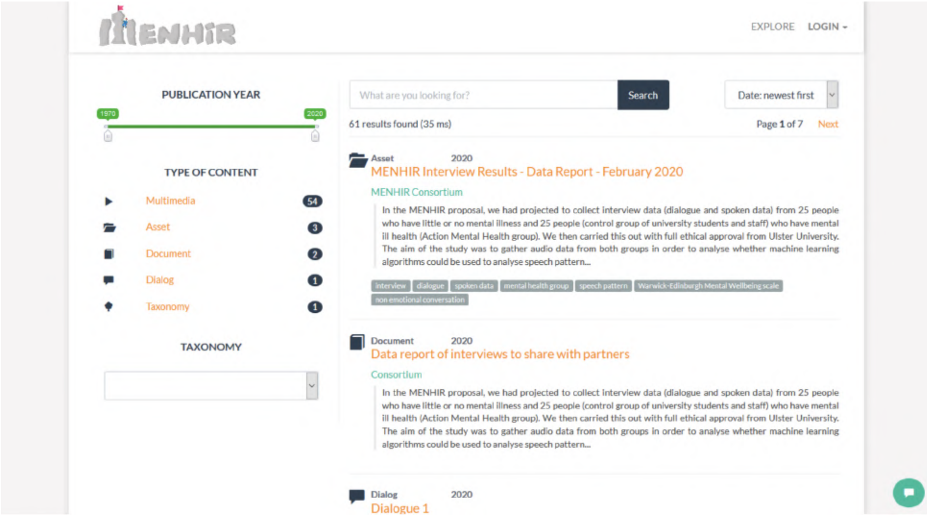


Figure 2.14.: menhir.projekt.onl portal.

The Media Archive (MA) component manages all multimedia objects in the platform, including audio, video, images, and presentation slides. Users can create, persist, manage, and classify different types of multimedia objects with their metadata. For example, the first recordings of interviews containing conversational audio data and the manual transcriptions created by English speakers from Northern Ireland were saved in this repository.

The Asset Manager (AM) component collects and combines related data, metadata, analysis results, and classifications into packages. For instance, for the Emotion Detection Server, a cronjob is developed and scheduled to run regularly, with three tasks: (1) searching for new audio files and their metadata and adding them into a new asset, (2) sending the new audio files and their metadata to the emotion detection server for analysis, and (3) receiving and adding analysis results into its package.

The AM also features an Emotion Audio Player that allows users to play audio files and perceive the subject's current emotional state in the audio. This feature allows users to edit, delete, and classify their packages using a user interface. Both audio recordings and individual analysis results inside a package can be classified using scientific content, emotion annotation, and other taxonomies.

5.2.1 | STORING EMOTIONAL CONTENT

In order to store all the metadata in a MPEG-7 format, a representation of affective taxonomies was proposed (Figure 2.15). This representation was chosen for its compatibility with multimedia content and ability to adequately capture and describe various emotional states. The use of MPEG-7 allows for a standardised format that can be quickly processed and analysed, facilitating the creation of affective models and other related applications.

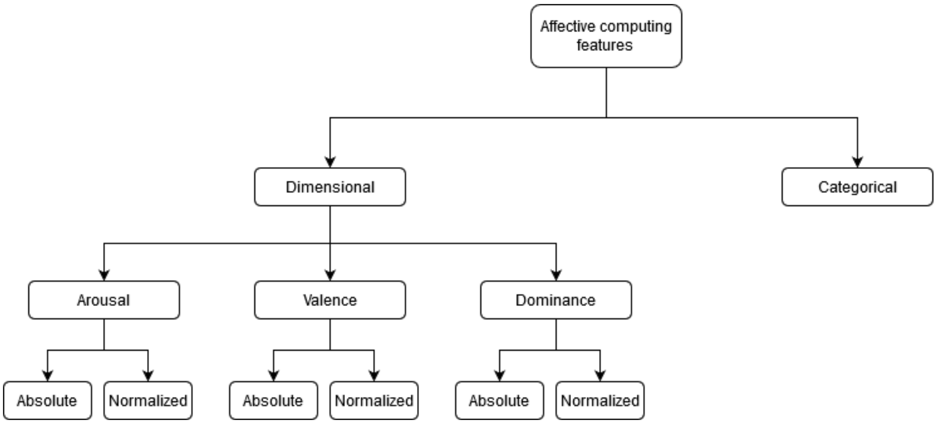


Figure 2.15.: Graphical representation of taxonomies.

After considering different options for representing emotions, we used the Emotion Markup Language (EmotionML[10]) as it was the most suitable for our requirements. This markup language implements several models, including the *category* element, the *dimension* element, the *appraisal* element, and the *action-tendency* element. Given that we were dealing specifically with the arousal and valence dimensions, we determined that the *dimension* element would be the most appropriate option for our use case.

The *dimension* element is used to jointly describe an emotion or a related state in terms of a set of emotion dimensions. When using the *dimension* element, it is necessary to declare a dimension-set vocabulary. In this case, we opted to use Mehrabian's PAD dimensions[11], which propose a three-dimensional description of emotions in terms of Pleasure, Arousal, and Dominance.

Annotated arousal and valence values (see questionnaire of Table 2.10 in section 5.1) were aligned with the scale values (ranging from 0 to 1) required by EmotionML.

[10]https://www.w3.org/TR/emotionml/
[11]https://www.w3.org/TR/emotion-voc/#pad-dimensions

An example of the resulting data representation in EmotionML looks like this:

Listing 2.1: EmotionML sample

```
<emotion dimension-set="http://www.w3.org/TR/
    emotion-voc/xml#pad-dimensions">
     <dimension name="arousal" value="0.3"/>
     <dimension name="pleasure" value="0.9"/>
     <dimension name="dominance" value="0.8"/>
</emotion>
```

Creation of MPEG-7 XML File

The XML Schema was developed to provide a clear and organised representation of the information contained in the XML file for each audio recording. The schema includes four key elements:

- The "info" element contains information about the audio location, participant information, interview duration, and responses.
- The "sections" element lists the different parts of the interview, including the introduction, main interaction, reading, and questions.
- The "users_turns" element contains information about the turns of both the counsellor and the participant.
- The "emotions" element contains emotion-related information in the EmotionML format.

With the XML schema as a guide, we created the MPEG-7 XML file, which follows the described structure and incorporates all relevant information.

Listing 2.2: MPEG-7 XML file

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<mpeg7 xmlns:xsi="http://menhir.projekt.onl/" xsi:
    noNamespaceSchemaLocation="menhir.xsd">
 <info> <audio ref="http://localhost/interview1.wav"/>
  <participant> <AnonID>CL1</AnonID>
   <group>AMH</group> <age>47</age>
   <gender>Male</gender> <depression>true</depression>
   <anxiety>true</anxiety>
   <other_mental_problems>PTSD</other_mental_problems>
   <social_media>
    <facebook>true</facebook> <twitter>true</twitter>
    <snapchat>false</snapchat> <smartphone>true</smartphone>
    <google_alexa>false</google_alexa>
   </social_media>
  </participant>
  <interview_duration>
   <seconds>1945</seconds> <time>0:32:25</time>
  </interview_duration>
  <q text="Has been informed about...">Yes</q>
  <q text="Has signed the informed consent">Yes</q>
  <q text="Has filled the DASS21">No answer</q>
  <q text="Belongs to the AMH_MG subset...">No answer</q>
  <final_remarks>N/A</final_remarks>
 </info>
 <sections>
  <sec id="sec1"> <start>6.32</start> <end>534.83</end>
   <info> <section>Introduction</section> </info>
  </sec>
  <sec id="sec2"> <start>534.83</start> <end>1788.16</end>
   <info> <section>Main Interaction</section>
    <q text="How are you feeling now?">OK</q>
    <q text="verbosity level">High</q>
    <q text="pers. inf. anonymized">No Answe</q>
   </info>
  </sec>
  <sec id="sec3"> <start>1788.16</start> <end>1917.81</end>
   <info> <section>Reading phase</section>
    <q text="Reading comfort level">High</q>
   </info>
  </section>
  <sec id="q1"> <start>544.97</start> <end>653.78</end>
   <info> <section>Q1</section> </info>
  </sec> ...
 </sections>
 <user_turns>
  <turn id="turn1"> <start>2.08</start> <end>6.27</end>
   <info> <user>counselor</user> </info>
  </turn>
  <turn id="turn2"> <start>6.27</start> <end>7.2</end>
   <info> <user>participant</user> </info>
  </turn> ...
 </user_turns>
 <emotions dimension-set="http://www.w3.org/TR/emotion-voc/xml#pad-
     dimensions">
  <emotion>
   <dimension name="pleasure"  value="0.50"/>
   <dimension name="arousal"  value="0.33"/>
   <reference uri="#turn1"/>
  </emotion> ...
 </emotions> </mpeg7>
```

5.3 | ANALYSIS OF THE ANNOTATIONS

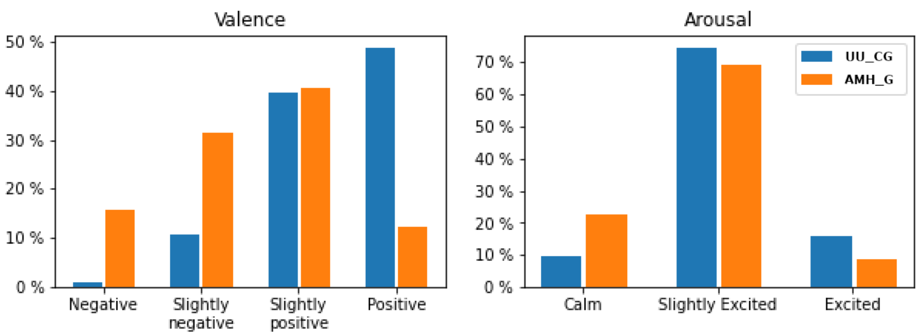
This emotional corpus comprises audio recordings from 41 users, with 21 users from the AMH_G group and 20 from the UU_CG. The total number of speaker interventions or turns is 1854 (4 hours, 15 minutes, and 45 seconds). The distribution of speakers and interventions associated with each mental disorder is shown in Table 2.11.

Table 2.11.: Distribution of Anxiety/Depression in AMH and Control.

	No. speakers	No. interventions
Depression	3	276
Anxiety	2	140
Both	16	824
Control	20	614

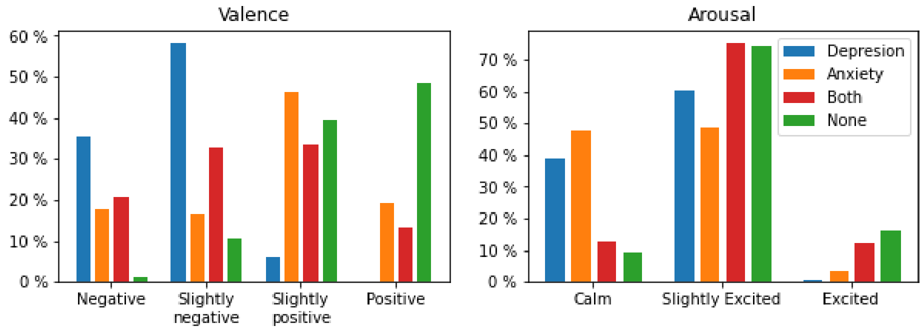
The emotional annotations of the counsellors were analysed, and a histogram was generated for the interventions annotated with different values of Valence and Arousal from the AMH_G and UU_CG groups, as shown in Figure 2.16. As expected, the AMH_G group had lower values of Valence, indicating a generally more negative emotional state compared to the UU_CG group, whose values were more positive. The same trend was observed in the Arousal levels, where AMH_G members exhibited lower excitement levels. However, the difference was less significant in this case, suggesting that Valence might be a more informative feature for detecting anxiety and depression.

Figure 2.16.: Percentage of interventions per group.



The analysis in Figure 2.17 focused on specific soft mental disorders and their differences. This figure displays the percentage of interventions from each group (Depression, Anxiety, Both, None) annotated with an emotional label for Valence and Arousal. Participants with depression were the most negative, while those without any disorder were the most positive when considering Valence. Participants with anxiety who were not diagnosed with depression were in between, indicating that Valence could be used to differentiate them from those with depression. Regarding Arousal levels, the differences were minor when comparing depression and anxiety.

Figure 2.17.: Percentage of interventions per illness.



Chapter 3

Methods and Models

This chapter provides a comprehensive overview of the methods developed to recognise emotions from speech analysis. It first covers different speech signal representations, such as acoustic features, spectrograms, and wave embeddings. Then, it introduces several classification and regression models for emotion recognition. Lastly, the chapter explores some practical issues of emotion recognition tasks.

1 | Speech signal representations

This section presents various digital representations of the audio signal. Firstly, subsection 1.1 discusses different sets of acoustic features. Next, subsection 1.2 deals with spectrograms. Lastly, recent representations of audio which employ neural networks to extract the features, such as wave embeddings, are explored in subsection 1.3.

1.1 | Acoustic features

In work López-Zorrilla et al. (2018) (in P2), the following set of 72 parameters were considered: Pitch, Zero Crossing Rate (ZCR), Energy, Entropy of the energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral flux, Spectral Rolloff, Chroma vector (12 coefficients), Chroma deviation, MFCC coefficients (12), LPC coefficients (16), Bark features (21). However, some of these parameters provide the same or similar information. For instance, LPC, Bark, and MFCC coefficients provide information about the frequency distribution of vocal tract resonances. In the same work, the following subsets of the parameters mentioned above were selected (**Preliminary Sets**) and explored in deVelasco et al. (2018) (in P3) and deVelasco Vazquez et al. (2019) (in P7):

- **Set A:** Pitch and Energy.
- **Set B:** Pitch, Energy and Spectral Centroid.
- **Set C:** Pitch, Energy, Spectral Centroid, ZCR and Spectral Spread.
- **Set D:** Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 12 MFCC coefficients.
- **Set E:** Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 16 LPC coefficients.
- **Set F:** Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 21 Bark features.

The first set was selected according to the studies performed by Williams and Stevens (1981), where the excitement related to the speaker affects the overall energy and pitch. In addition to time-dependent acoustic features such as pitch and energy, spectral features were selected for Sets B and C as a short-time representation of speech signals (Nwe et al., 2003). Different Cepstral-based features were added for Sets D, E, and F to represent stress in speech signals (Bou-Ghazale and Hansen, 2000).

Later, a comparison of different feature sets was conducted by deVelasco et al. (2022b) (in P15). One of the tested sets was a simplified version of Set D, called the **Baseline Set**, which only included pitch, energy, the entropy of energy, and the 13 MFCCs. The Low-Level Descriptors (LLDs) of **GeMAPS** (Eyben et al., 2015) and the Mel-spectrogram were also evaluated. The Mel-spectrogram emerged as the most effective way of representing audio for the tasks under consideration.

1.2 | MEL-SPECTROGRAM

In addition to the previously mentioned feature sets, spectrograms were explored to represent the acoustic signal. Spectrograms offer a time-frequency 2D acoustic representation that enables the visualization of the acoustic properties of sound events (see Figure 3.1). Essentially, spectrograms visually represent the spectro-temporal properties of sound, with the intensity usually depicted as a heat map using varying colours or brightness. Such a representation enables using 2D convolutional neural networks, commonly employed for image processing.

We first use the Mel-frequency Spectrogram in deVelasco et al. (2021) (in P10). This study generates the mel-frequency using sequences of 128 Fast Fourier Transform (FFT) components with a step size of 2.66 ms and a window size of 42.66 ms, resulting in significant overlapping. The process involved computing the squared Short-time Fourier transform of the audio wave, scaling it through a Mel filter bank, and then taking the logarithm of the intensities to convert the power spectrogram to decibel units for easier processing. In later works,

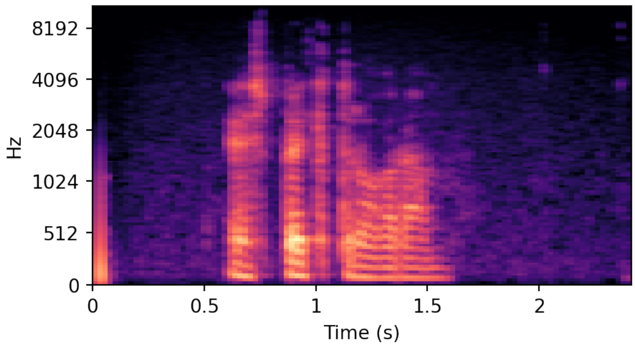


Figure 3.1.: Mel-Spectrogram of my voice saying specrtrogram.

such as deVelasco et al. (2022a) (in P16) and deVelasco et al. (2022b) (in P15), a normalization technique was applied to the Mel-Spectrogram in order to make it easier for machine learning models to process the data.

1.3 | WAVE EMBEDDINGS

In recent years, the interest in developing automatic data representations has increased. The idea of achieving rich representations through neural networks was explored with image autoencoders (Bank et al., 2020) in the field of Image Processing and also in the domain of Natural Language Processing (NLP) with word vectors (Church, 2017) and attention-based language models (Galassi et al., 2020).

Shortly afterwards, these techniques were adopted in audio analysis (Latif et al., 2023), and several models capable of creating rich representations from the audio waveform were published. In the work of deVelasco et al. (2022a) (in P16), some preliminary experiments were carried out with one of the first published models by Facebook AI Research: Wav2Vec 2.0 (Baevski et al., 2020). This model is the evolution of the one described in (Schneider et al., 2019), and it was first used for speech emotion recognition in English by (Luna-Jiménez et al., 2022) and in Spanish by deVelasco et al. (2021) (in P10). This representation has 1024 features plus the time dimension (250-time samples for 5 seconds).

TThe progress made with these models was so impressive that many new and interesting models appeared quickly. Among the most interesting ones are the Hubert model (Hsu et al., 2021) developed at Facebook AI Research and used in (Pastor et al., 2022a) or UniSpeech (Wang et al., 2021), UniSpeech-SAT (Chen et al., 2022b) and WavLM (Chen et al., 2022a) models developed at Microsoft. A comparison of all these models is made in (Ribas et al., 2023) and in Section 3 of the Appendix.

2 | CLASSIFICATION AND REGRESSION MODELS

The previous section describes the representations that can be used to subsequently perform machine learning tasks when trying to identify the emotional state of the speech. These audio representations consist of sequences of vectors of variable length depending on the duration of the audio. In machine learning, using functionals, such as means and standard deviations, is common to extract specific information from acoustic features. The goal is to smooth and reduce the time dimension. Functionals can be used with different types of acoustic features, including LLDs (Eyben et al., 2015), Mel-Spectrogram (Section 1.2), and Wave Embeddings (Section 1.3). Traditional classifiers, such as decision trees, Bayesian networks, Support Vector Machines (SVMs), and classical regressors, including linear regressions and Support Vector Regression (SVR), can benefit from this approach. The MLP (Figure 3.2) is the most similar architecture to a traditional classifier among Deep Neural Networks (DNNs) and can also benefit from fixed-length inputs for classification or regression tasks. This approach is advantageous when dealing with smaller datasets.

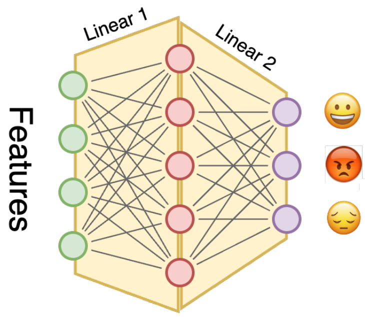


Figure 3.2.: A representation of MLP Network

Thanks to the versatility of neural networks, different network architectures can be employed to optimise the utilisation of temporal and spatial information. Computer vision has made significant progress since the beginning of the 21st century. One particular network architecture that brought about a revolution in computer vision is the Convolutional Neural Network (CNN). As a result, several studies have begun to explore the application of these advances to audio representations (Hou et al., 2018; Wyse, 2017). When dealing with spectrograms that exhibit both temporal and frequency dimensions, CNNs are a powerful tool as they can analyse 2D spaces effectively. This type of architecture has been used in several works, like deVelasco et al. (2021) (in P10), deVelasco et al. (2022b) (in P15) and deVelasco et al. (2022a) (in P16).

CNNs are widely used to process and analyse input data, generating a new representation of it (embedding), which can then be further analysed using another network structure, such as MLPs, as shown in Figure 3.3 and employed in deVelasco et al. (2021) (in P10). When the convolutional layers do not entirely reduce the temporal dimension, an alternative approach is to employ MLPs to evaluate each time step and then average the predictions, as deVelasco et al. (2022a) (in P16) demonstrated.

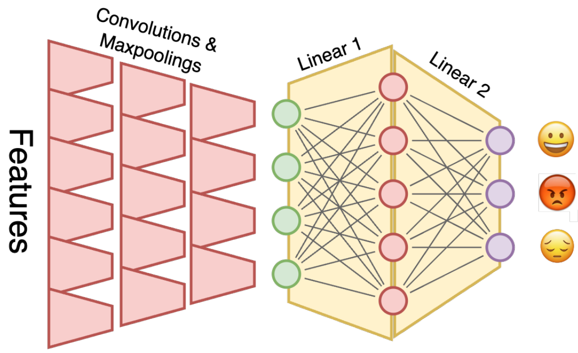


Figure 3.3.: An architecture of the emotion recognition model, consisting of a convolutional layer to analyse both temporal and spectral content and a MLP for emotion prediction.

Convolutional networks were not the only network architecture used to leverage temporal information. Recurrent Neural Networks (RNNs), particularly Long-Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997; Graves, 2012), were also widely used in emotion recognition (Wang et al., 2020; Tao and Liu, 2018) and in work deVelasco et al. (2018) (in P3), as shown in Figure 3.4.

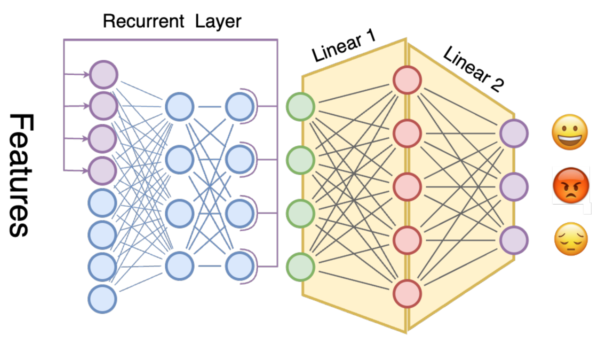


Figure 3.4.: An architecture of the emotion recognition model, consisting of a recurrent layer to analyse temporal content and a MLP for emotion prediction.

Shortly after the introduction of LSTMs, Gated Recurrent Units (GRU) (Cho et al., 2014), another type of RNN, were developed, capable of temporarily storing content with fewer computations and operations. This type of architecture was utilised in the model presented in deVelasco et al. (2022b) (in P15), which leverages CNN layers to extract feature information and RNN layers to process temporal information, as shown in Figure 3.5.

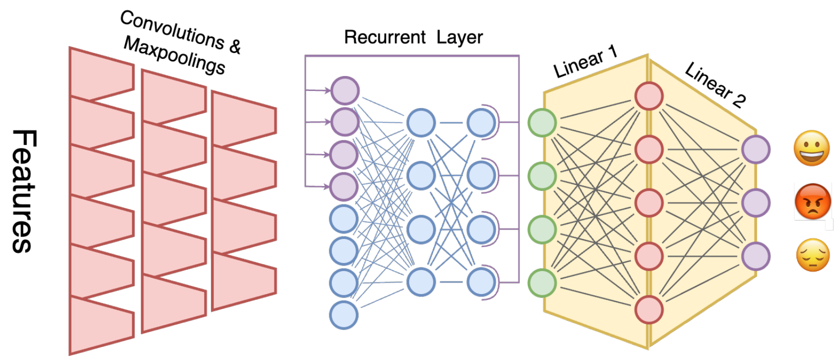


Figure 3.5.: An emotion recognition model architecture consisting of convolutional layers that generate a new input embedding, recurrent layers to analyse temporal content, and a MLP for emotion prediction.

Finally, it is important to highlight the impact of attention networks such as Transformers in recent years. These architectures started making significant progress in NLP (Sutskever et al., 2014; Brown et al., 2020; Galassi et al., 2020), moving through computer vision (Parmar et al., 2018; Carion et al., 2020; Dosovitskiy et al., 2020) to audio processing (Dong et al., 2018; Gulati et al., 2020; Chen et al., 2021), where the impact on the recognition of emotions from the speech is being analysed. An attractive characteristic of this network architecture is that all time slices contain the same importance in the output, avoiding the problem of forgetting the first inputs in recurrent networks. Moreover, all-time events are compared with all the others, avoiding choosing a specific context as in convolutional networks. Figure 3.6 shows an attention network next to an MLP. These two networks can be jointly trained for speech emotion prediction, or a pre-trained attention model can be utilised to extract embeddings (Wave Embeddings) which can then be used to train an MLP.

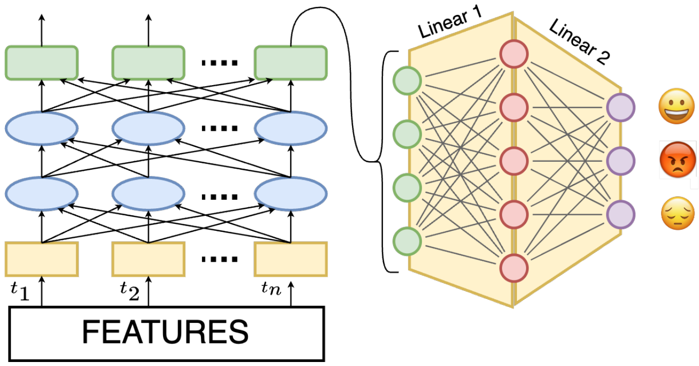
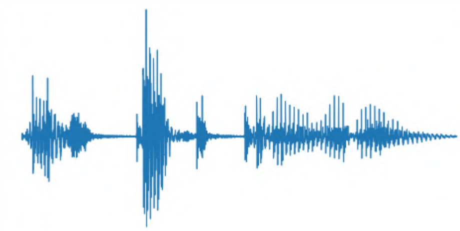
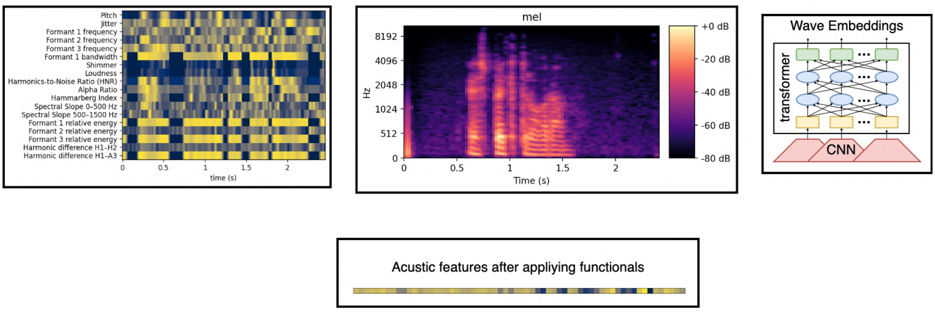


Figure 3.6.: An emotion recognition model architecture consisting of an attention-based transformer network for feature analysis and an MLP for emotion prediction.

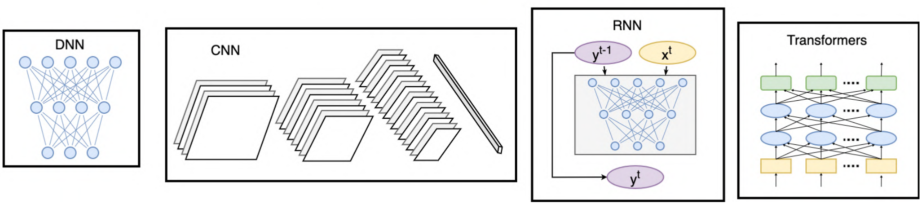
To sum up, Figure 3.7 shows the speech signal's process until the emotional state is detected. A desired acoustic representation (Figure 3.7b) is extracted from the raw audio signal (Figure 3.7a). Using a classical classifier or an MLP, some functionals of the extracted features are also computed. Finally, one or some combination of the network structures shown in Figure 3.7c extract the emotional content.



(a) Raw audio file.



(b) Different speech signal representations. Acoustic features on the left, Mel-Spectrogram on the middle and Wave Embeddings on the right. Some functionals can be applied to any representation to achieve a single vector representing the speech.



(c) Different network architectures. The simple MLP on the left, CNN on the middle left, RNN on the middle right and Attention Networks on the right.

Figure 3.7.: Different stages of the representation of the speech signal in a machine learning approach.

3 | Tackling Practical Issues

The practical challenges associated with emotion recognition tasks can impact the performance and accuracy of the models developed. In this section, two common issues encountered in these tasks are addressed: class unbalance in classification tasks (section 3.1) and discretisation in regression tasks (section 3.2). These issues are common across many machine learning applications, and addressing them is crucial to improve the developed models' effectiveness and applicability.

3.1 | Unbalance in classification tasks

One of the most widespread tasks in the machine learning framework is classification, which fits perfectly with the aforementioned categorical model for emotional representation. In this particular case, the task consists of building a model capable of distinguishing between the different emotional states, using only speech as input.

However, spontaneous emotions and real scenarios show a significant imbalance between emotional categories. Corpora gathered in spontaneous environments frequently display a significantly higher number of samples in the Neutral (Calm) category, while very few samples represent the remaining categories. This can be observed in Table 3.1, which is excerpted from deVelasco et al. (2022a) (in P16) and is also present in deVelasco et al. (2021) (in P10).

Table 3.1.: Frequency of the different categories in the corpora. Both the TV Debates and Empathic VA datasets are unbalanced. The majority class is the neutral emotion (*calm/indifferent* and *calm/bored/tired*), with more than 70% of the samples, which fits with real scenarios and spontaneous emotions.

TV Debates		**Empathic VA**	
Category	% audios	Category	% audios
Calm/Indifferent	73.64	Calm/Bored/Tired	79.47
Annoyed/Tense	14.32	Happy/Amused	13.55
Enthusiastic	4.72	Puzzled	3.11
Satisfied/Pleased	3.23	Annoyed/Tense	2.83
Worried	2.12	Sad	1.04
Interested.	1.57		
Others	0.40		

There are different ways to tackle the problem of unbalance in the literature. The most common procedures are under-sampling and over-sampling (Mohammed et al., 2020). These techniques balance the corpus by removing samples from the predominant category or repeating samples from unfrequented categories. Another more complex but also widely used technique is SMOTE (Chawla et al., 2002; Letaifa and Torres, 2021), which tries to create synthetic samples of the minority categories by making interpolations of the input data. However, it is important to note that these methods of addressing class imbalance create a false representation of emotions in reality and may affect machine learning models, causing them to predict less frequent categories as frequently as the dominant category.

The corpora employed in this thesis have a very limited number of samples (less than 5000 in the best case). Given this limitation, under-sampling is not a viable method for addressing class unbalance. Instead, the oversampling method is a valid option. It considers all samples in the corpus, with some categories appearing more frequently than others during training due to their higher frequency in the corpus.

From the beginning of the research, oversampling was identified as a practical approach to tackling the unbalance issue, as shown in deVelasco Vazquez et al. (2019) (in P7). However, various works have employed different configurations of the oversampling method. The most basic approach involves repeating the less frequent categories several times. For example, in deVelasco et al. (2021) (in P10), each sample was repeated five times, while in deVelasco et al. (2022a) (in P16), each sample was repeated only twice. The number of times the samples are repeated is determined by assessing the unbalance of the corpus, considering the frequency ratio and selecting a coherent and adjusted number that does not cause the model to predict all emotions with the same probability. Recently, deVelasco et al. (2022b) (in P15) proposed an individual adjustment per category, resulting in improved outcomes. Although the SMOTE technique was also tested, the experiment section concluded that oversampling and SMOTE generated similar results, with oversampling being simpler and easier to implement.

3.2 | Discretisation in regression tasks

On the other hand, the dimensional model is a regression problem as first approached in the work deVelasco et al. (2018) (in P3). In regression tasks, the network's last layer usually contains just one neuron, and the activation function used is typically ReLU or Sigmoid. The most commonly used cost function is Mean Square Error (MSE), which can also work as an evaluation metric. However, when most samples are in a narrow range, using MSE can generate a model that always predicts the same value, which is not very useful. In order to address this issue, we use the coefficient of determination metric (R^2), which evaluates the quality of the model. There is also a differentiable version of this metric that can be used as a loss function to improve results, used in deVelasco Vazquez et al. (2019) (in P7) and deVelasco et al. (2022b) (in P15).

In this case, each emotional axis (valence, arousal, and dominance) could be addressed as an independent regression problem, and thus with three regressor models, a dimensional emotional state could be inferred as in deVelasco et al. (2018) (in P3). However, as previously mentioned, the limited emotional range in natural environments results in a lack of granularity in the dimensional space, with few distant points, making it difficult for models to differentiate between emotional states. This issue arises from the annotations obtained from crowdsourcing, which are inherently discrete and lack the continuous nature of VAD dimensions. In this situation, the previously discussed balancing techniques cannot be used directly. This is because, based solely on the dimensional model, the samples do not belong to a group that allows us to determine how common or unusual they are.

One of the most straightforward solutions to address this problem is to discretise each dimension and transform the task into multiple classification problems, one for each dimension, as proposed in the works deVelasco et al. (2022b) (in P15) and deVelasco et al. (2022a) (in P16). This approach simplifies the problem because a regression task requires accurate predictions of specific points and can be effective in certain scenarios. In addition, the work deVelasco et al. (2022b) (in P15) uses the discretised categories to determine the optimal number of sample repetitions treating the problem as a regression task. Nevertheless, the results were better with the classification problem than with the regression ones.

Chapter 4

Experimental Evaluation

This chapter provides a comprehensive summary of all the conducted experiments. Sections 1, 2 and 3 present the contributions of all the published research works, divided by how they treat the emotion recognition task. The proposed approaches and achieved results with the categorical model for classification tasks are analysed in Section 1. Then, the dimensional model is studied, initially addressing the problem as a regression task (Section 2) and then as a classification task (Section 3). Each research work addresses some aspects of automatic emotion recognition, but the experimental conditions vary from work to work. Therefore, the experiments are not always comparable, making extracting meaningful conclusions difficult.

To avoid this issue, we performed a final round of experiments, where most of our proposals were analysed in comparable and statistically rigorous conditions. Section 4 summarises these results, while the Appendix contains all the details. These experiments and their analysis represent one of the fundamental contributions of this thesis.

1 | Categorical Model: Classification task

Classifying emotion categories is the most intuitive way to identify the speaker's emotional state. The first experiments related to the categorical model were carried out with the L6N dataset (Section 3 of Chapter 2) in the work deVelasco Vazquez et al. (2019) (in P7). This work attempted to discern between three emotional categories using SVMs and simple MLPs, obtaining a maximum performance of 0.4 F1 Score. To this end, the previously described feature sets devoted to detecting the focus on speech were used; see López-Zorrilla et al. (2018) (in P2).

An analysis of these results revealed that the majority class was predicted most of the time regardless of the input, resulting in F1 scores of around 0.33 for the three classes. An oversampling technique was introduced in the work deVelasco et al. (2021) (in P10) to address this issue. Additionally, the Mel-Spectrogram was used as input, and computer vision techniques, specifically convolutional networks, were employed, resulting in a significant performance improvement, with an F1 Score of 0.51.

TThis approach was also tested using the categorical model on the Empathic emotional corpus (Section 4.3 of Chapter 2). However, a performance of 0.27 F1 was achieved. The decrease in performance can be attributed to the nature of the problem, which involves defining four new emotion categories and results in a more significant data imbalance.

In the work deVelasco et al. (2022b) (in P15), a comparison of several audio representations (Baseline, GeMAPS, and Mel-Spectrogram explained in Section 1 of Chapter 3) was conducted to determine the best-performing set on the L6N dataset. For this purpose, a new network composed of both CNN and RNN layers was implemented, resulting in a superior performance with the Baseline sets and the Mel-Spectrogram (0.56 and 0.61 F1 Score, respectively). The improvement with the Mel-Spectrogram features may be attributed to the network architecture change and a more refined oversampling technique.

In a later study (deVelasco et al., 2022a; P16), the effectiveness of the Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) technique was evaluated. SWA is a generalisation technique that averages multiple points along the trajectory of Stochastic Gradient Descent (SGD) and is expected to produce better results. However, it led to worse performance for both corpora, with scores of 0.56 for L6N and 0.26 for Empathic. One possible explanation is that SWA may not work well when the model has converged to a local minimum, which is likely to happen in small datasets like the ones used in this study. In addition, preliminary experiments with Wave Embeddings were conducted, which did not achieve high-performance levels.

Finally, the behaviour and performance of both emotional models together were analysed in deVelasco et al. (2023) (in P19). In this work, a multitasking network was trained to simultaneously infer the categorical and dimensional representations of the emotional state. The categorical emotional state was predicted in two different manners. First, based on the representations generated by the neural network (similar to all the previous works); and second, using the scalar predictions of the dimensional model. The second approach yielded slightly better results (0.59 compared to 0.58).

A multimodal approach has been tested in an ongoing study (Palmero et al., 2023) carried out with the expert annotations of the EMPATHIC corpus (see Section 4.2 of Chapter 2). The study aims to predict the emotional state from audio signals (using WavLM wave embedding) and video and gaze features. Using the audio alone, a 0.65 F1 Score has been achieved. Adding video features improves the performance, resulting in a 0.67 F1 Score. However, including gaze features does not provide any relevant additional information, as no improvement is observed (F1 score of 0.65 with audio and gaze and 0.67 with audio, video, and gaze).

Table 4.1 provides a comprehensive summary of the performance of each evaluated model, as measured by the F1 score, for comparison purposes and assessments.

Table 4.1.: Summary of model performance measured by F1 score.

Article	Dataset	Features	Network[1]	Results (F1)	Notes
P7	L6N (3cat) 5500 audios	SetA	Avg./Std. + SVM	0.34	70%/30% split No oversampling
		SetD	Avg./Std. + MLP	0.35	
		SetB	Avg./Std. + SVM	0.29	70%/30% split Oversampling
		SetB	Avg./Std. + MLP	0.40	
P10	L6N (3cat) 4118 audios	Mel-Spectrogram	3xConv 2D + Flatten + 2xFC	0.51	10fold CV Oversampling (x5)
P15	L6N (3cat) 4118 audios	Baseline	2xConv (1D/2D) + biGRU[2] + FC	0.56	10fold CV Oversampling (x4/x9)
		GeMAPS		0.38	
		Mel-Spectrogram		**0.61**	
P16	L6N (3cat) 4118 (1266)[3] audios	Mel-Spectrogram	3xConv 2D + 3xFC	0.56	10fold cv SWA Oversampling (x2)
		Wav2Vec	3xConv 1D + 3xFC	0.27	
P18	L6N (3cat) 1525 (1096) audios full interventions	Mel-Spectrogram	3xConv 2D + 3xFC	0.49	10fold CV
		UniSpeechSAT	2xFC	0.59	
		DistilBERT	-	0.52	
		MultiModal	UniSpeechSAT DistilBERT 2xFC	**0.61**	
P19	L6N (3cat) 4118 audios	Mel-Spectrogram	3xConv2D + 3xFC	0.58	70/30 + SWA Oversampling (x2) Multitask[4]
				0.59	
			VGG16	0.57	
				0.57	
P10	EMP cowd (4cat) 2000 audios	Mel-Spectrogram	3xConv 2D + Flatten + 2xFC	**0.27**	10fold CV Oversampling (x5)
P16	EMP cowd (4cat) 4525 (4023)[5] audios	Mel-Spectrogram	3xConv1D + 3xFC	0.26	10fold CV + SWA Oversampling (x2)
		Wav2Vec		0.22	
ongoing	EMP expert 3 cat	WavLM	MLP	0.65	10fold CV Multimodality
		WavLM+gaze		0.65	
		WavLM+video		**0.67**	
		WavLM+video+gaze		**0.67**	

2 | DIMENSIONAL MODEL: REGRESSION TASK

Regarding the dimensional model, one of the most intuitive ways to tackle the problem is to approach it as a multi-dimensional regression problem. Three sets of experiments were carried out to explore this problem in different conditions.

The work deVelasco et al. (2018) (in P3) presents the results obtained with a preliminary version of the L6N dataset containing only 120 labelled samples. Regarding the acoustic representations used as the input to the regressor, the six Preliminary Sets of acoustic features presented in López-Zorrilla et al. (2018) (in P2) were compared. The regressor was a simple LSTM-RNN trained to predict the three VAD dimensions simultaneously.

Later, deVelasco Vazquez et al. (2019) (in P7) expanded the results with the same sets of features by creating an independent model for each VAD dimension, using a simple MLP network and Support Vector Regression (SVR). In this case, the full L6N dataset was used, which consists of 5500 labelled audios. The R^2 evaluation metric (Equation 4.1) was used in this case, as it considers the variance of the data and provides a more reliable measure of the goodness of the fitted model than the MSE. This metric was also used as a cost function to train the neural networks:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{4.1}$$

where y_i represents each predicted value produced by the model, f_i each reference, and $\bar{y}$ is the mean of all predicted values.

Finally, deVelasco et al. (2022b) (in P15) compare different input representations: the Baseline set, GeMAPS, and MelSpectrogram. In this case, the experiments were carried out with a subset of the L6N corpus; the samples labelled with very low inter-annotator agreement were discarded as they introduced noise in the learning process. As for the neural network architecture, a combination of convolutional layers and GRU recurrent layers was used. Oversampling was employed to improve the results, and the best results were achieved with the Mel-Spectrogram representation.

Table 4.2 summarises the results of the three experiments. In general, the performance of the models is relatively low for all the emotional dimensions

[1]The network architectures are defined in terms of layers, specifying the number of layers of each type (e.g., 2x) and concatenating each type of layer with the symbol "+".
[2]biGRU: Bidirectional Gated Recurrent Units (GRU)
[3]The L6N dataset including only those samples with a 60% or higher level of agreement.
[4]Multitask: Trained for both categorical and dimensional models simultaneously.
[5]The Empathic dataset including only those samples with a 60% or higher level of agreement.

(arousal, valence, and dominance). Therefore, the problem was simplified and treated as a classification task in Section 3. Note that comparing results across the experiments is not straightforward due to the differences in the size of the employed corpora.

Table 4.2.: Summary of model performance for the regression task in the dimensional model, measured by R^2 and MSE score.

Article	Dataset	Features	Network	Val	Aro	Dom	Notes
P3	L6N (120 audios)	Set F	RNN	0.14 MSE			SplitTest 70%/30% Training together MSE loss
		Rest of the sets	RNN	0.16 MSE			
		Fastext	DNN	0.12 MSE			
P7	L6N (5500 audios)	Sets F/F/E [6]	SVR	0.215	0.086	**0.119**	SplitTest 70%/30% R2 loss
		Sets E/D/D	DNN	**0.357**	0.116	0.095	
P15	L6N (4118 audios)	Baseline	2xConv (1D/2D) + biGRU + FC	0.08	0.20	0.02	10fold CV Oversampling R^2 loss
		GeMAPS		0.02	0.10	0.03	
		Mel-Spectrogram		0.11	**0.30**	0.06	

3 | Dimensional Model: Classification task

Identifying the user's emotional state through the dimensional model can also be approached by discretising each dimension (see section 3.3 of Chapter 2) and dealing with it as a classification problem.

This approach was first explored in deVelasco et al. (2022b) (in P15), where the Baseline, GeMAPS and Mel-Spectrogram were explored as the audio input representation. As in the case of the categorical model, the Mel-Spectrogram is the representation that offers the best performance, with F1-Scores of 0.52 in Valence, 0.70 in Arousal, and 0.60 in Dominance. Next, in deVelasco et al. (2022a) (in P16), the SWA technique and some Wave Embeddings (Wav2Vec) were tested, but in both cases, worse results were achieved in the L6N dataset.

With similar conclusions, these input representations were also tested with the EMPATHIC corpus (deVelasco et al., 2022a; P16). The best performance was again obtained using the Mel-Spectrogram: 0.38 F1 Score in Valence, 0.54 in Arousal, and 0.59 in Dominance.

[6] Sets selected for each task: F for both Valence and Arousal and E for Dominance

In deVelasco et al. (2023) (in P19), we attempted to predict the categories and dimensional VAD values using the same network, in a multitask fashion, using the same experimental procedure as in deVelasco et al. (2022a) (in P16). In this experiment, the network first predicted the dimensions' scalar value, and then each dimension's category was inferred based on this prediction. A 0.42 F1 Score in Valence, 0.67 in Arousal, and 0.57 in Dominance were obtained. Even if the task is much more challenging, the results are only slightly worse than the best ones, which ratifies that multitask learning is a valuable option for future research.

The latest tests with the L6N corpus were presented in Zubiaga et al. (2022a) (in P18). This study analysed the advantages of jointly using text and audio information in emotion recognition. A slightly different version of the L6N dataset was used in this case, defined in section 3.1 of Chapter 2. The Mel-Spectrogram and Wave Embeddings (UniSpeech) representations were compared for the audio processing part. These were processed with CNN and DNN networks. Regarding the text processing part, the DistilBERT pretrained transformer was used. The Wave Embeddings model (UniSpeechSAT) achieved superior performance than the Mel-Spectrogram, obtaining a 0.46 F1 Score in Valence, 0.73 in Arousal, and 0.57 in Dominance. The text-based model performed slightly worse than the acoustic model. However, combining both models into a multimodal model achieved similar performance with 0.47 in Valence, 0.70 in Arousal, and 0.56 in Dominance.

Finally, some experiments were conducted with the two-dimensional VA model within the MENHIR project, as reported in Zubiaga et al. (2022b) (in P17). These tests involved a comparison of the acoustic features GeMAPS and Wave Embeddings (HuBERT). The results suggest once again the superiority of Wave Embeddings.

Table 4.3 shows the results obtained when addressing the prediction of the dimensional model as a classification task.

Table 4.3.: Summary of model performance for the classification task in the dimensional model, measured by F1 score.

Article	Dataset	Features	Network	Val	Aro	Dom	Notes
P15	L6N (3V/2A/2D) 4118 audios	baseline	2xConv (1D/2D) + biGRU + FC	0.44	0.66	0.56	10fold CV Oversampling (x3)
		GeMAPS		0.41	0.63	0.58	
		Mel-Spectrogram		**0.52**	0.70	**0.60**	
P16	L6N (3V/2A/2D) 4118 (1266)[7] audios	Mel-Spectrogram	3xConv2D + 3xFC + AVG timesteps	0.47	0.70	0.58	10fold CV SWA Oversampling (x2)
		Wav2Vec	3xConv1D + 3xFC	0.27	0.55	0.42	
P19	L6N (3V/2A/2D) 4118 audios	Spectro	3xConv2D + 3xFC + AVG timesteps	0.42	0.67	0.57	10fold CV SWA Oversampling (x2) Multitask
			VGG16	0.41	0.67	0.56	
P18	L6N (3cat) 2964 audios full interventions	Mel-Spectrogram	3xConv2D + 3xFC	0.36	0.63	0.57	10fold cv
		UniSpeechSAT	2xFC	0.46	**0.73**	0.57	
		DistilBERT	-	0.46	0.59	0.56	
		MultiModal	UniSpeechSAT + DistilBERT + 2xFC	0.47	0.70	0.56	
P16	EMP (3V/2A/2D) 4525 (4023)[8] audios	Mel-Spectrogram	3xConv1D + 3xFC	**0.38**	**0.54**	**0.59**	10fold CV & SWA Oversampling (x2)
		Wav2Vec		0.26	0.49	0.52	
P17	MENHIR (4V/3A/-)	GeMAPS	DNN	0.35	0.41		SplitTest 90%/10%
		HuBERT		**0.41**	**0.57**		

[7] The L6N dataset including only those samples with a 60% or higher level of agreement.
[8] The Empathic dataset including only those samples with a 60% or higher level of agreement.

4 | Experiments in comparable conditions: Highlighted Results

All the experiments presented correspond to research works addressing a specific research question. Thus, the experimental conditions varied from work to work. These experiments have allowed us to understand how some factors affect the performance of emotion recognition systems. However, many of these experiments are not comparable, and extracting meaningful and strong conclusions is difficult.

Thus, we repeated most of our experiments, including some of the most novel trends in speech-processing technologies in our comparison. We compare the performance of different feature sets, neural network architectures and other hyper-parameters for each task and identify the most effective approach. The L6N and Empathic corpora are used in different classification problems: emotional category and arousal, valence, and dominance. The details of the experiments presented in this section can be found in the Appendix. This section highlights the most important findings. However, we encourage the reader to consult the Appendix for a more comprehensive understanding of how the experiments were carried out and the results obtained. Additionally, it is important to note that all the experiments presented in this section are yet to be published.

The experimentation was divided into two phases. The first phase aimed to find the most suitable hyperparameters for each network structure using the "La Sexta Noche" dataset and the categorical emotion classification problem (section 4.1). In the second phase, we analysed which network architecture and acoustic features performed best in each emotion recognition task and corpus (section 4.2). In addition, different approaches to address the unbalance problems were also tested in both phases.

4.1 | Grid Search for Categorical Model on L6N Dataset

During the first phase of experimentation, different structures of **Convolutional Neural Networks** were analysed, including both 1D and 2D networks. Different optimisers, learning rates, batch sizes, and activation functions were tested. The results showed that although 2D networks obtained more stable results, 1D networks offered the best peak performance. Moreover, the Adam optimiser provided much better than the classic SGD, and regarding the learning rate, the best option was $1e^{-3}$. Increasing the batch size offered slightly higher performance (64 was the best). Surprisingly, not including an activation function between the convolutional layers of the network resulted in higher performance compared to including ReLU or Sigmoid activation functions. All the explored hyperparameters are listed below, with the best option highlighted in bold.

- Type: **Conv1D** and Conv2D.
- Optimizer: SGD, **Adam**, and **AdamW**.
- Learning Rate: **0.001** and 0.0001.
- Batch Size: 32 and **64**.
- Activation function: **None**, ReLU and Sigmoid.

Besides CNNs, **Transformer network architectures** were also explored to process the speech representations. With this network architecture, a wide range of hyperparameters was tested, including learning rate, batch size, optimisers, number of layers and number of heads, how to add positional embedding and how to link the transformer output to the classification layer. The results showed that the best choice for the learning rate was $1e^{-4}$, for the batch size was 64, for the optimisers was AdamW with beta parameters of 0.9 and 0.98, and for the number of layers was 6. Furthermore, it was found that the best way to add positional embedding was to concatenate it. Additionally, the best way to join the transformer's output with the classifier layer was to average all the time instances. As for the number of heads, no clear conclusion was reached, as all values gave very similar performances. Below, all the values for the hyperparameters explored are shown, with the best option for each one in bold.

- Optimizer: Adam, and **AdamW**.
- Beta parameters: (0.9, 0.999) and **(0.9, 0.98)**.
- Learning Rate: 0.001, **0.0001** and 0.00001.
- Batch Size: 8, 16, 32 and **64**.
- Number of layers: 2, 4, **6**.
- Number of heads: 2, 3, 6, 9 and **15**.
- Positional embedding: sum and **concat**.
- Encoder to linear: first, **mean** and learnableSpecialToken.

Finally, different **Wave Embeddings** audio representations were tested. The objective was to choose which Wave Embedding network architecture and which model provided the best results and analyse the influence of the optimiser and batch size. Several pretrained models were analysed, including Hubert, UniSpeech, UniSpeechSAT, Wav2Vec2 and WavLM. It was found that several Hubert and WavLM checkpoints performed very well, but one of the UniSpeech-SAT checkpoints performed the best. As for the hyperparameters, the results showed that Adam and AdamW optimisers performed equally well, with a batch size of 64, getting slightly better results.

- Wave Embedding models: **Hubert (large-ll60k)**, UniSpeech (large-multilingual-1500h-cv), **UniSpeechSAT (sat-large)**, Wav2Vec2 (base) and **WavLM (large)**.
- Optimizer: **Adam** and AdamW.
- Batch Size: 32 and **64**.

These results are shown in the first rows of Table 4.4, which includes more results and is therefore included in the next section.

4.2 | Extending the Experiments: Empathic Dataset and Dimensional Model

During the second experimental phase, we included the multi-language Empathic dataset with both crowd and expert annotations and the dimensional models for both datasets (*L6N* and *Empathic*). In this section, only the results obtained from crowd annotations are presented. The rest of the results are shown in the Appendix.

Both datasets were tested on categorical and dimensional problems. We explored which acoustic representations were the most appropriate, including Mel-Spectrogram, the three best Wave Embedding models, GeMAPS, and eGeMAPS. Additionally, a simple MLP was tested for each feature, along with a 1D Convolutional Network and a Transformer Network. The results showed that all the models working with Wave Embedding features obtained the best results. Additionally, it was found that Transformers and MLP offered similar performances, outperforming convolutional networks.

Tables 4.4 and 4.5 display the bests results obtained for each dataset and task using a range of network architectures and acoustic characteristics. Across all cases, the Wave Embeddings representation consistently yielded the highest performance. Furthermore, the MLP architecture was the most effective and computationally efficient option, demonstrating superior performance in most cases while requiring fewer computations.

Table 4.4.: Best F1 scores achieved for the La Sexta Noche crowd dataset.

	Architecture	Mel-Spectrogram	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
Cat	M. Perceptron	0.599	**0.717**	0.709	0.693	0.549	0.543
	Conv. 1D	0.608	0.692	0.684	0.649	0.497	0.519
	Transformer	0.614	0.664	0.658	0.622	0.503	0.502
Aro	M. Perceptron	0.660	**0.726**	0.723	0.725	0.654	0.661
	Conv. 1D	0.654	0.714	0.718	0.718	0.657	0.648
	Transformer	0.677	**0.729**	0.711	0.707	0.625	0.626
Val	M. Perceptron	0.454	**0.518**	0.515	0.494	0.410	0.422
	Conv. 1D	0.442	0.486	0.487	0.479	0.394	0.397
	Transformer	0.462	0.486	0.488	0.470	0.426	0.406
Dom	M. Perceptron	0.555	0.613	**0.625**	0.604	0.544	0.568
	Conv. 1D	0.553	0.604	0.610	0.598	0.519	0.535
	Transformer	0.571	0.600	0.603	0.588	0.553	0.558

Table 4.5.: Best F1 scores achieved for the Empathic crowd dataset.

	Architecture	Mel-Spectrogram	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
Cat	M. Perceptron	0.284	0.401	0.396	**0.404**	0.295	0.301
	Conv. 1D	0.301	0.401	0.392	0.391	0.283	0.280
	Transformer	0.282	0.379	0.378	0.367	0.269	0.264
Aro	M. Perceptron	0.568	0.589	0.594	**0.596**	0.560	0.536
	Conv. 1D	0.552	0.556	0.565	0.563	0.523	0.535
	Transformer	0.550	0.561	0.566	0.568	0.513	0.514
Val	M. Perceptron	0.388	0.433	0.445	0.440	0.375	0.368
	Conv. 1D	0.387	0.433	**0.448**	0.435	0.364	0.365
	Transformer	0.382	0.441	0.438	0.432	0.369	0.361
Dom	M. Perceptron	0.573	**0.621**	0.619	0.619	0.565	0.564
	Conv. 1D	0.569	0.604	0.600	0.601	0.536	0.533
	Transformer	0.560	0.598	0.600	0.594	0.537	0.538

Chapter 5

XAI: Analysis of system behaviour

This chapter is aimed to understand the decision-making process involved in our emotion identification models. Specifically, we focus on “La Sexta Noche" TV show as a realistic scenario to investigate spontaneous emotions. To this end, we build a simple convolutional neural model that uses the Mel-Spectrogram as input to represent the audio.

We first analyse the model’s performance in predicting the categorical and VAD dimensional models of emotions. Then, we examine the behaviour of the model layer-by-layer. Then, we propose a visual representation of our class model (Deep dream) as a straightforward method of interpretation (Simonyan et al., 2014) that we evaluate employing methods involving filters and transformations. Our main objective is to understand the model’s emotion identification behaviour better, enhancing our comprehension of the computational methods and decision-making involved. The work presented here can be found in deVelasco et al. (2023) (in P19).

1 | Proposed CNN architecture

This work proposed a lightweight convolutional neural network architecture with only 43K parameters allowing the easy analysis and interpretation of the system behaviour. The proposed CNN was compared to the VGG16 architecture providing similar or superior results, as reported in deVelasco et al. (2023) (in P19). However, Table 5.1 does not include the results of the VGG16 architecture because only the proposed convolutional model is analysed.

The proposed network architecture is shown in Figure 5.1. It took the Mel-Spectrogram as input and was designed to classify emotions in terms of categories and VAD dimensions jointly. The VAD dimensions were computed by convolutional and pooling layers, followed by three linear layers with a sigmoid activation function, resulting in three scalar values at point A converted to discrete VAD predictions at point B. Then two methods were proposed to infer the labels associated with the categorical model. The first method predicted the categories based on the CNN output at point C. In contrast, the second method used the scalar predictions of the VAD model at point A to get the predictions at point D.

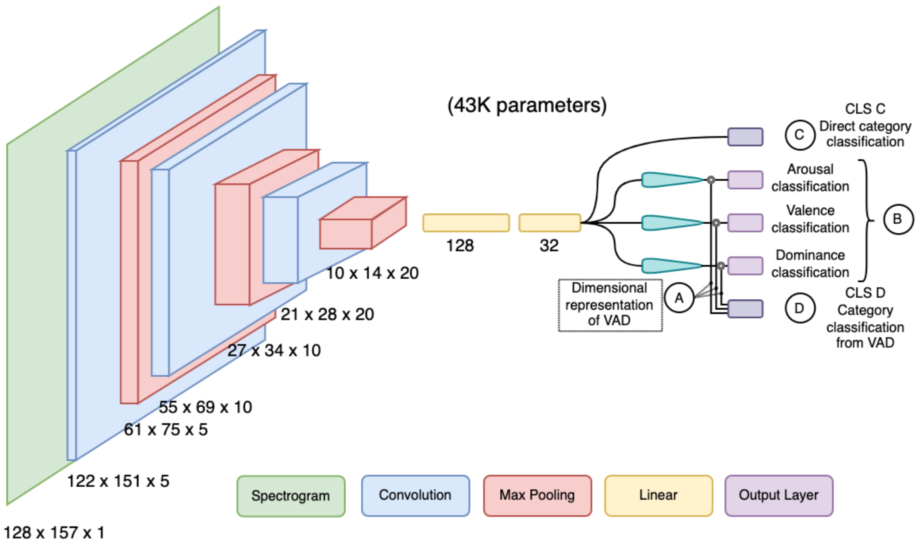


Figure 5.1.: Simple Convolutional Neural Network employed to get the joint classification of both emotional models from the Mel-Spectrogram.

All the configurations and hyper-parameters used to train the CNN can be found in the work of deVelasco et al. (2023) (in P19). It is worth noting that the cross-entropy loss function is employed for all classification tasks. The final loss is obtained by averaging the five losses from the two categorical and three VAD models with different weights. Empirical findings conclude that the optimal performance was achieved when the weights for VAD model losses were two times the weights for the categorical losses.

2 | Performance Analysis

The performances of the proposed classification system for both emotion categories and VAD dimensions are shown in Table 5.1. All results were obtained after a 10-fold cross-validation procedure. For the categorical model, two results are given, one corresponding to the direct categorical classification at output C (CLS C in Figure 5.1) and another using the predicted VAD floating point values at output D (CLS D in Figure 5.1). The average and standard deviation of five metrics, commonly employed to evaluate emotion classification systems (Letaifa and Torres, 2021; Pastor et al., 2022b) are reported: F1 score, Unweighted Accuracy (UA, also known as balanced accuracy or unweighted average recall), average precision, Matthews Correlation Coefficient and Area Under the ROC Curve (AUC).

Table 5.1.: Classification performance for the categorical and dimensional models classification task. Two ways of predicting the emotion categories were tested.

	Arousal	**Valence**	**Dominance**	**CLS C**	**CLS D**
F1	0.67 ± 0.11	0.42 ± 0.03	0.57 ± 0.05	0.58 ± 0.04	$\mathbf{0.59 \pm 0.05}$
UA	0.67 ± 0.09	0.45 ± 0.04	0.57 ± 0.03	0.57 ± 0.04	$\mathbf{0.58 \pm 0.04}$
Average precision	0.69 ± 0.13	0.44 ± 0.05	0.58 ± 0.07	$\mathbf{0.60 \pm 0.05}$	0.60 ± 0.06
Matthews corr. coef.	0.35 ± 0.17	0.14 ± 0.04	0.15 ± 0.06	0.39 ± 0.06	$\mathbf{0.41 \pm 0.05}$
AUC	0.74 ± 0.11	0.65 ± 0.03	0.63 ± 0.07	0.80 ± 0.03	$\mathbf{0.81 \pm 0.02}$

According to Table 5.1, the classifier CLS D, which first predicted VAD values, may perform slightly better than CLS C, which predicted categories directly. This table also shows that the best results were obtained for the VAD model in the Arousal dimension. However, it should be noted that Arousal only has two different values, i.e. Excited and Neutral, while Valence has three, namely Positive, Neutral, and Negative.

To better understand the VAD results, Figure 5.2 shows the VAD predicted values vs the annotated values. Straight lines in the figure show the borderlines learnt to discretise the problem, i.e., to transform the regression problem into a categorisation one. This figure shows the good performance of such a simple network. The low R2 scores obtained in the model's prediction of Arousal (-0.183), Valence (0.257) and Dominance (-0.637) can be attributed to the high level of noise present in the crowd-annotated corpus, as evidenced by the significant disagreement among annotators. Despite this, the model's ability to classify discrete emotions with moderate accuracy suggests that it can still capture meaningful patterns in the data. As a remark, opposite diagonals are very low-density regions.

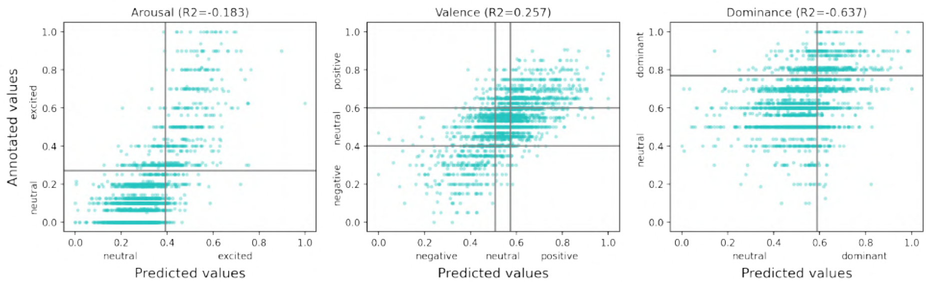


Figure 5.2.: Comparison of predicted and annotated values for Arousal, Valence, and Dominance. The dotted lines represent the splitting boundaries for each category, producing a confusion matrix with points for each data point.

Regarding Arousal, it seems easier to accurately predict higher values than lower ones since they are more scattered in the lower part of Figure 5.2. In Valence, it can be concluded that positive and negative categories are sometimes mixed with Neutral, but rarely among each other (see secondary diagonal in the figure). Finally, Dominance shows a lower correlation between the predicted and annotated values.

3 | Analysis of classification results

Figure 5.3 shows three projections of VAD values in a 2D space, in which points are coloured according to the annotated category. In the first row, both the VAD points and the categories are the ones perceived by annotators. The VAD points in the second row are the predicted ones but they are still coloured according to the perceived emotions. In the third row, the network predicts the VAD points and their category colours.

The first row of the figure shows mixed VAD points due to the variability of human emotions, but some patterns can be observed. For example, "Angry" samples show higher Arousal than "Calm" and "Happy", and in terms of Valence, "Happy" is the most positive emotion, followed by "Calm", which is neutral, and "Angry" which indicates negative Valence. The dominance axis does not show clear boundaries between the categories.

The second and third rows of Figure 5.3 show a transformation of the emotional space, where points correspond to the model's predictions and the boundaries between categories are more clearly defined. This suitable transformation may be attributed to the collaboration between the VAD and categorical models in decision-making. Interestingly, the annotated categories in the second row are even better separated than in the first row, where the annotated VAD val-

Figure 5.3.: Correlation between the VAD and categorical models, according to both the annotated and predicted data.

ues define the points. Similar relationships between the categories and the VAD axes can be noticed in the last two rows. In addition, the dominance axis shows that "Calm" is less dominant than "Happy" and "Angry", which is consistent with our expectations.

4 | Layer-by-Layer Analysis of the Model

In this section, we perform a visual analysis of the output of each network layer in order to understand how the network layers process data and how they contribute to the overall classification performance. In addition, patterns and trends in data representation that may not be noticeable otherwise can be observed by visualising the output of each layer in a two-dimensional space.

The dimensionality reduction techniques, like PCA, allow us to represent the output of each layer of the trained network in a two-dimensional space, simplifying the interpretation and visualisation. Accordingly, Figure 5.4 shows the progression of data representation through the layers, which results in effective classification outcomes. The dots of this visualisation indicate the Principal Component decomposition of the outputs of each convolutional layer. The colour of each dot indicates the corresponding category, with the first row representing the ground truth, i.e. the annotated categories, and the second row displaying the predicted categories.

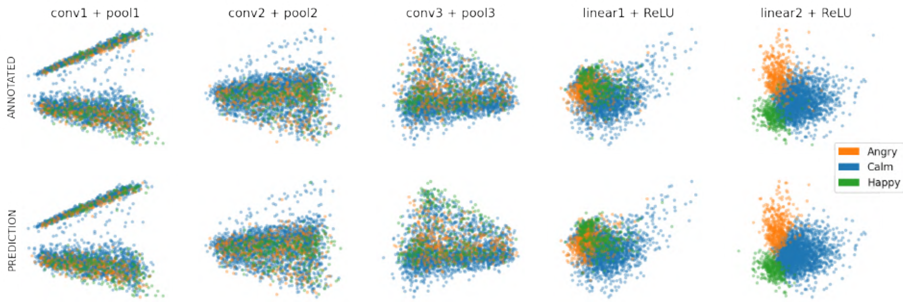


Figure 5.4.: The two-dimensional representation for each sample over different network layers using the PCA technique. The colour of each point represents the annotated categories (ground truth) in the first row and the predicted categories in the second row

The analysis of the plots in Figure 5.4 shows that as the network progresses, the samples become better defined and separated into categories. This observation is evident by the clear boundaries formed around each category in the final stages of the network, as seen on the right-hand side. It is worth noting that the reduction from 3416 to 2 in the "conv1 + pool1" stage (left image) is significantly higher than the reduction from 32 to 2 in the "linear2 + ReLU" stage (right image).

Furthermore, let us compare the predicted categories (second row) with the annotated values (first row) in the right image of Figure 5.4. We can observe similar boundaries for the annotated and predicted categories, which exhibits the network's capacity to provide reasonable predictions. These images also show that the predicted categories tend to include more *Angry* and *Happy* emotions that could be attributed to the oversampling method, which makes the minority classes more prominent.

5 | DeepDream in Speech Emotion Recognition

Class Model Visualisation is a technique within the Explainable Artificial Intelligence (XAI) framework that aims to generate image visualisations of the classes or categories the system attempts to predict (Simonyan et al., 2014). We decided to employ this method because it allows us to visually understand the features that the system considers when making predictions. A network f and a class of interest c are considered to generate these images. The goal is to generate an image visualisation I', which is representative of c, by maximising the class probability score $S_c(I)$ as described in Equation 5.1:

$$I' = \arg\max_{I} S_c(I) - \lambda \|I\|_2^2 \tag{5.1}$$

where λ stands for the regularisation weight.

The generated images, called Deep Dream images, provide information on what the neural model has learned for a specific class or category (Das and Rad, 2020). In this study, the Deep Dream images associated with different classes for both the categorical and VAD models are presented in Figure 5.5. This figure shows different frequency distributions for each category and dimension and, thus, different behavioural patterns. However, interpreting the obtained Deep Dream images is challenging because they encode several variables, such as the speaker's characteristics, the content of messages, and the environmental conditions, such as noise and channel.

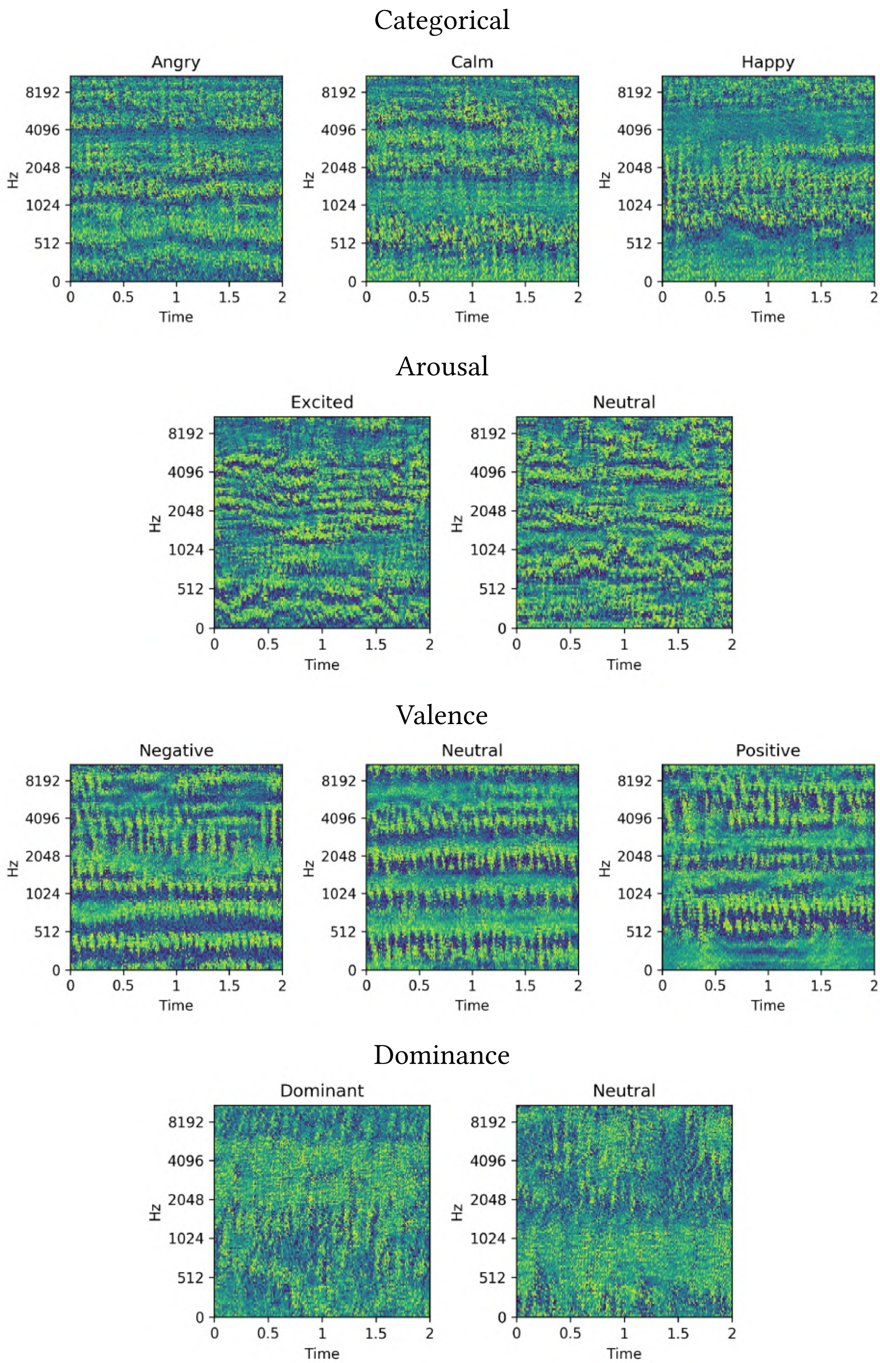


Figure 5.5.: Extraction of the suitable spectrogram that maximises the classification output for each class using the DeepDream technique.

In order to evaluate the capacity of the Deep Dream image to represent different categories, we propose the following experiment. We focus on the band above 512 Hz that shows low intensity in the Happy category but higher intensity in Calm, as Figure 5.5 shows. A sample labelled as Happy, whose spectrogram also showed low-intensity values in the band above 512 Hz, was selected. Then, we gradually increased the intensity values in that band until achieving the spectrogram shown in Figure 5.6. During this process, the values of the network's last layer responsible for making predictions were recorded and presented in Figure 5.7. This figure shows how the prediction changes from Happy to Calm as the intensity in the frequency band above 512 Hz increases.

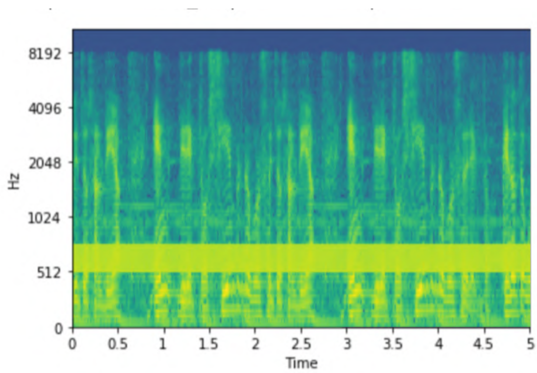


Figure 5.6.: Spectrogram modified to alter the network prediction from "Happy" to "Calm", intensifying frequencies above 512 Hz.

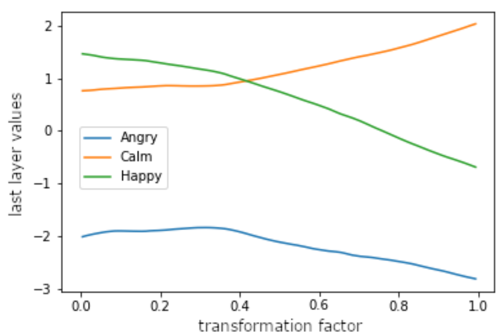


Figure 5.7.: Representation of how the values of the network prediction change when applying changes by a factor in the spectrogram in Figure 5.6.

Thus, a qualitative comparison of the Deep Dream images can provide insight into the achieved results, helping to understand why the system's predictions are sometimes inaccurate and some classes are mixed.

5.1 | Profile Analysis of Spectrograms

Thanks to the promising results in the trial described in the previous section (Section 5), an attempt was made to transform examples from one class to another and observe if the network could adapt its decision-making and correctly classify the transformed samples within the targeted emotional category.

This transformation process was executed through a methodology that involved the deletion of the initial frequency profile of the predicted category by the neural network, followed by the addition of the desired frequency profile of the target category. This methodology is explained in-depth in deVelasco et al. (2023) (in P19). Furthermore, a confusion matrix (Table 5.2) was employed to evaluate the performance of this approach. This confusion matrix shows how accurate the predictions are after the transformation process.

Table 5.2.: Confusion matrix with the percentage of correctly classified samples after profile transformation for each category (all test samples). Each sample has been transformed to the profile of each class and therefore each row sums up to 100%.

		Obtained category		
		Angry	Calm	Happy
Target category	Angry	97.67 %	2.33 %	0 %
	Calm	0 %	89.33 %	10.66 %
	Happy	8.00 %	0 %	92.00 %

The transformation may introduce noise that might lead to peak values that the system could misinterpret, leading to errors. However, the good results suggest that the deep dream images could be good representations of what the neural network learns for each category. Furthermore, for misclassified samples in this experiment (8% of “Happy” that were classified as “Angry”, 10% of “Calm” that were classified as “Happy”, and 2% of “Angry” that were classified as “Calm”), a re-annotation process might be considered in order to see whether the noise comes from the subjectivity associated with the annotation procedure. In this way, we could consider this procedure as a metric to evaluate the annotation quality.

Chapter 6

Conclusions

This research mainly focuses on improving our understanding of human-human and human-machine interactions by analysing paricipants' emotional status. For this purpose, we have developed and enhanced Speech Emotion Recognition (SER) systems for both interactions in real-life scenarios, explicitly emphasising the Spanish language. In this framework, we have conducted an in-depth analysis of how humans express emotions using speech when communicating with other persons or machines in actual situations. Thus, we have analysed and studied the way in which emotional information is expressed in a variety of true-to-life environments, which is a crucial aspect for the development of SER systems.

This study aimed to comprehensively understand the challenge we wanted to address: identifying emotional information on speech using machine learning technologies. Neural networks have been demonstrated to be adequate tools for identifying events in speech and language. However, the variability and intrinsic subjectivity of the proposed task, as well as the large number of possible configurations and network hyperparameters, make it difficult to take adequate decisions for the design and improvement of SER systems. As a consequence, we have carried out a considerable amount of experiments (more than 1700) to analyse, develop, assess and evaluate different models of emotions in a heterogeneity of communication scenarios, different ways to understand the labels, a variety of speech representations and network architectures, training configurations and a diversity of involved hyperparameters.

The process of experimentation has involved the gradual acquisition of knowledge, which has been documented in the works presented in this thesis. Most of them aimed to make local comparisons between some specific aspects; thus, the experimental conditions were tailored to each particular analysis. The experiments across different articles (from P1 to P19) are hardly comparable due to our continuous learning of dealing with the difficult task of identifying emotions in speech. In order to make a fair comparison, additional unpublished results are presented in the Appendix. These experiments were carried out under identical and rigorous conditions. This general comparison offers an overview of the advantages and disadvantages of the different methodologies for the automatic recognition of emotions in speech. Indeed, they allow us to analyse the effect of the aforementioned variability and diversity.

1 | Summary of Achievements

The following achievements summarise our contributions:

- We found **crowd annotation methodology** particularly useful for creating emotional speech corpora when considering real scenarios such as human debates or interviews. This method allows for capturing a wider range of emotional expressions since different people may perceive and label emotions differently based on their experiences. Therefore, the crowd annotation methodology can provide a more diverse and representative corpus of emotions in speech. Alternatively, we also used **expert annotation methodology** based on knowledge and experience. While an expert annotation can be very valuable, it may also be limited by their individual biases and perspectives. However, the expert annotation analyses the whole interaction and thus gets a better understanding of the context where the emotion appears. In contrast, analysing individual segments does not allow us to consider that context.

- We **developed**[1] and **analysed different emotional speech databases**, consisting of human-machine interactions between older adults and virtual agents, human-human interactions in a TV debate program and interactions with individuals suffering from depression or anxiety. Furthermore, the design and analysis of emotional speech databases have provided valuable insights into the relationship between emotions and communication frameworks. Our analysis revealed differences in emotions across these interactions, highlighting the importance of considering the specific context when analysing emotional speech. For example, in the TV debate corpus, emotions tended to be more intense and polarised, while in interactions with older adults, emotions were more calm and positive.

[1] The development of the databases has been carried out in collaboration with different partners in the projects already mentioned in Chapter 2.

- We explored different **speech representations** that could be used for emotion detection. Specifically, we considered acoustic features commonly used for speech processing, such as pitch, intensity, and spectral information. We also examined spectrograms, representing the speech signal in a convenient way to input Convolutional architectures. We finally choose Wave Embeddings, which encode the speech signal into a low-dimensional vector space. After an in-depth analysis, we found that Wave Embeddings, which capture both the temporal and spectral characteristics of the speech signal, were the most effective approach for detecting emotions.

- We explored different architectures for emotion recognition using neural networks, including MLP, Convolutional Networks, and Transformers as Attention-Based Networks. By doing so, we could identify each approach's strengths and weaknesses.
  - The **Multilayer Perceptron (MLP)** is a widely used neural network for emotion recognition, as it is relatively simple and efficient to train. Our research demonstrates that MLPs effectively handle small datasets, such as the spontaneous emotion corpora developed in our study, particularly with the Wave Embeddings representations (see the Appendix).
  - **Convolutional Networks** have shown great success in image and speech recognition tasks. They are beneficial when the input data are two-dimensional, such as the frequency and time dimensions of the speech Mel-Spectrogram, automatically learning and extracting relevant features for further classification.
  - **Transformers** are a more recent type of neural network that include attention-based networks and have shown to be highly promising in natural language processing tasks. They are particularly useful when the input data has a sequential structure, such as text or speech signals. Transformers can capture the relationships between different parts of the input data through self-attention mechanisms. Evidence of the effectiveness of these models lies in the fact that Wave Embeddings models are a combination of convolutional and transformer architectures.

- We deeply **analysed numerous configurations** of network architectures and acoustic representations, testing additional options for each network and audio representation and a variety of hyper-parameters such as learning rate, batch size, optimiser and activation functions or oversampling methods. The study also considered variety in communication scenarios on two different corpora (L6N and Empathic) and the classification tasks addressed in this report (categorical and dimensional models). According to the analysis, Wave Embeddings performed better than any other method on all tasks (see the Appendix). Despite being the simplest model, the MLP can obtain all necessary information from the speech representation to achieve the highest level of performance.

- We proposed using **XAI techniques** to analyse and understand the decision-making process in our emotion recognition models. We analysed the behaviour of the model layer-by-layer and proposed a class model visualisation using deep dream techniques as a simple interpretation method. This work aimed to understand the model's decision-making in emotion identification, improving our comprehension of computational methods in this task.

2 | Conclusions and Future Research

The methodologies developed in this research provide an opportunity to explore the role of emotions in human-machine interactions. Developing emotional speech recognition systems can contribute to robotics, customer service, and healthcare. For example, robots could be developed to recognise and respond appropriately to emotional cues, creating more empathic and natural interactions with humans.

In addition, the techniques and methodologies developed in this study can be applied to other fields beyond emotional speech recognition. For instance, methodologies in this thesis can also be applied to analyse speech to get indicators of some pathologies such as Parkinson's, Alzheimer's and COVID-19 diseases.

It also opens up new opportunities for research to explore the integration of emotion detection into natural language processing and the development of conversational agents. The ability to analyse conversations in context, including emotional context, could lead to a deeper understanding of human interactions and improve the development of conversational agents.

While this study contributes to the field of emotion recognition from speech, there are still several limitations and opportunities for future research. Based on the findings and conclusions of this study, some potential future research directions are:

- Investigation of new neural network models: While this study analysed several neural network models for emotion recognition in speech, new models are constantly being developed. New methods and architectures for processing time- and frequency-related connections in speech might be explored.
- Addressing the challenge of cross-cultural emotion recognition: This study mainly focused on emotion recognition in Spanish speech, but there are

significant differences in emotional expression across cultures and languages. Future research could investigate methods for cross-cultural emotion recognition, including developing multilingual models.

- Researching real-time emotion recognition in natural conversations: While this study mainly focuses on offline emotion recognition methods, future research could investigate methods for real-time emotion recognition in live conversations. The field of human-machine interaction, such as robotics and human monitoring, could benefit from this kind of research.
- Exploring the potential of emotional speech recognition in detecting and monitoring soft mental health conditions: The connection between emotions and mental health is evident, and the ability to recognise and track changes in emotional expression through speech could aid in the early detection and treatment of mental health disorders.

By addressing these areas, future studies can build on the contributions of this research and further advance the field of emotion identification in speech.

Part II.

Additional Content

Appendix A

Experiments in comparable conditions

All the experiments in Chapter 4 (except those in Section 4, which summarises the information in this appendix) are connected to papers that have already been published. Therefore, a chronological issue related to the experimental research process is translated into different conditions. These experiments had minor variations in different aspects, making it difficult to compare them directly. Conversely, the experiments in this appendix change several aspects of the used models, such as the network structure, the optimiser, or the oversampling method, keeping the rest of the hyper-parameters static for comparison.

The first step was determining which partitions must be used in the cross-validations for each corpus. Ten new partitions were defined in the case of *La Sexta Noche*. In *Empathic*, the partitions were agreed upon with different consortium members that developed the project.

All the models were trained for a maximum of 5000 iterations in all the experiments using the CrossEntropyLoss as the cost function. However, the type of input data, the balancing method, the network structure, or the optimiser changed from one experiment to another.

Additionally, two independent annotators were asked to annotate the first validation partition of the *La Sexta Noche* corpus to assess the feasibility of an independent annotator acting as a model for recognising emotions in speech, given the subjective nature of the task. The result was that the annotators came up with F1 Scores of 0.529 and 0.527, respectively. So far, the results obtained with this corpus outperformed these results, which indicates that we were right on track. From now on, when referring to the cross-validation partition annotated by independent annotators, we will refer to it as “cv0”.

1 | Transformers

Given the success of transformers in different areas, such as Natural Language Processing (NLP) (Sutskever et al., 2014; Brown et al., 2020; Galassi et al., 2020), Computer Vision (CV) (Parmar et al., 2018; Carion et al., 2020; Dosovitskiy et al., 2020), or even audio processing (Dong et al., 2018; Gulati et al., 2020; Chen et al., 2021), the group of experiments presented in this section explore the feasibility of using this type of structure in emotion recognition.

1.1 | Description of the experiments

The first set of experiments was carried out with *La Sexta Noche* corpus. The Mel-Spectrogram[1] was used as input. The transformer implemented is the same as the one used as an encoder in Devlin et al. (2018), but the number of layers and heads were modified. In addition, different oversampling strategies were tested. We also evaluated the effect of the batch size and looked for the most suitable learning rate for these network architectures with the Adam optimiser.

For these experiments, the optimiser being utilised is Adam with its default settings, except for the learning rate. Furthermore, some positional embeddings, which are also learned, are added to the input (the dimensions are not altered). Finally, we considered only the first output of the encoder's last layer for the classifier.

The following list shows how each of the above-mentioned hyper-parameters were changed:

- Learning Rate: **1e-3**, **1e-4** and **1e-5**.
- Batch Size: **8**, **16** and **32**.
- Oversampling methods:
  - **None**: No oversampling
  - **x2:** Duplicate minority category samples
  - **x4:** Quadruplicate minority category samples
  - **Max:** Increase the number of samples of non-majority categories to have the same number in all categories.
- Number of layers in the Transformer: **2**, **4** and **6**.
- Number of heads in the Transformer: **2**, **3**, **6**, **9** and **15**.

A total of 540 experiments were gathered, which will be analysed step by step in the following subsection.

[1]For the representation of the Mel-Spectrogram, a 128 filters bank has been extracted. For these experiments, only the 90 lowest filters in the bank are considered because the sampling frequency of some audios was lower than 16 kHz.

1.2 | Results

First, the best results in Figure A.1 are close to those achieved with the already published experiments. With the model configured with a 1e-4 learning rate, a batch size of 16, the oversampling method of x2 and a network configuration of 4 layers and 9 heads, an F1 Score performance of 0.57 was achieved. However, further analysis can give more information on which configurations affect the results.

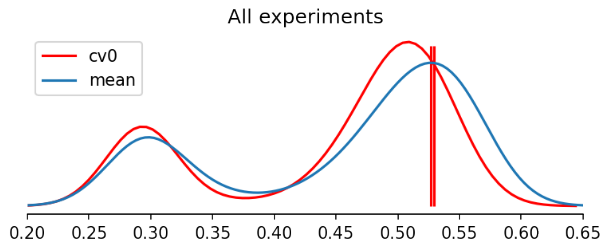


Figure A.1.: Density plot showing the F1 Score performance of all the models in the first set of experiments. The red line represents the F1 Score obtained only with the first iteration of the cross-validation to make it comparable with the independent annotators. The blue line represents the F1 Score obtained on average over all cross-validation iterations.

Foremost, we analyse the performance range of the results obtained in Figure A.1. The figure shows that the results with the first iteration generally performed slightly worse than those obtained with the independent annotators (red vertical lines), although some models outperformed them. The results with the average overall cross-validations (blue line) outperformed the ones obtained with the first cross-validation (red line). This probably indicates that this first cross-validation is one of the most challenging.

Two subgroups of results with F1-score values around 0.3 and 0.55 can be observed in Figure A.1. Bearing that the classification problem worked on is a 3-class problem, it is clear that the models that failed to distinguish between the different categories and consistently predict the same category are around 0.3, while those that succeeded in making a more significant distinction are around 0.55.

Figure A.2 shows that most models trained using the 1e-3 learning rate failed to achieve the desired performance and instead predicted a single category. Instead, models trained using 1e-4 and 1e-5 learning rates yielded promising results. However, the 1e-4 learning rate led to slightly better and more consistent results than 1e-5. Based on the previous results, we excluded the experiments with 1e-3 learning rate from the density subplots of Figure A.3.

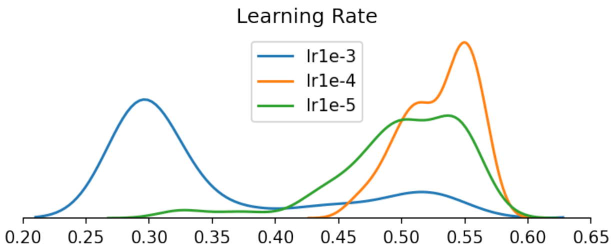
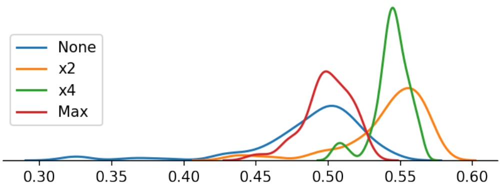
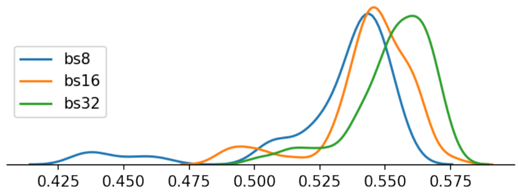


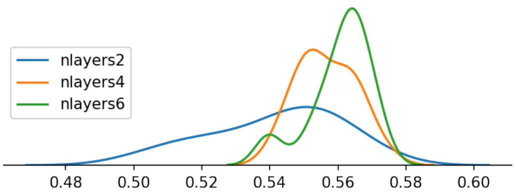
Figure A.2.: Density plots showing the F1 Score performance of all the models in the first set of experiments for three different learning rates.



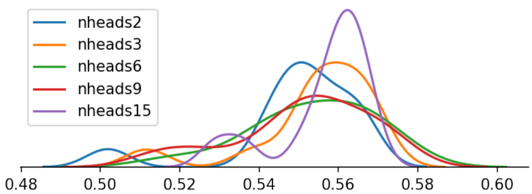
(a) Oversampling



(b) Batch Size



(c) Number of layers



(d) Number of heads

Figure A.3.: Density plot showing the F1 Score performance of all the models in the first set of experiments showcasing the separation of models by hyper-parameters: oversampling (a), the batch size (b), number of layers (c), and the number of heads (d).

The oversampling subplot (A.3a) shows that not oversampling (None in blue) or leaving all the categories with the same number of instances (Max in red) provided a slightly lower performance than those tests performed with a controlled oversampling that duplicate or quadruplicate the number of samples in minority classes (orange and green respectively). As a result, in the remaining subplots (batch size in A.3b, number of layers in A.3c, and the number of heads in A.3d), a new filter that eliminates samples that belong to the None and Max oversampling strategies was applied in addition to the one that removes experiments with 1e-3 learning rate.

Regarding the batch size (Figure A.3b), there were no significant differences, although there seemed to be a correlation that the larger the batch size, the better the results. However, every batch size value was kept for the analysis since all of them gave outstanding results.

According to Figure A.3c, a larger number of layers is a better choice in the network structure. The number of heads (Figure A.3d) did not directly correlate with the results, although the best performance was achieved when 15 heads were considered. Therefore, the choice of the number of heads was not crucial.

Table A.1.: The performance of the best model for each hyper-parameter, along with the 10 top-performing experiments from the first round of experiments.

#	Learning Rate	Batch Size	Oversampling	N. Layers	N. Heads	F1-Score
1	**1e-4**	**16**	**x2**	**4**	**9**	0.574
2	1e-4	**32**	x2	4	9	0.573
3	1e-4	32	x2	**6**	**6**	0.572
4	1e-4	32	x2	**2**	6	0.570
5	**1e-5**	32	x2	6	9	0.570
6	1e-4	32	x2	6	**3**	0.570
7	1e-4	32	x2	6	**15**	0.567
8	1e-4	32	x2	4	15	0.567
9	1e-4	32	x2	4	3	0.566
10	1e-4	32	x2	2	3	0.566
11	1e-5	32	x2	6	**2**	0.565
14	1e-4	32	**x4**	6	6	0.564
43	1e-4	**8**	x4	6	3	0.556
82	1e-4	32	**None**	6	6	0.548
84	**1e-3**	32	x2	2	9	0.547
173	1e-4	16	**Max**	2	15	0.529

Table A.1 displays the best-performing models for each hyper-parameter and the top 10 models. It allows for better visualisation of which hyper-parameter configuration performed the best. For the best cases of 1e-3 learning rate, not oversampling (None) or maximum oversampling (Max) approaches, the results were in positions 84, 82 and 173, respectively, out of a total of 540 experiments carried out, which leads us to conclude that they are not the best options. The performance of the best case for batch size 8 was in a good position (43) but far from the other options (1st place for 16 and 2nd place for 32). The remaining top configurations for each hyper-parameter were in the top 15, suggesting they are also suitable choices.

2 | Transformers and Convolutional Networks

In the second set of experiments, transformer-based attention networks were explored and compared to other network structures based on 1D and 2D convolutions.

2.1 | Description of the experiments

The second set of experiments used the same corpus as the first one, *La Sexta Noche*, and the input was the Mel-Spectrogram for all experiments. Oversampling was fixed (x2) to avoid an exponential increase in the number of experiments.

The best option (32) and a larger one (64) for the batch size of the previous experiments were considered. In addition, different optimisers were tested, including the Adam and AdamW (Loshchilov and Hutter, 2017) optimisers, which perform well on attention networks. For convolutional models, the classical SGD optimiser was also evaluated.

A 6-layer, 8-head network structure was kept in the transformer-based attention networks; including more layers seemed to be a good idea because there was a slight correlation between size and performance. However, fewer heads were chosen to speed up the experimentation process since the number of heads seems independent. As for the learning rate, 1e-4 was chosen because it was one of the best-performing values.

A grid search of 5 hyperparameters was carried out for the transformer-based attention networks. First, the aforementioned batch size and optimisers were tested. In addition, different values for the beta[2] parameters of the optimisers were compared with the standard option (0.9, 0.999) and the recommended values for the transformers (0.9, 0.98). Then, an attempt was made to investigate how the performance changed according to the treatment of positional embeddings. To this end, we added them to the input (Sum) without affecting the dimension of the input as in the previous experiments, or concatenated them (Concat) changing the input dimension. Finally, we also explored three options for the connection between the output of the transformer and the last classifier layer. These options were: taking only the first instant of the output sequence (First) as done in the previous experiments, averaging over all time instants

[2]The hyper-parameters β_1 and β_2 of Adam were initial decay rates used when estimating the first and second moments of the gradient, which were multiplied by themselves (exponentially) at the end of each training step (batch).

(Mean), or learning a new token to be added to the input sequence (LearnableSpecialToken). In total, 48 models were tested for the transformer models.

- Batch Size: **32** and **64**.
- Optimiser: **Adam** and **AdamW**.
- Beta parameter: **(0.9,0.999)** and **(0.9,0.98)**.
- Positional embedding: **Sum** and **Concat**.
- Encoder to Linear: **First**, **Mean** and **LearnableSpecialToken**.

We performed grid searches of the same five parameters for the unidimensional (1D) and bidimensional (2D) convolutional models. First, the batch size and optimisers were tested, as mentioned earlier. In addition, for each optimiser configuration, different learning rates were also evaluated since the appropriate learning rate may depend on the network structure. Then, attempts were made to include either a portion or the entire Mel-Spectrogram, with a 90-filters bank covering 3500Hz and a 128-filters bank covering 8000Hz, respectively. Finally, different activation functions for the convolutional layers were considered. All this makes a total of 72 experiments for each type of network structure (1D and 2D), which means 144 experiments. In short, these are all the hyper-parameters tested for the convolutional networks:

- Optimiser: **SGD**, **Adam** and **AdamW**.
- Convolution type: **Conv1D** and **Conv2D**.
- Learning Rate: **1e-3** and **1e-4**.
- Batch Size: **32** and **64**.
- Activation Function of Convolutional layers: **None**, **ReLU** and **Sigmoid**.
- Mel-Spectrogram, number of filters in bank: **90** and **128**.

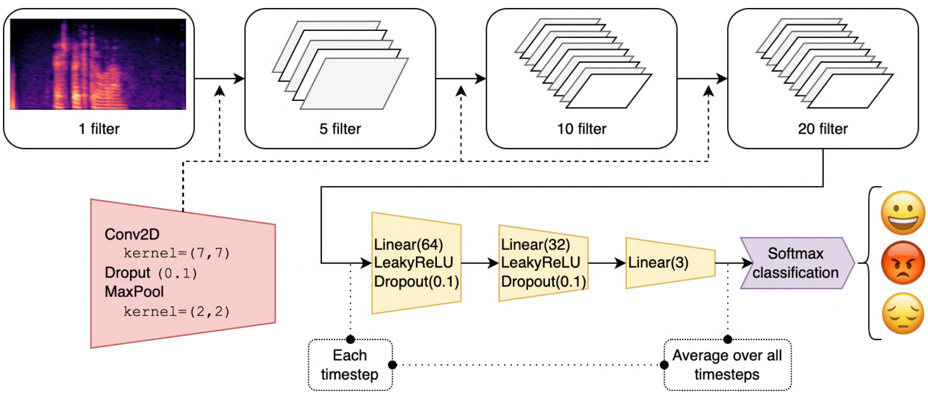


Figure A.4.: 2D convolutional network used in the final sets of experiments.

The convolutional networks were structured based on the approach used in the work deVelasco et al. (2022a) (in P16). The bidimensional model included a 2D convolutional network that processed the Mel-Spectrogram as an image (Figure 20.4). This network sequentially applied three 2D convolutional layers, each with a 7x7 kernel. The number of filters or channels increased while reducing both the temporal dimension and the Mel space with the help of a max-pooling layer, aiming to extract emotional features from the Mel-Spectrogram. The emotional state was then evaluated using three linear layers of 64, 32, and 3 neurons for each temporal state, and the results were averaged for the final prediction.

The unidimensional convolutional network (Figure A.5), on the other hand, employed two 1D convolutional layers with a kernel size of 15 and 5, respectively, to reduce the number of filters or channels in the original Mel-Spectrogram filters bank from 90/128 filters to 20 filters while also decreasing the temporal dimension. To further reduce the temporal dimension, 1D max-poolings with a kernel size of 5 were applied. The resulting output was then averaged across the temporal dimension and passed through the classification network.

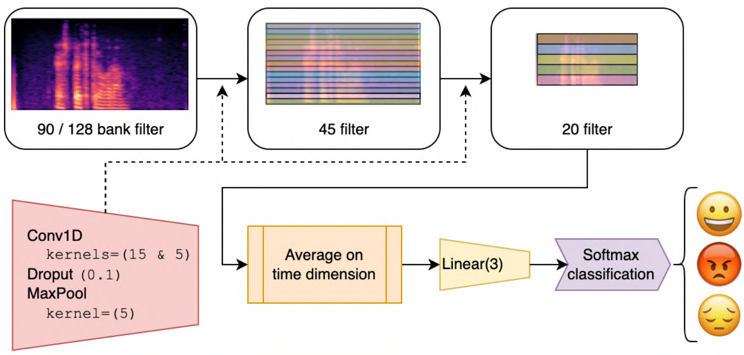


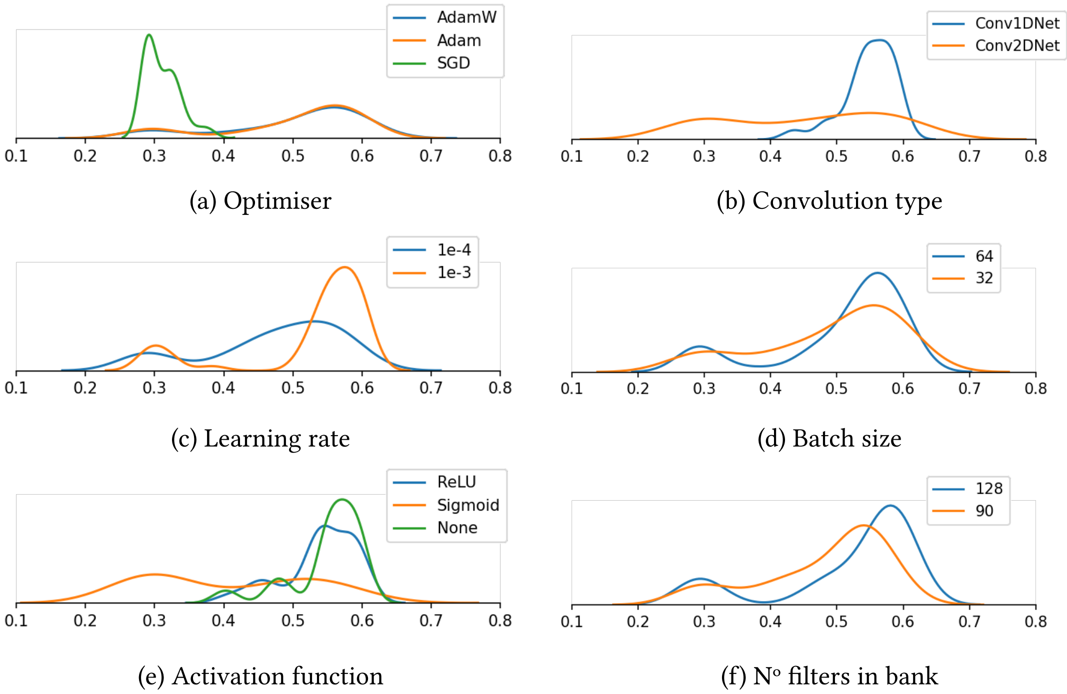
Figure A.5.: 1D convolutional network used in the final sets of experiments.

2.2 | Results

Before comparing the transformer networks with the convolutional ones, it is important to note that the transformer networks were already tested in Section 1. As a result, configurations that yielded poor performance were removed. Therefore, a separate analysis of both network types will be carried out to compare the remaining configurations.

2.2.1 | Convolutional Networks

We first show experiments carried out with 1D and 2D convolutional networks. Figure A.6 shows two subgroups of results: those that failed to make a correct classification and achieved approximately 0.3 F1 Score and those that obtained better results, around 0.55.



(a) Optimiser

(b) Convolution type

(c) Learning rate

(d) Batch size

(e) Activation function

(f) Nº filters in bank

Figure A.6.: Density plot showing the F1 Score performance of convolutional models in the second set of experiments per set of hyper-parameters: optimiser (a), convolution type (b), learning rate (c), batch size (d), activation function (e), and number of Mel-frequency filters in bank (f).

Figure A.6a shows that the SGD optimiser got an F1 Score of 0.3. Instead, Adam or AdamW optimisers usually led to better results. Thus, results obtained with the SGD optimiser were not included in the other subplots in Figure A.6.

1D convolutional network structures yielded more consistent results than 2D ones. However, as we will show later, the 2D convolutional network achieved higher maximum F1 Score values than its 1D counterpart. Therefore, while the 1D convolutional network is more stable, the 2D network can perform better in some scenarios.

Concerning the learning rate and batch size, higher values (1e-3 in the learning rate and 64 in the batch size) resulted in more stable results, making them both preferable choices. Among the activation functions used, the sigmoid was the one that gave the worst results, obtaining similar performances both using the ReLU function and without using any activation function. Finally, the 128-filters bank seems to be a more suitable option.

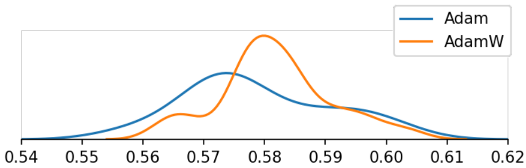
Table A.2.: The experiments corresponding to the best option for each convolution network hyper-parameters and the 10 best performance experiments.

#	Optimizer	Conv. Type	Learning Rate	Batch Size	Act. func.	Nº Mel-freq	F1-Score
1	**AdamW**	**Conv2**	**1e-3**	**32**	**None**	**128**	0.608
2	AdamW	Conv2	1e-3	**64**	None	128	0.604
3	AdamW	Conv2	1e-3	64	**ReLU**	128	0.603
4	**Adam**	Conv2	1e-3	64	None	128	0.602
5	Adam	Conv2	1e-3	32	ReLU	128	0.600
6	AdamW	**Conv1**	1e-3	64	ReLU	128	0.599
7	Adam	Conv1	1e-3	64	ReLU	128	0.597
8	Adam	Conv2	1e-3	32	None	128	0.596
9	Adam	Conv2	1e-3	32	None	**90**	0.593
10	Adam	Conv2	1e-3	64	ReLU	128	0.593
12	Adam	Conv1	**1e-4**	64	None	128	0.590
17	AdamW	Conv1	1e-3	64	**Sigmoid**	128	0.585
82	**SGD**	Conv1	1e-3	64	None	128	0.379

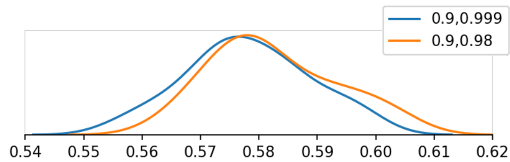
Table A.2 presents the best configuration for each hyper-parameter and the top 10 performing convolutional networks' results. The table confirms that the SGD optimiser could not outperform Adam and AdamW, as its highest-ranked option was placed in position 82 out of 144. Regarding network structure, although the 1D convolutional network showed more stability in the previous analysis (Figure A.6b), the 2D convolutional network obtained the best performance in this case. The 1e-3 learning rate was slightly more effective, while the batch size did not impact the results significantly. The activation function can be either ReLU or None. Finally, it is suggested that the Mel-Spectrogram with a 128-filters bank is the adequate input.

2.2.2 | Transformer Networks

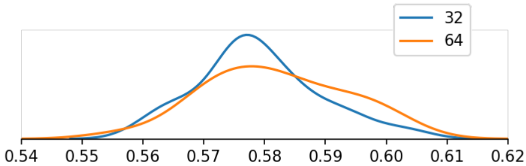
As described in Section 2.1, this set of experiments considers the best hyper-parameter configurations obtained from the experiments in Section 1. It then analyses other hyper-parameters, resulting in a reduced set of 48 experiments that yielded promising results. Looking at any subplot of Figure A.7, it can be observed that all of them fell within the range of outstanding results (between 0.55 and 0.61), outperforming the results of the first group of experiments.



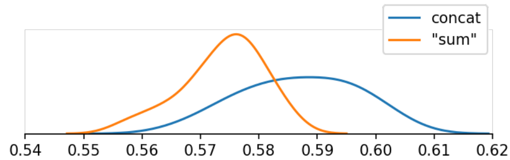
(a) Optimiser



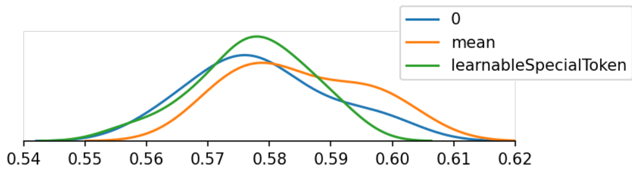
(b) Beta parameter



(c) Batch size



(d) Positional embedding



(e) Encoder to linear

Figure A.7.: Density plot showing the F1 Score performance of transformers models in this set of experiments for different hyper-parameters: optimizer (a), the beta parameter of the optimiser (b), the batch size (c), positional embedding (d), and the encoder (e).

In spite of these promising results, we evaluated the effect of each hyper-parameter on the performance. Attending to the optimiser, the results are similar to those obtained with convolutional networks, although the AdamW optimiser seems to have more stable results. As for the beta parameters of the optimiser, the recommended value for transformers performs slightly better, although the default values are not far behind. There are no significant differences in batch size, even though larger batch values seem to perform slightly better. The most significant differences are found in how positional embeddings are introduced. According to the results shown in Figure A.7d, the concatenation of the positional embeddings (concat) yields better results than their addition to the Mel-Spectrogram (sum) as done in the previous experiments. Finally, according to Figure A.7e, averaging over all the time instants is the best option for the transformer output used as input to the classifier network. This option slightly outperforms the results obtained, taking the first instant of the transformer output (done in previous experiments) and adding a new token to the input to be learned and taking the output of that time instant.

Table A.3.: The best option for each hyper-parameter of the transformers network architecture on this round of experiments as well as the experiments corresponding to the 10 best experiments.

#	Optimiser	Beta Param.	Batch Size	Pos. Emb.	Encoder to Linear	F1-Score
1	**AdamW**	**0.9,0.98**	**32**	**concat**	**mean**	0.603
2	**Adam**	0.9,0.98	**64**	concat	mean	0.601
3	Adam	0.9,0.98	64	concat	**0**	0.600
4	AdamW	**0.9,0.999**	64	concat	0	0.597
5	Adam	0.9,0.98	32	concat	mean	0.595
6	AdamW	0.9,0.999	64	concat	mean	0.595
7	Adam	0.9,0.999	64	concat	mean	0.595
8	Adam	0.9,0.98	32	concat	0	0.592
9	Adam	0.9,0.98	32	concat	**learnableSpecialToken**	0.592
10	AdamW	0.9,0.98	64	concat	mean	0.591
15	AdamW	0.9,0.98	32	**sum**	mean	0.585

In conclusion, Table A.3 displays the results of the transformer networks. It is evident from this table that the two hyper-parameters with the most significant impact on the results are the method of introducing positional embeddings and selecting the part of the transformer output to be used as input to the classifier.

2.2.3 | Comparing both architectures

After analysing the two structures independently, Figure A.8 compares their performances. The figure displays the density plot of the average of all cross-validations for each structure in a straight line and the density plot of the first cross-validation partition (cv0) in a dashed line, making it comparable with the annotation result by independent annotators (vertical dashed line).

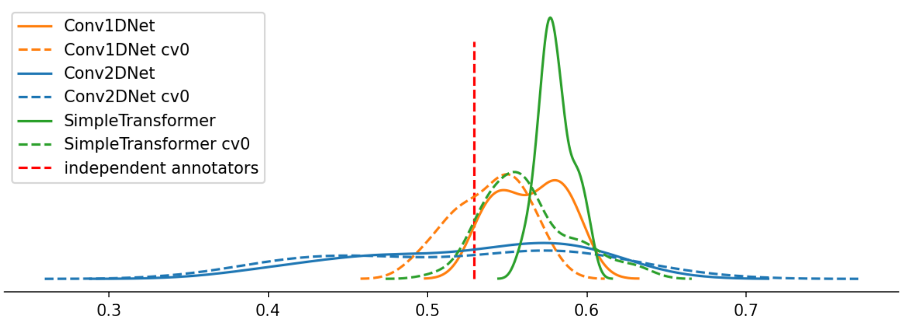


Figure A.8.: F1-Score performance comparison of convolutional and transformers network structures.

As Figure A.8 shows, and as shown in Tables A.2 and A.3 in Sections 2.2.1 and 2.2.2 respectively, all the structures achieve the maximum performance of 0.6 in F1 Score (0.608 with the 2D convolutions, 0.599 with the 1D convolutions and 0.603 with the transformers). However, the transformers (green in the graph) succeed in being the most stable and offer very similar performance in all experiments. In the case of the 1D convolutional networks, they do not achieve the stability of the previous ones but remain in an adequate range. In contrast, the 2D convolutions range from results close to 0.3 to the maximum achieved.

3 | Wave Embeddings

In this set of experiments, the Wave Embeddings representations extracted using neural networks seen in section 1.3 of Chapter 3 were tested.

3.1 | Description of the experiments

These experiments were also carried out with *La Sexta Noche* corpus. However, the Wave Embeddings were used for speech representations instead of the Mel-Spectrogram.

The experiments aim to explore the most appropriate embeddings and model versions for feature extraction before a separate neural model is developed to predict the emotional state. The following network architectures to get Wave Embeddings and model versions were evaluated in this experimentation:

- Hubert (Hsu et al., 2021):
  - facebook/hubert-**base-ls960** (768)
  - facebook/hubert-**large-ls960-ft** (1024)
  - facebook/hubert-**large-ll60k** (1024)
  - facebook/hubert-**xlarge-ls960-ft** (1280)
  - superb/hubert-**base-superb-er** (768)
  - superb/hubert-**large-superb-er** (1024)
  - superb/hubert-**large-superb-sid** (1024)
  - RamiEbeid/hubert-**base-ser** (768)
- UniSpeech (Wang et al., 2021):
  - patrickvonplaten/unispeech-**large-1500h-cv-timit** (1024)
  - microsoft/unispeech-**1350-en-168-es-ft-1h** (1024)
  - microsoft/unispeech-**large-multi-lingual-1500h-cv** (1024)
  - hf-internal-testing/tiny-**random-unispeech** (16)

- UniSpeech-SAT (Chen et al., 2022b):
  - microsoft/unispeech-**sat-base-100h-libri-ft** (768)
  - microsoft/unispeech-**sat-large** (1024)
  - microsoft/unispeech-**sat-base** (768)
  - microsoft/unispeech-**sat-base-plus** (768)
- Wav2Vec2 (Baevski et al., 2020):
  - facebook/wav2vec2-**base-960h** (768)
  - facebook/wav2vec2-**base** (768)
  - facebook/wav2vec2-**large-960h** (1024)
  - superb/wav2vec2-**base-superb-er** (768)
  - ehcalabres/wav2vec2-lg-xlsr-en-**speech-emotion-recognition** (1024)
- WavLM (Chen et al., 2022a):
  - patrickvonplaten/wavlm-**libri-clean-100h-base-plus** (768)
  - microsoft/wavlm-**large** (1024)
  - microsoft/wavlm-**base** (768)
  - hf-internal-testing/tiny-**random-wavlm** (16)

The model versions include information on how they have been trained, such as size (*base*, *large*, *xlarge*), the database used for training and its duration (e.g. ls960 for training on Libri Speech for 960 hours, *ll60k* for training on Libri Light for 60,000 hours), the task for which they were trained (*er* for emotion recognition, *ser* for speech emotion recognition, and *sid* for speaker identification), or if they were fine-tuned from another model (*ft*). These models can be found in the HuggingFace repository and are freely accessible. Only the bolded part of the model's name will be used when referring to a model throughout this study.

The Wave Embeddings were extracted, and the mean was applied as a functional[3]. A simple single-layer MLP was then used to obtain the emotional state. The number in brackets after the model name stands for the hidden layer's size.

In addition, the grid search also incorporated variations in optimisers, batch sizes, and oversampling configurations:

- Optimisers: **Adam** and **AdamW**.
- Batch Size: **32** and **64**.
- Oversampling methods:
  - **None**: No oversampling
  - **x2**: Duplicate minority category samples
  - **x4**: Quadruplicate minority category samples

[3]The mean is extracted in the training process at the batch level and is divided by the maximum batch length

– **Max**: Increase the number of samples of non-majority categories to get the same number in all categories.

3.2 | Results

Before analysing each grid search hyper-parameter, each checkpoint's performance is examined and shown in Figure A.9.

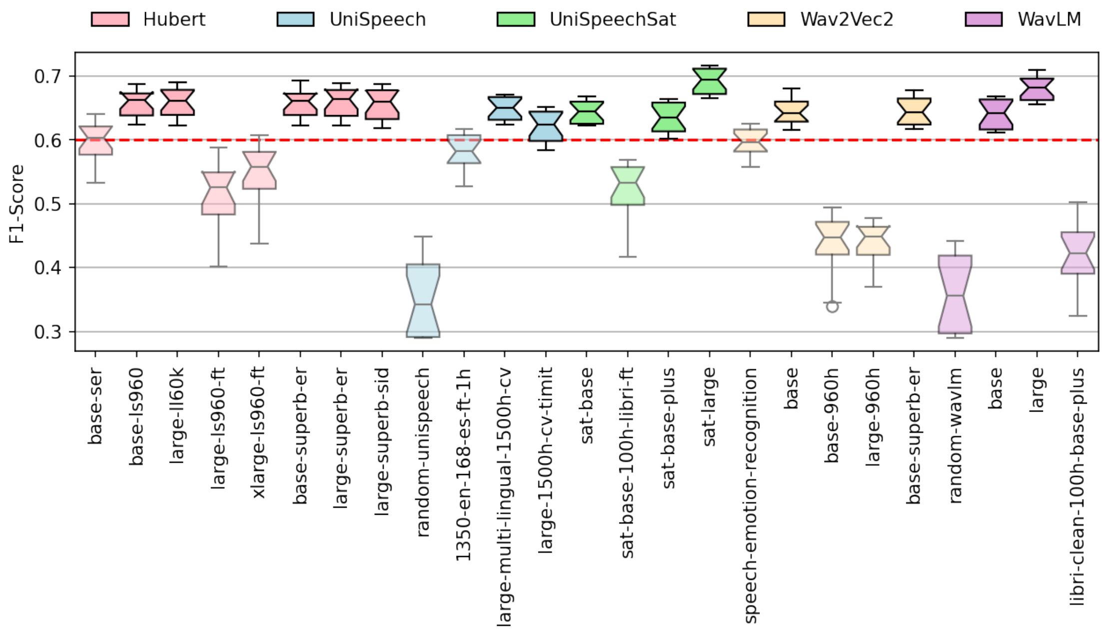
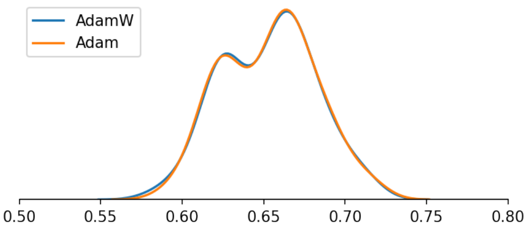


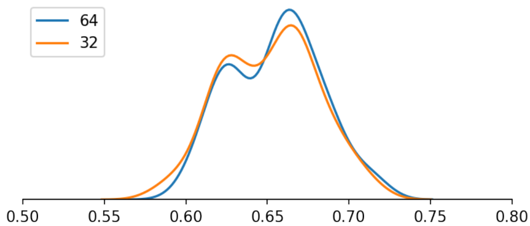
Figure A.9.: Boxplot of the F1-Score for each checkpoint. Each model is shown in a different color. A line has been drawn over the F1-Score of 0.6 to discard those checkpoints that do not reach it on average.

As shown in Figure A.9, most results are above the maximum obtained with the previous experimental sets. For this reason, all models that did not achieve an average F1-Score of 0.6 (red line) were excluded from the following analysis. It is worth mentioning that all models pre-trained for emotion recognition (those whose model names contain *er*, *ser* or *speech-emotion-recognition* of Wav2Vec2) either outperformed or were very close to outperforming the benchmark. However, none are among the top choices, including large models.

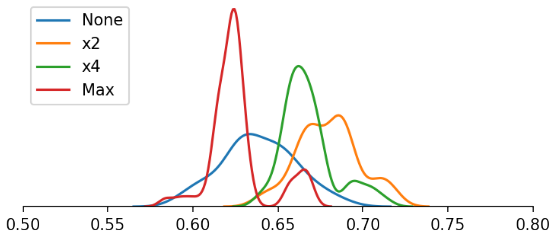
Once the models that did not meet the proposed minimum were discarded, we analysed each hyper-parameter of the grid search (Figure A.10). In this case, both *Adam* and *AdamW* optimisers perform almost identically. Regarding the batch size, it is still preferable to use a large one (64), although the difference is not very significant. The oversampling hyper-parameter, as expected, has the most significant impact on the results in this case. Repeating each sample twice achieves the best results while repeating each four times obtains slightly inferior performance, which is coherent with the previous results.



(a) Optimizer



(b) Batch size



(c) Oversampling

Figure A.10.: Density plot showing the F1-Score performance of the third set of experiments for different hyper-parameters: optimizer on top-left, the batch size on top-right, and oversampling on the bottom.

Table A.4 shows the best combination of each hyper-parameter and the F1-Score of the ten best combinations of parameters for these experiments. The UniSpeechSat embeddings are among the best choices with its *sat-large* model, followed closely by the WavLM ones with the *large* model, both occupying the top 10 choices.

Regarding the grid search hyper-parameters, the conclusions are similar to the ones presented previously. Oversampling performs best with an *x2* and an *x4* repetition, leaving the other options far behind. However, batch size and optimiser barely matter when looking at performance. Let us note that UniSpeechSat embeddings with the *sat-large* model appear in all the best hyper-parameter combinations.

For future experiments, the three best network architectures were chosen with their best models based on the average performance, which was plotted in Figure A.9 and is shown in detail in Table A.5.

Table A.4.: The experiments corresponding to the best option for each hyperparameter of the Wave Embeddings round of experiments as well as the 10 best performing experiments.

#	Embeddings	Model	Overs.	Batch S.	Optimizer	F1-Score
1	**UniSpeechSat**	**sat-large**	**2**	**64**	**Adam**	0.717
2	UniSpeechSat	sat-large	2	64	**AdamW**	0.716
3	UniSpeechSat	sat-large	2	**32**	Adam	0.715
4	UniSpeechSat	sat-large	**4**	64	Adam	0.713
5	UniSpeechSat	sat-large	2	32	AdamW	0.711
6	**WavLM**	**large**	2	64	Adam	0.709
7	WavLM	large	2	64	AdamW	0.709
8	WavLM	large	2	32	AdamW	0.708
9	UniSpeechSat	sat-large	4	64	AdamW	0.706
10	UniSpeechSat	sat-large	4	32	AdamW	0.704
15	**Hubert**	**base-superb-er**	2	32	AdamW	0.693
20	Hubert	**large-ll60k**	2	32	Adam	0.690
21	Hubert	**large-superb-er**	2	64	AdamW	0.689
24	Hubert	**base-ls960**	2	64	AdamW	0.688
25	UniSpeechSat	sat-large	**None**	64	Adam	0.688
27	Hubert	**large-superb-sid**	2	64	Adam	0.688
39	**Wav2Vec2**	**base**	2	64	Adam	0.680
41	Wav2Vec2	**base-superb-er**	2	64	AdamW	0.678
57	**UniSpeech**	**large-multi-lingual-1500h-cv**	2	64	AdamW	0.672
62	UniSpeechSat	sat-large	**Max**	64	Adam	0.669
64	UniSpeechSat	**sat-base**	2	64	AdamW	0.669
65	WavLM	**base**	2	32	AdamW	0.669
86	UniSpeechSat	**sat-base-plus**	2	32	Adam	0.664
131	UniSpeech	**large-1500h-cv-timit**	2	64	AdamW	0.651

Table A.5.: Mean and standard deviation of the performance of each network architecture and model.

N. Architecture	Model	F1-Score (avg. ± std.)
UniSpeechSat	sat-large	0.692 ± 0.020
WavLM	large	0.681 ± 0.020
Hubert	large-ll60k	0.659 ± 0.024
Hubert	large-superb-er	0.658 ± 0.024
Hubert	base-superb-er	0.658 ± 0.024
Hubert	base-ls960	0.658 ± 0.023
Hubert	large-superb-sid	0.657 ± 0.025
UniSpeech	large-multi-lingual-1500h-cv	0.649 ± 0.018
Wav2Vec2	base-superb-er	0.645 ± 0.023
Wav2Vec2	base	0.645 ± 0.022
UniSpeechSat	sat-base	0.644 ± 0.018
WavLM	base	0.641 ± 0.023
UniSpeechSat	sat-base-plus	0.635 ± 0.025
UniSpeech	large-1500h-cv-timit	0.622 ± 0.024
Wav2Vec2	speech-emotion-recognition	0.596 ± 0.022
Hubert	base-ser	0.596 ± 0.034
UniSpeech	1350-en-168-es-ft-1h	0.582 ± 0.027
Hubert	xlarge-ls960-ft	0.544 ± 0.054
UniSpeechSat	sat-base-100h-libri-ft	0.518 ± 0.050
Hubert	large-ls960-ft	0.511 ± 0.062
Wav2Vec2	large-960h	0.437 ± 0.038
Wav2Vec2	base-960h	0.434 ± 0.053
WavLM	libri-clean-100h-base-plus	0.420 ± 0.060
WavLM	random-wavlm	0.359 ± 0.064
UniSpeech	random-unispeech	0.354 ± 0.063

4 | Extending the Experiments: VAD Model and Crowd-Annotated *Empathic* Corpus

This experiment series aims to compare the previous findings across two different corpora and for each classification task.

4.1 | Description of the experiments

In this series of experiments, we analysed the VAD and the categorical model and the model performance in two different kinds of interactions. *La Sexta Noche* corpus contains human-human interactions, whereas the *Empathic* corpus is based on human-machine interactions. Once again, a grid search of different speech signal representations, types of network architectures, and types of oversampling methods was carried out.

For speech signal representation, the Mel-Spectrogram (with a 128 filters bank) and the three best Wave Embeddings representations (UniSpeechSat with the *sat-large* model, WavLM with the *large* model, and Hubert with the *large-ll60k* model) were used. Additionally, two commonly used acoustic feature sets in emotion recognition, GeMAPS and its extended version eGeMAPS (Eyben et al., 2015), were also included in the analysis.

Additionally, three network architectures were considered for modelling emotions. First, a simple MLP model was implemented as a very simple option. The temporal average was calculated for the Mel-Spectrogram and Wave Embeddings audio representations in this case. This network architecture consisted of a hidden layer of 64 neurons with the ReLUactivation function and the last classifier layer.

Then, the best configurations obtained from previous experiments were used to implement the 1D convolutional model and the transformer. LLDs were used as audio representations for GeMAPS and eGeMAPS in these two network architectures. For the 1D convolutional, no activation function was used in the convolutions. Moreover, in the case of the transformers, a 6-layer architecture was used with the minimum number of heads (2 or 5 depending on the input size), the positional embeddings were concatenated (concat), and for the last classifier layer, the average of all the time slices was considered (mean).

Finally, different oversampling methods were studied, including the common repetition (x4) experimented with in previous sections. However, SMOTE (Chawla et al., 2002) was also explored. This method creates new synthetic samples instead of repeating the existing ones.

To this end, SMOTE identifies which samples are borderline and then creates synthetic samples. Two methodologies were used to determine which samples were borderline. The first method (epsilon) defines a sample as borderline if the distance between the sample and the centre of the majority category is within 15% of the closest distances. In the second method (neighbour), a sample is borderline when at most 2 of the 6 nearest neighbours are from the same category. Once the borderlines are identified, the synthetic samples are generated considering the 4 nearest neighbours. If the sample is borderline, SMOTE creates 4 synthetic samples, otherwise 2. The SMOTE technique was applied only to the MLP and 1D Convolution networks, given the high requirement of computational resources of the transformers.

The remaining hyper-parameters (batch size, learning rate, optimiser, and specific configurations of each network structure) are fixed by the best results in the previous experiments. To summarise, we analysed the following hyper-parameters:

- **10-fold cross-validation**
- Corpora: ***La Sexta Noche*** and ***Empathic***
- Different problems: **Categorical**, **Arousal**, **Valence**, and **Domiannce**,
- Speech Signal Representations:
  - **Mel-Spectrogram**
  - Wave Embeddings: **UniSpeechSat**, **WavLM**, and **Hubert**.
  - Acoustic Features: **GeMAPS** and **eGeMAPS**.
- Oversampling method: **Repetition**, **SMOTE epsilon**, and **SMOTE neighbour**.
- Network Architecture with specific training hyper-parameters:

	Multilayer Perceptron	**Convolutional 1D**	**Transformers**
Batch Size	64	64	64
Learning Rate	1e-3	1e-3	1e-4
Optimiser	Adam	Adam	AdamW
beta parameter	(0.9,0.999)	(0.9,0.999)	(0.9,0.98)

4.2 | Results

Before analysing each hyper-parameter, it is helpful to see the performance obtained by each classifier (categorical, valence, arousal and dominance).

The boxplots shown in Figure A.11 show that the performance for the *La Sexta Noche* corpus is better than for the *Empathic* corpus. Additionally, the plot provides insights into which classification tasks are more challenging. The categorical model performs much better on *La Sexta Noche* because *Empathic*

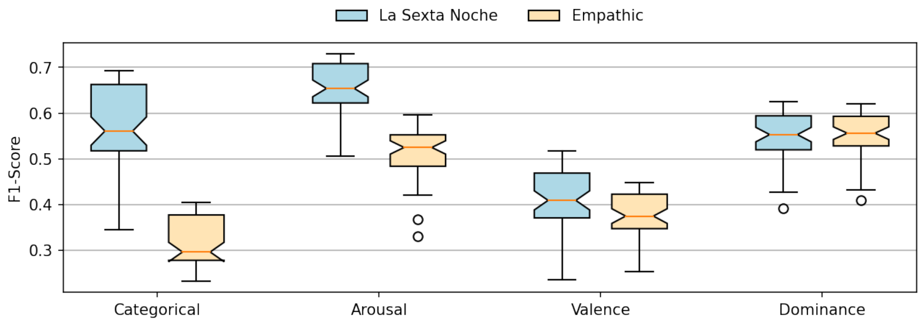


Figure A.11.: Boxplot of the F1-Score performance over both corpora in each classification task.

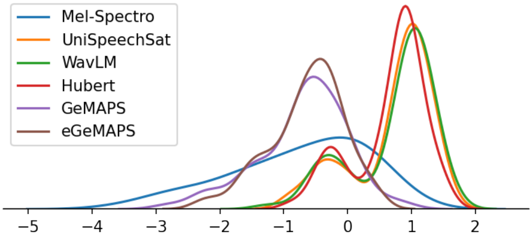
presents more emotion categories, and the emotions in the interaction are more subtle. However, the dimensional model was discretised similarly for each corpus, resulting in the same number of categories for each task: 2 for arousal, 3 for valence, and 2 for dominance. The performance of each classifier is related to the nature of the problem, and part of the challenges can be attributed to the class imbalance.

The experiments presented in Figure A.12 were standardised per task. Thus, results lower than 0.0 indicate worse results than the mean, and those higher than 0.0 indicate that they are better than the mean and are considered for future experiments.

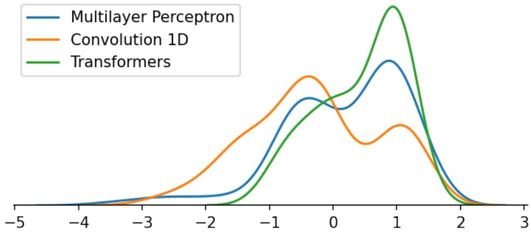
Figure A.12a shows that all audio representations generated with Wave Embeddings outperformed Mel-Spectrogram and GeMAPS. In spite of the no significant differences observed among the Wave Embeddings models, UniSpeech-Sat and WavLM performed slightly better in some particular cases.

Figure A.12b indicates that the transformer network obtained the most promising results regarding network architecture. However, the MLP produces surprising results despite its limited parameters. Note that time dependence was removed by using functionals. In tables A.6 to A.13, it can be observed that the MLP outperformed the best transformer network in many cases. Conversely, the 1D convolution architecture performed worse in most cases.

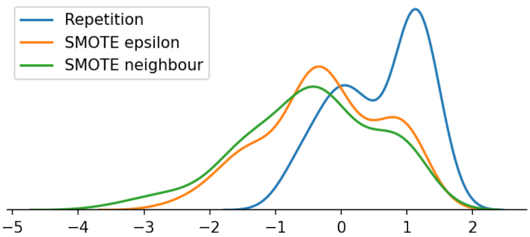
Finally, Figure A.12c indicates that common repetition yields the best results for oversampling. Furthermore, the two versions of SMOTE offer similar performance. However, the SMOTE technique is highly customisable and may produce better results if different configurations are implemented.



(a) Speech Signal Representation



(b) Network architecture



(c) Oversampling method

Figure A.12.: Density plots showing the performance of the fourth set of experiments divided by different hyper-parameters: speech signal representation (a), the network architecture (b), and oversampling (c).

The results for the categorical, arousal, valence, and dominance classifiers for *La Sexta Noche* corpus are presented in Tables A.6, A.7, A.8, and A.9, respectively. Similarly, Tables A.10, A.11, A.12, and A.13 show all the results obtained with *Empathic* corpus. The experiments highlighted in bold represent the best-performing combination of input features and the oversampling method for each network architecture.

Table A.6.: The F1-Score performance of all experiments carried out with the categorical model in *La Sexta Noche* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.5819	**0.6893**	0.6875	0.6540	0.5173	0.5414
M. Perceptron	SMOTE Epsilon	0.5987	0.6676	0.6731	0.6709	0.5237	0.5428
M. Perceptron	SMOTE Neighbour	0.5757	0.6727	0.6747	0.6502	0.5489	0.5065
Conv. 1D	oversampling	0.5629	**0.6918**	0.6840	0.6486	0.4966	0.5185
Conv. 1D	SMOTE Epsilon	0.3985	0.5375	0.5130	0.5417	0.3546	0.3568
Conv. 1D	SMOTE Neighbour	0.3963	0.5276	0.5392	0.5247	0.3458	0.3519
Transformer	oversampling	0.6138	**0.6639**	0.6582	0.6215	0.5032	0.5015

Table A.7.: The F1-Score performance of all experiments carried out with the arousal dimension of VAD model in *La Sexta Noche* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.6598	**0.7255**	0.7233	0.7248	0.6433	0.6610
M. Perceptron	SMOTE Epsilon	0.6405	0.7063	0.7088	0.7109	0.6255	0.6441
M. Perceptron	SMOTE Neighbour	0.6087	0.7082	0.6987	0.7030	0.6538	0.6555
Conv. 1D	oversampling	0.6540	0.7138	**0.7176**	0.7174	0.6566	0.6475
Conv. 1D	SMOTE Epsilon	0.6030	0.6193	0.6270	0.6285	0.5431	0.5599
Conv. 1D	SMOTE Neighbour	0.5632	0.6104	0.5913	0.6098	0.5069	0.5866
Transformer	oversampling	0.6772	**0.7293**	0.7111	0.7073	0.6245	0.6264

Table A.8.: the F1-Score performance of all experiments carried out with the valence dimension of VAD model in *La Sexta Noche* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.4542	**0.5179**	0.5148	0.4939	0.4105	0.4221
M. Perceptron	SMOTE Epsilon	0.3679	0.4703	0.4691	0.4496	0.3721	0.4083
M. Perceptron	SMOTE Neighbour	0.3317	0.4635	0.4686	0.4484	0.3708	0.3615
Conv. 1D	oversampling	0.4421	0.4859	**0.4865**	0.4790	0.3937	0.3973
Conv. 1D	SMOTE Epsilon	0.2366	0.3927	0.3932	0.4038	0.3349	0.3127
Conv. 1D	SMOTE Neighbour	0.2365	0.4062	0.3806	0.3988	0.2986	0.3036
Transformer	oversampling	0.4621	0.4861	**0.4876**	0.4696	0.4263	0.4057

Table A.9.: The F1-Score performance of all experiments carried out with the dominance dimension of VAD model in *La Sexta Noche* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.5552	0.6129	**0.6247**	0.6035	0.5436	0.5683
M. Perceptron	SMOTE Epsilon	0.5367	0.5955	0.6011	0.5818	0.5419	0.5257
M. Perceptron	SMOTE Neighbour	0.3919	0.5855	0.5945	0.5877	0.5272	0.5081
Conv. 1D	oversampling	0.5533	0.6038	**0.6103**	0.5982	0.5188	0.5351
Conv. 1D	SMOTE Epsilon	0.4763	0.5429	0.5436	0.5578	0.4751	0.4907
Conv. 1D	SMOTE Neighbour	0.4804	0.5019	0.5233	0.5370	0.4765	0.4702
Transformer	oversampling	0.5707	0.6002	**0.6027**	0.5879	0.5528	0.5580

Table A.10.: The F1-Score performance of all experiments carried out with the categorical model in *Empathic* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.2730	0.4006	0.3955	**0.4044**	0.2908	0.3006
M. Perceptron	SMOTE Epsilon	0.2836	0.3844	0.3756	0.3756	0.2948	0.2859
M. Perceptron	SMOTE Neighbour	0.2715	0.3742	0.3774	0.3746	0.2780	0.2916
Conv. 1D	oversampling	0.3006	**0.4010**	0.3919	0.3905	0.2827	0.2799
Conv. 1D	SMOTE Epsilon	0.2503	0.2900	0.2988	0.2985	0.2352	0.2374
Conv. 1D	SMOTE Neighbour	0.2348	0.2902	0.2929	0.2884	0.2356	0.2330
Transformer	oversampling	0.2817	**0.3789**	0.3780	0.3673	0.2690	0.2635

Table A.11.: The F1-Score performance of all experiments carried out with the arousal dimension of VAD model in *Empathic* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.5683	0.5886	0.5938	**0.5958**	0.5497	0.5356
M. Perceptron	SMOTE Epsilon	0.4203	0.5443	0.5368	0.5391	0.5598	0.4822
M. Perceptron	SMOTE Neighbour	0.3308	0.5370	0.5332	0.5395	0.4814	0.4991
Conv. 1D	oversampling	0.5520	0.5558	**0.5646**	0.5622	0.5299	0.5349
Conv. 1D	SMOTE Epsilon	0.4783	0.5201	0.5066	0.5423	0.4495	0.4691
Conv. 1D	SMOTE Neighbour	0.4814	0.5176	0.5070	0.5002	0.4414	0.4482
Transformer	oversampling	0.5503	0.5607	0.5658	**0.5682**	0.5134	0.5135

Table A.12.: The F1-Score performance of all experiments carried out with the valence dimension of VAD model in *Empathic* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.3875	0.4333	**0.4445**	0.4397	0.3751	0.3679
M. Perceptron	SMOTE Epsilon	0.3000	0.4231	0.4249	0.4219	0.3526	0.3424
M. Perceptron	SMOTE Neighbour	0.2873	0.4197	0.4211	0.4223	0.3445	0.3537
Conv. 1D	oversampling	0.3866	0.4321	**0.4478**	0.4352	0.3643	0.3649
Conv. 1D	SMOTE Epsilon	0.2969	0.3839	0.3767	0.3640	0.2990	0.2960
Conv. 1D	SMOTE Neighbour	0.2678	0.3641	0.3644	0.3751	0.2949	0.3050
Transformer	oversampling	0.3822	**0.4406**	0.4381	0.4318	0.3688	0.3606

Table A.13.: The F1-Score performance of all experiments carried out with the dominance dimension of VAD model in *Empathic* corpus.

Network	Oversampling	Mel-Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
M. Perceptron	oversampling	0.5733	**0.6205**	0.6192	0.6192	0.5653	0.5638
M. Perceptron	SMOTE Epsilon	0.4604	0.5881	0.5954	0.5840	0.5409	0.5278
M. Perceptron	SMOTE Neighbour	0.4101	0.5846	0.5939	0.5863	0.5391	0.5334
Conv. 1D	oversampling	0.5688	**0.6042**	0.6000	0.6005	0.5356	0.5330
Conv. 1D	SMOTE Epsilon	0.5253	0.5303	0.5520	0.5659	0.5164	0.5115
Conv. 1D	SMOTE Neighbour	0.5058	0.5299	0.5309	0.5434	0.4644	0.4967
Transformer	oversampling	0.5603	0.5976	**0.5996**	0.5940	0.5368	0.5381

5 | Extended Experiments with Expert Annotated *Empathic* Corpus for Multiple Languages

In addition to the crowd-annotated emotional corpus, the *Empathic* project also produced an expert-annotated corpus in the three languages of the project: Spanish, French, and Norwegian. This section presents an extension of the experiments performed in Section 4, where we evaluate the performance of various audio representations and network architectures on the expert-annotated *Empathic* corpus. Note that the *La Sexta Noche* corpus used in the previous experiments was only annotated by crowd workers, and therefore, no expert annotation is available for this dataset.

5.1 | Description of the experiments

In this final set of experiments, we repeated the experiments in Section 4 using the expert-annotated *Empathic* corpus for each language and task. We also included an additional task for detecting the presence of speech.

During the experiments, an analysis was carried out on different audio representations and network architectures tested in the previous experiments.

The same audio representations used in the previous experiments were considered, including the Mel-Spectrogram (with a 128 filters bank), the top three performing Wave Embeddings models (UniSpeechSat with the *sat-large* model, WavLM with the *large* model, and Hubert with the *large-ll60k* model), and the two sets of acoustic features, GeMAPS and eGeMAPS.

Three different types of network architectures have been considered, as in Section 4. First, a simple MLP model was implemented. Similar to previous experiments, the temporal average was calculated for the Mel-Spectrogram and Wave Embeddings audio representations. This network architecture consisted of a hidden layer of 64 neurons with the ReLU activation function and a final classifier layer. Then, a 1D convolutional model and a transformer with the same configurations as the experiments in Section 4 have also been implemented. LLDs were used as audio representations for GeMAPS and eGeMAPS in these two network architectures. In the case of the 1D convolutional, no activation function was used in the convolutions. Moreover, in the case of the transformers, a 6-layer architecture was used with the minimum number of heads (2 or 5 depending on the input), the positional embeddings were concatenated (concat), and for the last classifier layer, we considered the average of all the time instants (mean).

It should be noted that all hyper-parameters not modified by the grid search, that is, oversampling, batch size, learning rate, optimiser, and specific configurations to each network structure, are fixed by the best results in previous experiments. A summary of the conditions and hyper-parameters of these extended experiments is listed below:

- **10-fold cross-validation**
- Corpora: ***Empathic (experts)***
- Different problems: **Categorical**, **Arousal**, **Valence**, **Domiannce**, and **Speech**
- Speech Signal Representations:
  - **Mel-Spectrogram**
  - Wave Embeddings: **UniSpeechSat**, **WavLM**, and **Hubert**.
  - Acoustic Features: **GeMAPS** and **eGeMAPS**.
- Oversampling method: **Repetition** (x4).
- Batch Size: **64**.
- Network Architecture with specific training hyper-parameters:

	Multilayer Perceptron	**Convolutional 1D**	**Transformers**
Learning Rate	1e-3	1e-3	1e-4
Optimizer	Adam	Adam	AdamW
beta parameter	(0.9,0.999)	(0.9,0.999)	(0.9,0.98)

5.2 | RESULTS

As we have shown previously, the complexity of the task (degree of imbalance and number of categories) directly impacted performance.

Figure A.13 shows the performance obtained for each classifier in the corpus *Empathic (experts)*: categorical, arousal, valence, dominance, and speech. In addition, each classification task was developed for the three target languages of the EMPATHIC project (Spanish, French and Norwegian).

This figure shows that training a model for all three languages together resulted in slightly better performance than training them separately. Although the performance was generally moderately consistent across languages, there were instances, such as in the valence task, where the performance difference was significant. Thus, valence, arousal and even categorical classifications benefited from the mix of samples in all different languages. Therefore, we can conclude that the emotion polarity and level of excitement have some common features across cultures in their speech-based development. This discrepancy may be attributed to using different annotators for the samples, resulting in changes in perception and imbalance across different languages.

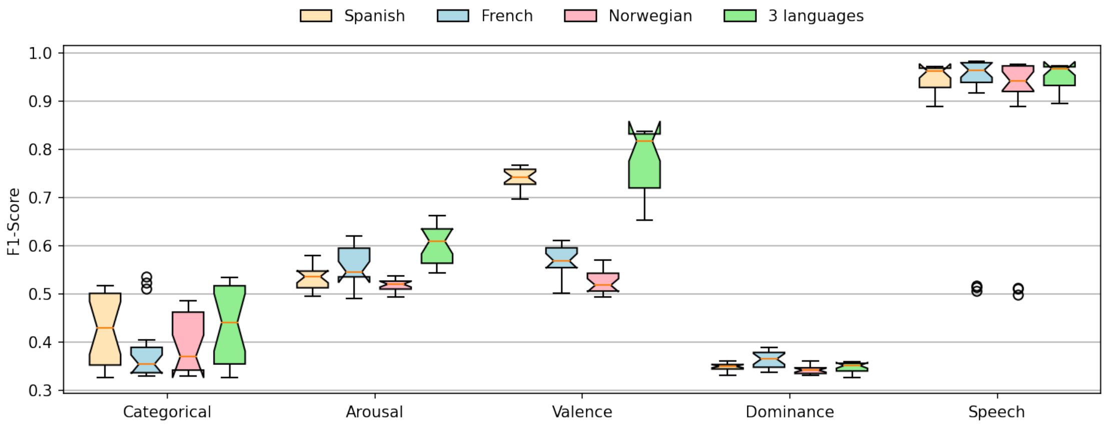
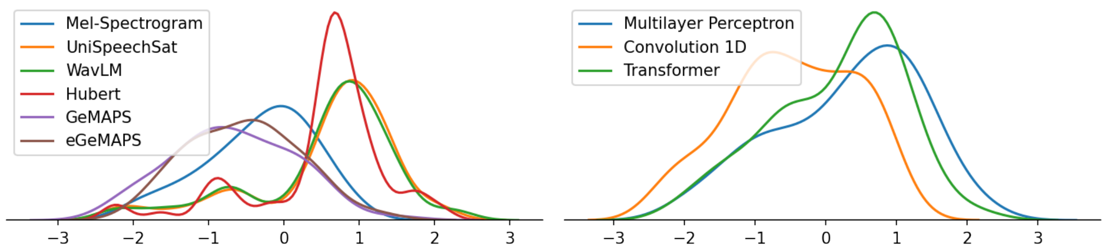


Figure A.13.: Boxplot of the F1-Score performance over different classification tasks and languages for *Empathic (experts)* experiments.

The results presented in Figure A.14 were standardised. As a result, values below 0.0 represent worse results than the mean, while values above 0.0 indicate better results.

Figure A.14a highlights that all audio representations generated with Wave Embeddings performed better, leaving Mel-Spectrogram and GeMAPS far behind. As far as Wave Embeddings models are concerned, we can observe that Hubert obtained slightly worse results.



(a) Speech Signal Representations

(b) Network Architecture

Figure A.14.: Density plots showing the performance in *Empathic (experts)* corpus for different hyper-parameters: speech signal representation on top-left, and the network architecture on top-right.

Regarding network architecture, the MLP network generated outstanding results, slightly superior to the transformer. The convolutional architecture was not able to outperform MLP and Transformers.

Finally, Tables A.14, A.15, A.16, A.17, and A.18 present the results obtained using the *Empathic (experts)* corpus for each task: categorical, arousal, valence, dominance, and speech.

Table A.14.: The F1-Score performance of all experiments carried out with the categorical model in the Expert-Annotated *Empathic* corpus.

Language	Network	Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
es	M. Perceptron	0.3255	0.5052	0.5098	0.5008	0.3596	0.3534
es	Conv. 1D	0.3699	0.4794	0.4776	0.4689	0.3500	0.3516
es	Transformer	0.3891	0.5175	0.5089	0.5005	0.3313	0.3326
fr	M. Perceptron	0.3298	0.5235	0.5357	0.5101	0.3753	0.4048
fr	Conv. 1D	0.3782	0.3303	0.3355	0.3342	0.3646	0.3673
fr	Transformer	0.3923	0.3337	0.3387	0.3426	0.3393	0.3379
no	M. Perceptron	0.3397	0.4680	0.4632	0.4578	0.3743	0.3655
no	Conv. 1D	0.3747	0.3459	0.3354	0.3400	0.3290	0.3292
no	Transformer	0.3844	0.4662	0.4745	0.4851	0.3535	0.3494
es fr no	M. Perceptron	0.3259	0.5203	0.5341	0.5177	0.3668	0.3649
es fr no	Conv. 1D	0.3815	0.4896	0.4881	0.4883	0.3479	0.3507
es fr no	Transformer	0.3934	0.5263	0.5256	0.5130	0.3315	0.3370

Table A.15.: The F1-Score performance of all experiments carried out with the arousal dimension in VAD model in the Expert-Annotated *Empathic* corpus.

Language	Network	Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
es	M. Perceptron	0.5003	0.5794	0.5670	0.5508	0.5127	0.5050
es	Conv. 1D	0.5374	0.5342	0.5429	0.5286	0.5124	0.5318
es	Transformer	0.5375	0.5369	0.5586	0.5485	0.4960	0.4956
fr	M. Perceptron	0.5407	0.6019	0.6105	0.6207	0.5728	0.6026
fr	Conv. 1D	0.5493	0.5083	0.5069	0.5210	0.5381	0.5410
fr	Transformer	0.5538	0.5340	0.5385	0.6053	0.5532	0.4900
no	M. Perceptron	0.4927	0.5373	0.5294	0.5198	0.5183	0.5200
no	Conv. 1D	0.5058	0.5126	0.5088	0.5153	0.5270	0.5310
no	Transformer	0.5234	0.5365	0.5204	0.5230	0.4927	0.5059
es fr no	M. Perceptron	0.5986	0.6544	0.6625	0.6495	0.5494	0.5480
es fr no	Conv. 1D	0.6004	0.6234	0.6356	0.6182	0.5561	0.5850
es fr no	Transformer	0.5888	0.6180	0.6404	0.6314	0.5446	0.5434

Table A.16.: The F1-Score performance of all experiments carried out with the valence dimension in VAD model in the Expert-Annotated *Empathic* corpus.

Language	Network	Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
es	M. Perceptron	0.7228	0.7632	0.7647	0.7599	0.7281	0.7317
es	Conv. 1D	0.7270	0.7537	0.7484	0.7475	0.6963	0.7104
es	Transformer	0.7369	0.7672	0.7600	0.7524	0.7239	0.7319
fr	M. Perceptron	0.5674	0.6107	0.6073	0.5959	0.5643	0.5532
fr	Conv. 1D	0.5570	0.5010	0.5056	0.5098	0.5266	0.5632
fr	Transformer	0.5691	0.5977	0.5936	0.5984	0.5692	0.5819
no	M. Perceptron	0.4936	0.5704	0.5661	0.5678	0.5176	0.5206
no	Conv. 1D	0.5150	0.4997	0.5055	0.4989	0.5179	0.5132
no	Transformer	0.5271	0.5427	0.5495	0.5420	0.4936	0.5058
es fr no	M. Perceptron	0.7861	0.8351	0.8374	0.8311	0.6529	0.7103
es fr no	Conv. 1D	0.7770	0.8222	0.8192	0.8147	0.6660	0.6892
es fr no	Transformer	0.8260	0.8319	0.8373	0.8321	0.6937	0.7474

Table A.17.: The F1-Score performance of all experiments carried out with the dominance dimension in VAD model in the Expert-Annotated *Empathic* corpus.

Language	Network	Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
es	M. Perceptron	0.3311	0.3591	0.3525	0.3537	0.3603	0.3446
es	Conv. 1D	0.3487	0.3505	0.3475	0.3532	0.3436	0.3377
es	Transformer	0.3407	0.3561	0.3527	0.3529	0.3303	0.3472
fr	M. Perceptron	0.3476	0.3866	0.3835	0.3867	0.3667	0.3703
fr	Conv. 1D	0.3625	0.3365	0.3366	0.3457	0.3475	0.3420
fr	Transformer	0.3768	0.3786	0.3892	0.3747	0.3546	0.3562
no	M. Perceptron	0.3315	0.3484	0.3428	0.3468	0.3448	0.3400
no	Conv. 1D	0.3435	0.3369	0.3351	0.3344	0.3315	0.3322
no	Transformer	0.3437	0.3468	0.3597	0.3593	0.3315	0.3399
es fr no	M. Perceptron	0.3293	0.3573	0.3519	0.3562	0.3505	0.3493
es fr no	Conv. 1D	0.3438	0.3581	0.3564	0.3518	0.3395	0.3341
es fr no	Transformer	0.3402	0.3584	0.3529	0.3540	0.3264	0.3274

Table A.18.: The F1-Score performance of all experiments carried out with the speech presence task in the Expert-Annotated *Empathic* corpus.

Language	Network	Spectro	UniSpeechSat	WavLM	Hubert	GeMAPS	eGeMAPS
es	M. Perceptron	0.8907	0.9720	0.9708	0.9685	0.9145	0.9266
es	Conv. 1D	0.9332	0.9679	0.9671	0.9605	0.8886	0.8948
es	Transformer	0.9695	0.9675	0.9683	0.9648	0.9342	0.9380
fr	M. Perceptron	0.9616	0.9821	0.9828	0.9814	0.9626	0.9752
fr	Conv. 1D	0.9638	0.5136	0.5063	0.5163	0.9174	0.9314
fr	Transformer	0.9765	0.9790	0.9797	0.9798	0.9610	0.9648
no	M. Perceptron	0.9314	0.9746	0.9760	0.9736	0.9300	0.9545
no	Conv. 1D	0.9412	0.5121	0.5108	0.4978	0.8894	0.9166
no	Transformer	0.9714	0.9739	0.9679	0.9734	0.9404	0.9418
es fr no	M. Perceptron	0.9024	0.9731	0.9739	0.9719	0.9319	0.9356
es fr no	Conv. 1D	0.9340	0.9704	0.9693	0.9666	0.8979	0.8956
es fr no	Transformer	0.9705	0.9712	0.9732	0.9687	0.9316	0.9403

Bibliography

A. Russell, J. (1983). Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45:1281–1288.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Averill, J. R. (2015). Emotion and anxiety: Sociocultural, biological, and psychological determinants. In *Emotions and Anxiety (PLE: Emotion)*, pages 99–142. Psychology Press.

Aïsha, S., Elisabeth, P., Ouriel, G., and Bruno, B. (2017). Predictive mechanisms are not involved the same way during human-human vs. human-machine interactions: A review. *Frontiers in Neurorobotics*, 11.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bank, D., Koenigstein, N., and Giryes, R. (2020). Autoencoders. *arXiv preprint arXiv:2003.05991*.

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161.

Blanco, R. J., Alcaide, J. M., Torres, M., and Walker, M. A. (2018). Detection of sarcasm and nastiness: New resources for spanish language. *Cogn. Comput.*, 10(6):1135–1151.

Bou-Ghazale, S. E. and Hansen, J. H. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on speech and audio processing*, 8(4):429–442.

Brinkschulte, L., Mariacher, N., Schlögl, S., Torres, M., Justo, R., Olaso, J. M., Esposito, A., Cordasco, G., Chollet, G., Glackin, C., Pickard, C., Petrovska-Delacrétaz, D., Hmani, M. A., Mtibaa, A., Fernández, A., Kyslitska, D., Fernández-Ruanova, B., Tenorio-Laranga, J., Aksnes, M., Korsnes, M. S., Reiner, M., Lindner, F., Deroo, O., and Gordeeva, O. (2021). The EMPATHIC project: Building an expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. *CoRR*, abs/2104.13836.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Callejas, Z., Akhlaki, K. B., Noguera, M., Torres, M., and Justo, R. (2019). MENHIR: mental health monitoring through interactive conversations. *Proces. del Leng. Natural*, 63:139–142.

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.

Calvo, R. A. and Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., and Kwok, K. (2022). Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Chakraborty, R., Pandharipande, M., and Kopparapu, S. K. (2017). *Analyzing emotion in spontaneous speech*. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022a). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., et al. (2022b). Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6152–6156. IEEE.

Chen, X., Wu, Y., Wang, Z., Liu, S., and Li, J. (2021). Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5904–5908. IEEE.

Chiba, Y., Nose, T., and Ito, A. (2017). Analysis of efficient multimodal features for estimating user's willingness to talk: Comparison of human-machine and human-human dialog. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 428–431. IEEE.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.

Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR*, abs/2006.11371.

deVelasco, M., Justo, R., Antón, J., Carrilero, M., and Torres, M. (2018). Emotion Detection from Speech and Text. In *Proc. IberSPEECH 2018*, pages 68–71.

deVelasco, M., Justo, R., Letaifa, L., and Torres, M. (2021). Contrasting the emotions identified in spanish tv debates and in human-machine interactions. pages 51–55.

deVelasco, M., Justo, R., López Zorrilla, A., and Torres, M. I. (2023). Analysis of deep learning-based decision-making in an emotional spontaneous speech task. *Applied Sciences*, 13(2):980.

deVelasco, M., Justo, R., and Torres, M. I. (2022a). Automatic identification of emotional information in spanish tv debates and human-machine interactions. *Applied Sciences*, 12(4).

deVelasco, M., Justo, R., Zorrilla, A. L., and Torres, M. I. (2022b). Automatic analysis of emotions from the voices/speech in spanish tv debates. *Acta Polytechnica Hungarica*, 19(5).

deVelasco Vazquez, M., Justo, R., Zorrilla, A., and Torres, M. (2019). Can spontaneous emotions be detected from speech on tv political debates? In *10th IEEE International Conference on Cognitive Infocommunications*, pages 289–294.

deVelasco Vázquez, M., López Zorrilla, A., and Justo Blanco, R. (2019). Iruzurrezko portaeren detekzioa crowd motako etiketazioan.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L., Xu, S., and Xu, B. (2018). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Eskimez, S. E., Imade, K., Yang, N., Sturge-Apple, M., Duan, Z., and Heinzelman, W. (2016). Emotion classification: how does an automated system compare to naive human coders? In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2274–2278. IEEE.

Esposito, A., Marinaro, M., and Palombo, G. (2004). Children speech pauses as markers of different discourse structures and utterance information content. In *Proceedings of the International Conference: From sound to sense*, volume 50, pages 10–13.

Esposito, A., Stejskal, V., and Smékal, Z. (2008). Cognitive role of speech pauses and algorithmic considerations for their processing. *Int. J. Pattern Recognit. Artif. Intell.*, 22:1073–1088.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Firdaus, M., Ekbal, A., and Cambria, E. (2023). Multitask learning for multilingual intent detection and slot filling in dialogue systems. *Information Fusion*, 91:299–315.

Galassi, A., Lippi, M., and Torroni, P. (2020). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99.

Gurney, E. (1884). What is an emotion? *Mind*, 9(35):421–426.

Hayette Hadjar, B. V., Maier, D., Mayer, G., Mc Kevitt, P., and Hemmje, M. (2021). Video-based emotion detection analyzing facial expressions and contactless vital signs for psychosomatic monitoring.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Horvat, M., Stojanović, A., and Kovačević, Ž. (2022). An overview of common emotion models in computer systems. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1008–1013. IEEE.

Hou, J.-C., Wang, S.-S., Lai, Y.-H., Tsao, Y., Chang, H.-W., and Wang, H.-M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In *Proc. of the ACM SIGKDD*, pages 64–67, New York, USA.

Irastorza, J. and Torres, M. I. (2016). Analyzing the expression of annoyance during phone calls to complaint services. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000103–000106. IEEE.

Irastorza, J. and Torres, M. I. (2019). Tracking the expression of annoyance in call centers. In *Cognitive Infocommunications, Theory and Applications*, pages 131–151. Springer.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B.-E., Corvol, J.-C., Vidailhet, M., Lehéricy, S., and Benali, H. (2021). X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech. *Frontiers in Neuroinformatics*, 15:578369.

Justo, R., Letaifa, L. B., Olaso, J. M., López-Zorrilla, A., deVelasco, M., Vázquez, A., and Torres, M. I. (2021). *A Spanish Corpus for Talking to the Elderly*, pages 183–192. Springer Singapore, Singapore.

Justo, R., Letaifa, L. B., Palmero, C., Fraile, E. G., Johansen, A., Vazquez, A., Cordasco, G., Schlogl, S., Ruanova, B. F., Silva, M., Escalera, S., deVelasco, M., Laranga, J. T., Esposito, A., Kornes, M., and Torres, M. (2020). Analysis of the interaction between elderly people and a simulated virtual coach. *Journal of Ambient Intelligence and Humanized Computing*, 11:6125–6140.

Kim, J. C. and Clements, M. A. (2015). Multimodal affect classification at various temporal lengths. *IEEE Transactions on Affective Computing*, 6(4):371–384.

Kiss, G. and Vicsi, K. (2017). Comparison of read and spontaneous speech in case of automatic detection of depression. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000213–000218. IEEE.

Krippendorff, K. (2004). *Content analysis: An introd. to its methodology.* Sage.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H. S., and Cambria, E. (2023). A survey on deep reinforcement learning for audio-based applications. *Artificial Intelligence Review*, 56(3):2193–2240.

Letaifa, L. B. and Torres, M. I. (2021). Perceptual borderline for balancing multi-class spontaneous emotional data. *IEEE Access*, 9:55939–55954.

Li, W., Li, Y., Pandelea, V., Ge, M., Zhu, L., and Cambria, E. (2022). Ecpec: emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*.

Li, W., Zhu, L., Mao, R., and Cambria, E. (2023). Skier: A symbolic knowledge integrated model for conversational emotion recognition.

Lleida, E. and Rodriguez-Fuentes, L. J. (2018). Speaker and language recognition and characterization: Introduction to the csl special issue. *Computer Speech & Language*, 49:107–120.

López Zorrilla, A., deVelasco Vázquez, M. d., Irastorza, J., Olaso Fernández, J. M., Justo Blanco, R., and Torres Barañano, M. I. (2018). Empathic: Empathic, expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. *Procesamiento del Lenguaje Natural*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J., and Fernández-Martínez, F. (2022). A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12.

López-Zorrilla, A., deVelasco Vázquez, M., Cenceschi, S., and Torres, M. (2018). Corrective focus detection in italian speech using neural networks. *Acta Polytechnica Hungarica*, 15:109–127.

Mesquita, B. and Leu, J. (2007). *The cultural psychology of emotion.* The Guilford Press.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

Moors, A. (2012). Comparison of affect program theories, appraisal theories, and psychological construction theories. In *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell*, pages 257–278. John Benjamins Amsterdam, The Netherlands.

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

Ortega Giménez, A., Lleida Solano, E., San Segundo Hernández, R., Ferreiros López, J., Hurtado Oliver, L. F., Sanchís Arnal, E., Torres Barañano, M. I., and Justo Blanco, R. (2018). Amic: affective multimedia analytics with inclusive and natural communication= amic: análisis afectivo de información multimedia con comunicación inclusiva y natural. *Procesamiento del Lenguaje Natural*, 1(61):147–150.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, 97(3):315.

Palmero, C., deVelasco, M., and et al (2023). Ongoing work. *IEEE Transactions on Affective Computing*.

Pappas, D., Androutsopoulos, I., and Papageorgiou, H. (2015). Anger detection in call center dialogues. In *2015 6th IEEE international conference on cognitive infocommunications (CogInfoCom)*, pages 139–144. IEEE.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR.

Pastor, M., Ribas, D., Ortega, A., Miguel, A., and Solano, E. (2022a). Cross-corpus speech emotion recognition with hubert self-supervised representation. *Proceedings of the IberSPEECH*, pages 76–80.

Pastor, M., Ribas, D., Ortega, A., Miguel, A., and SOLANO, E. L. (2022b). Cross-corpus speech emotion recognition with hubert self-supervised representation. *IberSPEECH 2022*.

Prinz, J. (2004). Which emotions are basic. *Emotion, evolution, and rationality*, 69:88.

Raquel Justo, M.Inés Torres, J. M. A. (2017). Measuring the quality of annotations for a subjective crowdsourcing task. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 58–68. Springer.

Ribas, D., Pastor, M. A., Miguel, A., Martínez, D., Ortega, A., and Lleida, E. (2023). Automatic voice disorder detection using self-supervised representations. *IEEE Access*, 11:14915–14927.

Riviello, M. T., Esposito, A., and Vicsi, K. (2012). A cross-cultural study on the perception of emotions: How hungarian subjects evaluate american and italian emotional expressions. In *Cognitive behavioural systems*, pages 424–433. Springer.

Rothwell, S., Elshenawy, A., Carter, S., iraga, D., Romani, F., Kennewick, M., and Kennewick, B. (2015). Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proc. of Interspeech 2015, Dresden, Germany, September 6-10, 2015*, pages 2784–2788.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Scherer, K. R. (1999). Appraisal theory. *Handbook of cognition and emotion*, pages 637–663.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Schlögl, S., Doherty, G., Karamanis, N., and Luz, S. (2010). Webwoz: a wizard of oz prototyping framework. In *Proceedings of the 2nd acm sigchi symposium on engineering interactive computing systems*, pages 109–114.

Schlögl, S., Milhorat, P., Chollet, G., and Boudy, J. (2014). Designing language technology applications: A wizard of Oz driven prototyping framework. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 85–88, Gothenburg, Sweden. Association for Computational Linguistics.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of experimental psychology*, 44(4):229.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53(9-10):1062–1087.

Schuller, B., Weninger, F., Zhang, Y., Ringeval, F., Batliner, A., Steidl, S., Eyben, F., Marchi, E., Vinciarelli, A., Scherer, K., et al. (2019). Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge. *Computer Speech & Language*, 53:156–180.

Scibilia, A., Pedrocchi, N., and Fortuna, L. (2022). Human control model estimation in physical human–machine interaction: A survey. *Sensors*, 22(5):1732.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Statharakos, N., Alvares, A. J., Papadopoulou, E., and Statharakou, A. (2022). Psychology of emotions. In *The Psychology of Anger*, pages 21–50. Springer.

Susanto, Y., Livingstone, A. G., Ng, B. C., and Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Sztahó, D., Tulics, M. G., Vicsi, K., and Valálik, I. (2017). Automatic estimation of severity of parkinson's disease based on speech rhythm related features. In *Proceedings of 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017) Debrecen, Hungary*, pages 11–16.

Tao, F. and Liu, G. (2018). Advanced lstm: A study about better time dependency modeling in emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910.

Torres, M. I., Olaso, J. M., Glackin, N., Justo, R., and Chollet, G. (2019a). A spoken dialogue system for the empathic virtual coach. In D'Haro, L. F., Banchs, R. E., and Li, H., editors, *9th International Workshop on Spoken Dialogue System Technology*, pages 259–265, Singapore. Springer Singapore.

Torres, M. I., Olaso, J. M., Montenegro, C., Santana, R., Vázquez, A., Justo, R., Lozano, J. A., Schlögl, S., Chollet, G., Dugan, N., Irvine, M., Glackin, N., Pickard, C., Esposito, A., Cordasco, G., Troncone, A., Petrovska-Delacretaz, D., Mtibaa, A., Hmani, M. A., Korsnes, M. S., Martinussen, L. J., Escalera, S., Cantariño, C. P., Deroo, O., Gordeeva, O., Tenorio-Laranga, J., Gonzalez-Fraile, E., Fernandez-Ruanova, B., and Gonzalez-Pinto, A. (2019b). The empathic project: Mid-term achievements. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '19, pages 629–638. ACM.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 3–10, New York, NY, USA. ACM.

Vea, T. (2020). The learning of emotion in/as sociocultural practice: The case of animal rights activism. *Journal of the Learning Sciences*, 29(3):311–346.

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., Cristani, M., Fuhrmann, F., Gilmartin, E., Hammal, Z., Heylen, D., Kaiser, R., Koutsombogera, M., Potamianos, A., Renals, S., Riccardi, G., and Salah, A. A. (2015). Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7:397–413.

Vu, B., deVelasco, M., Mc Kevitt, P., Bond, R., Turkington, R., Booth, F., Mulvenna, M., Fuchs, M., and Hemmje, M. (2021). *A Content and Knowledge Management System Supporting Emotion Detection from Speech*, pages 369–378. Springer Singapore, Singapore.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., and Huang, X. (2021). Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*, pages 10937–10947. PMLR.

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., and Tarokh, V. (2020). Speech emotion recognition with dual-sequence lstm architecture. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478.

Williams, C. E. and Stevens, K. N. (1981). Vocal correlates of emotional states. *Speech evaluation in psychiatry*, pages 221–240.

Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*.

Zubiaga, I., Justo, R., Torres, M., and deVelasco, M. (2022a). Speech emotion recognition in Spanish TV Debates . In *Proc. IberSPEECH 2022*, pages 186–190.

Zubiaga, I., Menchaca, I., deVelasco, M., and Justo, R. (2022b). Mental Health Monitoring from Speech and Language. In *Proc. Workshop on Speech, Music and Mind*, pages 11–15.

Part III.

Contributions: A Compilation of Published Articles

List of Journal Publications

deVelasco, M., Justo, R., López Zorrilla, A., & Torres, M. I. (2023). Analysis of deep learning-based decision-making in an emotional spontaneous speech task. *Applied Sciences*, 13(2), 980. https://doi.org/10.3390/app13020980. JCR (2021) 2.838 Q2

deVelasco, M., Justo, R., & Torres, M. I. (2022a). Automatic identification of emotional information in spanish tv debates and human-machine interactions. *Applied Sciences*, 12(4). https://doi.org/10.3390/app12041902. JCR (2021) 2.838 Q2

deVelasco, M., Justo, R., Zorrilla, A. L., & Torres, M. I. (2022b). Automatic analysis of emotions from the voices/speech in spanish tv debates. *Acta Polytechnica Hungarica*, 19(5). https://doi.org/10.12700/APH.19.5.2022.5.8. JCR (2021) 1.711 Q3

Justo, R., Letaifa, L. B., Palmero, C., Fraile, E. G., Johansen, A., Vazquez, A., Cordasco, G., Schlogl, S., Ruanova, B. F., Silva, M., Escalera, S., deVelasco, M., Laranga, J. T., Esposito, A., Kornes, M., & Torres, M. (2020). Analysis of the interaction between elderly people and a simulated virtual coach. *Journal of Ambient Intelligence and Humanized Computing*, 11, 6125–6140. https://doi.org/10.1007/s12652-020-01983-3. JCR (2020) 7.104 Q1

López-Zorrilla, A., deVelasco Vázquez, M., Cenceschi, S., & Torres, M. (2018). Corrective focus detection in italian speech using neural networks. *Acta Polytechnica Hungarica*, 15, 109–127. https://doi.org/10.12700/APH.15.5.2018.5.7. JCR (2018) 1.286 Q3

List of Conference Publications

Ben Letaifa Zouari, L., deVelasco Vázquez, M., Justo Blanco, R., & Torres Barañano, M. I. (2019). First steps to develop a corpus of interactions between elderly and virtual agents in spanish with emotion labels.

deVelasco, M., Justo, R., Antón, J., Carrilero, M., & Torres, M. (2018). Emotion Detection from Speech and Text. *Proc. IberSPEECH 2018*, 68–71. https://doi.org/10.21437/IberSPEECH.2018-15

deVelasco, M., Justo, R., Letaifa, L., & Torres, M. (2021). Contrasting the emotions identified in spanish tv debates and in human-machine interactions. 51–55. https://doi.org/10.21437/IberSPEECH.2021-11

deVelasco Vazquez, M., Justo, R., Zorrilla, A., & Torres, M. (2019). Can spontaneous emotions be detected from speech on tv political debates? *10th IEEE International Conference on Cognitive Infocommunications*, 289–294. https://doi.org/10.1109/CogInfoCom47531.2019.9089948

deVelasco Vázquez, M., López Zorrilla, A., & Justo Blanco, R. (2019). Iruzurrezko portaeren detekzioa crowd motako etiketazioan. https://doi.org/10.26876/ikergazte.iii.03.17

Justo, R., Letaifa, L. B., Olaso, J. M., López-Zorrilla, A., deVelasco, M., Vázquez, A., & Torres, M. I. (2021). *A Spanish Corpus for Talking to the Elderly*, 183–192. Springer Singapore. https://doi.org/10.1007/978-981-15-8395-7_13

Olaso, J. M., Vázquez, A., Ben Letaifa, L., deVelasco, M., Mtibaa, A., Hmani, M. A., Petrovska-Delacrétaz, D., Chollet, G., Montenegro, C., López-Zorrilla, A., Justo, R., Santana, R., Tenorio-Laranga, J., González-Fraile, E., Fernández-Ruanova, B. n., Cordasco, G., Esposito, A., Gjellesvik, K. B., Johansen, A. T., Kornes, M. S., Pickard, C., Glackin, C., Cahalane, G., Buch, P., Palmero, C., Escalera, S., Gordeeva, O., Deroo, O., Fernández, A., Kyslitska, D., Lozano, J. A., Torres, M. I., & Schlögl, S. (2021). The empathic virtual coach: A demo. *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, 848–851. https://doi.org/10.1145/3462244.3481574

Vu, B., deVelasco, M., Mc Kevitt, P., Bond, R., Turkington, R., Booth, F., Mulvenna, M., Fuchs, M., & Hemmje, M. (2021). *A Content and Knowledge Management System Supporting Emotion Detection from Speech*, 369–378. Springer Singapore. https://doi.org/10.1007/978-981-15-8395-7_28

Zubiaga, I., Justo, R., Torres, M., & deVelasco, M. (2022a). Speech emotion recognition in Spanish TV Debates . *Proc. IberSPEECH 2022*, 186–190. https://doi.org/10.21437/IberSPEECH.2022-38

Zubiaga, I., Menchaca, I., deVelasco, M., & Justo, R. (2022b). Mental Health Monitoring from Speech and Language. *Proc. Workshop on Speech, Music and Mind*, 11–15. https://doi.org/10.21437/SMM.2022-3

Additional Publications

López Zorrilla, A., deVelasco Vázquez, M., & Justo Blanco, R. (2019). Euskaraz hitz egiten ikasten duten makina autodidaktak. https://doi.org/10.26876/ikergazte.iii.03.16

López Zorrilla, A., deVelasco Vázquez, M., & Justo Blanco, R. (2020). Euskarazko elkarrizketa sistema automatikoa sare neuronalen bidez. https://doi.org/10.1387/ekaia.20987

López Zorrilla, A., deVelasco Vázquez, M., & Torres, M. I. (2021). *A Differentiable Generative Adversarial Network for Open Domain Dialogue*, 277–289. Springer Singapore. https://doi.org/10.1007/978-981-15-9323-9_24

Overview of the publications

In Figure A, the workflow and knowledge transfer among publications are depicted. Each point (P1, P2, P3, ...) corresponds to one of the attached publications hereafter in this section of the thesis. Additionally, each point is colored according to its associated project or corpus. The points are connected by arrows indicating the transfer of knowledge. Furthermore, each publication is accompanied by a brief text explaining its content.

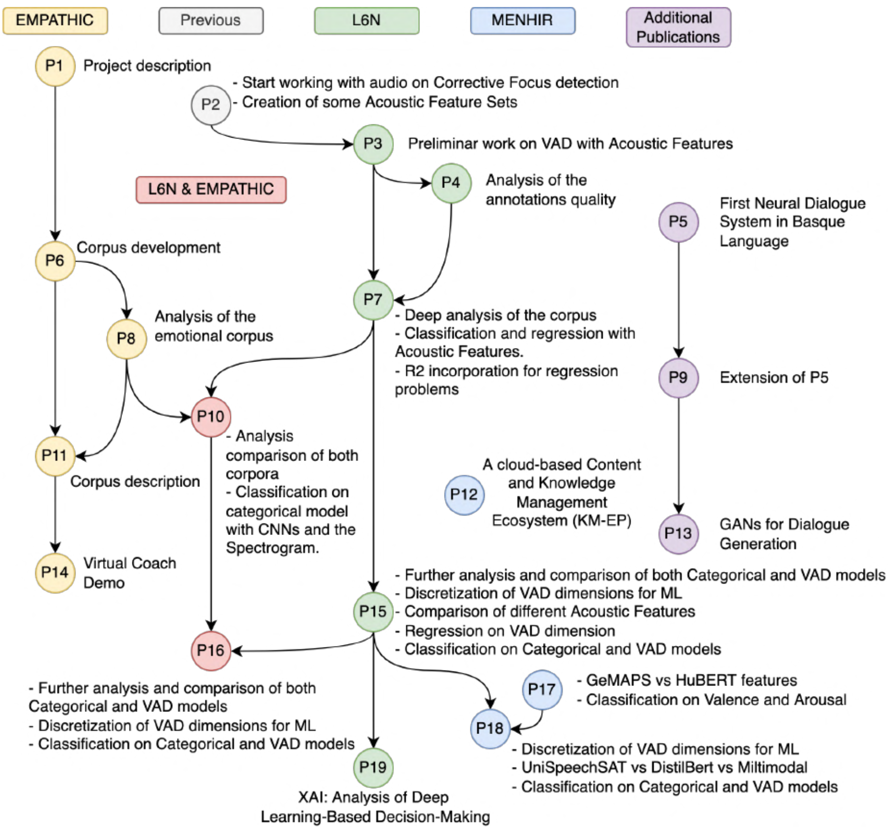


Figure A.: Diagram depicting the chronological order of publications related to the thesis, with connections and explanations of the progress made in each publication. Points are colored with the corpus associated and arrows indicate the transfer of knowledge from one publication to another.

Publication 1

EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly

Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 167-170 recibido 03-04-2018 revisado 26-04-2018 aceptado 04-05-2018

EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly

EMPATHIC: Coach virtual empático, expresivo y avanzado para mejorar el bienestar de las personas de edad avanzada sanas e independientes

Asier López Zorrilla, Mikel de Velasco Vázquez, Jon Irastorza Manso
Javier Mikel Olaso Fernández, Raquel Justo Blanco, María Inés Torres Barañano
Universidad del País Vasco (UPV/EHU)
{asier.lopezz, mikel.develasco, jon.irastorza,
javiermikel.olaso, raquel.justo, manes.torres}@ehu.eus

Abstract: The EMPATHIC project will research, innovate, explore and validate new paradigms and platforms, laying the foundation for future generations of Personalised Virtual Coaches to assist elderly people living independently at and around their home. Innovative multimodal face analytics, adaptive spoken dialogue systems and natural language interfaces are part of what the project will research and innovate, in order to help dependent aging persons and their carers. The project will use remote non-intrusive technologies to extract physiological markers of emotional states in real-time for online adaptive responses of the coach, and advance holistic modelling of behavioural, computational, physical and social aspects of a personalised expressive virtual coach. It will develop causal models of coach-user interactional exchanges that engage elders in emotionally believable interactions keeping off loneliness, sustaining health status, enhancing quality of life and simplifying access to future telecare services
Keywords: Multimodal dialogue systems, virtual coach

Resumen: El proyecto EMPATHIC tiene como misión investigar, explorar, innovar y validar nuevos paradigmas y plataformas, sentando las bases para las futuras generaciones de *coach* virtuales personalizados para ayudar a personas de avanzada edad que viven de forma independiente en su hogar. Además, EMPATHIC investigará e innovará con el objetivo de ayudar a las personas de avanzada edad dependientes y a sus cuidadores mediante análisis facial multimodal, sistemas de diálogo adaptables e interfaces de lenguaje natural. El proyecto utilizará tecnologías remotas no intrusivas para extraer marcadores fisiológicos de estados emocionales en tiempo real, las cuales influenciarán el comportamiento del coach. También se modelarán los aspectos computacionales, físicos y sociales de los coach virtuales expresivos desde un punto de vista holístico. Finalmente, se desarrollarán modelos que permitan la interacción entre el coach y el usuario, buscando así involucrar a éstos en interacciones que contribuyan a evitar la soledad, mantener el estado de salud, mejorar la calidad de vida y simplificar el acceso a futuros servicios de teleasistencia.
Palabras clave: Sistemas de diálogo multimodal, asistente virtual

ISSN 1135-5948. DOI 10.26342/2018-61-24

1 Project Founding and Consortium

The EMPATHIC project has been founded by the European Commission H2020-SC1-2017-RIA grant number 769872: *"Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly"*.

The consortium brings together 10 partners from 7 EU and associated countries (Norway and Israel). Among these partners, we find 4 universities or research centres, 3 healthcare centres related to institutions, and 3 companies. The organisations involved are:

- Universidad del País Vasco UPV/EHU (coordinator).
- OSATEK S.A.
- Oslo University Hospital.
- e-Seniors Association.
- Tunstall Healthcare (UK) Ltd.
- Technion - Israel Institute of Technology.
- Intelligent Voice Ltd.
- Acapela Group S.A.
- Institut Mines-Télécom.
- Seconda Università degli Studi di Napoli.

2 Introduction

Without undermining the important degree of development that public healthcare services in Europe have achieved, in terms of coverage and intensity, much care for people with limitations in autonomy is still provided in the private sector of the family, i.e. informal care. According to the Survey of Health, Age and Retirement in Europe (SHARE, 2002 - 2004), more than seven out of ten dependent elderly people in Italy, Germany, and Sweden receive informal care exclusively. In Spain, Austria and Holland it is more than half. In Belgium, Denmark and France, despite more coverage of formal services, more than one third receive such informal care. It must be recognised that informal care for the family is unpaid work that provides economic savings to welfare systems. However, the predictions point to less availability of informal caregivers in the future. This situation has led European countries to become interested in policies to support the informal care network, and caregiver support programs have become a priority area of action in the Union.

Studies suggest that attention to the lifestyle of the elderly can help them to maintain independent life (Willcox, Scapagnini, and Willcox, 2014a) (Davies, 2011). However, elder psychological obstacle, lack of knowledge and interpersonal and structural obstacles make it difficult for them. Therefore, besides promoting a healthy lifestyle, attention should be paid to the internal and external difficulties of the elderly through facilities and arrangement of activities. Here, virtual coaching is a very interesting solution (Ding et al., 2010) (Prescott et al., 2012) (Cavanagh and Millings, 2013).

Virtual Coach refers to a coaching program or device aiming to guide users through tasks for the purpose of prompting positive behaviour or assisting with learning new skills. Virtual coaches monitor how the user performs activities, provides situational awareness and gives feedback and encouragement matched to their cognitive state and circumstances at the same time. Further, a virtual coach matches its level of support to the user as his/her abilities change, and so the user can upload new options to the virtual coach as desired, and can define new well-being goals without even an office visit.

EMPATHIC will offer a personalised virtual coaching program in three languages (Spanish, French and Norwegian), ranging from low to high intensity levels. Such increased physical activity will help to decrease stress and depressive symptoms, and increase satisfaction with body function. An intervention to improve fruit and vegetable consumption based on the Dietary Guidelines for elders, and a review of practices, will also be applied, since proper nutrition plays a vital role in the health and wellbeing of older adults (Knoops et al., 2004) (Willcox, Scapagnini, and Willcox, 2014)b). EMPATHIC will also offer coaching to focus on brief behavioural techniques like social activation and pleasant event planning.

EMPATHIC will cater to healthy people of 65 years or more, and virtually coach them in order to help increase the years of independent active living and to improve health and slower deterioration. Contact with nature will be encouraged. Recommendations will include also healthy food (and shopping), having enough sun exposure, vegetables and protein intakes. All counselling will be given in a positive, optimistic manner. The device will include also social activities information, like films, operas, theatres, and conferences in their city.

To this end, this project will develop innovative multimodal face and speech analytics, adaptive spoken dialogue systems, intelligent computational models and natural human-computer interfaces, resulting in an emotionally-expressive virtual coach, designed to help aging users and their carers. Building upon neuroscience research, the project will use unobtrusive remote technologies to extract physiological markers of emotional states in real-time. The virtual coach will monitor facial cues and speech style that underpin the user's basic neural function, and will formulate online adaptive responses, facilitating interaction through mimicking, in turn promoting empathy and support with the user.

3 EMPATHIC objectives

3.1 Multidisciplinary research

EMPATHIC proposes multidisciplinary research and development, involving:

- Geriatrician, Neuroscientist, Psychiatric, health and social work specialists to implement the individual coaching goals.
- Psychologist, Neuroscientist and Computer Science experts for detection and identification of the emotional status of the user.
- Engineers and Computer Scientists in speech and language technologies, biometrics, image analysis, and machine learning.
- Telecare services, a senior association and a hospital interested in testing and validating EMPATHIC.
- Companies interested in providing and developing technology for the project and commercialising the products and derived services.

3.2 Main objectives

The six main objectives of this multidisciplinary research are listed next:

- Design a virtual coach, to engage the healthy-senior user and reach preset benefits, measured through project-defined metrics, to enhance well-being through awareness of personal physical status, by improving diet and nutritional habits, by developing more physical exercise and by social activity.
- Involve end-users and to reach a degree of fit to their personalised needs and requirements, derived by the coach, which will enhance their well-being.
- Supply the coach with non-intrusive, privacy-preserving, empathic, expressive interaction technologies.
- Validate the coach efficiency and effectiveness across 3 distinct European societies (Norway, Spain, and France), with 200 to 250 subjects who will be involved from the start.
- Evaluate/validate the effectiveness of EMPATHIC designs against relevant user's personalised acceptance and affordance criteria (such as the ability to adapt to users' underlying mood) assessed through well defined Key Performance Indicators.
- Drive the developed methodology and tools to industry acceptance and open-source access identifying appropriate evaluation criteria to improve the "specification-capture-design implementation" software engineering process of implementing socially-centred ICT products.

3.2.1 Scientific goals and research actions

These six objectives will be accomplished through the following sets of goals and research/development actions:

- Implement health-coach goals and actions through an intelligent computational system, intelligent coach and spoken dialogue system adapted to users' intentions, emotions and context.
- Provide automatic personalised advice guidance (through the coach) having a direct impact in empowering elder users into a wide of advanced ICT keeping improving their quality of life and level

their independent independency living status of the people as the age.

- Identifying non-intrusive technologies to detect the individual's emotional and health status.
- Provide the virtual coach with a natural, empathic, personalised and expressive communication model.

3.2.2 Technological goals and actions

- Develop a simulated virtual coach and acquire an initial corpus of dialogues through a Wizard-of-Oz to fulfil the initial end-users and data requirements of the Scientific Goals.
- Integrate and provide a proof-of-concept of the technology running on different devices.
- Validation through Field trials in the aforementioned three languages.

3.2.3 Exploitation goals

- Define a plan for the exploitation of the results by the consortium as a whole and by particular partners.

Acknowledgements

This work has been founded by the European Commission H2020-SC1-2017-RIA grant number 769872: "Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly".

References

Cavanagh, K. and A. Millings. 2013. Interpersonal computing: the role of the therapeutic relationship in e-mental health. *Contemporary Psychotherapy*, (43):197–206.

Davies, N. 2011. Promoting healthy ageing: the importance of lifestyle. *Nursing Standard (through 2013)*, 25(19):43.

Ding, D., H. Y. Liu, R. Cooper, R. A. Cooper, A. Smailagic, and D. Siewiorek. 2010. Virtual coach technology for supporting self-care. *Physical medicine and rehabilitation clinics of North America*, (21(1)):179–194.

Knoops, K. T., L. C. de Groot, D. Kromhout, A. E. Perrin, O. Moreiras-Varela, A. Menotti, , and W. A. Van Staveren. 2004. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly european men and women: the hale project. *Jama*, (292(12)):1433–1439.

Prescott, T., T. Epton, V. Evers, K. McKee, M. Hawley, T. Webb, D. Benyon, S. Conran, R. Strand, M. Buning, P. Verschure, P. Dario, and T. Group. 2012. Robot companions for citizens: Roadmapping the potential for future robots in empowering older people. *In BRAID (Bridging Research in Ageing and ICT Development) Final Conference.*

SHARE. 2002 - 2004. Survey of Health, Ageing and Retirement in Europe. http://www.share-project.org/.

Willcox, D. C., G. Scapagnini, and B. J. Willcox. 2014a. Healthy aging diets other than the mediterranean: a focus on the okinawan diet. *Mechanisms of ageing and development*, 136:148–162.

Willcox, D. C., G. Scapagnini, and B. J. Willcox. 2014)b. Healthy aging diets other than the mediterranean: a focus on the okinawan diet. *Mechanisms of ageing and development*, (136):148–162.

Publication 2

Corrective Focus Detection in Italian Speech Using Neural Networks

Corrective Focus Detection in Italian Speech Using Neural Networks

Asier López-Zorrilla[1], Mikel deVelasco-Vázquez[1], Sonia Cenceschi[2], M. Inés Torres[1]

[1] Speech Interactive Research Group, Universidad del País Vasco UPV/EHU Barrio Sarriena s/n, 48940, Leioa, Spain
asier.lopezz@ehu.eus, mikel.develasco@ehu.eus, manes.torres@ehu.eus

[2] ARCSLab, Dep. of Electronics, Information and Bioengineering, Politecnico di Milano. Piazza Leonardo da Vinci 32, 20133, Milan, Italy
sonia.cenceschi@polimit.it

Abstract: The corrective focus is a particular kind of prosodic prominence where the speaker is intended to correct or to emphasize a concept. This work develops an Artificial Cognitive System (ACS) based on Recurrent Neural Networks that analyzes suitable features of the audio channel in order to automatically identify the Corrective Focus on speech signals. Two different approaches to build the ACS have been developed. The first one addresses the detection of focused syllables within a given Intonational Unit whereas the second one identifies a whole IU as focused or not. The experimental evaluation over an Italian Corpus has shown the ability of the Artificial Cognitive System to identify the focus in the speaker IUs. This ability can lead to further important improvements in human-machine communication. The addressed problem is a good example of synergies between Humans and Artificial Cognitive Systems.

Keywords: Focus; Stress; Prosodic prominence; Neural networks

1 Introduction

The stress prominence in speech is a phenomenon clearly related to human communication. Speakers usually focus acoustically one or more syllables of their speech in order to express emotions, which allows to position this work in the field of Affective Computing [1], or to introduce a new topic/concept into the dialog. Corrective focus is a particular kind of prosodic prominence where the speaker is intending to correct or to emphasize a concept. Thus, hereinafter we will refer to *focus* instead of citing the more general concept of prominence. The focus is a clearly cultural phenomenon, which is very dependent of the language and additional cultural facts. Thus, it is more frequent in some languages such as

English and Italian than in Spanish or French, that are very strong syllable-timed languages. The focus fits into the list of paralinguistic [2] and suprasegmental characteristics of human speech defined as prosody, involved in the cognitive processes of communicating and understanding. As a consequence, the automatic recognition of the occurrence of a prosodic prominence [3], or a focus in particular, in human speech is interesting for many different fields of study, Linguistics, Cognitive Sciences, etc. Moreover, it takes an important role in Human-Machine Communication.

In summary, the problem addressed in this work is the analysis of the intra-cognitive communication [4, 5] between a set of speakers who emphasized a word according to their communicative intention and a set of listeners aimed at detecting the focus in order to properly decode the message emitted by the sender. In this framework this work develops an Artificial Cognitive System (ACS) that plays the role of the listener resulting in inter-cognitive infocommunications [4, 5] between each speaker and the artificial system, thus using just the audio as the only CogInfoCom channel [6]. The ACS is based on Recurrent Neural Networks (RNNs) that analyzed suitable features of the audio channel. The capacities of such an artificial system are compared to the ones of the humans listeners allowing to analyze the synergies between Humans and artificial cognitive systems, i.e. between Engineering and Cognitive Sciences [7]. The results of our experiments showed the ability of the artificial cognitive system to identify the focus in the speaker IUs, which can result in further important improvements in human-machine communication [8].

The main novelty of this work lies in addressing the automatic focus detection with RNNs. This choice is based on the concept that the human speech is a continuous signal in the temporal domain where each syllable (focused or not) keeps a clear relation with the previous and following ones. In particular, we propose two different approaches to build the ACS. The first one is aimed at detecting focused syllables within a given utterance or Intonational Unit (IU), as explained in [9]. The second one identifies a whole IU as focused or not, so each of them address a different goal. Additional contributions refer to the proposed network structures that are powered only by the acoustic part of the message. Hence, the textual input is not required and as a consequence many technical problems can be bypassed allowing the methodology be improved and adapted to deal with other languages.

The experimental evaluation of the proposal was carried out over a subset of Italian Intonational Units based on the CALLIOPE Corpus [10, 11]. This corpus aims at cataloging IUs from an acoustic point of view, which agrees with our goal to investigate the prosody. Thus, we go beyond the analyses based on linguistic and language related contents, and consider the speech from a phonological and psychoacoustic point of view, as proposed in [12].

Section 2 deals with the pragmatic role and automatic detection of the corrective focus and includes some related works. Section 3 describes the two proposed approaches for the automatic corrective focus detection that are intended to reproduce the mechanism of understanding the focus normally unconsciously implemented at the cognitive level. Experiments carried out are fully described in Section 4. Section 4.2 shows the experiments carried out under the syllable-based approach whereas Section 4.3 deals with the experiments achieved at IU level. Section 4.4 includes a perceptual test concerning the focus recognition by Italian native speakers, allowing a comparison between the prediction ability of humans and ACSs. Finally, some concluding remarks are reported in Section 5.

2 Related Work

The stress prominence in speech [13] is a phenomenon that is easily and naturally produced and perceived by humans during a conversation. It is mainly produced with communicative purpose, but it is also related to the emotional status. Among the different kind of stress prominences, the corrective focus [14] is the main subject of research in this work. It consists in an acoustic stress applied to a syllable or entire word, in order to correct a content or a concept cited by the previous speaker.

Prosodic and paralinguistic cues have been largely explored in Natural Language Processing (NLP) [4], as well as the particular topic of the automatic detection of prominence [15]. Although textual information has been used in addition to acoustic features for the automatic focus detection [16], we are interested in working only with acoustic features because it simplifies the ACS and also makes it more language-independent. In this framework, [17] proposes a free-of-text automatic detection of stress on the Hungarian language at syllable level based on peaks of prosodic features.

If we consider Neural Networks methodologies in this area, the number of researches decrease considerably, and it is really limited narrowing down to the Italian language [18, 19]. Multiple types of stresses have been studied and classified with standard Feedforward Neural Networks [20, 21, 22] and with Convolutional Neural Networks [23, 24] with more success than other machine learning techniques. However, to the best of our knowledge RNNs have never been applied to detect the focus yet.

Another topic of interest regards the acoustic feature selection involved in focus characterization. Several studies have been carried out to determine which features are the most informative [15, 25, 26]. These seem to converge on variants of the same features: the duration of the focused syllable, the energy, the fundamental frequency contour, and the spectral emphasis. We report our own conclusions throughout the Section 4, where we show that the optimal feature selection depends on how the focus detection problem is addressed.

3 Automatic Corrective Focus Detection

The automatic recognition of the focus occurrence has a direct application for forensic or NLP purposes, where there is a need to identify new topics as well as a pragmatic and emotional discontinuities of the speaker on large amount of data. In such a case a procedure that works well at sentence level is needed. Distinctively, linguistic and phonology subjects, such as the characterization of dialects or the learning of a language, might require a more refined system allowing to get the time position of the focus into a word.

As a consequence we propose to formalize two different pattern recognition tasks to be solved. In the first task a given syllable in an IU has to be classified as *focused* or *not focused*. To this end the acoustic features of the given syllable as well as its previous and following temporal context will be considered. This task was named as the *focus in syllables classification problem* (FSP).

The second task will deal with whole IUs. In this case, the ACS will predict if any syllable in the sentence has been uttered with a corrective focused or not. Therefore, the acoustic features will be calculated at regular time windows in the whole IU. This task will be referred as the *focus in IUs classification problem* (FIUP).

3.1 The Focus in Syllables Classification Problem (FSP)

This Section describes the FSP approach aimed at detecting focused syllables in given IUs. The section first includes some details of the feature extraction methodology for this problem, then it explains two ways to combine these features in order to build the input of the classifiers, and it finally describes the structure of the proposed Neural Networks.

Feature extraction. The feature extraction procedure was based on a short-term analysis of the speech signal over 25 ms windows overlapping each 10 ms. For each frame we extracted: Pitch, Zero-Crossing Rate (ZCR), Energy, the Spectral Centroid, Spectral Spread, 13 Mel-frequency Cepstral Coefficients (MFCCs), 16 Linear Predictive Coefficients (LPCs) and 29 Bark features[1]. Additionally, we

[1] Pitch, LPCs, and Bark features were extracted with the Praat Speech Analysis Tool [27] whereas ZCR, energy, spectral centroid, spectral spread and MFCCs with the PyAudioAnalysis library.

computed the first and second derivatives of these 63 features, which increased the number of available features per frame to 189.

Then this number was increased again to 378 by adding the long-term smoothed features of the short-term ones. The smoothing was carried out by calculating the average value of the short-term features centered on the given frame. The number of feature vectors involved in that average were 23 (the central one and 11 previous and following vectors). This time interval is very close to the mean syllable duration in Italian: (0.235 ± 0.1) s[2]. This makes sense because the problem to be solved is the detection of focused syllables which are quasi-stable during their duration. Finally, every feature vector was normalized so that its mean and standard deviation per IU are 0 and 1 respectively.

Building the input to the classifiers. In order to build the input vector to be supplied to the classifier we assume that each IU in the corpus is segmented into syllables, i.e. that we know when each syllable starts and ends. Thus, given a syllable in a IU the input vector will consist of the feature vector corresponding to the center of the syllable under consideration along with some additional feature vectors representing the syllable context as well as the duration (in seconds) of the syllable. At this point two different methods to get such a context were proposed: a fixed frame distance and a context size related to the syllable duration.

> **Fixed frame distance.** In this approach both the context size and the frame distance are fixed. The first refers to the number of left and right context feature vectors that will be selected, whereas the second to the distance between consecutive context vectors. As an example, if the context size is fixed to 2 and the frame distance equals 3, the input would be built as in the Figure 1.

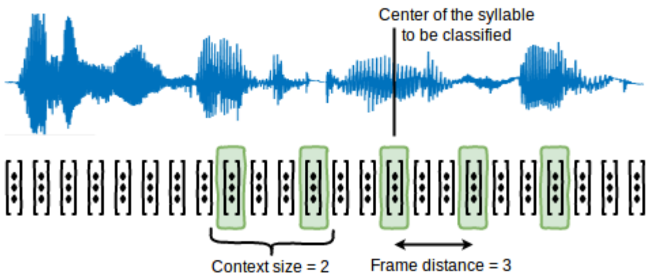


Figure 1

An example of how to build the input with fixed frame distance. In this case, 5 vectors were taken in total: the central one and two left and right context vectors, according to the context size. The frame distance was set to 3.

[2] This value was computed after an automatic syllable segmentation process of our corpus.

Beginning, center and ending of neighbor syllables. In this approach the context feature vectors are selected among the ones representing the beginning, the center and the end of the neighbor syllables, according to the segmentation of the IU into syllables. Hence, in this case we only need to specify the context size. Figure 2 shows an example of the input for a context size set to 3.

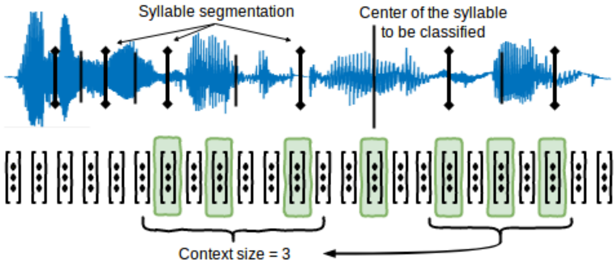


Figure 2
An input built using feature vectors corresponding to the beginning, center and ending of neighbor syllables. Since the context size was set to 3, the vectors corresponding to the end of the central syllable (which is also the one that corresponds to the beginning of the next syllable), to the center of the next syllable and to the end of the next syllable were selected as the right context. Symmetrically, the left context consists of the vectors corresponding to the beginning of the central syllable, to the center of the previous syllable and to the beginning of it.

Classifiers. The previous methods allow the generation of training examples that can be used by common machine learning algorithms. Once the specific set of acoustic features are selected and the methodology to build the input is chosen, all the feature vectors can be concatenated to form a fixed-dimensional input vector representing each syllable in the corpus. Then, classifiers such us Naive Bayes, Support Vector Machines (SVM) and conventional Feedforward Deep Neural Networks can be directly trained. These classifiers were used for the experiments shown in Section 4.2. However, the temporal relationship between the feature vectors that compose the input of each training example is not considered enough by these classifiers. Thus, more complex neural networks based on recurrent layers might be more suitable. In this framework we propose RNNs with two parallel sets of recurrent layers. The first one processes the left (previous in time) context vectors forward, i.e., it takes first the farthest context vector in the left-side and sequentially all the left context vectors until the central vector is processed. Symmetrically, the other set of recurrent layers processes the right context vectors backwards. Additionally, our architecture includes another parallel set of feedforward layers, which processes the scalar corresponding to the duration of the syllable we want to classify. Finally, the three sets are merged and the network ends with a set of feedforward layers. Figure 3 shows a graphical representation of the proposed Bidirectional RNNs. These networks led to the best system performance when dealing with the FSP according to the experiments carried out (see Section 4.2).

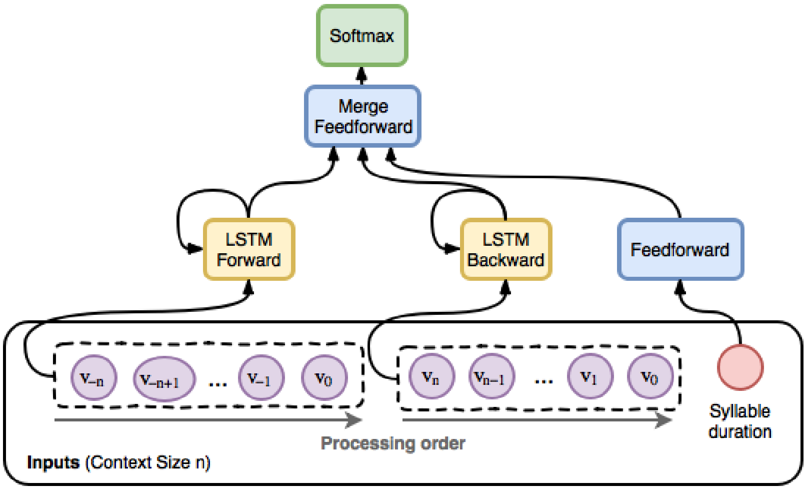


Figure 3

An example architecture of a neural network used in the FSP. The two sets of recurrent layers consist of a single LSTM layer each. The duration of the syllable is also processed with a single feedforward layer. Then the output of these three layers are merged into a feedforward layer followed by a softmax layer of two outputs, one per class.

3.1 The Focus in IUs Classification Problem (FIUP)

This Section describes the FIUP approach aimed at classifying a whole IU as focused or not. The feature extraction methodology for this problem is the one used to deal with FSP problem. Thus, this section just explains the way to combine these features in order to build the input of the classifiers, and then it describes the structure of the proposed Neural Networks.

Building the input to the classifiers. We propose two different ways to build the input of the classifiers: the first one is based on regular sampling of the sequence of feature vectors whereas the second one is based on the output of the networks classifying syllables (FSP) as focused or not focused.

Fixed frame distance. If we use a fixed sampling rate from the beginning to the end of the IU to select the feature vectors that will be involved in the classification process, more than one training example per IU can be generated. More precisely, if the frame distance was set to *n*, we can generate *n* examples, just alternating the vector from where the sampling starts.

From the FSP to the FIUP. In this approach we take advantage of the classifiers trained to solve the FSP. Each given IU can be automatically segmented into (pseudo-)syllables. Then, the input corresponding to each of these pseudo-syllables can be propagated across an already trained classifier. Afterwards, these predictions can be used as an alternative input to train a classifier to deal with FIUP. This approach is specially interesting if the classifier trained to solve the FSP is a Neural Network, since not only its output can be used, but also the output of the penultimate layer, which contents more features about the syllable.

Classifiers. An additional difference between the FSP and the FIUP approaches is that common classifiers cannot directly be trained. In fact, Naive Bayes and SVMs classifiers as well as Feedforward Neural Networks require the dimension of input vector to be fixed for all the examples. However, such a condition will certainly not be met due to the variable length of the IUs (if we are using the first way to build input), and/or because of the variable number of syllables in the IUs.

RNNs, though, are still directly trainable in this scenario. These are able to sequentially process any sequence of vectors of arbitrary length, which makes them really suitable for this task. In particular, we propose bidirectional RNNs. One set of layers processes the whole sequence of feature vectors forwards, from the first vector to the last. Another set of layers processes the sequence in the inverse order, backwards from the last vector to the first. Figure 4 shows a graphical representation of the proposed structure. Note that the proposed RNN is able to deal with inputs obtained under the two building methodologies proposed.

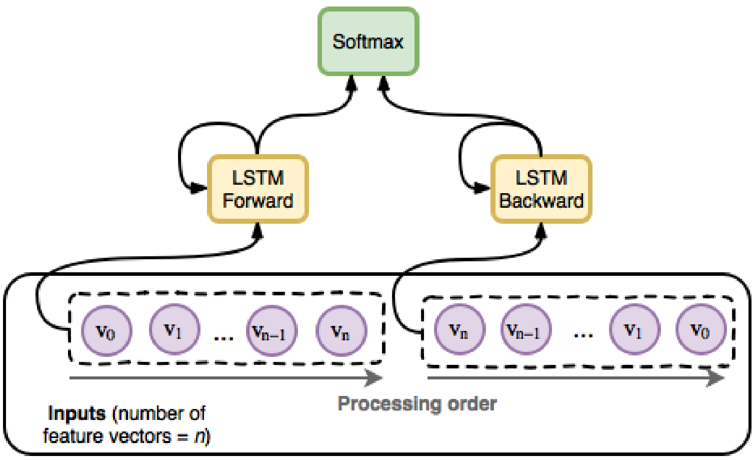


Figure 4
An example RNN used in the FIUP. A LSTM layer processes the input forwards and another forward. The network ends in a softmax layer of two outputs.

4 Experimental Study

Two series of experiments were carried out to evaluate the performance of the ACS. The first series aims to validate the proposals described in Section 3.1 when dealing with the FSP whereas the second one focus in the FIUP under the approaches proposed in Section 3.2. An additional set of experiments allow to analyse the human perception abilities for the same data collection. A subset of the Standard Italian Corpus (SIC) described in Section 4.1 was used for all the experiments.

4.1 The Standard Italian Corpus

Italian is a romance, iso-syllabic and free-stress language [19]. Then, the position of a contrastive focus is just a communicative choice of the speaker. The presence of focus has been related to the duration of the syllable, or to the distance between peaks of energy (syllable nuclei). In fact, the duration of a focused syllable is typically higher than the one of not focused syllables of the same speaker. However, it is unrelated to the tonic/tonic syllables alternations providing the rhythm [26].

The Corpus selected to carry out the proposed series of experiments is based on CALLIOPE (Combined and Assessed List of Latent Influences On Prosodic Expressivity), a conceptual model created within the LYV project[3] aiming at categorizing all IUs. Each IU is thus associated to a "point" into this space and associated to a tuple composed of 12 labels (detailed descriptions in [10]). In this multidimensional space each dimension represents a characteristic influencing the vocal paralinguistic components of the speech assuming values in a set of labels.

Table 1
List of the CALLIOPE dimensions

Group	**Dimensions (F_i)**
Dialogic	Structure (F_1), Linguitic modality (F_2), Intonational focus (F_3), Rhetorical form (F_4), Motivational state (F_5), Speech mood (F_6), Spontaneity (F_7), Punctuation forms (F_8), Emotions (F_9)
Background	Expressiveness skill (F_{10}), Social context (F_{11}), Launguage (F_{12})

Each IU has a subjective correspondence with a specific prosodic unit. Starting from this conceptual model a database of Italian standard speech has been defined and created. CALLIOPE dimensions are divided into two groups as shown in Table 4.1. The Dialogic group contains characteristics directly related with the

3 LYV is a project of the Polisocial program 2016-2017, http://www.polisocial.polimi.it focused on the improvement of prosodic and expressive skills of Italian speakers with cognitive disabilities, through the use of technology in complex contexts [28].

communication context, where the corresponding sets of labels are fully defined. The second group contains background dimensions, i.e, characteristics that exist regardless of whether or not there is an interaction.

The selected corpus concerns a subspace of the CALLIOPE model, obtained narrowing the field of recordings by setting 6 dimensions as follows. The language (F_{12}) is the Standard Italian [29], recited by able-bodied (F_{10}) actors (F_7) and the contents concern daily situations (F_{11}) and absence of particular motivational states (F_5) emotions (F_9). The corpus considers 13 Calliope's labels (among the remaining 6 dimensions) an includes the Corrective Focus, which was validated by a perceptive test performed on about 200 Italian native-speakers. Audio files were recorded in WAV format (44.1 kHz 16 bit) with different modes and microphones to obtain a model as independent as possible from the technical apparatus. 14 speakers (7 men and 7 women) aged between 33 and 48 were recorded. Each speaker recorded 278 IUs (139 with meaning and 139 pseudo-sentences [30] with equal prosody) so that the corpus contains 1946 sentences with meaning and 1946 pseudo-sentences. Considering both real and pseudo sentences, 2884 IUs do not contain any prosodic prominence while 1008 contains one or more corrective focuses.

This database is ready for the experimental evaluation of the proposals to solve the FIUP through the second series of experiments. However, the FSP needs a segmentation of each IU into syllables that have to be labelled. To this end we proposed an automatic syllable segmentation procedure that was based on the syllable positions provided by Praat [27], i.e. the beginning and end of each syllable. Some few errors appeared for long syllables that were sometimes split into two subsegments. Then, we manually labeled each of these (pseudo-)syllables as *focused* or *not focused*. In total, the resulting corpus consists of 44923 pseudo-syllables; 1867 focused and 43056 not focused. This corpus is highly unbalanced and includes one focused pseudo-syllable per 22 non focused ones, approximately.

4.2 Study of the FSP

Preliminar experiments. The initial experiments included the parametric Naïve Bayes classifier and the geometric SVM one as well as Feedforward Neural Networks. The average F1-score between the two classes in our dataset was used to evaluate the performance of each classifier. This measure was computed after a 7-fold cross-validation process. In each iteration the instances of 2 of the 14 speakers in the corpus were left as the test partition. All the neural networks were implemented with the WBNN toolkit[4], while the Scikit-learn toolkit was chosen to

[4] The first and second authors of this work are the main developers of this open source toolkit,. which is still under development. It can be found at https://github.com/develask/White-Box-Neural-Networks.

train the Naïve Bayes and the SVM classifiers. Columns 2 to 4 in Table 2 show the results of these experiments and confirm that Neural Networks outperform both SVM and Naive Bayes classifiers in terms of the average F1-score.

Table 2
Average F1-score obtained by different classifiers

	Best RNN	**Best feedforward NN**	**Best SVM**	**Best Naïve Bayes**
Average F1-score	0.693	0.618	0.576	0.512

Experiments with the proposed Recurrent Neural Networks. We then focused on bidirectional RNNs due to their ability to process sequences of variable length. In particular, we explored several RNN architectures and hyperparameters as well as several ways to build the input to the network and its parameters. First column, in Table 2 shows the best results that were achieved with RNN that clearly outperfomed the ones obtained by Feedforward NN. The structure of this best RNN is very similar to the one previously shown in Figure 3. Each recurrent layer consists of 10 LSTM cells[5], the layer that processes the syllable duration is made of 8 sigmoidal units, the layer after merging the three sets of parallel layers consists of 20 sigmoidal units, and the network ends in a softmax layer of two units, one per class. Results in column one in Table 2 were obtained when the set B of features (pitch, energy and spectral centroid without any derivative) was selected. Finally, a fixed frame distance of 11 and a context size of 9 vectors resulted to be the best configuration to build the RNN input. The RNNs were trained by stochastic gradient descent with an exponentially decaying learning rate during a fixed number of epochs. The best choice for theses parameters was to reduce the learning rate from 0.5 to 0.1 throughout 75 epochs.

This is the configuration for the ACS achieving the higher system performance shown in Table 2, i.e. the best RNN. To get these results we had previously explored two techniques to deal with the imbalance of the data set. We first included a classical variable decision threshold to determine the confidence level[6] required by the RNN to predict that the input corresponds to a focused syllable. An exhaustive search of this parameter was carried out to maximize the average F1-score between the two classes in the training partition. As an alternative we proposed to apply an increasing imbalance schedule in the training data [32]. To this end the network was trained with different data each epoch, starting from a not very unbalanced subset of the training data and slowly adding more examples from the majority class. The best schedule was to increase the imbalance from 5 (5 non-focused syllables per each focused one) to the real imbalance (around 22),

5 We implemented the LSTM version proposed in [31].

6 The confidence level is the output of the neuron of the softmax layer that corresponds to focused syllables.

with a scaled hyperbolic tangent function. Table 3 shows how the performance was improved with the use of these techniques.

Table 3
Average F1-score obtained with the proposed techniques to deal with unbalanced data

	RNN with threshold and imbalance schedule	**RNN with threshold**	**Baseline RNN**
Average F1-score	0.693	0.618	0.576

Effect of the sets of features. We explored a variety of features as well as several ways to combine them. Then, the six sets of features listed below were selected. Additionally, we also experimented with sets that added the first derivatives of the proposed features on the one hand or the first and the second derivatives on the other hand. Note that all the features correspond to the long-term smoothed version.

Set A. Pitch and energy.
Set B. Pitch, energy and spectral centroid.
Set C. Pitch, energy, spectral centroid, ZCR and spectral spread.
Set D. Pitch, energy, spectral centroid, ZCR, spectral spread and 13 MFCCs.
Set E. Pitch, energy, spectral centroid, ZCR, spectral spread and 16 LPCs.
Set F. Pitch, energy, spectral centroid, ZCR, spectral spread and 29 Bark features.

Figure 5 shows the performance of the described best model when different sets of features were used. First and second derivatives led to a decrease of performance for all the feature sets. i.e. they did not add any information. Pitch, energy and spectral centroid resulted to be the most informative features for this problem. The high performance obtained by the ACS when a so reduced set of features was used outlines the capability of the proposed RNN structure and configuration.

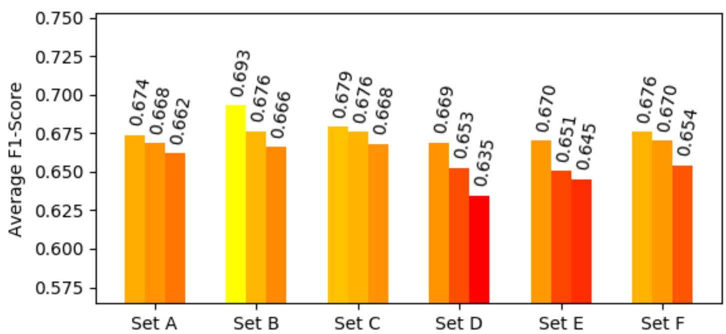


Figure 5
Average F1-score obtained with the best network trained with different sets of features. The three columns showed per set indicate the performance when no derivatives are added (left column), when the first derivative is added (central column) and when the first and second derivatives are added (right column).

Effect of the context. Figure 6 shows the ACS performance of the described best model and best set of features for different values of the context size and frame distance as defined in Section 3.1. Figure 6 evidences that a lack of information, i.e. a small context size, drastically worsens the system's performance. However, big context sizes do not significantly reduce the classification capacity of the proposed ACS. Thus, the ability of the LSTMs to *forget* non relevant events appear to pay off but the computation time is clearly much higher. On the other hand, the analysis of the frame distance shows an optimal range between 5 and 15 frame distance where the performance does not significantly depend on the value of this parameter. However, the average F1-score clearly decays out of this range. Thus, low frame distances considers a few context but very big ones seem to lead to a loss of important events.

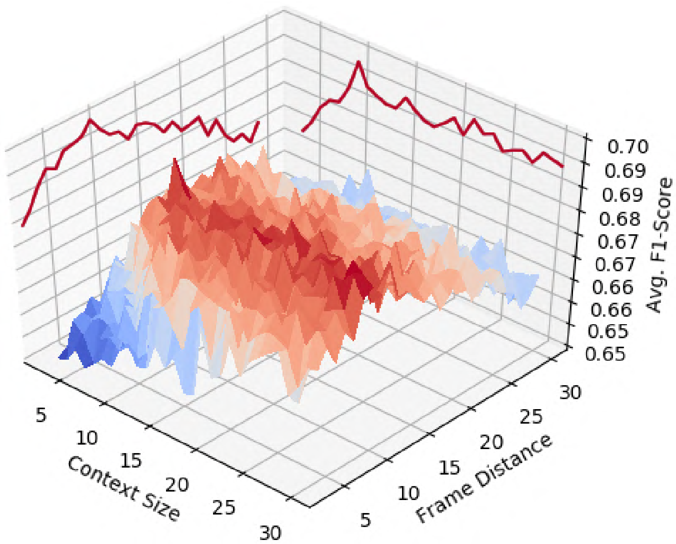


Figure 6

Average F1 score in the FSP of the described best network and best set of features for different values of the context size and frame distance as defined in Section 3.1

4.3 Study of the FIUP

A second series of experiments were carried out with the Standard Italian Corpus in order to deal with the FIUP. The sets of parameters defined in Section 4.2 were also considered for these experiments.

Experiments with the proposed RNNs. The RNNs proposed to solve the FIUP are based on the architectures described in Figure 4. The best results were obtained when 60 LSTM cells per recurrent layer were considered and the RNN was trained during 40 epochs. The best learning rate schedule was still an exponentially decaying one from 0.5 to 0.1. In addition, a variable decision threshold was included to optimize the average F1-score in the training partition. However, the use of a schedule throughout the epochs to deal to the imbalance at training time did not lead to any improvement in this case. This is probably due to the fact that the imbalance is not so high in the FIUP (around 3 IUs without focus per each IUs with focus).

Effect of the sets of features. Figure 7 shows the performance of the described best RNN when different sets of features were used. Unlike the FSP problem the first derivatives seem to be significant mainly for set F. In fact, the size window analysis is now bigger so that the information provided by derivatives is meaningful. Moreover, Set F, which consists of the pitch, the energy, the spectral centroid, ZCR, the spectral spread and 29 Bark features, led to the higher ACL performance for this problem achieving a great average F1-score of 0.826. In the same way spectral changes seem also to be more significant for larger windows.

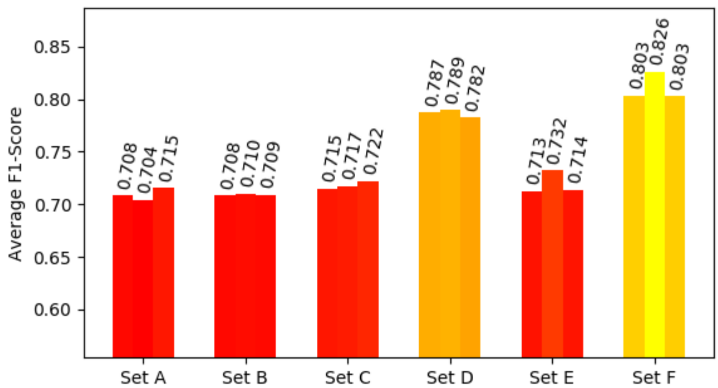


Figure 7

Average F1-score obtained with the best network trained with different sets of features for the FIUP. As before, the columns represent the addition of no derivatives, the addition of the first derivatives, and the addition of the first and second derivatives.

Effect of the context. When dealing with the FIUP the context is just represented by the frame distance at which input vectors at subsampled. Figure 8 shows the ACS performance of the described best model and best set of features for different values of the frame distance as defined in Section 3.2. Figure 8 evidences a similar effect of the frame distance in system performance than the one analyzed for FSP. In fact, Figure 8 still shows an optimal range where the performance does not significantly depend on the value of this parameter and a very strong decrease of F1-score out of this range. Thus, once again big frame distances seem to lead to a loss of important information.

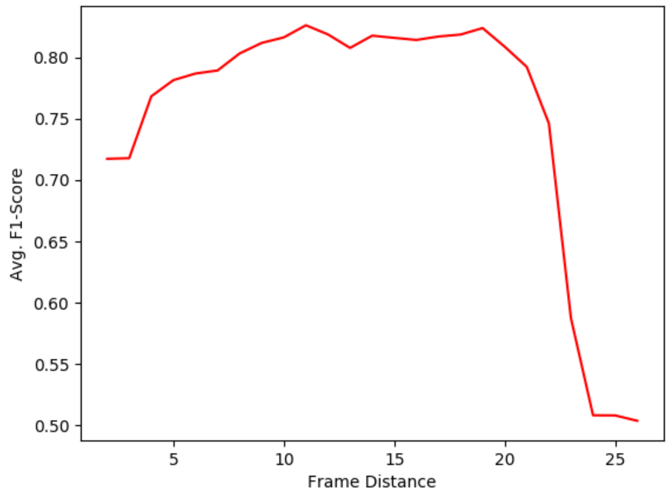


Figure 8

Average F1 score of the described best network and best set of features for different values of the frame distance in the FIUP as defined in Section 3.2

From the FSP to the FIUP. Figure 9 shows the results when predictions from previous FSP classifier were used as inputs for RNN proposed in Section 4.2 to deal with FIUP. Figure 9 evidences that the ACS performances are now lower than the ones got by the previous direct approach. However, let us note that the best result (a F1-score of 0.756) was obtained with an RNN trained on top of the outputs (of the last and penultimate layers) of a network that processes the Set F of features, with no derivatives. Thus, the spectral information seem to be also meaningful with this approach when dealing with the FIUP.

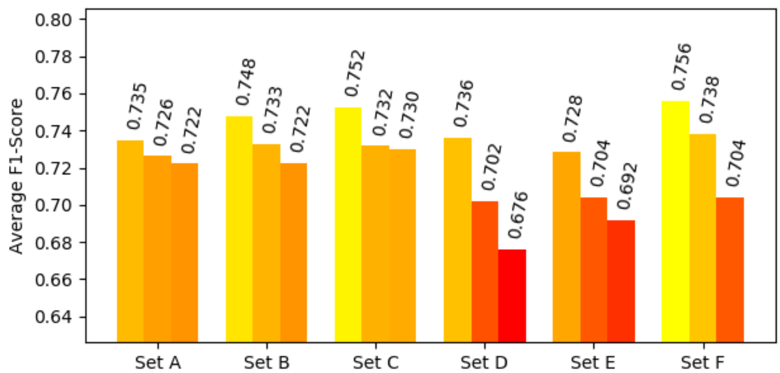


Figure 9

Average F1-score got when training RNNs on top of the features extracted with FSP classifiers

4.4 Human Perception Tests for the FIUP

A series of Human Perception Tests was also carried out with the Italian Corpus. To this end a set of 203 adults, Italian native-speakers, were asked to recognize the 13 labels mentioned in Section 4.1. They classified all the sentences and pseudo-sentences in the corpus without repetitions, i.e., only one speaker per IU. In this work we just considered the question related to the presence of a corrective focus. The average F1-score between the two classes was 0.444, which is much lower than the performance got by both approaches of the ACS dealing with the same FIUP, i.e. 0.826 and 0.756 respectively in terms of F1 scores.

The low perceptive rates may be due to the listener's need of the context provided by a previous interaction of other speaker. It seems that one single IU is not enough to ensure the human focus recognition. In contrast, with the narrow context preferred by the ACS, the human auditory apparatus seems to require a very broad one, extending to other parts of the dialogue.

Concluding Remarks

The corrective focus is a particular kind of prosodic prominence where the speaker is intended to correct or to emphasize a concept. This work has developed an Artificial Cognitive System (ACS) that played the role of the listener resulting in inter-cognitive infocommunication between a speaker and the artificial system, thus using just the audio as the only CogInfoCom channel. The ACS is based on Recurrent Neural Networks that analyze suitable features of the audio channel. Two different approaches to build the ACS has been developed. The first one addressed the detection of focused syllables within a given Intonational Unit whereas the second one identify a whole IU as focused or not. For the first problem the proposed RNN achieved an F-score of 0.693 with a reduced set of acoustic features whereas the RNN were able to get a really high F1-score of 0.826, with a larger set of acoustic features that also includes variations. Experimental results showed the need of context to detect the focus. However, this context is reduced to neighbor syllables. On the other hand, human perception experiments showed that Humans were able to get just an F1 score of 0.444 probably due to the lack of broad contexts including previous dialog turns.

The results of our experiments showed the ability of the Artificial Cognitive System to identify the focus in the speaker IUs, which can lead to further important improvements in human-machine communication. The behavior of the ACS to identifies the focus in speech that can be interpreted, to some extend, as an estimation, optimistic in this case, of the human cognitive load when dealing with the same problem, showing synergies between Humans and Artificial Cognitive Systems.

Acknowledgements

The research leading to the results in this paper has been conducted in the project EMPATHIC (Grant N: 769872) that received funding from the European Union's Horizon 2020 research and innovation programme.

Additionally, this work has been partially funded by the Spanish Minister of Science under grants TIN2014-54288-C4-4-R and TIN2017-85854-C4-3-R, by the Basque Government under grant PRE_2017_1_0357, and by the University of the Basque Country UPV/EHU under grant PIF17/310.

References

[1] Lisetti, C. L. (1998) Affective Computing

[2] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013) Paralinguistics in Speech and Language—State-of-the-Art and the Challenge. Computer Speech & Language, 27(1) 4-39

[3] Terken, J. (1991) Fundamental Frequency and Perceived Prominence of Accented Syllables. The Journal of the Acoustical Society of America, 89(4) 1768-1776

[4] Baranyi, P., & Csapó, Á. (2012) Definition and Synergies of Cognitive Infocommunications. Acta Polytechnica Hungarica, 9(1) 67-83

[5] Baranyi, P., Csapó, Á., & Sallai, G. (2015) Cognitive Infocommunications (CogInfoCom) Springer

[6] Fulop, I. M., Csapó, Á., & Baranyi, P. (2013, December) Construction of a CogInfoCom ontology. In Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on (pp. 811-816) IEEE

[7] Irastorza, J., & Torres, M. I. (2016, October) Analyzing the Expression of Annoyance during Phone Calls to Complaint Services. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000103-000106) IEEE

[8] Torok, A. (2016, October) From Human-Computer Interaction to Cognitive Infocommunications: A Cognitive Science Perspective. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000433-000438) IEEE

[9] Cresti, E. (2000) Spoken Italian Corpus: an Introduction [Corpus di italiano parlato: Introduzione] (Vol. 1) Accademia della Crusca

[10] Cenceschi, S., Sbattella, L., & Tedesco, R. (2018) Towards Automatic Recognition of Prosody. In Proceedings of 9th International Conference on Speech Prosody 2018 (pp. 319-323)

[11] Sbattella, L., Tedesco, R., & Cenceschi, S. (2017) The Definition of a Descriptive Space of Italian Prosodic Forms: The CALLIOPE Model. In XIII Convegno Nazionale AISV (pp. 1-3) ITA

[12] Dominguez, M., Farrús, M., & Wanner, L. (2016) An Automatic Prosody Tagger for Spontaneous Speech. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 377-386)

[13] Werner, S., & Keller, E. (1995, May) Prosodic Aspects of Speech. In Fundamentals of Speech Synthesis and Speech Recognition (pp. 23-40) John Wiley and Sons Ltd.

[14] Gussenhoven, C. (2008) Types of Focus in English. In Topic and focus (pp. 83-100). Springer, Dordrecht

[15] Tamburini, F. (2003) Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System. In Eighth European Conference on Speech Communication and Technology

[16] Beke, A., & Szaszák, G. (2014, November) Combining NLP Techniques and Acoustic Analysis for Semantic Focus Detection in Speech. In 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom) 2014 (pp. 493-497) IEEE

[17] Tündik, M. Á., Gerazov, B., Gjoreski, A., & Szaszák, G. (2016, October) Atom Decomposition-based Stress Detection and Automatic Phrasing of Speech. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000025-000030) IEEE

[18] Tamburini, F., Bertini, C., & Bertinetto, P. M. (2014) Prosodic Prominence Detection in Italian Continuous Speech using Probabilistic Graphical Models. In Proceedings of Speech Prosody (pp. 285-289)

[19] Kori, S., Farnetani, E., & Cosi, P. (1987) A Perspective on Relevance and Application of Prosodic Information to Automatic Speech Recognition in Italian. In European Conference on Speech Technology

[20] Jenkin, K. L., & Scordilis, M. S. (1996, October) Development and comparison of three syllable stress classifiers. In Spoken Language, 1996 ICSLP 96 Proceedings, Fourth International Conference on (Vol. 2, pp. 733-736) IEEE

[21] Li, K., Qian, X., Kang, S., & Meng, H. (2013) Lexical Stress Detection for L2 English Speech Using Deep Belief Networks. In Interspeech (pp. 1811-1815)

[22] Shahin, M. A., Ahmed, B., & Ballard, K. J. (2014, December) Classification of Lexical Stress Patterns Using Deep Neural Network Architecture. In Spoken Language Technology Workshop (SLT) 2014 IEEE (pp. 478-482) IEEE

[23] Heba, A., Pellegrini, T., Jorquera, T., André-Obrecht, R., & Lorré, J. P. (2017, October) Lexical Emphasis Detection in Spoken French Using F-BANKs and Neural Networks. In International Conference on Statistical Language and Speech Processing (pp. 241-249) Springer, Cham

[24] Stehwien, S., & Vu, N. T. (2017) Prosodic Event Recognition using Convolutional Neural Networks with Context Information. arXiv preprint arXiv:1706.00741

[25] Streefkerk, B. M. (1997) Acoustical Correlates of Prominence: A Design for Research. In Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam (Vol. 21, pp. 131-142)

[26] Giordano, R. (2008, May) On the Phonetics of Rhythm of Italian: Patterns of Duration in Pre-planned and Spontaneous Speech. In Proceedings of the 4th Speech Prosody Conference, Campinas, BR

[27] Boersma, P. (2006) Praat: Doing Phonetics by Computer. http://www. praat. org/

[28] Sbattella, L. (2006) La mente orchestra. Elaborazione della risonanza e autismo. Vita e Pensiero

[29] Canepari, L. (1980) Italiano standard e pronunce regionali. Cooperativa libraria editrice degli studenti dell'università di Padova

[30] Cibelli, E. (2012) Shared Early Pathways of Word and Pseudoword Processing: Evidence from High-Density Electrocorticography

[31] Graves, A., & Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures. Neural Networks, 18(5-6) 602-610

[32] López-Zorrilla, A., de Velasco-Vázquez, M., Serradilla-Casado, O., Roa-Barco, L., Graña, M., Chyzhyk, D., & Price, C. C. (2017, June) Brain White Matter Lesion Segmentation with 2D/3D CNN. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 394-403) Springer, Cham

Publication 3

Emotion Detection from Speech and Text

IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Emotion Detection from Speech and Text

Mikel deVelasco, Raquel Justo, Josu Antón, Mikel Carrilero, M. Inés Torres

Universidad del Pais Vasco UPV/EHU

mikel.develasco@ehu.eus, raquel.justo@ehu.eus, josuantonsanz@gmail.com, mcarrilero001@ikasle.ehu.eus, manes.torres@ehu.eus

Abstract

The main goal of this work is to carry out automatic emotion detection from speech by using both acoustic and textual information. For doing that a set of audios were extracted from a TV show were different guests discuss about topics of current interest. The selected audios were transcribed and annotated in terms of emotional status using a crowdsourcing platform. A 3-dimensional model was used to define an specific emotional status in order to pick up the nuances in what the speaker is expressing instead of being restricted to a predefined set of discrete categories. Different sets of acoustic parameters were considered to obtain the input vectors for a neural network. To represent each sequence of words, a models based on word embeddings was used. Different deep learning architectures were tested providing promising results, although having a corpus of a limited size.

Index Terms: Emotion Detection, Speech, Text transcriptions

1. Introduction

The emotion recognition is the process of identifying human emotions, a task that is automatically carried out by humans considering facial and verbal expressions, body language, etc. However, this is a challenging task for an automatic system. In recent years, the great amount of multimedia information available due to the extensive use of the Internet and social media, along with new computational methodologies related to machine learning, have led to the scientific community to put a great effort in this area [1, 2].

Emotion recognition from speech signals relies on a number of short-term features such as pitch, vocal tract features such as formants, prosodic features such as pitch loudness, speaking rate, etc. Surveys on databases, classifiers, features and classes to be defined in the analysis of emotional speech can also be found in [3]. Regarding methodology, statistical analysis of feature distributions has been traditionally carried out. Classical classifiers such as the Bayesian or Super Vector Machines (SVM) have been proposed for emotion features from speech. The model of continuous affective dimensions is also an emerging challenge when dealing with continuous rating emotion labelled during real interaction [1, 4]. In this work, recurrent neural networks have been proposed to integrate contextual information and then predict emotion in continuous time using a three-dimensional emotional model.

Speech transcripts have also been demonstrated to be a powerful tool to identify emotional states [5]. Over the last decade, there has been considerable work in sentiment analysis [6]. Moreover, the detection of emotions such as anger, joy, sadness, fear, surprise, and disgust have also been addressed [7]. However, spoken language is informal and provides information in an unstructured way so that developing tools to select and analyse sentiments, opinions, etc. is still a challenging topic [8]. In early systems dealing with emotion detection in text, knowledge-based approaches were applied making use of emotion lexicons, such as Sentiwordnet [9]. Other methods, employ machine learning based approaches [10], where statistical classifiers are trained using large annotated corpora and the emotion detection can be seen as a multi-label classification problem. In this work we propose to use neural networks to solve the regression problem given the 3-dimensional emotional model.

The main contribution of this work is the appropriate selection of a neural network architecture for emotion detection considering both acoustic signals and their corresponding transcription. Additionally, the proposed architecture has been adapted to the 3-dimensional VAD (Valence, Arousal and Dominance) emotional model [11].

Section 2 describes the two proposed approaches for the automatic emotion recognition from used features to the common network architectures basics. Experiments carried out are fully described in Section 3. Section 3.1 aims to explain the difficulties to find out a Spanish corpus and it has led us to create our own corpus. Section 3.2 mentions the used baselines methods and which measure has been used for testing. Section 3.3 shows the experiments carried out under the regression problem of emotional status with acoustic features whereas Section 3.4 deals with the experiments achieved at the regression problem of emotional status with language features. Finally some concluding remarks are reported in Section 4.

2. Emotion Detection from Speech and Language

Emotion detection from speech is based on the extraction of relevant features from the acoustic signal, that can be seen as hints of the emotional status. That is, a numerical vector that represent the specific information related to emotional status and embedded in an acoustic signal is needed. There are numerous acoustic parameters that can be obtained using the free software *Praat*[1] or the free python library *pyAudioAnalysis*[2]. Considering these tools the following set of 72 parameters could be considered: Pitch, Zero Crossing Rate (ZCR), Energy, Entropy of the energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral flux, Spectral Rolloff, Chroma vector (12 coefficients), Chroma deviation, MFCC coefficients (12), LPC coefficients (16), Bark features (21). However, some of these parameters provide the same or very similar kind of information. For instance, LPC, Bark and MFCC coefficients provide similar information about the phonemes without considering the vocal tract. Thus, such a big set of parameters is useless since it will complicate the learning procedure requiring more training data. In this work different subsets of the aforementioned parameters were explored:

[1] http://www.fon.hum.uva.nl/praat/
[2] https://pypi.org/project/pyAudioAnalysis/

 10.21437/IberSPEECH.2018-15

- Set A: Pitch and Energy.
- Set B: Pitch, Energy and Spectral Centroid.
- Set C: Pitch, Energy, Spectral Centroid, ZCR and Spectral Spread.
- Set D: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 12 MFCC coefficients.
- Set E: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 16 LPC coefficients.
- Set F: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 21 Bark features.

The first set was selected according to the studies performed in [12] where the arousal state of the speaker affects the overall energy and pitch. In addition to time-dependent acoustic features such as pitch and energy, spectral features were selected for Sets B and C as a short-time representation for speech signal [13]. For Sets D, E and F different Cepstral-based features were added, proven that they are good for detecting stress in speech signal [14].

When regarding emotion detection from language the same procedure has to be carried out. First of all a vectorial representation of the transcribed text is needed. In this case, we hope to capture some meaning of the utterance that might help in the detection of specific emotional status. An appropriate representation should consider some semantic information like the word embeddings *word2vec*[15], *doc2vec* [16] or *GLOVE* [17]. *Word2vec* embeddings, the most simple model, are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. *Word2vec* takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. However, this technique represents each word of the vocabulary by a distinct vector, without parameter sharing. In particular, they ignore the internal structure of words, which is an important limitation for morphologically rich languages. For example, in Spanish, most verbs have more than forty different inflected forms and this leads to a vocabulary where many word forms occur rarely (or not at all) in the training corpus, making it difficult to learn good word representations. Thus, [15] proposes to learn representations for character n-grams, and to represent words as the sum of the n-gram vectors. The model (known as *FastText*) can be seen as an extension of the continuous skip-gram model [18] which takes into account subword information.

Additionally, a way of representing the emotional status is needed in order to establish a machine learning problem. A categorical emotion description (e.g. six basic emotions) is an easy way to procedure but it provides a quite constrained model. Affective computing researchers have started exploring the dimensional representation of emotion [19] as an alternative. Dimensional emotion recognition, aims to improve the understanding of human affect by modelling affect as a small number of continuously valued, continuous time signals. It has the benefit of being able to: (i) encode small changes in affect over time, and (ii) distinguish between many more subtly different displays of affect, while remaining within the reach of current signal processing and machine learning capabilities [20]. In our work, we represent the problem of dimensional emotion recognition as a regression one, where each emotional status is represented as three-dimensional real-valued vector. The dimensions of this vector correspond to Valence (corresponding to the concept of polarity), Arousal (degree of calmness or excitement), and Dominance (perceived degree of control over a situation): the VAD model.

In order to solve the regression problem of emotional status detection, we propose to use deep learning. When considering emotion detection from speech Long-Short Term Memory (LSTM) neural networks were tested. The underlying idea is to be able to learn the relationship among present and past information although existing a big distance among them. That is, they have memory and they can manage with temporal sequences of data like the sequence of vectors extracted from an acoustic signal. When regarding emotional status detection from text a classical feedforward network was considered because of simplicity. Such networks have proven to be efficient for problems in similar tasks, like sentiment analysis [21].

3. Experiments

We have carried out two series of experiments for the evaluation of the regression processes. In the first one we present the most interesting results related to the emotion detection from speech, and in the second we show the most interesting results on emotion detection from text.

3.1. Corpus

As far as we know there is no Spanish three-dimensional corpus within the literature, so for the experiments, we have created a small corpus using the VAD model. The corpus consists of 120 fragments between 3 and 5 seconds taken from the Spanish TV program "La Sexta Noche". This TV program consist of political debate, news and events, and discussions commonly appear. Each fragment has been transcribed manually and tagged using crowdsourcing (the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people) techniques. In this case each fragment has been labeled by 5 different annotators, following the next questionnaire:

1. In order to address the Valence: "How do you perceive the speaker?"
   - Excited
   - Slightly Excited
   - Neutral
2. In order to address the Arousal: "His mood is . . . "
   - Positive (nice / constructive)
   - Slightly Positive
   - Slightly Negative
   - Negative (unpleasant / non colaborative)
3. In order to address the Dominance: "How do you perceive the speaker about the situation in which he or she is in?"
   - More dominant / controlling the situation / . . .
   - He or she does not dominate the situation neither is he or she cowed.
   - More coward / defensive / . . .

Once the tags were generated by crowdsourcing, the answers collected from all the annotators were transferred to the three-dimensional model, making the average of each answer

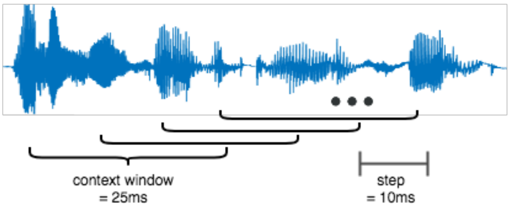


Figure 1: *Schematic diagram of speech production.*

for all fragments where the first answer of each question was assigned the value 0, the last answer was assigned the value 1, and the rest of the answers a midpoint. Then the corpus was split into two sets, 70% of the fragments were used for training purposes and the remaining 30% for test.

3.2. Baselines models and Evaluation Metrics

Both emotion detections problems, from speech and from text, has been tested first with Linear Regression (LR) [22] and Super Vector Regression (SVR) [23] (with three different types of kernels *linear*, *poly*, and *rbf*), in order to compare with Neural Networks.

Regarding the input of these baselines models, two different approaches have been analysed to fix the problem of time sequence. On the one hand, we fit the models with full information, considering each feature in on each time-step independent (*full* models), and on the other hand, calculating the mean of each feature over time-steps (*mean* models).

In relation to the evaluation metric, the Mean Square Error (MSE) has been used, because it seems to provide a good interpretation of how far the prediction and the true label are. In this problem, MSE can be described as the mean of the distances between the points of the true label on the three-dimensional model and the predicted points on the same three-dimensional model.

3.3. Experiments with acoustic features

In order to obtain the acoustic parameters, each audio has been divided into individual frames using a context window of 25 milliseconds and a step of 10 milliseconds (as shown in Figure 1), obtaining 300 frames per audio. A vector made up of the selected acoustic features was associated to each frame. Different experiments were carried out using the different feature sets (A, B, C, D, E, F) described in Section 2. Additionally, for each set, different experiments were also performed including both the first and the first and the second derivatives.

The network proposed to address the regression problem of emotional status with acoustic features is a Recurrent Neural Network (RNN). The network is composed with an LSTM layer (the architecture proposed by [24]) of 10 cell memory blocks, to get a representation of the audio along the time. Subsequent layers are two Dense layers which aim to infer the VAD model from the representation given by the LSTM. The fist Dense layer consist of 15 units and *ReLU* activation function while the second and last consist of 3 units and *sigmoid* activation function.

The output layer contains a sigmoidal activation function to take advantage of the output limitation benefits, it is bounded between 0 and 1. On the other hand, the hidden layer contains the *ReLU* activation function because it provides good results and avoids the vanishing gradient problem [25].

The layers of the proposed network consist of a small number of units or cells not to build a large network architecture, since we have a limited sized training corpus.

Table 1: *Best result obtained with baseline models and Recurrent Neural Network (RNN) in the regression problem of emotional status with acoustic features. Each pair of set and model has been tested with acoustic features itself, with first derivatives and with first and second derivatives, but only best performance is shown. MSE error has been used in order to compare.*

	LR	SVR linear	SVR poly	SVR rbf	RNN
Set A	0.1682	**0.1660**	0.1661	0.1691	0.1670
Set B	0.1686	0.1653	0.1685	0.1710	**0.1565**
Set C	0.1690	0.1679	0.1718	0.1709	**0.1576**
Set D	0.1742	0.1894	0.1703	0.1710	**0.1665**
Set E	0.1699	0.2007	0.1842	0.1711	**0.1664**
Set F	0.1733	0.2256	0.1898	0.1710	**0.1413**

As shown in Table 1, the proposed network slightly improves the results of the baseline models in almost all the sets in the corpus. It can also be concluded that by selecting a smaller set of parameters, better results are obtained with the baseline model. However, the set A seems to have insufficient information and Bark features are of great help in the case of networks providing the best results.

3.4. Experiments with language features

Regarding the word representation, *FastText* embeddings from SBWC[3] has been used in the experiments. The mentioned embeddings are a Skipgram model of 300 dimensions and 855380 different word vectors, trained with Spanish Billion Word Corpus[4] with more than 1.4 billion words.

The regression problem of emotional status with language features has been addressed with a small Deep Neural Network (DNN). This network consist of three similar layers; the first two layers are composed of 5 units, a sigmoidal activation function and followed by Dropout layer with 0.5 of keep-probability in order to prevent to the overfitting problem [26]; while the last layer, the output layer, is a Dense layer of 3 units and the sigmoidal activation function (same as the network proposed for the regression problem of emotional status with acoustic features).

Table 2: *Best result obtained with baseline models and Deep Neural Network (DNN) in the regression problem of emotional status with language features. MSE error has been used.*

	LR	SVR linear	SVR poly	SVR rbf	DNN
Mean	0.1906	0.1350	**0.1165**	0.1197	0.1203
Full	0.1356	0.1292	0.1199	0.1229	**0.1196**

As shown in Table 2, the proposed network achieves similar results when comparing it to the baselines models. It is an interesting result given the small size of the training set and the great

[3] https://github.com/uchile-nlp/spanish-word-embeddings/blob/master/README.md

[4] http://crscardellino.me/SBWCE/

impact it has when building neural networks. The obtained results suggest that increasing the annotated training corpus neural networks might improve the baseline models.

4. Conclusions

The main goal of this work was to develop an automatic emotion detection system from speech and language. The system acted over acoustic fragments extracted from a TV show and their corresponding transcriptions. Each fragment was annotated by means of a crowdsourcing platform using a 3-dimensional VAD model. Different neural networks architectures were tested and the obtained results show that RNN can outperform baseline systems when considering emotion detection from speech. Moreover, using a simple feedforward neural network with a very small training corpus (84 sentences) similar results to those obtained with baseline models can be achieved.

For further work we propose to get a bigger annotated corpus by using crowdsourcing tools to better train the proposed neural networks. Additionally, the two knowledge sources (acoustic and text) might be merged to provide a more accurate emotion detection system.

5. Acknowledgements

This work has been partially founded by by the Spanish Government (TIN2014-54288-C4-4-R and TIN2017-85854-C4-3-R), and by the European Commission H2020 SC1-PM15 program under RIA 7 grant 69872.

6. References

[1] J. Irastorz and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in *7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2016, p. 103106.

[2] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. B. Heinzelman, "Emotion classification: How does an automated system compare to naive human coders?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 2274–2278.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, pp. 1162–1181, 2006.

[4] A. Mencattini, E. Martinelli, F. Ringeval, B. W. Schuller, and C. D. Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Trans. Affective Computing*, vol. 8, no. 3, pp. 314–327, 2017. [Online]. Available: https://doi.org/10.1109/TAFFC.2016.2531664

[5] C. Clavel, G. Adda, F. Cailliau, M. Garnier-Rizet, A. Cavet, G. Chapuis, S. Courcinous, C. Danesi, A.-L. Daquo, M. Deldossi *et al.*, "Spontaneous speech and opinion detection: mining call-centre transcripts," *Language resources and evaluation*, vol. 47, no. 4, pp. 1089–1125, 2013.

[6] S. Mohammad, C. Dunne, and B. Dorr, "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 599–608.

[7] J. R. Bellegarda, "Emotion analysis using latent affective folding and embedding," in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 2010, pp. 1–9.

[8] R. Justo, T. Corcoran, S. M. Lukin, M. Walker, and M. I. Torres, "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web," *Knowledge-Based Systems*, vol. 69, pp. 124–133, 2014.

[9] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[10] S. Volkova and Y. Bachrach, "Inferring perceived demographics from user emotional tone and user-environment emotional contrast," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1567–1578.

[11] R. A. Calvo and S. Mac Kim, "Emotions in text: dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[12] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[13] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[14] S. E. Bou-Ghazale and J. H. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 8, no. 4, pp. 429–442, 2000.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[16] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 1188–1196.

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*. ACL, 2014, pp. 1532–1543.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[19] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. New York, NY, USA: ACM, 2014, pp. 3–10.

[21] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann." *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, 2013.

[22] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[23] C.-C. Chang, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology, 2: 27: 1–27: 27, 2011 "http://www. csie. ntu. edu. tw/˜ cjlin/libsvm"*, vol. 2.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

Publication 4

Iruzurrezko portaeren detekzioa crowd motako etiketazioan

Iruzurrezko portaeren detekzioa *crowd* motako etiketazioan

de Velasco Vázquez, Mikel; López Zorrilla, Asier eta Justo Blanco, Raquel

Euskal Herriko Unibertsitatea UPV/EHU

mikel.develasco@ehu.eus

Laburpena

Lan honek crowd motako etiketazioan agertu daitezkeen kalitate baxuko etiketak detektatzea du helburu. Proposatutako metodologia balioztatzeko, saiakuntzak ataza zail eta subjektibo batekin egin ditugu: emozioen detekzioarekin. Iruzurrezko langileak topatzeko zenbait neurri proposatu dira, etiketatze denboran, langileen arteko adostasunean eta langileen erantzunen banaketan oinarriturikoak. Neurri bakoitza baliagarria dela frogatu dugun arren, gure ondorio nagusia neurriak batzerakoan iruzurrezko langileak detektatzeko probabilitatea handitzen dela da.

Hitz gakoak: Metodo gainbegiratuak, Etiketazioa, Iruzurrezko jokaera, Jendetza, Subjektibotasuna

Abstract

This work aims at detecting low quality labels in crowdsourcing annotation tasks. We validate our proposal carrying out experiments in a difficult and subjective task: emotion recognition. We have developed several measures in order to detect fraudulent behaviour, including measures related to the labelling time, worker inter-agreement and the distribution of the answers. Not only do we show that each of the described measures is helpful but we also demonstrate that mixing them is the best way to go.

Keywords: Supervised learning, Annotation, Fraudulence behaviour, Crowdsourcing, Subjective

1. Sarrera eta motibazioa

Adimen artifizialaren barruan, metodo gainbegiratuek (*supervised methods*), etiketaturiko data kantitate handiak erabili ohi dituzte eredu sendoak sortzeko. Hala ere, etiketatze prozesua, askotan, esfortzu, denbora eta diruaren beharra duen lana da. Tradizionalki, lan espezifikorako heziak izan diren etiketatzaile adituak erabiltzen dira prozesu honetarako, baina halako etiketatzaileak aurkitzea nahiko zaila da, prozesu konplexua eta garestia bihurtuz. Gainera, ataza subjektiboetan, emozioen analisian adibidez, jende askoren iritzia etiketatzaile adituen iritzia baino interesgarriagoa izan daiteke.

Azken urteetan, *crowdsourcing*-a (jendetza bitartez sortutako etiketazioa) etiketazioak egiteko eraginkorragoa den alternatiba berria bilakatu da. Izatez, informazioaren berreskurapenean, hizkuntza naturalaren prozesamenduan, zein beste eremutara lotuta dauden arlo ezberdinetan erabili ohi da. Crowdsourcing plataformen artean, Amazoneko Mechanical Turk[1] edo SamaSource[2] aipa daitezke. Metodologia honek, etiketatze lana etiketatzaile anitz ezberdinekin burutzea ahalbideratzen du, etiketazio lan oso bat mikro-lan askotan zatituz. Gainera kasu askotan langile askoren iritzia kontuan hartzea onuragarria izan daiteke.

Lan honetan, oso subjektiboa den ataza batekin lan egin dugu, audio-hizketa segmentuetan emozioen detekzioan hain zuzen ere. Emozioak anbiguotasunik gabe definitu ezin direnez (Ortony eta Turner, 1990), jende anitzen iritzia oso interesgarria eta baliotsua da. Anbiguotasun horrek iritzien arteko desadostasuna sor dezake, entzulearen pertzepzioarengatik ala ingurune kulturalarengatik sortutakoa adibidez (Gurney, 1884; Scherer, 1999). Hori dela eta, etiketatze prozesua burutzeko metodorik aproposena crowdsourcing-ekin egitea da, ahal bezain beste iritzi kontuan hartzeko.

[1]`www.mturk.com`
[2]`www.samasource.org`

Dena den, ondoren azalduko diren arrazoiengatik, crowdsourcing-ak kalitate baxuko etiketak sor ditzake. Kalitate baxuko etiketen kausa nagusia langileen jarreraren ondorioa da. Horregatik, lan hau jarrera desegoki bat duten langileak detektatzera bideratuta dago. Iruzurrezko langilak topatzeko zenbait neurri proposatzen ditugu, etiketatze denboran, langileen arteko adostasunean eta langileen erantzunen banaketan oinarriturikoak.

2. Arloko egoera eta ikerketaren helburuak

Crowdsourcing-a nahiko hedatuta egon arren eta datu bilketa zein datu-etiketak balioztatzeko baliagarria izan arren, metodo honekin lortutako emaitzen kalitatea oraindik dudan dago. Izan ere, komunitateak arlo honetan ahalegin handiak egin arren (Eickhoff eta de Vries, 2011; Gennaro *et al.*, 2010; Gadiraju *et al.*, 2015), datuen fidagarritasuna neurtzea ez da bat ere erraza. Eickhoff autoreen (2013) lanaren arabera hainbat langile desegoki aurki daitezke: gutxiengo baldintzak betetzen ez dituzten langile ezgaituak alde batetik, erantzun okerrak emanez esperimentua baliogabetzen saiatzen diren langile maltzurrak bestetik, eta lanari beharrezko aditasuna jartzen ez dioten langile adigabeak azkenik.

Fidagarriak ez diren langileak antzemateko biderik arruntena *gold-standard* deritzon aurredefinitutako mikrolanak mikro-lan arruntekin nahastea da. Era honetan, erantzun ezaguneko galdera bat oker erantzuten duten etiketatzaileak baztertzen dira. Metodo hau lan askotarako fidagarria izan arren, ez da baliozkoa kasu guztietan, galdera irekiak duten lanetarako, adibidez. Era berean, langile maltzurrak gold-standard lana sahiesteko era berriak bilatu ditzakete, kontu berrietan informazioa bererabiltzeko testen galderak ikasiz (Rothwell *et al.*, 2015) besteak beste. Gainera, ataza subjektiboetan, gold-standard metodoak ez dauka zentzu askorik, anbiguotasun gabeko kasuak aukeratzea, berez, ariketa subjektiboa delako. Adibidez gure ataza eta audioaren transkripzioa erabat desberdiak dira. Audioaren transkripzioan gertatzen ez den bezala, gurean ezin da objektiboki jakin zein den egiazko etiketa.

Gold-standard metodologia desegokia den kasuetarako beste irtenbide batzuk garatu dira. Adibidez, Filatova autoreek (2012) lanean ironia eta sarkasmoa etiketatzean, gehiengoaren botazioan oinarritutako algoritmo bat erabili zuten kalitatea kontrolatzeko (Ipeirotis *et al.*, 2010). Dena den, gehiengoaren botazioa erabiltzeko eta kalitate kontrol ona ziurtatzeko, mikro-lan bakoitzeko etiketa ugari behar dira eta horrek etiketazio lana garestitzen du. Ipeirotis *et al.* autoreen (2010) lana etiketatzaileen arteko adostasunean oinarrituta dago, eta Dawid eta Skene autoreen (1979) lanean inspiratuta dagoen itxaropen-maximizatze algoritmo bat erabiltzen du. Hala eta guztiz ere, etiketatzaileen arteko adostasunak huts egin dezake langile maltzurrak detektatzerakoan, sistema automatikoen bidez erantzunen atzean dagoen banaketa ikasi dezaketelako. Arazo honi irtenbidea emateko, Raquel Justo autoreek (2017) laneko errore-tasa kalkuluan oinarritzen den algoritmo bat proposatzen dute.

Etiketatzaileen arteko adostasun neurketen artean bateratzeko lanak egon arren (Klaus, 2011), neurri desberdinen artean nahasmen handia dago. Dakigunaren arabera, ez dago ataza subjektiboko datuen kalitatea neurtzeko fidagarria den metodorik. Hori dela eta, lan honetan zehar iruzurrezko langileen detekziorako hainbat neurri proposatzen ditugu eta euren konbinazio lineala erabat ereaginkorra dela erakusten dugu.

3. Iruzurrezko langileen detekzioa

Atal honetan, hasteko iruzurrezko langileak topatzera bideratuta dauden neurriak azaltzen ditugu. Geroago, neurri hauek guztiak zein testuingurutan balioztatu diren azalduko dugu. Azkenik, saiakuntzetan lortutako emaitzak erakusten ditugu.

3.1. Data-kalitatearen neurketa

Crowdsourcing teknikak erabiliz lortutako etiketazioen osteko analisia egiteko, denboran oinarritutako neurketa simpleetatik etiketatzaileak beraien artean duten adosmena neurtzen duten neurrietara arteko neurrriak aztertu genituen. Neurri hauek adierazteko, hau da lanean zehar erabiliko dugun notazioa:

- **Etiketatzaile bat** edo **langile bat** adierazteko, l indizea erabiliko da. Adb. n_l, l langileak egindako mikrolanen kopurua izango zen.
- **Mikro-lan** bakoitza adierazteko, s indizea erabiliko da. Adb. n_s, mikro-lan bakoitzeko erantzunen kopurua izango zen. Gure ikerketaren barruan, segmentu bakoitza 5 aldiz erantzun da.
- **Galdera bakoitzaren erantzun posibleak** adierazteko, e indizea erabiliko da. Adb. n_e, e erantzun posiblea zenbat aldiz hautatu den.

3.1.1. Etiketatze denboran oinarritutako neurketak

Denborarekin erlazionatuta dauden bi informazio desberdin erabili ditugu. Alde batetik, langile bakoitzak etiketatzen eman duen gehiengo denbora (T_l), eta bestetik, mirko-lan bakoitzaren etiketa sortzeko behar izan duen denbora (t_e), hau da etiketa bakoitza sortzeko erabilitako denbora. Datu hauekin, langile bakoitzeko hiru neurketa desberdin kalkulatzen ditugu:

1. Lanaldian emandako denborarik luzeena (T_l). Neurri honek mikro-lanen artean 10 minutuko etenaldirik gabe lan egin duen gehienengo denbora adierazten du.
2. Mikro-lan bat burutzeko erabilitako bataz besteko denbora, (1) formulan azaltzen den bezala.

$$\overline{t_l} = \frac{\sum t_{ls}}{n_l} \tag{1}$$

3. Mikro-lan bat burutzeko erabilitako denboraren desbiderapen estandarra, (2) formulan azaltzen den bezala.

$$\sigma(t_l) = \sqrt{\frac{1}{n_l}\sum(t_{ls} - \overline{t_l})^2} \tag{2}$$

3.1.2. Etiketatzaileen arteko adostasunean oinarritutako neurketak

Crowdsourcing teknikak erabiliz ateratako datuen adostasuna kalkulatzeko hainbat lan egin dira, (Bennet *et al.*, 1954; Scott, 1955; Cohen, 1960; Krippendorff, 2004) adibidez. Neurri hauek etiketazio guztiak batera duten adostasuna neurtzen dute. Guk langileak deskribatzeko neurriak bilatzen ditugunez, adostasun neurri hauek ez dira zuzenenan aplikagarriak. Raquel Justo autoreen (2017) lana jarraituz langile mailako neurri bat definituko dugu adostatsun neurri batetik abiatuz.

Modu honetan, (3) ekuazioan agertzen den bezala langile mailako adostasun neurria kalkulatzen da, langile guztiekin kalkulatutako adostasuna (A) eta langile bat gabe kalkulatutako adostatuna (A_l) erabiliz.

$$\Delta A_l = \frac{A - A_l}{n_l} \tag{3}$$

Bi langileen arteko adostasuna neurtzeko modurik sinpleena ehuneko adostasuna ala antzemandako adostasuna da, (4) ekuazioan ikusten den bezala.

$$A_o = \frac{1}{n_I}\sum_{i\in I} agr(s_{l_1 i}, s_{l_2 i}) \tag{4}$$

non I bi langile etiketaturiko mikro-lanen arteko ebakidura ($S_{l1} \cap S_{l2}$) den; n_I bi langileak etiketaturiko mirko-lan kopurua; eta $agr(s_{l_1 i}, s_{l_2 i})$, l_1 langileak i mikro-lanerako emandako emaitzaren eta l_2 langileak i mikro-lan berdinerako emandako emaitzaren adostasun funtzioaren emaitza.

Galdera ezberdinetarako, adostasun funtzio ezberdinak izan ditzakegu, erantzun posibleen artean dauden erlazioen arabera. Adibiderik errezena hurrengoa da:

$$agr_{i,k} = \begin{cases} 1: \text{bi langileek etiketa bera jartzen badute} \\ 0: \text{bi langileek etiketa ezberdina jartzen badute} \end{cases} \tag{5}$$

Hala ere, erantzun posibleak ordenaren bat jarraitzen badute, adostasun funtzio egokiago bat sor dezakegu, gehiengo adostasuna 1 eta gutxiengo adostasuna 0 izanik. Adibidez, hiru erantzun posibleak eskala batean irudikatu ahal badira, “Asko”, “Gutxi” eta “Ezer ez” esate baterako, funtzio egoki bat hurrengoa liteke (6):

$$agr_{i,k} = \begin{cases} 1 & : \text{bi langileak etiketa bera jartzen badute} \\ 0.5 & : \text{etiketa bakar bat “Gutxi” bada} \\ 0 & : \text{beste kasuetan} \end{cases} \tag{6}$$

Antzemandako adostatsuna literaturan aurkezten den adostasun neurri hedatuena izan arren (Artstein eta Poesio, 2008), ez du adostasunak ausaz gertatu daitezkeela kontuan hartzen. Hori dela eta, antzemandako adostasuna ausazko adostasunarekin egokitu ahal da, 7 ekuazioan ageri den Krippendorff autoreek (2004) lanean azalduriko α neurria bezala.

$$\alpha = \frac{A_o - A_e}{1 - A_e} \tag{7}$$

$$A_e = \sum_{i \in I} P(i|l_1) \cdot P(i|l_2) \tag{8}$$

non A_e itxarondako ausazko adostasuna den. Itxarondako adostasuna, (8) ekuazioari jarraituz, ausaz erantzuten duten bi langileen artean (l_1 eta l_2) dagoen adostasuna neurtzen du, I bi langileek etiketaturiko mikro-lan berdinen multzoa izanik.

Krippendorff autoreen (2004) lanean ageri den moduan, α koefizientea (9) formularen arabera kalkulatu dezakegu. (10) formulan o_{e_c,e_k} balioak kalkulatzeko formuaren bidez, e_c eta e_k erantzun posibleen artean antzemandako desadostasuna kalkulatzen da. Bestalde, (11) formulan e_c erantzun posiblerako antzemandako desadostasuna kalkulatzen da. (9) formulan ikusten den agr_{e_c,e_k}-rako, (5) eta (6) adosmen funtzioak erabil daitezke, kasuaren arabera.

$$\alpha = 1 - (\sum_{c \in E} o_{e_c} - 1) * \sum_{c \in E} \sum_{k \in E} \frac{o_{e_c,e_k} * agr_{e_c,e_k}}{o_{e_c} * o_{e_k} * agr_{e_c,e_k}} \tag{9}$$

$$o_{e_c,e_k} = \sum_{s \in S} \frac{n_{e_c s} * n_{e_k s}}{n_s - 1} \tag{10}$$

$$o_{e_c} = \sum_{k \in E} o_{e_c,e_k} \tag{11}$$

Krippendorff autoreek (2004) lanean azalduriko α nola kalkulatzen den jakin eta gero, aurretik aipatutako ΔA_l kalkulatzen dugu α balioarekin ($\Delta\alpha_l = \frac{\alpha - \alpha_l}{n_l}$). $\Delta\alpha_l$-ren balio negatiboak adostasunaren hobekuntza adierazten du. Beraz $\Delta\alpha_l$ zenbat eta txikiagoa izan orduan eta handiagoa izango da adostasuna.

$\Delta\alpha_l$-k langile baten eragina neurtzen du, hala ere, beste adostasun eta elementu batzuk baita neurrian eragina dute. Hori dela eta, Raquel Justo autoreek (2017) lanean proposaturiko β_l neurria erabili dugu ere. β_l neurriak langile baten eta mikro-lan berberak etiketaturko beste langileen arteko antzemandako adostasuna neurtzen du, (12) formulan agertzen den bezala.

$$\beta_l = \frac{1}{|S_l|} \sum_{s \in S_l} \sum_{m \in L_s - l} \frac{agr_{m,l}}{|L_s| - 1} \tag{12}$$

non S_l, l langileak egindako mikro-lan multzoa den; L_s, s mikro-lana egin duten langileak eta $agr_{m,l}$, (5) zein (6) adostasun funtzioa izan daiteke, kasuaren arabera.

3.1.3. Erantzunen banaketetan oinarritutako neurketak

Beste alde batetik, oso interesgarria izan daiteke, ausazko erantzunak zein erantzun bakarra erantzuten duten langileak detektatzea. Horretarako langile bakoitzak etiketaturiko datuen banaketa (b_l), langile guztiek batera etiketaturiko datuen banaketarekin (B) konparatzen duen metodo bat proposatzen dugu. Metodo honek banaketen distantzia neurtzen du (13).

$$d = \text{distantzia}(b_l, B) \tag{13}$$

Gure lanean bi distantzia mota desberdinekin lan egin dugu. d_1 distantzia euklidear arrunta da (14) eta d_2 distantziak probabilitate banaketan magnitude orden desberdineko elementuak egotea gehiago zigortzen du (15).

$$d_1(b_l, B) = \sum_{e \in E} (\mathbf{l}_e - \mathbf{B}_e)^2 \quad (14) \qquad d_2(b_l, B) = \sum_{e \in E} \left(\frac{\max(\mathbf{l}_e, \mathbf{B}_e)}{\min(\mathbf{l}_e, \mathbf{B}_e)} - 1 \right) \quad (15)$$

non b_l langilearen probabilitate banaketa den eta $\mathbf{B}$ langile gusztien probabilitate banaketa den, $\mathbf{l}_e$, eta $\mathbf{B}_e$ haien probabilitate banaketaren elementu bat izanik.

3.2. Ataza definizioa eta etiketatze prozesua

Aurreko atalean azaldutako neurri guztiak emozioak etiketatzeko lan batean aztertu nahi ditugu. Horretarako, crowdsourcing teknikak erabili dira "La Sexta Noche" corpusa emozioz etiketatzeko. "La Sexta Noche" telebista sailean emozio naturalak agertzen direnez, nahiko arraroa da mutur-emozioak aurkitzea, zalantzazko etiketak sortuz. Beste alde batetik, sei orduko telesaiak direnez, emozio bakarra aurkitu daitekeen tarteetan segmentatu behar dira. Segmentu horien luzera emozio bat adierazteko bezain luzeak eta emozioa aldaketa bat ez agertzeko bezain laburrak izan behar dute. Hortaz, "La Sexta Noche" corpusa 3 eta 5 segundu arteko segmentuetan banatu dugu, crowdsourcing-en bidez etiketa emozionala jartzeko.

Banatu dugun segmentu bakoitzaren emozioa adierazteko hainbat eredu aurki daitezke. Alde batetik emozioen eredu kategorikoa daukagu (Ekman, 1992), eta bestetik eredu dimentzionala (Gunes eta Pantic, 2010; Russell eta Mehrabian, 1977). "La Sexta Noche" telesailaren mintzagaien artean, politika, gertaera eta berriak zein eztabaidak maiz agertzen direla kontuan hartuz eta esperimentu txiki batzuk aurre-eginez, 10 emozio ezberdineko bilduma sortu genuen corpus hau etiketatzeko. Sortutako bildumarekin eredu kategoriko eta beste galdera batzuekin eredu dimentzionalako galdetegia sortu genuen:

1. Nola hantzematen duzu hizlaria?
   - Aztoratuta
   - Zertxobait aztoratuta
   - Neutrala
2. Bere gogo egoera hurrengoa da:
   - Positiboa (atsegina / eraikitzailea)
   - Zertxobait positiboa
   - Neutrala
   - Zertxobait negatiboa
   - Negatiboa (desatsegina / kolaboragaitza)
3. Nola ikusten duzu hizlaria egoera horretan?
   - Egoera kontrolatzen
   - Ez du egoera kontrolatzen ezta jarrera defentziboan egoten
   - Egoera defensiboan
4. Aukeratu hizlariaren emozioa hoberen deskribatzen duen aukera:
   - Lasaia / Axolagabe
   - Azpertuta / Nekatuta
   - Hunkituta
   - Harrituta
   - Alaia
   - Interesdun
   - Kezkatuta
   - Lotsatia
   - Haserre
   - Tentsoa

Galdetegia definitu eta gero, ikerketa taldearen baliabideen artean dagoen CrowdZienzia[3] crowdsorcing plataformarekin (Justo *et al.*, 2016) etiketatu dugu. Plataforma hau erabiltzeko arrazoien artean: 1) Amazoneko Mechanical Turk plataforma bakarrik Estatu Batuetan dagoela erabilgarri; 2) beste plataforma batzuek, SamaSource bezala, gaztelania mintzatzen dutenen artean %20 baino gutxiago espainarrak direla, eta ataza honetarako Espainian jaio diren pertsonak baliagarriagoak dira Amerikako gaztelainarekin dagoen diferentziagatik; eta 3) gure arteko etiketatze probak egiteko etiketatzaile finko batzuk edukitzea ahalbidetzen digula daude.

[3]Komunitate zientifikorako erabilgarri, muga zehatz batzuen artean. `https://crowdzientzia.ehu.eus`

3.3. Saiakuntza eta emaitzak

Aurreko ataletan azaldutako neurriak kontuan hartuta, CrowdZientzia plataformarekin sortutako datuak aztertuko ditugu. Datu hauek corpusetik auzaz aukeratutako 5.500 segmentuetatik datoz eta etiketak sondoak izateko segmentu bakoitza 5 aldiz etiketatzea nahikoa zela erabaki genuen, 27.500 mikro-lan sortuz. Mikro-lan guztiak egiteko 129 langilek parte hartu zuten, baina langile guztiek ez zuten lanaren proportzio bera etiketatu. Gutxien etiketatu zuen langileak mikro-lan bakarra etiketatu zuen, 2081 mikro-lan etiketatu zituen lan gehien egin zuen langileak, eta batez bestekoa 213 mikro-lan izan ziren. Langileei Amazoneko txekeekin ordaintzea erabaki genuen.

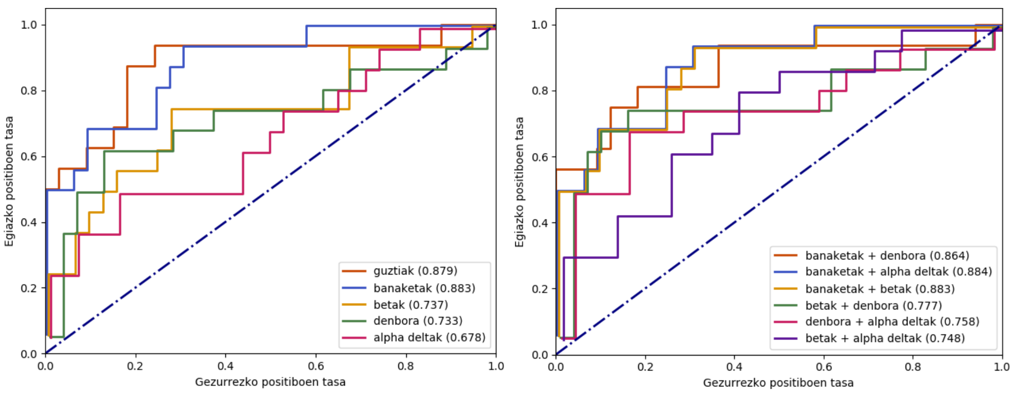
3.1 ataleko neurriak konparatzeko, aztertutako neurriren batean 20 langile txarren artean zeuden langileak hartu genituen (45 langile, langile gehienek neurri askotan txarrenetarikoen artean zeudelako) eta eskuz klasifikatu genituen iruzurrezko langile ala langile egoki bezala. Honetarako, aipatutako langile bakotzaren lana bi pertsonen artean aztertu eta iruzurrezko langilea ala langile egoki bezala sailkatzea eztabaidatu genuen. Ikerketa honetan, 6 iruzurrezko langile eta zalantzazko 10 iruzurrezko langile aurkitu genituen (3.2 atalean azaldutako galdetegiaren galdera batzuk txarto egiten zutela dirudielako).

Ondoren, aztertutako 45 langileak (6 iruzurrezko langile eta 39 langile egoki) zeinbait sailkatzaile linealekin doitu genituen. Sailkatzaileak neurri multzo ezberdinekin doitzea probatu genuen. Hau da, neurriak lau multzo ezberdinetan banatu genituen, multzo bakoitzean 3.2 atalean azaldutako galdera bakoitzeko eta 3.1.1, 3.1.2 eta 3.1.3 ataletan azaldutako neurri mota ezberdinetarako neurri bat jarri genuelarik. Lehenengo, bigarren eta hirugarren galderako neurrientzat (6) funtzioa erabili dugu eta laugarren galderarako (5) funtzioaren moldapen bat.

- Alde batetik **denbora**-multzoa, 3.1.1 atalean azaldu ditugun denbora luzeena T_l, bataz besteko denbora $\overline{t_l}$ eta erabilitako denboraren desbiderapen estandar $\sigma(t_l)$ neurriekin.
- Bestetik, 3.1.2 atalean, kalkulaturiko $\Delta\alpha_l$ balioa erabiliko dugu **alpha deltak**-multzorako.
- 3.1.2 ataletik ere, β_l neurria erabili da **betak**-multzoa sortzeko.
- Azkenik, 3.1.3 atalean azaldutako d_1 eta d_2 distantziak erabili dira **banaketak**-multzoa sortzeko.

Sailkatzaileen errendimendua neurtzeko, ROC kurba[4] eta ROC kurbaren azpialdeko azalera kontuan hartu dira.

1. irudia. Neurri multzo ezberdinekin ateratako ROC kurba



(a) Banaka (b) Binaka

1a irudiari begiratuz, ROC kurbaren azpialdeko azalerarik handiena lortutako multzoa *banaketak* multzoa izan da, 0.883 balioarekin. Beraren azpitik *betak* (0.737), *denbora* (0.733) eta azkenik *alpha deltak* (0.678) multzoak daude. Dena den, multzo guztiak batuz, marka hobetzen dugula konprobatu dezakegu, *guztiak* multzoa 0.879 markarekin. 1b irudian berriz, sortutako multzoak binaka batu ditugu, multzo berriak bakarka baino hobeagoak direla konprobatzeko.

[4]ROC kurba, erantzun egokien eta alarma faltsuen ehunekoen erlazioa irudikatzen duen erlazioa da.

4. Ondorioak

Lan honek aztertutako neurrien baliagarritasuna aztertzea du helburu, horregatik ez da sailkatzaile on bat doitzera denbora gehiegi eman.

Burututako lanaren arabera, erabilitako denbora neurriak, adostasun neurriak eta erantzunen banaketa neurriak, langileen iruzurrezko portaera antzemateko baliagarriak direla ondorioztatu dezakegu, neurri guztiak batera ROC kurbaren azpialdeko azalera handitzeaz gain, edozein neurri mutzo ezberdin batzerakoan bakarka baino ROC azalera handiagoa lortzen delako.

Honen salbuespen bakarra banaketak gehi betak multzoa da, banaketak bakarrik erabiltzean lortzen den marka berdina lortzen delako.

Neurrien interesa alde batera utzita, oso subjektibodun atazarekin lan egiterakoan crowd bidezko etiketazioak baliogarriak direla konturatu gara, etiketatze lan osteko iruzur detektatze lanak egin behar izan arren. Denbora tarte txiki batean mikro-lan asko bete daitezkelako, adostasun onargarri bat lortuz. Hala ere, deskribatutako neurrietaz lagunduta, detektatutako iruzurrezko langileak egindako lana ezabatzerakoan, etiketen arteko adostasuna igotzen dela konprobatu dugu, eta horren ondorioz, kalitate hobeagoko corpus bat sortu daiteke.

5. Etorkizunerako planteatzen den norabidea

1a eta 1b irudiei begira 4. ataleko ondorioak atera ditugu, baina hau izan daitekenaren susmo bakar bat da. Azken finean lan honetako datuak erabili ditugu bakarrik eta aztertutako neurriak bakarrik kasu honetan horrela erantzutea gerta daiteke. Horregatik, 3.1 atalean azaldutako neurri guztien azterketa sakon bat etorkizunean egiteko asmoa dugu, objektibodun zein subjektibodun etiketazio lan ezberdinekin. Baina subjektibodun lanetan arreta handiagoa jartzea espero dugu, gutxiago aztertu den eremu zail bat delako.

Beste alde batetik, iruzurrezko langileak detektatzea lortu arren, ezin da kalitatezko corpus bat sortu ahal dela ziurtatu. Hori dela eta, ateratako datuekin eredu konputazionalak sortzea ikerketa bide berri bat izango litzateke. Hala eta guztiz ere, corpusa sortzerakoan kontuan izan diren langile kopurua aztertzeko susmoa dugu. Iruzurrezko langileak kentzerakoan datuen arteko adostasuna igotzen den arren, etiketa gutxiago lortzen dira. Horregatik datu kopuraren eta datu kalitatearen arteko oreka lortu behar da.

Azkenik, azaldu ez den arren, gero eta gehien kolaboratu duten langileak gero eta iruzurrezko langileak izateko probabilitatea daukatela dirudi. Gertaera hau etiketatzea ataza neketsua delako izan daiteke, eta etiketa kopuru batetik pasatzerakoan langileek erantzun bera aukeratzerako joera dutela antza ematen du. Hori dela eta, hurrengo saiakeretan langile bakoitzari mikro-lan kopuru maximo bat egiteko ahalmena ezarriko diegu, nekatzeko denboraren probabilitatea murrizteko.

6. Erreferentziak

Artstein, Ron, eta Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34.555–596.

Bennet, E. M., R. Alpert, eta A. C. Goldstein. 1954. Communications through limited response questioning. *Public Opinion Quarterly* 18.303–308.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70.213–220.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20.37–46.

Davies, M., eta J. L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics* 38.1047–1051.

Dawid, A. P., eta A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* 28.20–28.

Dhall, Abhinav, Roland Goecke, Simon Lucey, eta Tom Gedeon. 2011. Acted facial expressions in the wild database.

Eickhoff, Carsten de Vries, Arjen P. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16.121–137.

Eickhoff, Carsten, eta Arjen P. de Vries. 2011. How crowdsourcable is your task? In *Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, Hong Kong, China.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition & emotion* 6.169–200.

Filatova, Elena. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proc. of LREC 2012, Istanbul, Turkey, May 23-25, 2012*, 392–398.

Fleiss, J.L., eta others. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76.378–382.

Gadiraju, Ujwal, Ricardo Kawase, Stefan Dietze, eta Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the ACM CHI 2015, Seoul, Republic of Korea*, 1631–1640.

Gennaro, Rosario, Craig Gentry, eta Bryan Parno. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Proc. of CRYPTO'10, Santa Barbara, CA, USA*, 465–482.

Gunes, Hatice, eta Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 1.68–99.

Gurney, Edmund. 1884. What is an emotion? *Mind* 9.421–426.

Ipeirotis, Panagiotis G., Foster Provost, eta Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proc. of the ACM SIGKDD*, 64–67, New York, USA.

Justo, Raquel, José M. Alcaide, eta M. Inés Torres. 2016. Crowdscience: Crowdsourcing for research and development. In *Proc. of IberSpeech'2016, Portugal*, 403–410.

Klaus, Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Krippendorff, Klaus. 2004. *Content analysis: An introd. to its methodology*. Sage.

——. 2007. Computing Krippendorff's Alpha Reliability. Technical report, University of Pennsylvania, Annenberg School for Communication.

Mohammad, Saif M, eta Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29.436–465.

Ortony, Andrew, eta Terence J Turner. 1990. What's basic about basic emotions? *Psychological review* 97.315.

Raquel Justo, M Inés Torres, José M Alcaide. 2017. Measuring the quality of annotations for a subjective crowdsourcing task. In *Iberian Conference on Pattern Recognition and Image Analysis*, 58–68. Springer.

Rothwell, Spencer, Ahmad Elshenawy, Steele Carter, Daniela iraga, Faraz Romani, Michael Kennewick, eta Bob Kennewick. 2015. Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proc. of Interspeech 2015, Dresden, Germany, September 6-10, 2015*, 2784–2788.

Russell, James A, eta Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality* 11.273–294.

Scherer, Klaus R. 1999. Appraisal theory. *Handbook of cognition and emotion* 637–663.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19.321–325.

Tarasov, Alexey, Sarah Jane Delany, eta Charlie Cullen. 2010. Using crowdsourcing for labelling emotional speech assets.

Valstar, Michel, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, eta Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 3–10. ACM.

7. Eskerrak eta oharrak

Egileok gure esker ona adierazi nahiko genioke Euskal Herriko Unibertsitateari, Espainako gobernuako TIN2017-85854-C4-3-R zenbakidun diru laguntzari eta H2020 Europako Batzordeko SC1-PM15 programako RIA 7 deialdiko 769872 zenbakidun laguntzari, hurrenez hurren, ikerketa hau babesteagatik.

Publication 5

Euskaraz hitz egiten ikasten duten makina autodidaktak

Euskaraz hitz egiten ikasten duten makina autodidaktak

López Zorrilla, Asier; deVelasco Vázquez, Mikel eta Justo Blanco, Raquel

Euskal Herriko Unibertsitatea UPV/EHU

asier.lopezz@ehu.eus

Laburpena

Lan honetan sare neuronalen bidez euskaraz hitz egiten ikasten duen elkarrizketa sistema automatikoa aurkezten dugu. Horretarako, Turingen testaren ideia era konputazionalean inplementatzen duten sare neuronal sortzaile aurkariak erabili ditugu. Normalean erabiltzen diren ingelesezko corpusak baino bi magnitude ordena txikiagoa den euskarazko corpus batekin halako sareak entrenatzea badagoela frogatzen dugu. Amaitzeko, euskararen morfologia kontuan hartzen duen aurreprozesamendua erabiltzea komenigarria dela erakusten dugu. Dakigunaren arabera, sare neuronaletan oinarrituta dagoen euskarazko lehen elkarrizketa sistema aurkezten dugu.

Hitz gakoak: elkarrizketa sistema automatikoak, sare neuronalak, sare neuronal sortzaile aurkariak, euskara

Abstract

This work presents a neural dialogue system capable of learning Basque. To this end, we build upon generative adversarial networks which implement the idea of the Turing test. We demonstrate that training such a dialogue system with corpora two orders of magnitude smaller than usual English corpora is feasible. Finally, we also found that preprocessing the Basque language according to its morphology helps training these neural models. To the best of our knowledge, this is the first attempt to develop a neural dialogue system in Basque.

Keywords: dialogue systems, deep learning, generative adversarial networks, Basque

1. Sarrera eta motibazioa

Elkarrizketa sistema automatikoek pertsona eta makinen arteko komunikazioa eta interakzioa ahalbidetzen dute, lengoaia naturalaren bidez. Oro har, bi motatako elkarrizketa sistema desberdintzen dira: helburudunak eta helbururik gabekoak edo eremu irekikoak. Lehenengo kategorian erabiltzailearen nahi espezifikoak asetzeko eraikitako sistemak sartzen dira. Adibidez, busen ordutegiak eta lineak kontsultatzeko (Olaso Fernández eta Torres, 2017) eta jatetxeetan edo hoteletan erreserbak egiteko (Bordes eta Weston, 2016) balio duten sistemak helburudunak dira. Hauetaz gain, azken urteotan hedatu diren laguntzaile birtualak, hala nola Siri, Cortana, Google Assistant edo Alexa, elkarrizketa sistema helburuduntzat ere har ditzakegu, normalean euren lana erabiltzailearen aginduak burutzea baita, esate baterako dei bat egitea edo Interneten biharko eguraldiaren iragarpena bilatzea.

Beste aldetik, eremu irekiko elkarrizketetan ez dago aldez aurretik definitutako helbururik ezta gairik. Hau da, erabiltzaileak eta makinak ez diote elkarri hitz egiten helburu espezifiko batekin; interakzioa bera naturala eta zentzuduna izatea da helburua. Horretarako, sistemak esaldi ahal bezain logiko, koherente eta informatzaileekin erantzun behar dio erabiltzaileak esaten duenari. Beste modu batean esanda, sistemak era gizatiarrean hitz egin behar du. Lan honetan elkarrizketa sistema mota horietan zentratuko gara.

Era gizatiarrean hitz egitearen ideiarekin lotuta, Alan Turing matematikariak 1950. urtean bere test famatua aurkeztu zuen: Turingen testa (Turing, 1950). Testaren ideia nagusia honakoa da: sistema automatiko bat kalitatezkoa edo adimenduna dela esateko, sistema hori eta pertsona bat elkar bereizezinak izan behar dute haiekin hitz egiterako orduan. Sistema batek halako propietatea betetzen duen egiaztatzeko, Turingek hainbat epaile zenbait makinekin hitz egiten jartzea proposatu zuen, makina batzuen atzean sistema automatikoak eta besteen atzean pertsonak daudelarik. Egoera horretan epaileek ehuneko altu[1] batean usteko balute sistema automatikoa pertsona bat

[1]Eztabaida handia dago sistema batek Turingen testa gainditzeko behar duen portzentajearen inguruan. Erreferentzia gisa, 2011. urtean Indian Institute of Technology Guwahati institutuan ospatutako Turingen test batean, epaileek pertsonak pertsona moduan sailkatu zituzten ebaluazioen % 63,3-an.

dela, orduan sistema hori erabat adimenduna dela esan liteke.

Denbora pasa ahala, Turingen testa gainditzearen ideiak gero eta ikerketa gehiago bultzatu zituen adimen artifizialaren arloan. Adibidez, 1966. urtean ELIZA programa (Weizenbaum, 1966) aurkeztu zuten MIT-eko ikerlariek. Programaren funtsa hitz gakoak detektatzean eta horien arabera aurredefinitutako esaldi bat aukeratzean datza. Algoritmo hori sinplea izan arren, hainbat epailek pertsonatzat hartzea lortu zuen.

Hurrengo hamarkadetan ikerketek aurrera jarraitu zuten arren, benetan Turingen testa gainditzeko gai zen sistemarik ez zen lortu. 2011. urtean Turingen test batean inoiz lortu diren emaitzarik onenak Cleverbot sistemak[2] lortu zituen, berarekin hitz egin zuten 1.334 epaileetatik % 59,3-ak pertsonatzat hartu zuenean. ELIZA-k ez bezala, Cleverbot-ek ez ditu aurredefinitutako esaldiak erabiltzen. Horren ordez urteetan zehar pertsonekin edukitako elkarrizketak erabiltzen ditu erantzuterako orduan. Hitz gutxiekin esanda, esaldi bati erantzuteko Cleverbot-ek esaldi hori edo antzeko bat esan duenean zein erantzun jaso duen bilatzen du, eta erantzun horretaz abiatuz sortzen du bere erantzuna. Ideia hau interesagarria bada ere, konputazionalki nahiko konplexua da, denbora zein memoriaren ikuspegitik, datu-base oso handi batean bilaketak egitea baitakar. Are gehiago, datu-basea gero eta handiagoa izan, orduan eta denbora eta memoria gehiago beharko du halako sistema batek erantzun bat lortzeko.

Eragozpen horiek saihesteko, baita adimen artifizialaren beste arloetan izan duten emaitzengatik, azken urteetan sare neuronalak elkarrizketa sistemak eraikitzeko teknologia nagusia bilakatu dira. Sare neuronalak datuetatik eredu konputazional konplexuak lortzeko balio duten paradigma konputazional bat dira, bereziki eraginkorra datuen kantitatea oso handia denean. Ulertzekoa da, beraz, arloko autore gehienek ingelesez dauden datu-baseekin lan egitea, normalean hauek baitira handienak, eta hortaz sare neuronalak hobeto funtzionatuko dutelako. Baina, zer gertatzen da baliabide gutxiagoko hizkuntzekin? Ba al dago sare neuronaletan oinarrituriko elkarrizketa sistema automatikoak eraikitzerik euskaraz?

Lan honetan erakusten dugu baietz, badagoela. Normalean erabiltzen diren datu-baseak baino bi magnitude ordena txikiagoak diren datu-baseak erabiliz modu koherente eta zentzudunean euskaraz hitz egiten duen elkarrizketa sistema automatikoa aurkezten dugu.

2. Arloko egoera eta ikerketaren helburuak

Sare neuronalen bidezko eremu irekiko elkarrizketa sistemak itzulpen automatikorako erabiltzen diren sareetan oinarritzen dira, hots, sekuentziatik-sekuentziarako sareetan (Sutskever *et al.*, 2014; Cho *et al.*, 2014) (*Sequence to sequence networks* ingelesez). Sare neuronal horiek luzera arbitrarioko bektore segida bat har dezakete sarrera moduan, eta era berean beste luzera arbitrarioko segida bat sortu. Hala, eremu irekiko elkarrizketak sortzearen problema transdukzio problema bat bezala planteatzen badugu, sare horiek erabili ditzakegu. Hori egiteko, sarrera erabiltzaileak esandako hitzen segida izango da, eta irteera sistemaren erantzunari dagozkion hitzen sekuentzia.

Sare horiek entrenatzeko, edo euren parametroak doitzeko, ikasketa metodo gainbegiratuak erabili ohi dira, aipatutako sarrera-irteera bikoteez osaturiko corpusen bat erabiliz. Adibidez, lan honetan filmen azpitituluak erabiliko ditugu corpus hau eratzeko: sarrera bakoitza aktore batek esandako esaldi bat izango da, eta dagokion irteera beste aktore batek emandako erantzuna. Metodologia hau erabiliz emaitza interesgarriak lortu ahal diren arren, askotan horrela entrenatutako sareek informaziorik gabeko erantzun orokorrak sortzeko joera dute, hala nola *I don't know* edo *I'm sorry*[3] (Sordoni *et al.*, 2015; Serban *et al.*, 2016). Tuan eta Lee autoreek (2019) adierazten duten moduan, ikasketa metodo gainbegiratuek irteera bakarra esleitzen diote sarrera bakoitzari, baina horrek ez ditu elkarrizketen propietateak behar bezala jasotzen. Izatez, hitz egiten dugunean, norbaitek esan duenari erantzuteko hamaika esaldi ezberdin erabili ahalko genituzke, guztiak onargarriak. Horrela, esaldi askoren erantzuna izan daitezkeen esaldi generikoak probabilitate handiarekin sortzen ditu sareak.

Arazo hori konpontzeko, ikasketa gainbegiratuaren ordez sare sortzaile aurkariak (*Generative adversarial networks* ingelesez) (Goodfellow *et al.*, 2014) erabiliko ditugu lan honetan. Sare sortzaile aurkariek Turingen testaren ideia era konputazionalean aplikatzea ahalbidetzen dute. Kasu honetan, erantzunak sortzen dituen sareari (sare sortzailea hemendik aurrera) ez zaio adieraziko zein irteera dagokion sarrera bakoitzari. Horren ordez, beste sare batek, sare diskriminatzaileak, ebaluatuko ditu sare sortzaileak emandako erantzunak, zein punturaino gizatiarrak diren esanez, Turingen testaren epaile batek egingo lukeen modu berean. Sare sortzailearen helburua sare diskriminatzaileak berari emandako ebaluazioa ahal bezain beste hobetzea izango da. Sare diskriminatzailearena, aldiz, pertsonek sortutako eta sare sortzaileak sortutako esaldien artean bereiztea izango da. Modu honetan, bi sareak

[2]`https://www.cleverbot.com/`, azken bisita 2019ko martxoaren 22an.

[3]Arazo hori deskribatzen duten lanek ingelesez egiten dituztenez saiakuntzak, ingelesez ere ipini ditugu haiek erakutsitako adibideak.

iteratiboki entrenatuko dira; sortzailea saiatuko da diskriminatzaileak hura pertsonatzat hartzen, diskriminatzaileak sortzailearen eta pertsonen artean bereizten ikasten duen bitartean.

Halako optimizazio prozesua burutzea, dena den, ez da sinplea, sareak entrenatzeko normalean erabiltzen diren gradienteetan oinarritutako optimizazio metodoak ez baitira zuzenean aplikagarriak. Xehetasunetan sartu gabe, sare diskriminatzailearen irteera ez da diferentziagarria sare sortzailearen parametroekiko, sortzaileak sortutako hitzak diskretuak dira eta (Yu *et al.*, 2017). Errefortzu bidezko ikasketa erabili daiteke gradienteetan oinarritutako metodoen ordez (Li *et al.*, 2017; Hori *et al.*, 2019), baina horrek entrenamenduaren konbergentzia zaildu dezake (Sutton *et al.*, 1998). Beste aukera bat *straight-through Gumbel-softmax* (Bengio *et al.*, 2013; Jang *et al.*, 2016) zenbateslearen bidez gradientearen hurbilketa bat egitea da, Lu *et al.* (2017) eta Shetty *et al.* (2017) autoreek erakusten duten moduan. Azkenik, lan honetako autoreek guztiz diferentziagarria den sare sortzaile aurkari bat aurkeztu berri dute (López Zorrilla *et al.*, 2019)[4], hitzen errepresentazio bektorial hurbilduak erabiltzen dituena, ondoren azalduko dugun moduan.

Testuinguru honetan, lan honen ekarpenak hiru dira: alde batetik, López Zorrilla *et al.* autoreek (2019) proposatutako sare sortzaile aurkaria balioztatzen dugu, ingelesez gain euskaraz ere eraginkorra dela frogatuz; bigarrenik, modu koherente eta zentzudunean hitz egiten duen sare neuronaletan oinarritutako elkarrizketa sistema automatikoa euskaraz eraikitzea badagoela frogatzen dugu; eta amaitzeko lematizazio prozesu baten bidez corpusaren tamaina txikiagoagatik sortutako desabantailak nola leundu daitezken erakusten dugu.

3. Ikerketaren muina

Atal honetan, hasteko, erabilitako sare neuronalen egiturak azalduko ditugu. Ondoren, bi sareen parametroak doitzeko erabilitako algoritmo iteratiboa aurkeztuko dugu. Jarraitzeko, euskaraz dagoen corpusa nola aurreprozesatu dugun deskribatuko dugu. Azkenik, burututako saiakuntzak eta lortutako emaitzak erakutsi eta aztertuko ditugu.

3.1. Sare sortzaile eta diskriminatzailea

Sare sortzailea sekuentziatik-sekuentziarako sare bat da, *long short-term memory* (LSTM) (Hochreiter eta Schmidhuber, 1997) kodetzaile eta deskodetzaile errekurrente independenteekin (Sutskever *et al.*, 2014) eta arreta modulu batekin (Bahdanau *et al.*, 2015; Luong *et al.*, 2015). Sare honek T luzera arbitrarioko hitzen errepresentazio bektorialen (Mikolov *et al.*, 2013) segida bat hartuko du sarrera moduan: $\mathbf{v} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T$. Sarrera hori prozesatu ostean, irteera moduan beste τ luzera arbitrarioko segida bat bueltatuko du, elementu bakoitza sareak sor ditzakeen hitz guztien arteko probabilitate-banaketa delarik: $\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_\tau$.

Bestalde, sare diskriminatzailea beste bi kodetzaile errekurrentez osaturik dago, guztiz konektatutako geruza batzuez jarraituak, Kannan eta Vinyals autoreen (2017) antzera. Kodetzaile bakoitzak esaldi bat hartzen du sarrera moduan. Batek erabiltzailearen mezua $\mathbf{v}$ prozesatuko du, eta besteak erantzuna $\mathbf{u}$. Sistemaren irteera 0 eta 1-en arteko zenbaki erreal bat, a, izango da, erabiltzailearen mezuari emandako erantzuna zein punturaino gizatiarra den adierazten duena. Irteera zenbat eta baxuagoa, orduan eta gizatiarragoa izango da erantzuna, sarearen irizpidearen arabera. Bi sarrerak, berriz ere, hitzen errepresentazio bektorialen moduan hartuko ditu sareak.

Bi sareen gainontzeko xehetasunak (López Zorrilla *et al.*, 2019) lanean aurkitu daitezke.

3.2. Ikasketa algoritmoa

Bi sareak entrenatzeko, hiru optimizazio prozesu era iteratiboan burutuko ditugu. Lehenago aipatu dugun bezala, alde batetik sare sortzailea entrenatuko dugu diskriminatzaileak hura pertsonatzat hartzeko, hau da, diskriminatzailearen irteera minimizatzeko. Bigarrenik, diskriminatzailea entrenatuko dugu sare sortzaileak sortutako erantzunak eta corpusetik hartutako erantzunak desberdintzeko. Amaitzeko, Li *et al.* autoreen (2017) legez, sare sortzailearen parametroak ikasketa metodo gainbegiratuen bidez doituko ditugu ere, prozedura guztiaren konbergentzia bermatzeko.

Hiru optimizazio prozesu hauek definitzeko, horietako bakoitzean gradienteetan oinarritutako optimizazio metodoekin minimizatuko ditugun galera-funtzioak zehaztuko ditugu.

[4] Apirilean argitaratuko da lana.

Sare sortzailearen parametroen egiantz handieneko estimazioa

Sare sortzailearen parametroak ikasketa gainbegiratuaren bidez doituko ditugu egiantz handieneko estimazio baten bidez. Hau da, corpuseko sarrera-irteera bikote bakoitzarentzat, sareak sarrera prozesatzean irteera desiratua sortzeko duen probabilitatea maximizatuko dugu. 1 ekuazioan agertzen den galera-funtzioa erabiliko dugu.

$$L_{EH} = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{v},s \in \mathcal{C}} \frac{1}{|s|} \sum_{t=1}^{|s|} -\log \mathbf{p}_t[s_t] \ , \tag{1}$$

non $\mathcal{C}$ $\mathbf{v}$ sarreretaz eta s irteera desiratuez osatutako corpusa den, s_t irteera desiratuaren t-garren hitzari dagokion indizea, eta $\mathbf{p}_t[s_t]$ sareak t-garren denbora unean s_t hitzari esleitutako probabilitatea den. Sarearen irteera, $\mathbf{p}$, $\mathbf{v}$ sarreraren funtzioa da noski, baina mendekotasun hori ez dugu esplizituki adierazi notazioa ez korapilatzeko.

Sare diskriminatzailearen galera-funtzioa

Sare diskriminatzailearen entrenamendua burutzeko lehenik eta behin corpus berri bat sortu beharko dugu, $\mathcal{C}_D$, $\mathcal{C}$ corpusetik abiatuz eta sare sortzailea erabiliz. Diskriminatzaileak pertsonek emandako eta sare sortzaileak sortutako erantzunen artean desberdintzen ikasi behar duenez, bi eratako laginak behar ditu euren artean diskriminatu ahal izateko. Horretarako, bi motatako hirukoteez osatuko dugu $\mathcal{C}_D$ corpusa. Hirukote bakoitza erabiltzaileak bidalitako mezu batez, erantzun batez eta 0 edo 1 izan daitekeen etiketa batez osatuta egongo da. Lehenengo motako hirukoteek gizakiek emandako erantzunak edukiko dituzte, eta beraz etiketa 0 izango da. Hirukote hauek lortzeko $\mathcal{C}$ corpuseko bikoteak erabili genituen zuzenean. Bestalde, bigarren motako hirukoteek sare sortzaileak sortutako erantzunak edukiko ditu, eta ondorioz etiketaren balioa 1 izango da. Hirukote hauek eratzeko, $\mathcal{C}$ corpusetik hartu dira erabiltzailearen mezuak, gero hauek sare sortzaileari pasa sarrera moduan, eta sarearen irteera erabili erantzun moduan. $\mathcal{C}_D$ eraiki ondoren, entropia gurutzatuko galera-funtzioa erabili dugu sare diskriminatzailearen parametroak doitzeko (2 ekuazioa).

$$L_D = \frac{1}{|\mathcal{C}_D|} \sum_{\mathbf{v},\mathbf{u},l \in \mathcal{C}_D} - \left[l \cdot \log a + (1-l) \cdot \log (1-a) \right] \ , \tag{2}$$

non $\mathbf{v}$ erabiltzailearen mezuaren hitzen errepresentazio bektorialen segida den, $\mathbf{u}$ erantzunarena, l erantzuna pertsona batena edo sare sortzailearena den adierazten duen eskalarra, eta a diskriminatzailearen irteera. Berriro ere, a-k $\mathbf{v}$ eta $\mathbf{u}$-rekiko duen mendekotasuna ez dugu esplizituki adierazi.

Sare sortzailearen galera-funtzio aurkaria

Azkenik, sare sortzailea diskriminatzailearen irteera minimizatzeko galera-funtzioa definitzea erraza da, diskriminatzailearen irteera bera baita, 3 ekuazioan ageri den bezala.

$$L_S = \frac{1}{|\mathcal{C}_S|} \sum_{\mathbf{v} \in \mathcal{C}_S} a \ , \tag{3}$$

non $\mathcal{C}_S$ corpusa $\mathcal{C}$ corpusean dauden sarrera mezuez osatuta dagoen, $\mathbf{v}$ horietako bakoitza delarik. a diskriminatzailearen irteera da.

3 ekuazioko galera-funtzioa gradienteetan oinarritutako optimizazio metodoekin minimizatu ahal izateko, a sare sortzailearen parametroekiko diferentziagarria izan behar du. Sare sortzaileak $\mathbf{v}$ sarrera $\mathbf{p}$ irteeran era guztiz diferentzialean transformatzen du. Era berean, sare diskriminatzaileak bere bi sarrerak, $\mathbf{v}$ eta $\mathbf{u}$, era guztiz diferentzialean transformatzen ditu a irteeran. Hortaz, diferentziagarritasuna ez galtzeko $\mathbf{p}$ $\mathbf{u}$-n transformatu behar da transformazio diferentziagarri baten bidez. $\mathbf{p}$-ko elementu bakoitza, hots, $\mathbf{p}_t$, sareak esan ditzakeen hitz guztien arteko probabilitate-banaketa bat da. Normalean $\mathbf{p}_t$-ko maximoaren argumentua hartuko genuke sareak t-garren denbora unean esan duen hitza bezala. Baina argmax operazioa ez da deribagarria.

Arazo horri irtenbidea emateko, López Zorrilla *et al.* autoreen (2019) prozedura berdina erabiltzen dugu lan honetan. $\mathbf{p}_t$-ri dagokion errepresentazio bektoriala, $\mathbf{u}_t$, lortzeko, $\mathbf{p}_t$-ko k elementurik handienak hartzen ditugu, *top-k* operazio baten bidez. Horrela elementu horien $\tilde{\mathbf{p}}_t$ balioak eta $\mathbf{k}_t$ indizeak lortzen ditugu. Jarraian $\widetilde{\mathbf{p}}_t$ normalizatzen dugu *softmax* normalizazio batekin, $\widehat{\mathbf{p}}_t$ lortuz. Azkenik, $\mathbf{u}_t$ kalkula dezakegu $\mathbf{k}_t$ indizeei dagozkien hitzen errepresentazio bektorialen arteko bataz besteko aritmetiko haztatua eginez, pisuak $\widehat{\mathbf{p}}_t$ direlarik.

Goi mailako optimizazio algoritmoaren deskribapena

Erabiliko diren hiru galera-funtzioak deskribatu ondoren, hauek iteratiboki minimizatzeko prozedura zehaztuko dugu. Hasteko, sare sortzailearen parametroak ez ditugu ausaz hasieratuko. Horren ordez, hainbat iteraziotan zehar doituko ditugu hasieran, 1 ekuazioko egiantz handieneko galera-funtzioa minimizatuz. Behin sare sortzaileak kalitate onargarriko esaldiak sortzen dituela, $\mathcal{C}_D$ corpusa bere erantzunekin eta $\mathcal{C}$-ko pertsonen erantzunekin hasieratuko dugu, eta sare diskriminatzailea lehenengo aldiz entrenatuko dugu.

Ondoren algoritmoaren begizta nagusia hasten da. Horretan, sare sortzailea eta diskriminatzailea iteratiboki entrenatzen dira. Sare sortzailea entrenatzeko galera-funtzio aurkaria (3 ekuazioa) eta egiantz handieneko galera-funtzioak (1 ekuazioa) txandakatzen dira. Prozedura osoan zehar, sare sortzailearen entrenamendu prozesu bakoitza amaitu ostean, hainbat sarrera ausaz aukeratzen dira $\mathcal{C}$-tik eta sare sortzaileak sortutako erantzunak $\mathcal{C}_D$-ra gehitzen dira, eta diskriminatzailea entrenatzen da hainbat iteraziotan zehar. Prozeduraren konbergentzia bermatzeko, diskriminatzailea entrenatzerakoan probabilitate handiagoarekin hartzen dira $\mathcal{C}_D$-n sartutako erantzun berriagoak.

3.3. Euskararen aurreprozesamendua eta lematizazioa

Esan dugunez euskarazko corpus batekin entrenatuko ditugu sareak. Ingelesa ez bezala, euskara hizkuntza eranskaria da egitura morfologikoaren aldetik. Hau da, euskarak monema independenteak elkartuz sortzen ditu hitzak. Horrela, askotan euskaraz hitz batekin esan daitekeena ingelesez hainbat hitz erabiliz adierazi behar da. Adibidez, ingelesezko "*to the cinema*" euskaraz "zinemara" bezala itzuliko litzateke, edo "*because of the baby*" "haurrarengatik" bezala. Sareen ikuspegitik hitz bakoitza token independente bat denez, sareak ez ditu ikusten euskaraz gertatzen diren hitzen arteko erlazioak, euskararen prozesamendu automatikoa zailduz. Hasiera batean behintzat, sarearentzat "haurrarengatik" eta "haurraren" hitzak "haurrarengatik" eta "daitezke" bezain ezberdinak dira.

Honek hitzen errepresentazioa zailtzen du bi sareen sarreran, baita sare sortzailearen irteeran ere. Sareen sarreretan, hitzen egituran arreta jartzen duten errepresentazio bektorialak erabiliko ditugu hitzen arteko erlazio horiek sortzeko, *Fastext* (Bojanowski *et al.*, 2016) esate baterako. Dena den, irteeran ezin da arazoa horrela konpondu, sare sortzaileak hitzen arteko probabilitate-banaketa bat sortzen duelako. Honi irteera ematen saiatzeko, hitzen lexemak kasu marketatik eta postposizioetatik banatzea proposatzen dugu. Zehazki, izen, izenordain, izenondo eta determinanteak bananduko ditugu lan honetan. Hitzen lexema eta kategoria gramatikala topatzeko, Agerri *et al.* autoreek (2014) eraikitako kode irekiko lematizadorea erabiliko dugu. Izen, izenordain, izenondo edo determinante baten lexema eta postposizioa banatuko diren ala ez erabakitzeko, baldintza sinple bat erabiliko da. Halako hitz baten bukaera postposizio baten berdina balitz, orduan hitza lexema eta postposizioan bananduko dugu. Adibidez, "zeruko" hitza "zeru" lexeman eta "-ko" postposizioan banatuko genuke. Hogeita bi postposizio hartu genituen kontuan: "-ri", "-ei", "-rekin", "-ekin", "-ren", "-en", "-n", "-tik", "-dik", "-rik", "-ra", "-tara", "-rengana", "-engana", "-rantz", "-raino", "-z", "-rako", "-ko", "-entzat", "-tzat" eta "-gatik".

Lematizazioaz gain, izen propioak <izen> tokenera bihurtuko ditugu, normalean pertsonen izenak baitira, eta beraz, funtzio berdina dutelako esaldietan. Era berean, zenbakiak <zenbaki> tokenera bihurtuko dira.

3.4. Saiakuntzak eta emaitzak

Orain arte azaldutako sareak, ikasketa algoritmoa eta euskararen aurreprozesamendua balioztatzeko, OpenSubtitles (Lison eta Tiedemann, 2016) corpusaren euskarazko bertsioarekin entrenatuko dugu deskribatutako elkarrizketa sistema. Corpus horretatik milioi bat sarrera-irteera bikote atera daitezke, ingelesezko bertsioan baino 420 aldiz gutxiago. Corpusa 3.3 atalean azaldutako metodologiarekin aurreprozesatuko dugu, ondorioz hitz desberdinen kopurua berrehun milatik ehun milara jaitsiz. Normalean egiten den bezala, hitz horietako azpimultzo bat baino ez dugu kontuan hartuko saiakuntzetarako: maiztasun handieneko 15.000 hitzak. Gainontzekoak corpusetik kenduko dira. Aurreprozesamenduaren efektua erakusteko, corpus aurreprozesatua zein aurreprozesatu gabearekin entrenatuko ditugu sareak.

Kasu bietan, dena den, hiper-parametro berdinak erabiliko ditugu sareen arkitekturan eta baita ikasketa algoritmoan. Hiper-parametro horietako inportanteenak jarraian aipatzen ditugu. Sare errekurrente guztiak, hau da, sare sortzailearen kodetzailea, deskodetzailea, eta sare diskriminatzailearen bi kodetzaileak, bi LSTM geruzaz osatuta daude. Sare sortzailearen geruzak 1.028 zelda dituzte, eta diskriminatzailearenak 128. Adam optimizazio metodoa (Kingma eta Ba, 2014) erabiliko dugu 3.2 ataleko hiru galera-funtzioak minimizatzeko, 512 tamainako *batch*-ak erabiliz. Hitzen errepresentazio bektorialak *Fastext* metodologiarekin hasieratuko dira. Sare sortzailea 50.000 iteraziotan zehar entrenatuko dugu, ikasketa begizta hasi baino lehen. Hori 500 aldiz errepikatu dugu ondoren. Iterazio bakoitzean sare sortzailea zein diskriminatzailea 40 iterazioetan zehar entrenatuko da.

Liu *et al.* autoreek (2016) adierazten duten moduan, ebaluazio automatikoak ez dira komenigarriak elkarrizketa sistemen kalitatea neurtzeko. Hortaz, entrenatutako sistemen funtzionamendua erakusteko, zenbait sarrerako mezuen aurrean emandako erantzuna erakusten dugu emaitza modura. Erreferentzia bezala, metodologia berdinarekin [5] baina OpenSubtitles corpusaren ingelesezko bertsioarekin entrenatutako sareak emandako erantzunak ere erakusten ditugu. Erantzun guztiak 1 taulan ageri dira.

1. taula. Mezu ezberdinen aurrean elkarrizketa sistemek emandako erantzunak. A ingelesez entrenatutako sistema da, B euskaraz aurreprozesamenduarekin entrenatutakoa eta D aurreprozesamendurik gabekoa.

Sarrera mezua	Sistemaren erantzuna
A: hi !	hello , <person>.
B: kaixo !	kaixo , <izen> naiz , <izen>-ren semea naiz .
D: kaixo !	kaixo , kaixo.
A: I love you	you know , I think maybe we could talk ...
B: maite zaitut	nik ere maite zaitut .
D: maite zaitut	maite zaitut
A: who is she ?	she 's my girlfriend , <person> , and ...
B: nor da neska hori ?	<izen> da neska .
D: nor da neska hori ?	nor da neska ?
A: what's the color of the sky ?	it 's a beautiful blue sky
B: zein da zeru -ko kolorea ?	zeru urdin iluna da .
D: zein da zeruko kolorea ?	zein da zure kolore faboritoa ?
A: do n't worry	I do n't want to hear about it.
B: ez kezkatu	ez dut ulertzen zergatik egin duen hau zu -tzat , <izen> ?
D: ez kezkatu	ez dut ahaztuko
A: could you come with us there ?	I 'll come with you .
B: toki hori -tara lagunduko diguzu ?	bai , horixe uste dut nik ere .
D: toki horretara lagunduko diguzu ?	jakina baietz
A: I forgot it .	I did n't know you were coming .
B: ahaztu egin zait .	<izen> .
D: ahaztu egin zait .	zer ?
A: can you see anything ?	I see you in the mirror .
B: zerbait ikusten ?	ez , ez
D: zerbait ikusten ?	ez .
A: it was a plane crash .	I was in the hospital
B: hegazkin istripu bat izan zen .	<izen> larri zaurituta zegoen eta bere etxe -ra bueltatu zen .
D: hegazkin istripu bat izan zen .	sute bat izan zen .
A: did he wake up again ?	no he did not, but he did n't sleep .
B: jaiki al da berriz ?	zer esan nahi du horrek ?
D: jaiki al da berriz ?	astelehenean atera behar duzu .
A: yesterday I saw you in the mountains .	and now I see you
B: atzo mendi -tik ikusi zintudan .	ikusi zintudan <zenbaki> aldi -n .
D: atzo menditik ikusi zintudan .	zer ?
A: see you	see you later , <person> .
B: agur	agur , aita .
D: agur	agur , ene erregea

[5] Ingelesez corpusaren tamaina handiagoa denez, sareak ere handiagoak dira eta iterazio gehiagotan entrenatu dugu. Zehaztasunak (López Zorrilla *et al.*, 2019) erreferentzian ematen dira.

4. Ondorioak

1 taulan ikusi daitekeen moduan, sare neuronal sortzaile aurkarien bidez euskaraz era nahiko koherente eta zentzudunean hitz egiten duen elkarrizketa sistema automatikoa lortu dugu. Ingelesarekin konparatuz euskaraz dauden baliabideen tamaina askoz txikiagoa izan arren, sare neuronalen bidezko metodologiak erabiltzea badagoela frogatu dugu. Horretarako, euskararen morfologia kontuan hartzea inportantea dela erakutsi dugu ere. Izen, izenordain, izenondo edo determinanteak lexema eta postposizioetan banatzea komenigarria da, sareak era eraginkorrago batean prozesatzen baitu lengoaia. 1 taulari begira, aurreprozesu horrekin sareak esaldi konplexuagoak sortzeko joera duela esan dezakegu.

Amaitzeko, lan honekin proposatu berri dugun (López Zorrilla *et al.*, 2019) eta testuarekin era guztiz diferentzialean lan egin dezakeen sare sortzaile aurkarien arkitektura baliozkotzen dugu, elkarrizketa sistema automatikoak euskaraz eraikitzeko aproposa dela egiaztatuz.

5. Etorkizunerako planteatzen den norabidea

Dena den, lan honetan aurkeztutako metodologia eta ideiak asko garatu behar dira benetan pertsona baten moduan euskaraz hitz egiten duen sistema lortzeko. Izatez, ingelesez ere oraindik urrun gaude halako sistemak sortzetik. Oraingoz baliabide handiagoko eta txikiagoko lengoaiekin sortutako sistemak parekatzea da gure hurrengo helburua. 1 taulan ageri den moduan, ingelesez entrenatutako sare sortzaile aurkaria era zentzuduneagoan eta gizatiarreagoan hitz egiten du euskarazko sistemarekin konparatuz.

Diferentzia hauek murrizteko, ezagutzaren transferentzia (*transfer learning* ingelesez) egiteko teknikak erabiltzeko asmoa daukagu. Ezagutzaren transferentziaren ideia nagusia corpus handiagoekin baina eginkizun ezberdin baterako entrenatutako ereduak eredu berriak sortzeko erabiltzea da. Kasu honetan, beraz, ingelesez sortutako sarea euskarazko sistema hobetzeko erabiltzea izango da gure helburua.

6. Erreferentziak

Agerri, Rodrigo, Josu Bermudez, eta German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual NLP tools. In *LREC*, volume 2014, 3823–3828.

Bahdanau, Dzmitry, Kyunghyun Cho, eta Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Bengio, Yoshua, Nicholas Léonard, eta Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* .

Bojanowski, Piotr, Edouard Grave, Armand Joulin, eta Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Bordes, Antoine, eta Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683.

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, eta Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, eta Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hochreiter, Sepp, eta Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9.1735–1780.

Hori, Takaaki, Wen Wang, Yusuke Koji, Chiori Hori, Bret Harsham, eta John R Hershey. 2019. Adversarial training and decoding strategies for end-to-end neural conversation models. *Computer Speech & Language* 54.122–139.

Jang, Eric, Shixiang Gu, eta Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* .

Kannan, Anjuli, eta Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198* .

Kingma, Diederik P, eta Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, eta Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547* .

Lison, Pierre, eta Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, eta Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Lu, Jiasen, Anitha Kannan, Jianwei Yang, Devi Parikh, eta Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, 314–324.

Luong, Minh-Thang, Hieu Pham, eta Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

López Zorrilla, Asier, Mikel deVelasco Vázquez, eta M. Inés Torres. 2019. A differentiable generative adversarial network for open domain dialogue. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, eta Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Olaso Fernández, Javier Mikel, eta M. Inés Torres. 2017. User experience evaluation of a conversational bus information system in spanish. In *8th IEEE International Conference on Cognitive Infocommunications*.

Serban, Iulian V, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, eta Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Shetty, Rakshith, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, eta Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, eta Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses.

Sutskever, Ilya, Oriol Vinyals, eta Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Sutton, Richard S, Andrew G Barto, eta others. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Tuan, Yi-Lin, eta Hung-Yi Lee. 2019. Improving conditional sequence generative adversarial networks by stepwise evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* .

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* LIX.433–460.

Weizenbaum, Joseph. 1966. ELIZA— a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9.36–45.

Yu, Lantao, Weinan Zhang, Jun Wang, eta Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

7. Eskerrak eta oharrak

Lan honen egileok gure esker ona adierazi nahiko genioke Eusko Jaurlaritzari, Euskal Herriko Unibertsitateari eta baita Europar Batzordeari, PRE_2017_1_0357 eta PIF17/310 zenbakidun diru laguntzekin, eta H2020 SC1-PM15 programako RIA 7 deialdiko 769872 zenbakidun diru laguntzarekin, hurrenez hurren, ikerketa hau babesteagatik.

Publication 6

First Steps to Develop a Corpus of Interactions between Elderly and Virtual Agents in Spanish with Emotion Labels

Leila Ben Letaifa, Mikel de Velasco, Raquel Justo and M. Inés Torres. ***SLSP 2019: 7th International Conference on Statistical Language and Speech Processing SLSP, Ljubljana, Slovenia - October 14-16, 2019***

Universidad del País Vasco Euskal Herriko Unibertsitatea

First steps to develop a corpus of interactions between elderly and virtual agents in Spanish with emotion labels

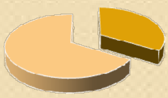
EMPATHIC
European H2020 project

Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly

Objectives

Audio emotion annotation
Emotion recognition

Database description



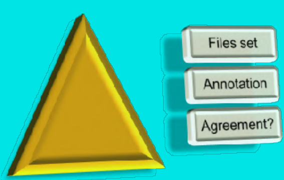
134 recording of audiovisual dialogs
79 elder people participants
Speech is extracted

Annotation protocol

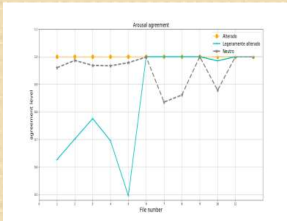
Labels

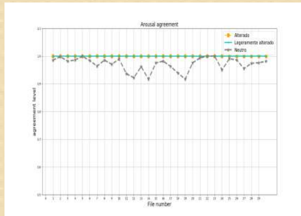
Category : calm sad happy puzzled tense
Arousal : excited / slightly excited / neutral
Valence : positive / slightly positive / slightly negative / negative
Dominance : rather dominant / neither dominant nor intimidated / rather intimidated

Procedure



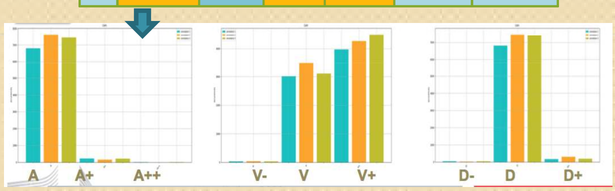
Inter-annotator agreement





Annotations' analysis

	calm	sad	hap	puzl.	tense	total
1	7017	17	260	347	12	7:17:56
2	7794	19	292	297	24	7:11:10
3	7655	21	244	360	20	7:02:10

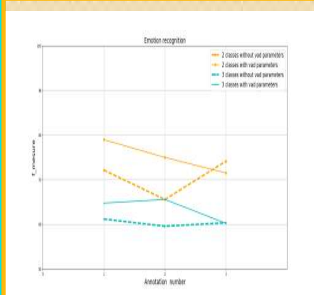


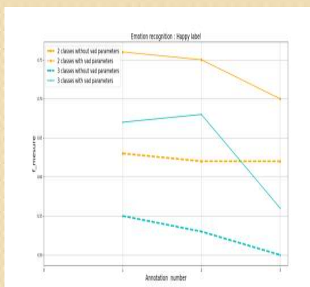
Emotion recognition

Acoustic parameters

Arousal,, valence, dominance
ZCR, Energy Entropy of energy
Spectral features :
Centroid | Spread | Entropy | Speed | rolloff

Classification system





Publication 7

Can Spontaneous Emotions be Detected from Speech on TV Political Debates?

Can Spontaneous Emotions be Detected from Speech on TV Political Debates?

Mikel deVelasco Vázquez
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: mikel.develasco@ehu.eus

Raquel Justo
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: raquel.justo@ehu.eus

Asier López Zorrilla
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: asier.lopezz@ehu.eus

María Inés Torres
Universidad del País Vasco UPV/EHU
Leioa, Spain 48940
Email: manes.torres@ehu.eus

***Abstract*—Decoding emotional states from multimodal signals is an increasingly active domain, within the framework of affective computing, which aims to a better understanding of Human-Human Communication as well as to improve Human-Computer Interaction. But the automatic recognition of spontaneous emotions from speech is a very complex task due to the lack of a certainty of the speaker states as well as to the difficulty to identify a variety of emotions in real scenarios.**
In this work we explore the extent to which emotional states can be decoded from speech signals extracted from TV political debates. The labelling procedure was supported by perception experiments where only a small set of emotions has been identified. In addition, some scaled judgements of valence, arousal and dominance were also provided. In this framework the paper shows meaningful comparisons between both, the dimensional and the categorical models of emotions, which is a new contribution when dealing with spontaneous emotions. To this end Support Vector Machines (SVM) as well as Feedforward Neural Networks (FNN) have been proposed to develop classifiers and predictors. The experimental evaluation over a Spanish corpus has shown the ability of both models to be identified in speech segments by the proposed artificial systems.

Index Terms: speech processing, emotion detection from Speech, human-AI, affective computing

I. INTRODUCTION AND CONTEXT

During the last years the Scientific Community has shown an increasing interest in affective computing and its potential capability to change the way in which Human-Machine interaction is carried out by getting a better understanding of Human-Human Communication. This is an artificial system able to analyze the intra-cognitive communication [1][2][3] between humans in order to develop ICT applications aimed to cooperate in Human-Machine communication. As a consequence, this is a good example of Cognitive infocommunications [2], that deals with the idea of cognitive processes and ICT applications working together in order to take benefit of each other and go beyond their isolated capabilities [3].

One of the goals of affective computing is the study and development of systems that can detect emotions from multimodal signals. In this work we deal with video recordings but focusing on speech, since it is inseparably intertwined with the emotional status during the cognitive process in human communication. Furthermore, it seems to be a good indicator of depression [4], very related to the emotional status, or even parkinson disease [5].

Most of the research on the identification of emotional features from video recordings has been carried out with a reduced set of acted emotions [6] [7] [8]. To this end the basic set of emotions defined by Eckman [9] has been broadly used, mainly for facial expressions. At this point it is important to note that the choice of acted emotions was just based on the easiness to get them rather than supported by any hypothesis [8]. In contrast, current research focuses more on the identification of emotions in scenarios that implements realistic tasks [10], which is the framework of our research.

Nevertheless, the automatic recognition of spontaneous emotions from speech is a very complex task. To begin with, the intensity of spontaneous emotions is generally lower than the one of acted emotions resulting in a smaller emotional space where emotions are closer [11]. In addition, researchers of emotions have established that ordinary communication involves a variety of complex feeling states that cannot be characterized by a reduced set of categories, which does not cover the wide range of affect states [10]. Therefore a number of researchers [12][10] [8] propose a dimensional [13] representation where each affect state is represented by a point in a two-dimensional space, namely valence and arousal, which some authors extend to three by also considering dominance.

Another important drawback is that the surface realizations of the underlying spontaneous emotions are different to those associated to acted emotions [11] [14], which complicates the direct application of the results of the investigations carried out with acted emotions as well as the use of acted data for training purposes. Furthermore, the set of emotions that appear in each specific real scenario is very task dependent and, thus, also the related automatic detection is. For example, the goal may just be to recognize anger through a simple anger/no anger classification in call centers [15] or to identify annoyance activation levels [16] [17] in customer assistance calls.

An additional weak point of spontaneous emotions is the labelling procedure, since the current emotion of the speaker

978-1-7281-4793-2/19/$31.00 ©2019 IEEE

cannot be unequivocally established. In fact, the emotional label assigned by a speaker to his own utterance might differ to the one assigned by a listener to the same utterance, being the first one closer to the current emotion [11]. However the speaker self annotation is not usually a realistic approach. As a consequence, the annotation of utterances in terms of spontaneous emotions is generally carried out through perception experiments, which are based on the particular judgement of every single annotator. Therefore, the disagreement amount annotators as well as the distance between the emotion expressed and the emotion perceived can be significant. In contrast, if emotions are expressed by professional actors, or just elicited, then the annotation procedure is not required [18]. Thus, the generated emotion is always labelled by the intent of the actor.

The previous framework shows spontaneous emotions generated and perceived to be very dependent of a variety of factors that make every data analysis and every automatic recognition task challenging and difficult for comparison. In this context, the main goal of this work is the analysis of the emotional content of speech produced by journalists and politicians on TV political debates. In summary, the problem addressed in this work is the analysis of the intra-cognitive communication between politicians and journalists during political debates. Additional contributions are the data annotation procedure and analysis as well as some baselines results of automatic detection of spontaneous emotions from the speech. The analysis carried out through different feature sets can be seen as an artificial cognitive capability that actually measures a human cognitive load, as defined in [16]. Section II describes the annotation procedure and data analysis in terms of both, categorical and dimensional models. Then some experiments aimed at the automatic detection of spontaneous emotions are shown in Section III. Finally Section IV present some concluding remarks.

II. Data Analysis

The specific task we are dealing with is described in this section. Then we summarize the labelling procedure along with the analysis of the annotated data.

A. Task

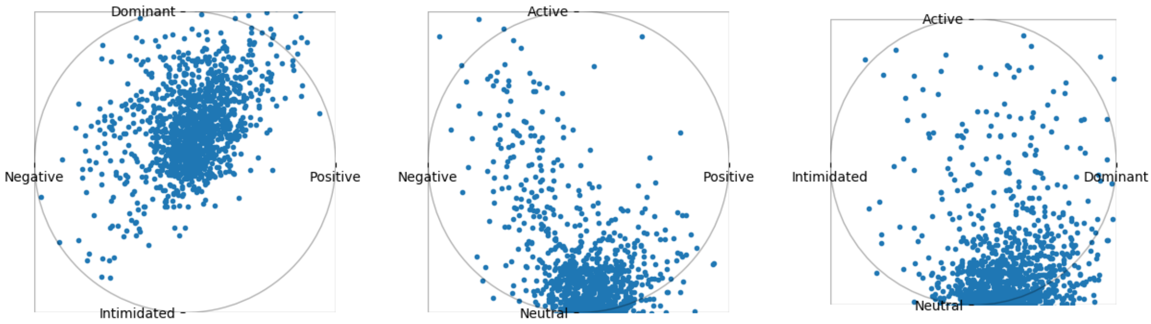
In contrast with acted emotions, spontaneous emotions show a high dependency of the specific environment in which they appear. As a consequence, the design of an appropriate corpus to develop automatic detectors of emotions as well as the choice of the specific set of emotions of interest are very linked to the particular task. In this work, the Spanish TV program "La Sexta Noche" was considered. This is a weekly broadcasted program in which a set of journalists and politicians talk about current issues. It is held as a round table discussion with a moderator that guides and conducts the debate. The program usually includes controversial topics to be discussed, so that emotional content could be expected. However, the participants are often used to speak in public, also to join TV debates, so it is not expected that they lose the control of the situation. In fact, we are in a realistic scenario in which emotions are subtle.

We first selected the broadcasts during the electoral campaign of the Spanish general elections in December 2015. Our goal was to design a corpus to train neural networks as well as to develop other data driven approaches of interest. To this end, the audios associated to the selected programs were split into smaller segments from two to five seconds. An algorithm was then designed to get audio chunks that included clauses, which are the smallest grammatical unit that can express a complete proposition. Accordingly, the algorithm uses silences and pauses, as well as the text transcriptions, to identify the compatible utterances. This procedure provided a set of 5500 audio chunks.

B. Representing Emotions by Categories and Dimensions

The audio chunks described in Section II-A, were labelled with emotional information. To this end, we considered both, the categorical model an the dimensional one, to represent the emotion associated to each chunk. The first one defines a set of discrete categories that ranks from the basic set defined by Ekman to larger sets defining more specific an realistic affect states. For this work we defined the set of categories of interest based on the selection provided in [19]. Then, it was adapted to the specific features of the task. For instance, *Sad* was not included since it is not expected to appear in political debates. The dimensional one is a psychological model that characterize affect states in terms of two or three dimensions, namely valence, arousal and dominance (VAD) [10], [20]. A crowd annotation using a crowdsourcing platform [21] was carried out to get emotional labels for both, VAD and categorical models. Each audio-clip was labeled by 5 annotators from the crowd, that were not previously trained. The goal was to pick up the diversity in people's perception in order to deal with the ambiguity associated to the interpretation of emotional information. Each annotator was asked to fill the following questionnaire for each audio:

- How do you perceive the speaker?
  - Excited
  - Slightly excited
  - Neutral
- His/her mood is:
  - Positive
  - Slightly positive
  - Neutral
  - Slightly negative
  - Negative
- How do you perceive the speaker in relation with the situation which he/she is in?
  - Rather dominant / controlling the situation
  - Rather intimidated / defensive
  - Neither dominant nor intimidated
- Select the emotion that you think describes better the speaker's mood:
  - Embarrassed
  - Bored/Tired
  - Disconcerted/Surprised
  - Angry
  - Interested
  - Satisfied/Pleased
  - Worried
  - Enthusiastic
  - Annoyed/Tense
  - Calm/Indifferent
- Quality of the audio:
  - Correct
  - Overlapping of several speakers
  - Advertisement
  - Other



(a) Valence & Dominance (b) Valence & Arousal (c) Dominance & Arousal

Fig. 1: Three projections of the data in the VAD space.

The first 3 questions are related to the VAD model and the fourth one to the categorical model. The fifth question was added to detect bad quality audios like music or overlappings. These audios were removed from the corpus.

The annotators' responses to the first three questions were used to provide a representation of each audio chunk according to the VAD model. Specifically, a conversion of the selected levels of valence, arousal and dominance into a real point in a 3D space was needed. To this end, a discrete value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited:1, Slightly excited: 0.5, Neutral: 0. Then the average value considering the 5 annotations was computed to represent each annotated chunk in the 3D space.

In order to analyse the annotated data, three projections of the points obtained in the 3D space were achieved and shown in Figures 1a, 1b and 1c, namely arousal/dominance, arousal/valence and dominance/valence. These figures show that in most of the audios speakers are neutral (not excited or not very active). Their mood is also neutral in terms of valence (no positive neither negative) but they look to be rather dominant. These results correlate well with the kind of audios we are dealing with, in which people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally they appear to be neutral with regarding their opinions (valence tends to be neutral or slightly positive).

For the categorical model, an agreement level of $>= 60\%$ was required in the annotations provided to each audio chunk to be considered. Thus, only five categories could be taken into account when the previous agreement thresholds were required, *Calm, Enthusiastic, Annoyed, Worried, Satisfied*. The rest of them were frequently mixed up with other ones so they were never associated to an audio chunk.

Then, the average of the valence and activation values of all the audios labeled with and specific category was computed. The corresponding point is represented in Figure 2 with a circle (each color represents a different category). The triangles

Fig. 2: Categorical average of the valence and arousal (dots) vs associated theoretical value [22], [23] (triangles).

are related to the position associated to the same category according to the map given in [22], [23]. This map shows the relationship among discrete categories and their representation in a valence/arousal space. From this figure it can be concluded that when comparing real vs. expected values the arousal is always lower in realistic emotions and the valence tends to be more neutral (more positive for annoyed and worried and more negative for calm, enthusiastic and satisfied). This means that most speakers in the proposed task tend to be more neutral than expected. In addition, their emotional expressions seem to be subtle, in accordance to spontaneous emotions and in contrast to simulated ones. Moreover, the whole space associated to realistic emotions (space occupied by circles) is smaller than the one related to the expected ones (space occupied by triangles).

Given that dominance is out of the previous representation the obtained dominance levels for each category were given on Figure 3. This figure shows that there are two different levels of dominance in the audios considered, a medium level for *Calm, Annoyed* and *Worried* and a higher level for *Enthusiastic* and *Satisfied*. Dominance seems to be relevant for this specific task since it is present in journalists' and politicians' speech and it is worth considering it. Moreover, it seems to be higher for positive emotions like *enthusiastic* and a bit lower for negative ones like *Annoyed*.

Fig. 3: Box-plot representation of most frequent categories over dominance dimension.

Finally, a confusion matrix with the number of audio chunks that follow a pattern of 3-2/2-3 annotations (3 annotations c_i/c_j and 2 annotations c_j/c_i) was built. From this matrix it was concluded that Annoyed and Worried are mixed up frequently and the same happens for Enthusiastic and Satisfied. This fact is also reflected in Figure 3, where their distributions are overlapped for dominance. Therefore, they were mixed and for the categorization experiments only three different categories were considered.

III. Exploring automatic detection of spontaneous emotions from speech

Two different and broadly used supervised learning paradigms were employed to carry out the automatic detection of spontaneous emotions from speech: Support-Vector Machines (SVM) [14] and Feedforward Neural Networks (FNN)[24], [25]. These models are capable of both classification and regression. Thus, they were used for classifying speech audios according to both the categorical and the dimensional VAD model, explained at Section II-B.

Our corpus consists in variable-length segments of speech, but SVMs can only process fixed-length inputs. This problem was overcome using the average and standard deviation of each acoustic feature over the whole audio chunk as input. Even though the FNN models can process variable-length audios using recurrent or convolutional layers, during our experiments only fully-connected layers have been used, in order to compare SVMs with FNNs fairly.

Six set of features were explored in this work according to previous experiments carried out with larger sets [26][27]:

- Set A: Pitch and Energy.
- Set B: Pitch, Energy and Spectral Centroid.
- Set C: Pitch, Energy, Spectral Centroid, ZCR and Spectral Spread.
- Set D: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 12 MFCC coefficients.
- Set E: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 16 LPC coefficients.
- Set F: Pitch, Energy, Spectral Centroid, ZCR, Spectral Spread and 21 Bark features.

A. Experimental set up

Four kind of SVMs and FNNs were trained, with little variation in their hyper-parameters. The first was a classifier that will predict an emotion from the categorical model. Each of the three remaining models were regressors devoted to predict the value associated to each of the three axis of the VAD model: valence, arousal and dominance.

All the SVMs for both categorical and dimensional models used the Radial Basis Function (RBF) kernel. Then, SVMs were trained until convergence. Furthermore, two-layered multilayer perceptrons FNN, with sigmoidal activation functions were also trained. However, FFN used different output layers and loss functions in the classification and regression tasks. A softmax activation function was selected for categorical classification whereas a sigmoidal activation function was applied to deal with the dimensional regressions. Then, the cross-entropy loss was employed for categorical classification, and the batch-level coefficient of determination (R^2) as the regression loss function. In addition, this coefficient was also proposed as evaluation metric in the test partition.

B. Experimental results

During a first series of experiments, the classifiers performed poorly and tended to predict the majority class, as shown in the first two rows of Table I. In order to avoid this problem, an oversampling method was employed (second row of Table I) to equalize the number of examples per category. The oversampling technique improves most of the results regardless of the set and the model used.

Table I shows that FNN models obtained slightly more accurate results than SVMs. The best FNN model was obtained using the oversampling method trained with the Set B of features.

		Set A	Set B	Set C	Set D	Set E	Set F
Without oversampling	Net	0.283	0.283	0.283	0.350	0.283	0.283
	SVM	0.336	0.283	0.324	0.279	0.281	0.283
With oversampling	Net	0.243	**0.400**	0.238	0.339	0.271	0.358
	SVM	0.284	0.290	0.221	0.280	0.287	0.214

TABLE I: Macro F1 Score on categorical experiments.

Independent models for each axis of the dimensional model were trained, because different sets of features might extract better the information required to build accurate models for each dimension. To this end, the same 6 set of features were used along with both FNN and SVM paradigms.

Two different metrics are proposed for evaluation purposes. The first and more frequently used in the literature is the mean squared error (MSE). However MSE scores as good models those models which predict the mean of the training data. Alternatively, we found out that R^2 (R squared), also known as the coefficient of determination, was instead a better metric. In fact, R^2 punishes harder regressors, which always predict the same value. Table II shows that those sets with low MSE (supposed to be best models) are not the ones with the highest R^2 score, which scores the strength of the relationship between the predictors and response.

Set	Model	Valence		Arousal		Dominance	
		MSE	R^2	MSE	R^2	MSE	R^2
A	Net	0.019	0.172	0.011	0.003	0.018	0.085
	SVM	0.020	0.111	0.011	0.016	0.018	0.089
B	Net	0.023	0.177	0.011	-0.021	0.019	0.033
	SVM	0.025	0.104	0.010	0.035	0.017	0.074
C	Net	**0.016**	0.086	0.010	0.005	**0.016**	0.088
	SVM	0.019	0.050	0.010	0.045	**0.016**	0.103
D	Net	0.020	0.275	0.012	**0.116**	0.018	0.095
	SVM	0.023	0.160	0.013	0.065	0.020	0.073
E	Net	0.022	**0.357**	**0.008**	0.095	0.019	0.069
	SVM	0.028	0.202	**0.008**	0.082	0.018	**0.119**
F	Net	0.022	0.330	0.012	0.077	0.018	0.070
	SVM	0.018	0.215	0.011	0.086	0.018	0.081

TABLE II: Results of 3 dimensional models tested with MSE and R^2 Score.

Table II shows that sets D, E and F achieve better results than sets A, B and C. This can be explained due to the information given by LPC, Bark and MFCC features, that provide similar information. This table also shows significantly better R^2 scores for valence whereas the ones for arousal looks lowers. These founds match the expectations according to Figure 1.

There is not a learning paradigm that seems to fit best the three dimensions. While the feed-forward nets are more suitable methodologies in the valence and arousal dimensions, SVM regressions seem to be more suitable for the dominance. The best feature set seems to be Set E, with the exception of the arousal dimension, where Set D is the best performing.

IV. Concluding Remarks

In this work we have analyzed the emotional content of speech produced by journalists and politicians on TV political debates. Spontaneous emotions have been labelled through perception experiments carried out over a crowdsourcing platform. These experiments reported a very reduced emotional map for this task where only a few emotions clearly appeared. The dimensional model showed distributions around neutral values, including some positive tendency towards dominance and positive valence. The dimension averages of identified categories define a reduced dimensional map where spontaneous emotions occupy a short space towards low values of arousal. The work has also explored the automatic detection of the spontaneous emotions for this task. The experiments carried out did not result in impressive accuracies, which matches the outcomes of the data analysis and also due to the fixed-length strong constraint for FNN. Better results are expected using both recurrent and convolutional networks due to the fact that some information is lost just after computing the average and standard deviation of each acoustic feature over the whole audio. Finally, it has to be outlined that R^2 score seems to be a more accurate evaluation metric for these tasks than the broadly used MSE, which provides very optimistic results when samples are close. This system develop an artificial cognitive system that represents the human decision process of the annotators when analyzing the intra-cognitive communication between politicians and journalists in political debates.

V. Acknowledgements

This work has been partially funded by the Spanish Government under grant TIN2017-85854-C4-3-R (AEI/FEDER, UE) and conducted in the project EMPATHIC (Grant n 769872) funded by the European Union's H2020 research and innovation program.

References

[1] P. Baranyi and A. Csapó, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 67–83, 2012.

[2] P. Baranyi, A. Csapó, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Springer International, 2015.

[3] P. Baranyi, "Special issue on cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 15, no. 5, pp. 7–10, 2018.

[4] K. Gbor and K. Vicsi, "Comparison of read and spontaneous speech in case of automatic detection of depression," in *8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 09 2017, pp. 213–218.

[5] D. Sztah, M. G. Tulics, K. Vicsi, and I. Vallik, "Automatic estimation of severity of parkinson's disease based on speech rhythm related features," in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017, pp. 000 011–000 016.

[6] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 371–384, Oct 2015.

[7] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: how does an automated system compare to naive human coders?" in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, March 2016, pp. 2274–2278.

[8] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[9] P. Ekman, *Handbook of Cognition and Emotion*. Sussex, U.K.: John Wiley and Sons, Ltd., 1999, ch. Basic Emotions.

[10] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010. [Online]. Available: http://dx.doi.org/10.4018/jse.2010101605

[11] R. Chakraboty, M. Pandharipande, and S. K. Kopparapu, *Analyzing Emotions in Spontaneous Speech*. Singapore: Springer Nature, 2017.

[12] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH*, 2008.

[13] J. A. Russel, "A circumflex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 116–1178, 1980.

[14] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech and Language*, vol. 53, pp. 156 – 180, 2019.

[15] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Anger detection in call center dialogues," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Oct 2015, pp. 139–144.

[16] J. Irastorza and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in *Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on*. IEEE, 2016, pp. 000 103–000 106.

[17] J. Irastorza and M. Torres, *Cognitive Infocommunications, Theory and Applications. Topics in Intelligent Engineering and Informatics*. Springer, 2019, ch. Tracking the Expression ofAnnoyance in Call Centers.

[18] T. Bnziger, M. Mortillaro, and K. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 156 – 180, 2012.

[19] A. S. Cowen and D. Keltner, “Self-report captures 27 distinct categories of emotion bridged by continuous gradients,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 38, p. E7900E7909, September 2017.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, “Avec 2014: 3d dimensional affect and depression recognition challenge,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC ’14. New York, NY, USA: ACM, 2014, pp. 3–10.

[21] R. Justo, J. M. Alcaide, and M. I. Torres, “Crowdzientzia: Crowdsourcing for research and development,” in *Proceedings of IberSpeech*, November 2016, pp. 403–410.

[22] K. R. Scherer, “What are emotions? and how can they be measured?” *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[23] J. A. Russell, “Pancultural aspects of the human conceptual organization of emotions,” *Journal of Personality and Social Psychology*, vol. 45, pp. 1281–1288, 12 1983.

[24] Z. Zhang, J. Han, E. Coutinho, and B. W. Schuller, “Dynamic difficulty awareness training for continuous emotion prediction,” *CoRR*, vol. abs/1810.05507, 2018.

[25] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5200–5204.

[26] M. de Velasco, R. Justo, J. Antn, M. Carrilero, and M. I. Torres, “Emotion Detection from Speech and Text,” in *Proc. IberSPEECH 2018*, 2018, pp. 68–71.

[27] A. Lopez-Zorrilla, M. deVelasco Vazquez, S. Cenceschi, and M. Ines Torres, “Corrective focus detection in italian speech using neural networks,” *ACTA POLYTECHNICA HUNGARICA*, vol. 15, no. 5, pp. 109–127, 2018.

Publication 8

Analysis of the interaction between elderly people and a simulated virtual coach

Journal of Ambient Intelligence and Humanized Computing (2020) 11:6125–6140
https://doi.org/10.1007/s12652-020-01983-3

ORIGINAL RESEARCH

Analysis of the interaction between elderly people and a simulated virtual coach

Raquel Justo[1] · Leila Ben Letaifa[1] · Cristina Palmero[2] · Eduardo Gonzalez-Fraile[3] · Anna Torp Johansen[4] · Alain Vázquez[1] · Gennaro Cordasco[5] · Stephan Schlögl[6] · Begoña Fernández-Ruanova[3] · Micaela Silva[4] · Sergio Escalera[2] · Mikel deVelasco[1] · Joffre Tenorio-Laranga[3] · Anna Esposito[5] · Maria Korsnes[4] · M. Inés Torres[1]

Received: 14 October 2019 / Accepted: 17 April 2020 / Published online: 22 May 2020

Abstract
The EMPATHIC project develops and validates new interaction paradigms for personalized virtual coaches (VC) to promote healthy and independent aging. To this end, the work presented in this paper is aimed to analyze the interaction between the EMPATHIC-VC and the users. One of the goals of the project is to ensure an end-user driven design, involving senior users from the beginning and during each phase of the project. Thus, the paper focuses on some sessions where the seniors carried out interactions with a Wizard of Oz driven, simulated system. A coaching strategy based on the GROW model was used throughout these sessions so as to guide interactions and engage the elderly with the goals of the project. In this interaction framework, both the human and the system behavior were analyzed. The way the wizard implements the GROW coaching strategy is a key aspect of the system behavior during the interaction. The language used by the virtual agent as well as his or her physical aspect are also important cues that were analyzed. Regarding the user behavior, the vocal communication provides information about the speaker's emotional status, that is closely related to human behavior and which can be extracted from the speech and language analysis. In the same way, the analysis of the facial expression, gazes and gestures can provide information on the non verbal human communication even when the user is not talking. In addition, in order to engage senior users, their preferences and likes had to be considered. To this end, the effect of the VC on the users was gathered by means of direct questionnaires. These analyses have shown a positive and calm behavior of users when interacting with the simulated virtual coach as well as some difficulties of the system to develop the proposed coaching strategy.

Keywords Human behavior analysis · Human–machine interaction · Spanish · Emotional analysis from speech · Language and face

1 Introduction

Despite advances in health care and technology, most of the eldercare is still provided by informal caregivers, i.e. friends and family members. According to predictions, however, this type of care will decrease in the future, for which studies encourage society to concentrate on improving the lifestyle of the elderly, helping them to remain independent for a longer period of time (Willcox et al. 2014). In particular, socio-behavioral and environmental conditions are seen a crucial factor affecting longevity (Kirkwood 2005), which to some extent explains variations found in the aging process, ranging from active and positive to feeble and dependent. We believe that four principles promote active aging, namely dignity, autonomy, participation, and joint responsibility. Information and Communication Technologies (ICT)

✉ M. Inés Torres
manes.torres@ehu.eus

1 Universidad del País Vasco UPV/EHU, Bilbao, Spain

2 Universitat de Barcelona and Computer Vision Center, Barcelona, Spain

3 Osatek/Osakidetza, Bilbao, Spain

4 Department of Old Age Psychiatry, Oslo University Hospital, Oslo, Norway

5 Università degli Studi della Campania Luigi Vinvitelli, Caserta, Italy

6 MCI Management Center Innsbruck, Innsbruck, Austria

are expected to make such principles possible, allowing the elderly to stay active members of the societal community while helping them remain independent and self-sufficient (Brinkschulte et al. 2018).

Consequently, the EMPATHIC (Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly) project (Montenegro et al. 2019; Torres et al. 2019a, b) aims to contribute to technological progress in this area by researching, innovating and validating new interaction paradigms and platforms for future generations of personalized virtual coaches (VC) to promote healthy and independent aging. The project is centered around the development of the EMPATHIC-VC, a non-obtrusive, emotionally-expressive virtual coach whose aim is to engage senior users in enjoying a healthier lifestyle concerning diet, physical activity, and social interactions. This way, they actively minimize their risk of potentially chronic diseases, which contributes to their ability to maintain a pleasant and autonomous life, while in turn helping their carers.

In this framework this paper aims to analyze the interaction between the EMPATHIC-VC and the users, mainly focusing on the analysis of the human behavior. Actually, the increasing pervasiveness of the computers in our society requires empirical studies of the human behavior during human–machine interaction that provides guidelines for the design of such interactive machines (Justo et al. 2008; Pedersen et al. 2018). As a consequence, one of the goals of the EMPATHIC project is to ensure an end-user driven design, involving senior users from the beginning and during each phase of the project, by considering their needs, by gathering initial data from them as well as their opinions regarding the technology to be developed, and by allowing them to use the personalized prototype from its first version to the final proof of concept. In order to keep them in the loop, senior users are planned to be involved in several sets of test sessions, the first set being the focus of this paper. In these sessions the seniors carried out an interaction with a Wizard of Oz (WOZ) driven, simulated system (Dahlbäck et al. 1993). That is, they believed they were interacting with an autonomous machine while actually the system was operated by an unseen human being. A coaching strategy based on the GROW model (Whitmore 2010) was used throughout these sessions, so as to guide interactions and engage the elderly with the goals of the EMPATHIC project. This way, the senior users were, on the one hand, given the chance to interact with what they thought was a final system (although the system was still not built) and, on the other hand, able to provide very valuable information as to its potential future developments. In addition, it allowed for the collection of an audiovisual data corpus which is currently used to train the machine learning models underpinning the different modules of the entire EMPATHIC system. Another important aspect to be considered for the design of the virtual coach is its visual aspect, which will have a direct impact on the user reaction. Thus, this paper also reports some studies aimed to design for elders' virtual agent acceptance.

Human behavior during the interaction with technical systems strongly depends on the goals and tasks to be developed by the interacting devices as well as on their ability for adaptation to individual user profiles and skills, preferences and emotional states (Irastorza and Inés Torres 2019; Siegert et al. 2013). Working on the aforementioned interaction framework, both the human and the system behavior were analyzed. The way the wizard implements the GROW coaching strategy is one of the key aspects to be analyzed to characterize the system behavior during the interaction with the seniors. The language used by the virtual agent, which is proposed by the natural language generator, as well as his or her physical aspect are also important cues that define the system behavior and that will be analyzed in this paper.

Regarding the user behavior, the vocal communication provides cues, which can be extracted from the speech and language analysis, that provide information about the speaker's feelings that are closely related to human behavior (Siegert et al. 2013). In the same way, the analysis of the facial expression, gazes and gestures can provide information on the non verbal human communication during the interaction even when the user is not talking. So the main focus of the analysis of the elderly behavior while interacting with the simulated virtual coach relies on their affective state, that might work as an indicator of the success of the Virtual Coach (VC). A set of perception experiments were carried out to identify and annotate the emotional status of the seniors. These experiments focused on the users' speech and also on their facial expressions recorded in the interactions. In addition, in order to engage senior users, their preferences and likes had to be considered. To this end, the effect of the VC on the users was gathered by means of direct questionnaires. These questionnaires were completed after the interactions, once the users had a better understanding of the intended system functionality.

The main contributions of this paper rely on the analysis of the behavior of Spanish[1] elderly people when interacting with a WoZ driven, simulated agent. This analysis is mainly based on the identification of the user emotional status as well as on their direct opinions of the system behavior provided through questionnaires. In addition, the behavior of the system will also be analyzed in terms of language and visual aspect.

[1] EMPATHIC project also runs human–machine interactions in France and Norway so that cross cultural analysis will also be carried out in the near future.

The paper is organized as follows: Sect. 2 describes the building procedure of the virtual coach interacting environment, which includes the WoZ platform, the coaching model and the preliminary studies for the agent acceptance. Section 3 describes the way in which the interaction sessions were designed and carried out and the way in which the end users were recruited. Then, in Sect. 4 the behavior of the wizard is analyzed through the language generated and the aspect of the virtual agent. Sections 5 and 6 provide a whole description of the emotional analysis of the user interactions regarding speech, language and facial expressions. Section 7 closes the work summarizing extracted conclusions and providing some cues for future research directions.

2 Building the virtual coach interacting environment

In order to involve seniors in the definition, development and consequent optimization of the EMPATHIC-VC it was necessary to employ various early stage prototyping methods (e.g. use case descriptions, sketches, scenarios, etc.). One of the used methods, which is particularly popular when building technology based on natural language (Schlögl et al. 2015) or other types of artificial intelligence driven applications (Dahlbäck et al. 1993), was Wizard of Oz (WOZ). The key principle of the WOZ method is that study participants believe they are interacting with an autonomous system while actually the system's actions are controlled by a human (i.e. the 'wizard'). In most cases this wizard is situated in a different room and connected to the study setting through a remote network connection. Consequently, WOZ sessions require a minimum of two researchers, i.e. the wizard controlling the technology and an additional facilitator dealing with all the participant related tasks (i.e. welcoming, informed consent, questionnaires, debriefing, etc.). For the EMPATHIC simulated VC both of these researchers received relevant training to prepare them for their tasks. The facilitator had to follow a strict procedural protocol when receiving participants and administrating questionnaires (cf. Sect. 3). The wizard received dedicated training concerning the used WOZ platform (cf. Sect. 2.1) as well as the dialogue structure which had to be followed.

2.1 The Wizard of Oz platform

Since decisions on the overall architecture of a virtual agent based application, such as the one envisioned by the EMPATHIC-VC, usually require extensive discussions, it was decided to use WebWOZ[2] (Schlögl et al. 2010a) as separate WOZ prototyping platform for early stage investigations. WebWOZ, which has been previously used by a number of research and development initiatives (e.g. Cabral et al. 2012; Milhorat et al. 2013; Sansen et al. 2016), offers an adjustable wizard interface which can be structured according to different dialogue stages (Schlögl et al. 2010b, 2011). For simulating interactions with the EMPATHIC-VC, the WebWOZ wizard interface was further extended by an audio/video transmission and recording function based on the WebRTC standard, a graphical representation of the dialogue to help guide the wizard, and the possibility to upload and consequently integrate text-based utterances. In addition, the WebWOZ client interface was integrated with five different virtual agents, which allowed participants to select their preferred interaction partner (Torres et al. 2019b).

2.2 The coaching scenarios implementing the GROW model

Coaching has been defined as a result-orientated systematic process. It generally uses strong questions in order to provide people the capacity of discovering their own abilities and draw on their own resources. In other words, the role of a coach is to foster change by facilitating a coaches' movement through a self-regulatory cycle (Grant 2003). One of the most common used coaching methodologies is the GROW Model (Whitemore 2009). This model provides a simple methodology and an adaptable structure for coaching sessions. Moreover, efficiency has been demonstrated in some Theoretical Behavior Change Models such as the Trans theoretical Model of Change (TTM) (Passmore 2011, 2012).

A GROW coaching dialogue consists of four phases which give the name to the model: Goals or objectives, Reality, Options and Will or action plan. During the first phase (Goal), the interaction aims at getting the specification of the objective that the user wants to achieve, for example, to reduce the amount of salt in order to diminish the related risk of hypertension. Then, this goal has to be placed within the personal context in which the user lives (Reality), and the potential obstacles which needs to be identified. In the next phase (Options), the agent's goal is to incite the user to analyze his/her options in achieving the objective within his/her reality. Then the final goal of the interaction is the specification of an action plan that the user will carry out in order to advance towards goals (Will). The EMPATHIC-VC is planned to deal with four coaching sub-domains: nutrition (Sayas 2018b), physical activity (Sayas 2018c), leisure (Sayas 2018a) and social and family engagement. A professional coach provided a set of handcrafted coaching sessions for each of these sub-domains. The GROW model uses Goal Set Questions (GSQ—e.g."Welcome Jorge, how can I help you?") to define the objective of the user, Motivational Questions

[2] https://github.com/stephanschloegl/WebWOZ.

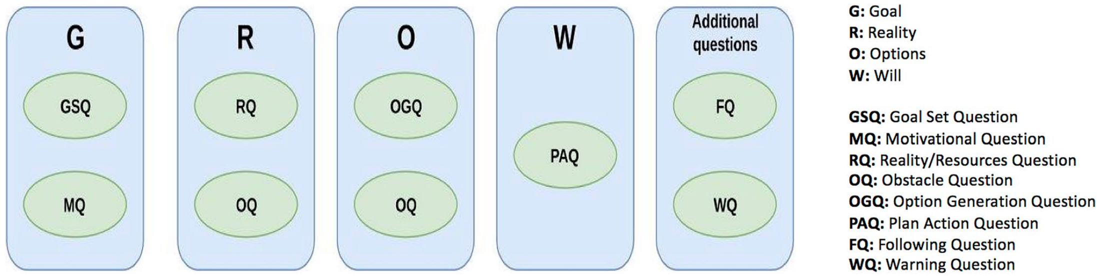


Fig. 1 The structure of the GROW model

Fig. 2 Handmade conversation created by a professional coach

E-To what extent would having regular meal times help you achieve your goal of eating the same or similar amounts of food in each of the main meals? (GSQ)

J- It would bring me much closer to my goal.

E- So what makes you want to have regular meal times? (MQ)

J- I´m clear about that: to manage to eat in a more orderly way, thereby distributing about the overall amount of food across the main meals.

(MQ—e.g."What would you achieve if you changed the way you eat?") to look for some sort of motivation which may help him/her achieve a set goal, Reality/Resources Questions (RQ—e.g."And what happens when you just eat bits?") to analyze the current situation of the user and establish resources, Obstacle Questions (OQ—e.g."And if you are out, what are you going to snack on?") to determine obstacles in the accomplishment of the goal, Option Generation Questions (OGQ—e.g."What small step could you take that would get you closer to your milestone of having meals planned?") to define possible actions a user has to perform in order to achieve the goal, Plan Action Questions (PAQ—e.g."What are you going to do to achieve your goal of adopting a more regular eating pattern?") to establish an action plan, Following Questions (FQ—e.g."How has your plan gone concerning the timing of your meals?") to ask a user about an ongoing plan, and Warning Questions (WQ—e.g."What is your blood pressure like?") to know if the user has any (other) health problems which may need to be considered. The GROW model structure is shown in Fig. 1.

An example of such a handcrafted session is shown in Fig. 2. Then, the wizard strategy was designed according to two different scenarios based on the conversations and indications provided by the professional coach. However, the Wizard had to develop and add new strategies to deal with real user interactions. Thus, a specific wizard profile was created defining a system behavior.

2.3 The user-centered iterative design for elders' virtual agent acceptance

While aiming at implementing a virtual coach devoted to assist the elderly population in their independent living, the goal was to abandon the human–machine interaction techno-centric paradigm and focus on the needs and intentions of the relevant elder end-users, their abilities, aptitudes, preferences, and desires. As for its implementation, the EMPATHIC-VC had as initial requirements accessibility and usability by a wide variety of elderly users, ranging from field experts, practitioners, persons with different knowledge (culture, instruction and occupations), needs (impaired and communicatively disordered individuals) age, and preferences.

To this aims, we have taken a user-centered iterative design, assessing users' interactions in context so that (a) trustworthy human–agent relationships are build, (b) emotional states and negative moods such as depression are reliably detected (Buendia and Devillers 2014; Cavanagh and Millings 2013; DeSteno et al. 2012; Parker and Hawley 2013), and (c) appropriate advice on actions is provided. This was built upon several theoretical experiments, to collect a substantial quantity of data assessing seniors' willingness and interest in initiating and retaining conversations with an agent upon different qualitative agent features (such as gender and voice) in comparison to differently aged populations such as adults and adolescents.

With this research, we have acquired a deeper understanding of how to design emotionally-aware interactive agents that exhibit coherent visual, vocal and gestural affordances, and adapt to the user's underlying intentional and emotional states in a cooperative and ethically sound manner. All the executed experiments were driven by the key idea that any intelligent social ICT interface should be capable of establishing an *empathic relationship*; hence the emphasis of the investigations was on mood enhancement linked to use-cases in e-mental health and support for older/vulnerable people.

The rich repertoire of theoretical results acquired is summarized below, in particular for agent's gender and voice.

A first pilot experiment, focusing on user requirements and expectations with respect to participants' age and familiarity with technological devices (such as smartphones, laptops, and tablets) showed that, as for gender, elders prefer to be assisted by female agents (Esposito et al. 2018b). In this context, an ad-hoc questionnaire was developed to assess senior's preferences, expectations and requirements, in order to customize the consequently developed EMPATHIC-VC to the needs of the targeted end-user population, i.e. elders.

It has to be noted that starting with this pilot, the questionnaire has been gradually modified, in an attempt to incorporate the Theory of Acceptance Model (TAM) proposed by Davis (1989) and the pragmatic and hedonic dimensions proposed by Hassenzahl (2004). The result has been given the name Virtual Agent's Acceptance Questionnaire (VAAQ) and may count as a direct outcome of the Empathic project.

For the above mentioned pilot investigation using an early version of the VAAQ it was further learned that seniors' preference for female agents was significantly higher than for male agents for all the questionnaire dimensions, independently of seniors' genders and technology savviness.

In order to remove the biases introduced by differences in agent's personalities, a second set of experiments was conducted (Esposito et al. 2018a). In these trials, the four proposed agents (two males and two females) were endowed of a "neutral" personality, and their facial expressions were neither smiling, saddening, nor worrying. This test definitively confirmed seniors' preferences to be assisted by female agents which scored significantly better than male agents in all the questionnaire subsections.

In order to assess whether seniors' preferences toward female speaking agents were a specific requirement of the elder population, we defined another set of tests involving adolescents, adults, and seniors for a total of 316 participants split in 7 groups, each composed of approximately 45 subjects, equally balanced for gender (Esposito et al. 2019a). There were two groups of adolescents (mean age = 14.5, SD = ± 0.5 years), two of adults (mean age = 25.1, SD = ± 3.5 years), and two of seniors (mean age = 71.4, SD = ± 6.5 years). It was found that elders' willingness to interact was significantly higher for speaking than mute agents, and, in the speaking context, it was significantly higher for female speaking than male speaking agents. In addition, for elders in the speaking context, female agents were judged significantly more positive than male agents for attractiveness, pragmatic, and hedonic (identity and feeling) qualities. None of these significant differences was observed for adolescents and adults administered with mute and speaking agents and elders administered with mute agents.

When the three elder groups were compared on their enjoyment/acceptance scores for mute, speaking and only voice interfaces, elders' preferences were significantly higher for female speaking agents and only female voice interfaces.

The discussed experiments suggest that the successful incorporation of assistive social technologies in everyday life is strongly depending on the user's perception and acceptance of them (de Graaf et al. 2015). In particular, robots, virtual agents, and generally, interactive assistive user interfaces, need to be specifically tailored to people's needs, and personalized according to their specific requirements and expectations (Seiki et al. 2017),

3 Description of the interaction sessions and user studies

The potential participants for the following interactive study were defined as "healthy seniors" for which the inclusion criteria was: (i) female or male older than 65 years, (ii) living independently (not institutionalized), (iii) being able to read, write and speak fluently in Spanish. For the recruitment of the sample different strategies such as advertising posters, informative notes, mailing and flyers spread in the local areas were used. The consequent study setting employed the previously described WOZ method in order to observe and systematically record both participants' behavior and system operation. In this setting, the first step for participants was to sign an informed consent form before enrolling in the study. Then, the experimental protocol included three steps:

1. The completion of two health questionnaires: Participants were asked to fill in the Geriatric Depression Scale (GDS) and the World Health Organization Quality of Life (WHO-QoL-BREF). The GDS is a dichotomous

Fig. 3 Setup with a participant during a session

("yes" or "no") 30-item (10 negatively worded and 20 positively worded) self-report scale aimed at rating depression (Yesavage et al. 1982). Total scores range from 0 to 30 points, where higher scores mean higher probability of having a depression diagnose. The WHO-QoL-BREF is an abbreviated (26 items) generic quality of life scale developed by the World Health Organization (Who-QoL Group[3]) which assesses four domains: physical health, psychological health, social relationships, and environment. These questionnaires were administered before interacting with the EMPATHIC-VC so as to provide unbiased scores.

2. The interaction with the VC: In this step we used laptops equipped with a webcam, a microphone and a mobile connection (4G/4G+). Participants were logged into a secured session (protected by username and password) with an individual alphanumeric ID code to keep their identity safe. Then, based on their personal preference, they chose one of five available visual representations, i.e. agents, for their VC. Each of these agents (3 female and 2 male) showed different characteristics with respect to their appearance. From that point on, in order to avoid potential impacts the supervisor may have on the dialogue or the actual interaction, participants were left alone with the VC. Two dialogues of 5–10 min each were completed. The first served as an introduction to the system and thus did not focus on any specific issues. The second revolved around a conversation related to nutrition/food (cf. Fig. 3). The structure of this second dialogue was based on the GROW coaching model (Whitmore 2010) presented above. As described earlier, a GROW coaching dialogue consists of four phases; i.e. **G**oals or objectives, **R**eality, **O**ptions and **W**ill or action plan. In the given setting the goal was set on participants' nutritional habits and respective objectives.

3. Finally, after the interaction with the VC, participants were asked to give feedback using a number of user-feedback questionnaires; i.e. the EMPATHIC Virtual Agent Acceptance Questionnaire (VAAQ) (Esposito et al. 2018a), the System Usability Scale (SUS) (Brooke 1996) and the Emotion auto-annotation form. The VAAQ was developed to explore participants' satisfaction in interacting with virtual agents. It contains three sections: (i) the socio-demographic status, (ii) the willingness to be involved in interactions with a Virtual Agent (VA) and (iii) the perceptions of the respective agent features. The SUS contains ten statements regarding a system's usability to which participants respond to on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". Finally, the emotion auto-annotation form was an ad-hoc questionnaire that asked participants about their two most intense emotions experienced during the contacts with the VC.

A total of 156 WOZ user studies (78 Spanish individuals in 2 sessions) were conducted. The following insights are based on the collected demographic data, the feedback provided by wizards who simulated the EMPATHIC-VC, facilitators and study participants.

3.1 Study set up

Experience has shown that at least two people were required to realistically conduct a WOZ user study, one who acts as a human simulator, i.e. the wizard and one who acts as a facilitator, greeting study participants, introducing them to the study purpose, administrating questionnaires, and helping the participants in case of confusion or technical issues. From a procedural point of view, we further found it imperative that, once the interaction with the VC started, the facilitator had to leave the room. Otherwise the participant tended to look at and talk to the facilitator instead of conversing with the actual agent. This behavior may be explained by a participant's lack of reassurance when interacting with a novel technology.

3.2 Study participants

With our sample size of 78 individuals, we found a higher concentration of users from the first age cohort. That is, 60% of participants were from the age group 65-70, with 69.23% identified as female; and 67.14% had higher education (HE).

In general, we found that the concept of a virtual agent seemed rather frightening to many people of the targeted age group (i.e. aged 65 or older). While we did use face-to-face meetings to overcome this fear as much as possible, it should be noted that for this type of technology anxiety poses a significant challenge, particularly when it comes to

[3] https://doi.org/10.1016/0277-9536(95)00112-K.

the recruitment of study participants. Consequently, recruitment via flyers/posters was difficult (even when conducted in senior centers or elderly homes). However, we found that recommendations coming from other participants who had already taken part and enjoyed the study, helped mitigate the problem. Still, a lot of personal coaching was usually required to make people feel comfortable. Here, our experience has shown that participants needed approx. Ten minutes interaction with the VC to 'lose their fear' regarding the technology—in particular, when studies took place somewhere away from peoples' homes or familiar living environments. With regard to the study inclusion criteria, the studies have shown that elderly people are rather pessimistic when evaluating their personal health status. That is, while initially we were searching for 'healthy' participants aged 65 or older, we had to realize that most representatives of this group would not include themselves due to minor health issues they perceived (e.g. minor hearing problems, minor vision impairments). As for the interaction, it seemed important that participants thought they would interact with a prototypical system. This helped keep the expectations regarding speed and accuracy low. In this context, the speed with which a simulated system responds may be seen a particular challenge. Especially in cases where the wizard could not use a pre-defined utterance and had to type a response. An additional challenge with this generation of on-the-fly utterances concerns the great potential for typos and other mistakes, which are forwarded to the text-to-speech module and, consequently, spoken out loud to a study participant. However, being aware of the prototypical status of the system, study participants were rather tolerant toward these types of issues.

3.3 The dialogues

The participants were usually pre-informed about some of the content to be addressed by the coach so that they could think about relevant topics in advance (e.g., they were told to think about certain goals they would like to achieve before starting the conversation). Such was necessary to keep the interaction going and reduce the number of "yes/no" answers. Still, in particular with respect to the nutrition scenario, it was difficult to keep the conversation flowing, as the scenario was looking for personal goals, yet people were often satisfied with their status-quo and, thus, did not find much to talk about.

Changing the conversational focus due to missing participant goals also caused some side effects. Finally, from a conversational point of view, we found that different types of back-channeling (i.e., approving a participant's input) had a significant influence on the 'smoothness' of the conversation. That is, while rather basic approval utterances such as "interesting" or "good" seemed to distort the conversation, other strategies which re-used participants' words or sentence structures (e.g., Participant: "I like to walk 2 hours every day"; Agent: "You walk 2 hours every day?") helped in keeping participants engaged and consequently the conversation flowing.

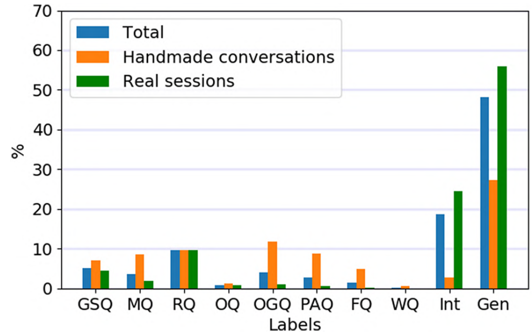


Fig. 4 Distribution of Dialogue Acts

4 Analysis of Wizard behavior

The analysis of the Wizard behavior was carried out over the system turns used in the WOZ sessions. These turns were initially established to define each scenario. However, the Wizard had to create new turns to go ahead with the GROW model, to get more developed answers from the user or even to resolve situations caused by the behavior of real users, as analyzed in the previous section. System turns were labeled according to the needs of the Natural Language Generation (NLG) module. They were annotated in terms of Dialogue Acts, Polarity and Entities. These annotations allow for a structured analysis of the VC behavior when addressing participants.

In addition to labeling the WOZ sessions, annotations were also assigned to the set of (handmade) coaching sessions proposed by the professional coach (see Fig. 1). A comparison between the interactions during the conversations with the simulated EMPATHIC-VC and the ones created by the professional coach is a way to test the behavior of the Wizard strategy (although it has to be noted that these VC interactions pose a higher level of artificiality and they might be far from real interactions conducted with a human coach).

4.1 Annotation in terms of Dialogue Acts

The set of Dialogue Acts (DAs) defining the turns of the coach consist of the eight questions used in the GROW mode, i.e. (GSQ), (MQ), (RQ), (OQ), (OGQ), (PAQ),

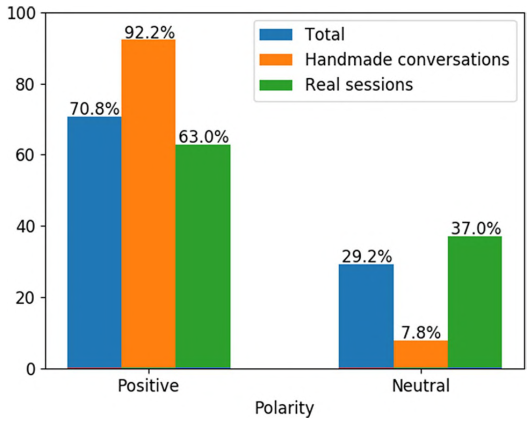


Fig. 5 Distribution of the polarity

(FQ), (WQ) extended by the Introduction label (Int) defining a typical sentence uttered by the coach during the first session with the user, and the General label (Gen) used for all the other interventions the coach performed during the conversations (greetings, agreements, etc.). This labeling structure allows to evaluate the wizard's alignment with the GROW model (cf. Fig. 2).

Figure 4 shows the distribution of DAs for three different sets: the interactions coming from the conversations created by the professional coach, the interactions of the WOZ, and a new set joining the previous two. In all the cases, the most used utterances were general sentences (Gen), which were related to the most common expressions in a conversation. This shows that the wizard was indeed following the instructions of the professional coach considering that a conversation should not be a succession of GROW questions, but it should rather follow a more natural dialogue structure resembling a bidirectional conversation.

The second most frequent label in the wizard data was related to the introduction session (Int). Looking at the overall data distribution it can be observed that there were few sentences in the handmade conversations annotated with this label. This is, however, due to the fact that the professional coach did not follow an entire introduction procedure but rather gave us some sample instructions on how such an introduction session should be mapped out. The wizard data shows that those instructions were followed.

Generally, we can see that the sessions with the human coach exhibit a quite balanced distribution of GROW labels, whereas in the sessions with the wizard a significantly higher number of GSQ, MQ and RQ labels appear, all of which are situated in the initial two phases of the GROW model. This fact suggests that the wizard managed to connect with participants but seemed to have difficulties advancing deeper into the coaching dialogue.

4.2 Annotation of polarity

A key quality defining a successful coach is not to show mood characteristics which may be perceived as negative. For the EMPATHIC-VC it was thus established that the system should express a positive attitude whenever the user's mood is perceived positive or neutral attitude whenever the user's mood is perceived negative.

Keeping a positive attitude was also a guideline expressed by the professional coach. In fact, less than one tenth of the utterances recorded in the handmade sessions were annotated with neutral mood (cf. Fig. 5). In the wizard sessions, there were less differences in polarity values. That is, even though the predominant mood was also positive, a significant number of mood stages were labeled as neutral. Such could be caused by the characteristics of the conversations. As was mentioned before, the sessions dealt mostly with the G and R phases of the GROW model, which focus on exposing the problems that users found in their life. In this context, users often express negative moods and the wizard, consequently, neutral behavior.

4.3 Annotation of entities

One way to make users feel that the machine understands them is to use the same or similar linguistic elements that they have used in their turns. Respective linguistic elements include named entities such as names, places, food, etc. These entities, which have to be identified in the user turns, can be added in the responses of the VA. Thus, in this phase, the annotation focused on identifying all the linguistic elements which can be interpreted as an entity and assigning them to the corresponding category. The following types of entities were defined: *Actions*, *Dates*, *Food*, *Frequency*, *Hobbies*, *Places*, *Quantities*, *Topic*, *User Name* and *Others*. The distribution of these entities in the different types of conversations are shown in Fig. 6.

As Fig. 6 shows, *Action* is the type of entity which was most identified in the sessions with the professional coach. These entities are related to objectives, obstacles, options and action plans of the user. The percentage is lower for the wizard session given that the wizard was mainly working in the first two phases of the GROW model, so some of the entities related to *Action* did never appear.

Another type of entity the professional coach tried to introduce in the conversations is the user's name. This is considered a way of increasing the perceived friendliness. Indeed, the professional coach did frequently use the user's name during a coaching session. The wizard, however, did not seem to include the name so frequently.

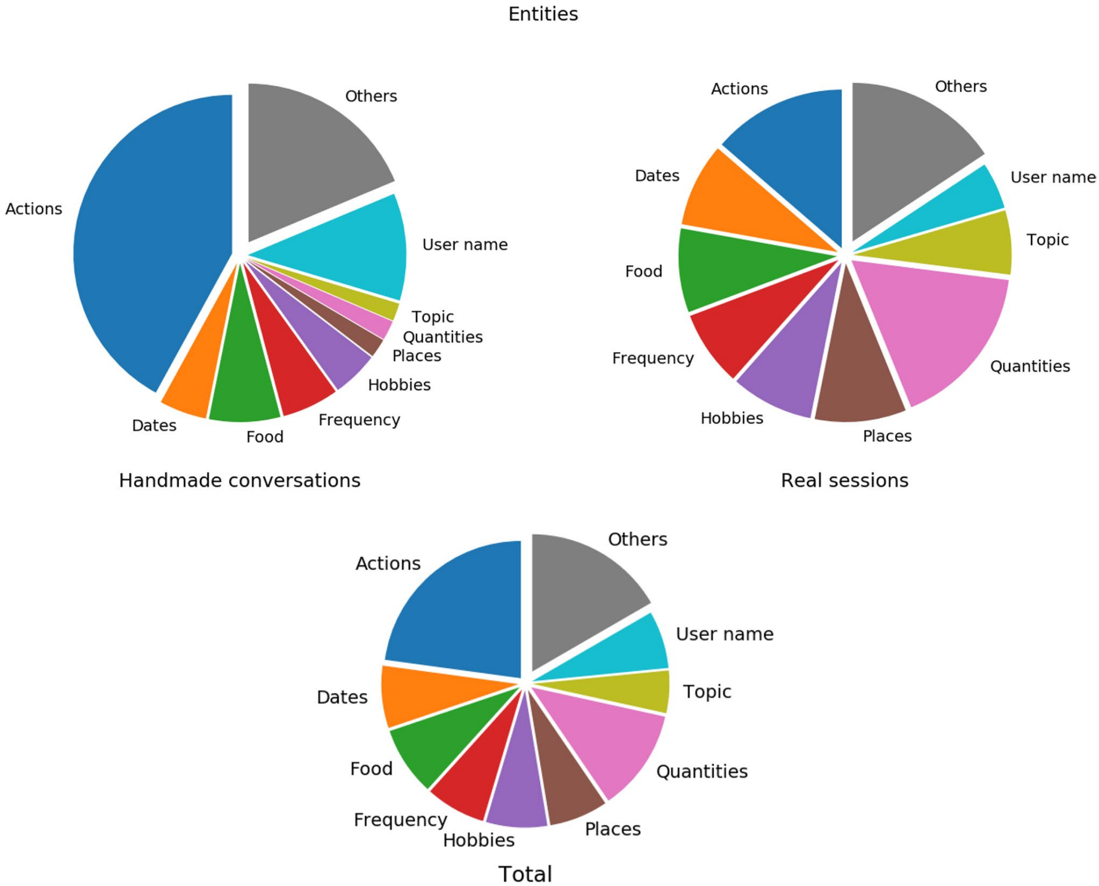


Fig. 6 Distribution of entities for the handmade conversations, for the real wizard sessions and for all the conversations together

As for the other labeled entities, hobbies, music and travel were the main elements users talked about during the first sessions. Thus, we can see that here the entities *Places* and *Hobbies* were among the most frequent ones. Similarly, *Quantities* and *Food* were often identified in the second sessions, related to nutrition.

4.4 Behavior profile of the Wizards

Based on these comparisons carried out between real and handmade sessions, we can thus conclude that the wizard behavior was very similar to the behavior described by the professional coach. Some differences have been found, but they seem to be more related to the progress of the conversation than to the wizard's strategy.

To sum up, we have found that the wizard tried to mix the use of GROW questions or introduction sentences with more general expressions so as to maintain a fluent and natural conversation, while focusing on the actual topic of each session. Furthermore, the wizard kept a positive mood when possible and the neutral mood was employed otherwise. Finally, an attempt was made to let the users lead the conversation without forcing them into achieving the final stages of the GROW model.

4.5 Agent preference

For their interactions, participants had to select one out of five different agents, shown in Fig. 7. The analysis of the agent preference shows that 66.7% of the participants selected a female agent, and that Natalie was the most popular one selected by 49.7% of the participants (remaining 17.9% Lena, 10.3% Alice, 15.4% Christian and 1.3% Adam).

Comparing female and male participants regarding agent gender preference, 79% of male and 70% of female participants

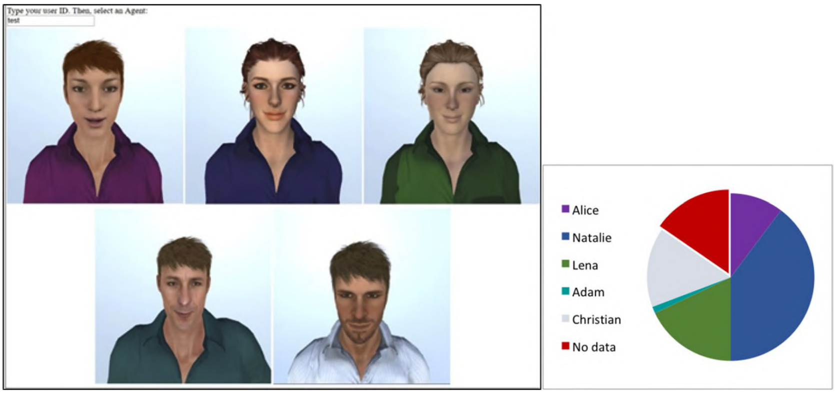


Fig. 7 The five agents that could be chosen by the users and a diagram representing the their selection using different colors

Fig. 8 Annotation in terms of categorical and VAD model of two generic audio segments

preferred a female agent and 21% of male and 8.5% of female participants preferred a male agent.

Concerning the agent's age, the majority of the participants (i.e. 69.2%) preferred an agent looking 29-48 years old; 25.6% preferred an age between 29–38 years, and 43.6% and age between 39–48 years. When asked to guess the age of the agent, participants perceived the agent to be on average 34.4 years old (SD of 5.57). This is a good indication that the 'looks' of the agent did correlate with the participants' preferred age.

Additional comments given by participants concerned the VC's general physical appearance and its latency with respect to responses and movements.

5 Emotional status from speech and language

In order to analyze the emotional status of participants, the conversations between the participants and the wizard were recorded and annotated in terms of emotions. The speech signal was manually labeled from scratch by three Spanish native annotators. Since emotion perception is gender dependent (Vidrascu 2007b), two men and one woman were selected for the annotation task.

The annotators determined manually the emotional state limits (i.e., segment) and also the emotional label associated to that segment. No particular instructions were given to them, except that they should annotate all the signal (no segment without annotation) and that a high agreement between annotators was desirable.

The annotation was made in terms of both a categorical and a dimensional model. The dimensional VAD model is a psychological model that characterizes affect states in terms of two or three dimensions, namely valence, arousal and dominance (VAD) (Gunes and Pantic 2010; Valstar et al. 2014). Thus, the annotators assigned four labels to each audio segment: one label related to a specific category and 3 additional labels, one related to the valence, another one to the arousal and a final one to the dominance as shown in Fig. 8.

The annotation procedure was organized in three steps. First, a set of files was chosen to be annotated by each annotator separately. Then, the inter-annotator agreement was computed. If the agreement was less than a predefined threshold, the annotators discussed and re-annotate the files as shown in Fig. 9. Following this procedure, we managed to reach an agreement level for the categorical model annotations that was greater than 90% for all emotions and even 100% for *sad* and *tense*.

5.1 Analysis of dimensional annotation

The labels assigned to the dimensional VAD model were:

- Valence: positive, neither positive nor negative, negative
- Arousal: excited, slightly excited, neutral

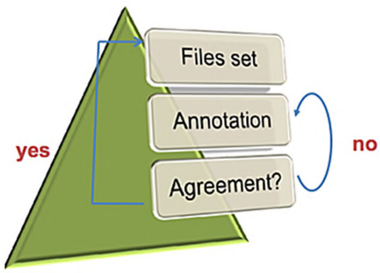


Fig. 9 Annotation procedure

- Dominance: dominant, neither dominant nor intimidated, defensive

The three labels assigned to each segment were converted to a real point in a 3D space where the axes correspond to valence, arousal and dominance. To this end, a discrete value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited: 1, Slightly excited: 0.5, Neutral: 0. Then, the average value of the annotations provided by the three annotators was computed to represent each annotated segment in the 3D space.

Figure 10 shows the probability density function of each variable (valence, arousal, dominance) estimated by using a Gaussian kernel density estimator. The results show that, in most of the cases, low values of Arousal along with positive values of Valence and neutral values of Dominance were obtained. This indicates that in the interaction with the wizard users did not achieve high excitation levels, which corresponds with our expectations, and means that the interaction with the system did not unsettle people. The dominance values were also quite neutral, neither dominant nor intimidated. This means that the behavior of the virtual coach was appropriate and did not make people feel intimidated.

Table 1 Number of segment annotated with each category label

Annotation	Calm	Sad	Amused	Puzzled	Tense
First	7017	17	260	347	12
Second	7794	19	292	297	24
Third	7655	21	244	360	20
Agreement	3368	12	100	90	13

Finally, the valence results show that users have positive feelings with regard to the interaction with the system.

5.2 Categorical model

The categorical labels assigned to each audio segment were: *calm/tired/bored*, *sad*, *amused/satisfied*, *puzzled* and *tense*. Let us note that some categories were intentionally combined into one label, (e.g. calm, tired and bored) because we saw in previous experiments that they were frequently mixed up in this task. Moreover, keeping a long list of categories would have increased the difficulty of the annotators' task, providing lower agreement values. From now on, the mixed categories will be referred to by their first label, that is *calm* for *calm/tired/bored* and *amused* for *amused/satisfied*.

The annotation of the database by the three annotators led to the categories shown in Table 1. Additionally, we decided to consider for our work only those segments where all the three annotators agreed on the given label. Since the segment limits were also defined by the annotators there could be a mismatch among them, thus, we selected the segment intersection with the same label from all the annotations, even if this led to a fragmentation of some segments into two or more different ones. This happened, for instance, with *tense*, where there were 12 segments from the first annotator and a higher number of segments (13) when agreement was required.

According to the obtained results, it can be concluded that the most frequent label was *calm*, suggesting that the

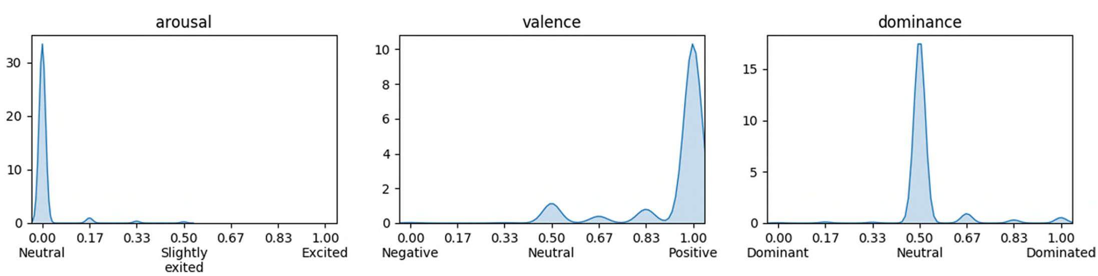


Fig. 10 The probability density function of Valence, Arousal and Dominance according to the data

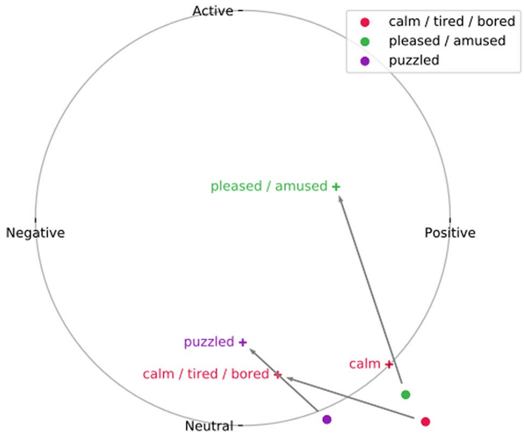


Fig. 11 Comparison between the theoretical and experimental values of arousal and valence associated to each emotion. Experimental data (circles) were achieved by computing the average of valence and arousal values of all the audio segments labeled with a specific emotion. The theoretical points (crosses) were extracted from the diagram given in Cambria et al. (2012)

dialogue system seems to have been perceived as friendly by the users. The labels *sad* and *tense* were quasi absent. In fact, when agreement between the annotators was required they almost disappeared, which means that the virtual coach did not provoke such negative feelings in users while they were interacting.

In addition, we compared the theoretical and experimental values of arousal and valence associated to each emotion. To do so, the average of valence and arousal values of all the audio segments labeled with a specific emotion was computed and the corresponding point represented in Fig. 11 using a circle. Then, a cross was used to represent the position of the same emotions in the ⟨ arousal, valence⟩ space according to the diagram given in Cambria et al. (2012). For instance, the purple circle in Fig. 11 was achieved by computing the average of valence and arousal values for all the samples labeled as *puzzled* in our database. The purple cross, instead, was extracted from the diagram given in Cambria et al. (2012) that represents where the values of valence and arousal should theoretically be for *puzzled*. Note that *Tense* and *sad* emotions were removed from Fig. 11 because there were not enough samples to represent them confidently. In order to theoretically represent the mixed classes, *calm/tired/bored* and *amused/pleased*, the mid point of the different classes was computed. That is, for *amused/pleased* the theoretical points of *amused* and *pleased* were considered and the mid point between them computed. According to this comparison, it can be concluded that in our experiments there were some differences regarding theoretical values. The differences might be due to the specific task, where real and not acted emotions were involved to some extent, as it also happened in deVelasco et al. (2019) where a different task was considered. Specifically, the real arousal values of our experiments were lower than the expected ones for all emotions. This could be due to the fact that real emotions seem to be more subtle than the acted ones and not so extreme. However, real values of valence were higher than the expected ones for all categories in our experiments. This might also due to the task. Participants that accept to take part in such a trial are usually curious and show a positive predisposition with regard to the situation. Furthermore, they interact with a system in a controlled environment with other people in the surrounding, not alone, so they usually do not allow negative feelings to rule their behavior. Looking at the vectors illustrating the differences between real and theoretical values, the one related to *calm* appears to be a bit different from the others. We may explain this by showing the theoretical value of calm. It seems that most annotators that used *calm/tired/bored* actually were labeling with calm and not with the other two emotions.

5.3 Analysis from text

Besides the acoustic information, there are other sources that can provide information about the emotional status of the users. The semantic meaning involved in a user utterance, for instance, can provide complementary information in some scenarios (Justo et al. 2018). Thus, we analyzed the users behavior focusing on the text obtained from the transcriptions of the user utterances when they were interacting with the system. Specifically, we consider the polarity of the text associated to the utterances.

Firstly, the transcriptions of the audio recordings were manually extracted, in terms of user turns, by professional annotators. Then, each transcription was manually labeled by Spanish native annotators. Although up to nine different annotators were involved in the process, only one annotator labeled each transcript. They were asked to consider each user turn and divide it into segments according to the topic they were dealing with and then to assign a polarity value to the corresponding segment. The possible polarity values were: *negative*, *neutral* and *positive*.

The histogram of the segments labeled with the different polarity values is shown in Fig. 12, where it can be observed that more than 60% of the segments were neutral and around 28% positive. Negative segments were almost absent, accounting for only 5% of the total segments. These results can be compared to the valence values obtained from the annotation of acoustic segments. In both cases the negative values are not significant, meaning that the users do not show negative feelings when regarding the interactions

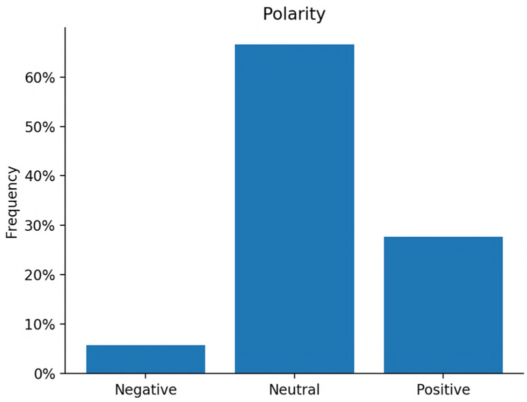


Fig. 12 Histogram with the text segments labeled with each polarity value

with the system, as mentioned above. However, in this case there are more segments labeled as neutral and less labeled as positive. This might be due to the bias associated to the specific annotators and also due to the information itself, because there might be utterances were a positive feeling can be perceived from the acoustic information but the semantic meaning of the message does not imply anything that denotes positiveness. For instance, the text *'I usually eat varied fruits and vegetables'* does not show a positive polarity but, depending on the way it is pronounced, it might be associated to an acoustic segment labeled with a positive valence.

6 What the participant facial expressions says

Following the analysis of Sect. 5, the visual modality was also manually labeled from scratch by two Spanish annotators to analyze facial expressions of emotion in user-wizard interaction videos. To guarantee that only visual information was taken into account, videos were muted throughout the annotation procedure.

6.1 Annotation protocol

As for speech, the annotators determined the emotional state limits and the emotional label associated to that segment using a categorical model. This time, however, the annotators were instructed with particular guidelines to follow so as to ensure a common annotation protocol. First, only facial expressions and head movements had to be taken into account to annotate emotions. Out-of-face information, such as body and hand movements, were out of the scope of the annotation. Second, some participants have a specific neutral expression according to their physiognomy. For instance, some people have facial features that can be perceived as happy, even though they are not trying to communicate a state of happiness at that time. To learn this baseline neutral face, annotators were requested to watch the whole video once before starting the annotation procedure.

The annotation procedure was similar to that of emotions from speech (see Fig. 9). First, a subset of 4 videos was selected to train the annotators. Once annotated, the inter-annotator agreement was computed. Then, annotators were requested to discuss the minimum level an expression must be perceived to label it with a specific category, to reach a consensus and re-annotate the files with their updated protocol. This process continued until all the videos reached a valid inter-annotator agreement.

The categorical labels assigned to each segment were: *sad, annoyed/angry, surprised, happy/amused, pensive* and *other*. As for speech, some categories were combined in one label due to the outcome of previous experiments. The first four are included in Ekman's universal expressions of emotion (Ekman and Keltner 1997). *Pensive* is not an emotion per se; however, it is included in our model as it has shown to be a frequent facial expression present in conversation and it is informative of our internal and cognitive states (El Kaliouby and Robinson 2005; Rozin and Cohen 2003). Annotators were instructed to annotate as one of the first 5 categories those segments in which it was clear for them that the expression was present. *Other* was used to denote either those segments in which one expression was taking place but which was not included in our expression list, or when more than one expression from the list was present. Finally, all non-labeled instances were considered to be a *neutral* expression, denoting the baseline face as well as calmed, quiet, or very subtle emotions which do not exceed the consensual expression thresholds.

6.2 Analysis of categorical annotations

The final inter-rater agreement level for our selected categorical annotations was high, above 80% on average. For the reminder of the section, we consider as gold standard those segments where both annotators agreed on a given label. Following the speech analysis, we also selected the segment intersection with the same label for both annotators. To do so, we included *neutral* as another label, even though it was not manually labeled. This intersection procedure caused some small segments to have no assigned label, which usually happens at the annotated segment limits when the onset/offset of a facial expression takes place.

Table 2 reports the number of segments for each category and annotator, as well as the gold standard (*Agreement*),

Table 2 Number of segments annotated with each category label

Annotation	Sad	Angry	Surprised	Happy	Pensive	Other	Neutral	Total
First	0	0	12	234	2033	0	2250	4529
Second	0	1	44	151	2060	3	2245	4504
Agreement	0	0	5	141	1825	0	2382	4353

Table 3 Percentage of each category label with respect to the total amount of annotated time

Annotation	Sad	Angry	Surprised	Happy	Pensive	Other	Neutral
Agreement	0	0	0.01	0.63	11.95	0	87.41

for all annotated videos. Table 3 shows the frequency of each category for the gold standard with respect to the total amount of annotated time. As we can observe, *pensive* is the most frequent manually-labeled expression, appearing 12% of the time, followed by *happy/amused*, present in 6% of the total annotated time. The lack of perceived negative emotions is in line with Sect. 5's results. This suggests that the interaction with the wizard was positive and that the users were engaged in the conversation. Despite such findings, the absence of facial expressions (*neutral*) clearly dominates over all categories. While it is on par with *pensive* with respect to number of segments, participants spent most of the interaction (around 87% of the time) showing no apparent emotion. This is expected, as users had to wait for the system responses for most part of the interaction. However, this is a good sign, as participants could have started to feel angry or sad due to such waiting times, but instead they tended to remain calm.

It is worth noting that, even though the speech and video results follow the same trend, the total annotated time for video is much higher than for speech, due to the fact that speech instances constitute just a fraction of the whole recorded interaction. Therefore, there is more emotional information from video than from audio in a user-wizard interaction. There are indeed some expressions of emotion that can be better perceived from video than from speech. *Pensive*, for instance, which appears frequently right before or while speaking, can only be inferred from the visual modality. However, what we say and how we say it are also informative of our emotional status. Hence, information from speech, language semantics and facial expressions should be combined in a multi-modal manner in order to better understand the emotional status of the user at a given time.

7 Conclusion and future work

This paper analyzed the sessions that a selected set of Spanish seniors carried out to interact with a Wizard of Oz driven, simulated system. A coaching strategy based on the GROW model was used throughout these sessions. In this interaction framework, both the human and the system behavior were analyzed. Regarding the system behavior, the analysis concluded that the wizard was not intended to implement a succession of GROW questions, but rather follow a more natural dialogue structure resembling a bidirectional conversation. Moreover, the sessions with the wizard show that the conversation stayed mainly at the initial two phases of the GROW model, that is, the wizard managed to connect with participants but seemed to have difficulties advancing deeper into the coaching dialogue.

On the other hand, the wizard mood was frequently labelled as positive (60%) when analyzing the language generated to be pronounced by the virtual agent, and neutral the remaining times. A higher percentage of positive content will probably be achieved when the sessions go beyond the G and R phases of the GROW model.

Regarding the user behavior, the probability density function of each dimension (valence, arousal, dominance) of the VAD model show low values of Arousal along with positive values of Valence and neutral values of Dominance, when analyzing the emotional labels extracted from speech. This indicates that users did not achieve high excitation levels when interacting with the Wizard. In fact, the behavior of the virtual coach seems to be appropriate and did not make people feel intimidated. Finally, the valence results show that users have positive feelings with regard to the interaction with the system. In terms of categories, the most frequent label was *calm*, suggesting also that the dialogue system seems to been perceived user friendly.

Regarding the analysis of the emotional labels associated to video segments the most frequent label was *pensive* followed by *happy/amused*. The lack of perceived negative emotions is in line with results obtained from the speech analysis. This also suggests that the interaction with the wizard was positive and that the users were engaged in the conversation. Despite such findings, participants spent most of the interaction (around 87% of the time) showing no apparent emotion, because users had to wait for the system responses for most part of the interaction.

It is worth noting that, even though the speech and video results follow the same trend, the total annotated time for video is much higher than for speech, due to the mute part of videos where users were waiting for the system's interaction. Some expressions of emotion can only be inferred from the visual modality since they are mainly associated to silent participant. Hence, information from speech, language semantics and facial expressions should be combined in a multi-modal manner in order to better understand the emotional status of the user at a given time.

Future work will include the analysis of on going interaction sessions carried out in Norway and France, which will allow a cross-cultural analyses of the user behavior. Moreover, cross-model analysis of the emotional analysis will be carried out in depth by also including the results of the questionnaires in the analysis.

Acknowledgements The research presented in this paper is conducted as part of the project EMPATHIC that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 769872.

European Commission

References

Anna E, Terry A, Marialucia C, M EA, Alda T, Inées TM, Stephan S, Gennaro C (2019) Seniors' acceptance of virtual humanoid agents. In: Leone A, Caroppo A, Rescio G, Diraco G, Siciliano P (eds) Ambient assisted living. Springer International Publishing, Cham, pp 429–443

Brinkschulte L, Mariacher N, Schlögl S, Torres MI, Justo R, Olaso JM, Esposito A, Cordasco G, Chollet G, Glackin C et al (2018) The empathic project: building an expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. In: SMARTER LIVES 2018: digitalisation and quality of life in the ageing society. Universität Inssbrück, pp 36–52

Brooke J (1996) SUS-A quick and dirty usability scale. Usability evaluation in industry. CRC Press, Boca Raton ISBN: 9780748404605

Buendia A, Devillers L (2014) From informative cooperative dialogues to long-term social relation with a robot. Natural interaction with robots, knowbots and smartphones. https://doi.org/10.1007/978-1-4614-8280-2_13

Cabral JP, Kane M, Ahmed Z, Abou-Zleikha M, Székely E, Zahra A, Ogbureke KU, Cahill P, Carson-Berndsen J, Schlögl S (2012) Rapidly testing the interaction model of a pronunciation training system via wizard-of-oz. In: Proceedings of the LREC international conference on language resources and evaluation, Istanbul

Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds) Cognitive behavioural systems. Springer, Berlin

Cavanagh K, Millings A (2013) (Inter)personal computing: the role of the therapeutic relationship in e-mental health. J Contemp Psychother 43(4):197–206. https://doi.org/10.1007/s10879-013-9242-z

Cordasco G, Esposito M, Masucci F, Riviello MT, Esposito A, Chollet G, Schlogl S, Milhorat P, Pelosi G (2014) Assessing voice user interfaces: the vassist system prototype. In: 2014 5th IEEE conference on cognitive infocommunications (CogInfoCom), pp 91–96

Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of oz studies—why and how. Knowl Based Syst 6(4):258–266

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340. http://www.jstor.org/stable/249008

de Graaf M, Ben Allouch S, Klamer T (2015) Sharing a life with harvey: exploring the acceptance of and relationship-building with a social robot. Comput Hum Behav 43:1–14. https://doi.org/10.1016/j.chb.2014.10.030

DeSteno D, Breazeal C, Frank RH, Pizarro D, Baumann J, Dickens L, Lee JJ (2012) Detecting the trustworthiness of novel partners in economic exchange. Psychol Sci 23(12):1549–56

deVelasco M, Justo R, López-Zorrilla A, Torres MI (2019) Can spontaneous emotions be detected from speech on tv political debates? In: Proceedings of 10th IEEE international conference on cognitive infocommunications **(in press)**

Ekman P, Keltner D (1997) Universal facial expressions of emotion. In: Segerstrale U, Molnar P (eds) Nonverbal communication: Where nature meets culture, pp 27–46

El Kaliouby R, Robinson P (2005) Real-time inference of complex mental states from facial expressions and head gestures. In: Real-time vision for human–computer interaction. Springer, Berlin, pp 181–200

Esposito A, Amorese T, Cuciniello M, Esposito AM, Troncone A, Torres MI, Schlögl S, Cordasco G (2018a) Seniors' acceptance of virtual humanoid agents. In: Italian forum of ambient assisted living. Springer, Berlin, pp 429–443

Esposito A, Schlögl S, Amorese T, Esposito A, Torres MI, Masucci F, Cordasco G (2018b) Seniors' sensing of agents' personality from facial expressions. In: Miesenberger K, Kouroupetroglou G (eds) Computers helping people with special needs. Springer International Publishing, Cham, pp 438–442

Esposito A, Amorese T, Cuciniello M, Riviello MT, Esposito AM, Troncone A, Cordasco G (2019a) The dependability of voice on elders' acceptance of humanoid agents. In: Proc. Interspeech 2019, pp 31–35. https://doi.org/10.21437/Interspeech.2019-1734

Esposito A, Amorese T, Cuciniello M, Riviello MT, Esposito AM, Troncone A, Torres MI, Schlögl S, Cordasco G (2019b) Elder user's attitude toward assistive virtual agents: the role of voice and gender. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-019-01423-x

Grant AM (2003) The impact of life coaching on goal attainment, metacognition and mental health. Soc Behav Pers 31(3):253–263

Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1(1):68–99. https://doi.org/10.4018/jse.2010101605

Hassenzahl M (2004) The interplay of beauty, goodness, and usability in interactive products. Hum Comput Interact 19(4):319–349

Hassenzahl M (2008) The interplay of beauty, goodness, and usability in interactive products. Hum Comput Interact 19(4):319–349. https://doi.org/10.1207/s15327051hci1904_2

Irastorza J, Inés Torres M (2019) Tracking the expression of annoyance in call centers. Springer International Publishing, Cham, pp 131–151. https://doi.org/10.1007/978-3-319-95996-2_7,

Justo R, Saz O, Guijarrubia V, Miguel A, Torres MI, Lleida E (2008) Improving dialogue systems in a home automation environment. In: Proceedings of the 1st international conference on ambient media and systems. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Ambi-Sys '08, pp 2:1–2:6. http://dl.acm.org/citation.cfm?id=1363163.1363165

Justo R, Manso JI, Pérez S, Torres MI (2018) Bi-modal annoyance level detection from speech and text. Procesamiento del Lenguaje Natural 61:83–89. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5647

Kirkwood TB (2005) Understanding the odd science of aging. Cell 120(4):437–447

Milhorat P, Schlögl S, Chollet G, Boudy J (2013) What if everyone could do it?: A framework for easier spoken dialog system design. In: Proceedings of the 5th ACM SIGCHI symposium on engineering interactive computing systems. ACM, New York, EICS '13, pp 217–222. https://doi.org/10.1145/2494603.2480325

Montenegro C, López Zorrilla A, Mikel Olaso J, Santana R, Justo R, Lozano JA, Torres MI (2019) A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. Multimodal Technol Interact 3(3):52. https://doi.org/10.3390/mti3030052

Parker SG, Hawley MS (2013) Telecare for an ageing population? Age Ageing 42(4):424–425. https://doi.org/10.1093/ageing/aft056. http://oup.prod.sis.lan/ageing/article-pdf/42/4/424/28849/aft056.pdf

Passmore J (2011) Motivational interviewing—a model for coaching psychology practice. Coach Psychol 7(1):35–39

Passmore J (2012) An integrated model of goal-focused coaching: an evidence-based framework for teaching and practice. Int Coach Psychol Rev 7(2):146–165

Pedersen T, Johansen C, Jøsang A (2018) Behavioural computer science: an agenda for combining modelling of human and system behaviours. Hum Centric Comput Inf Sci 8(1):7. https://doi.org/10.1186/s13673-018-0130-0

Rozin P, Cohen AB (2003) High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans. Emotion 3(1):68

Sansen H, Chollet G, Glackin C, Badii A, Torres MI, Petrovska-Delacrétaz D, Schlögl S, Boudy J (2016) The Roberta IRONSIDE Project: a humanoid personal assistant in a wheelchair for dependent persons. In: Proceedings of the ATSIP international conference on advanced technologies for signal and image processing, Monastir. https://doi.org/10.1109/ATSIP.2016.7523110

Sayas S (2018a) Dialogues on leisure and free time. Tech. Rep. DP3, Empathic project

Sayas S (2018b) Dialogues on nutrition. Tech. Rep. DP1, Empathic project

Sayas S (2018c) Dialogues on physical exercise. Tech. Rep. DP2, Empathic project

Schlögl S, Doherty G, Karamanis N, Luz S (2010a) Webwoz: a wizard of oz prototyping framework. In: Proceedings of the 2nd ACM SIGCHI symposium on engineering interactive computing systems. ACM, New York, EICS '10, pp 109–114. https://doi.org/10.1145/1822018.1822035

Schlögl S, Doherty G, Karamanis N, Scheider A, Luz S (2010b) Observing the wizard: in search of a generic interface for wizard of oz studies. In: Proceedings of the Irish HCI conference, Dublin, Ireland, pp 43–50

Schlögl S, Schneider A, Luz S, Doherty G (2011) Supporting the wizard: interface improvements in wizard of oz studies. In: Proceedings of the BCS HCI conference on human–computer interaction, Newcastle

Schlögl S, Doherty G, Luz S (2015) Wizard of oz experimentation for language technology applications: challenges and tools. Interact Comput 27(6):592–615. https://doi.org/10.1093/iwc/iwu016

Seiki ST, Tamamizu K, Saiki S, Nakamura M, Yasuda K (2017) Virtualcaregiver: personalized smart elderly care. Int J Softw Innov (IJSI) 5:1–14

Siegert I, Hartmann K, Philippou-Hübner D, Wendemuth A (2013) Human behaviour in hci: Complex emotion detection through sparse speech features. In: Proceedings of 4th international workshop on human behavior understanding, vol 8212. Springer, New York, pp 246–257. https://doi.org/10.1007/978-3-319-02714-2_21

Torres MI, Olaso JM, Glackin N, Justo R, Chollet G (2019a) A spoken dialogue system for the empathic virtual coach. In: D'Haro LF, Banchs RE, Li H (eds) 9th International workshop on spoken dialogue system technology. Springer Singapore, Singapore, pp 259–265

Torres MI, Olaso JM, Montenegro C, Santana R, Vázquez A, Justo R, Lozano JA, Schlögl S, Chollet G, Dugan N, Irvine M, Glackin N, Pickard C, Esposito A, Cordasco G, Troncone A, Petrovska-Delacretaz D, Mtibaa A, Hmani MA, Korsnes MS, Martinussen LJ, Escalera S, Cantariño CP, Deroo O, Gordeeva O, Tenorio-Laranga J, Gonzalez-Fraile E, Fernandez-Ruanova B, Gonzalez-Pinto A (2019b) The empathic project: mid-term achievements. In: Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments, ACM, New York, PETRA '19, pp 629–638. https://doi.org/10.1145/3316782.3322764

Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, Cowie R, Pantic M (2014) Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th international workshop on audio/visual emotion challenge. ACM, New York, AVEC '14, pp 3–10

Vidrascu L (2007a) Analyse et détection des émotions verbales dans les interactions orales. Ph.D. thesis, Paris11 University

Vidrascu L (2007b) Analysis and detection of emotions in real-life spontaneous speech. Theses, Université Paris Sud-Paris XI. https://tel.archives-ouvertes.fr/tel-00624085

Whitemore J (2009) Coaching for performance: growing human potential and purpose: the principles and practice of coaching and leadership. Nicholas Brealey Publishing, London

Whitmore J (2010) Coaching for performance: growing human potential and purpose: the principles and practice of coaching and leadership. People skills for professionals. Nicholas Brealey Publishing. https://books.google.es/books?id=eTZiP_8dqIYC

Willcox DC, Scapagnini G, Willcox BJ (2014) Healthy aging diets other than the Mediterranean: a focus on the okinawan diet. Mech Ageing Dev 136–137:148–162

Yesavage JA, Brink T, Rose TL, Lum O, Huang V, Adey M, Leirer VO (1982) Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 17(1):37–49

Publication 9

Euskarazko elkarrizketa sistema automatikoa sare neuronalen bidez

Ekaia, 2020, 37, 327-341
https://doi.org/10.1387/ekaia.20987

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Euskarazko elkarrizketa sistema automatikoa sare neuronalen bidez

(A neural dialogue system in Basque)

Asier López Zorrilla, Mikel de Velasco Vázquez, Raquel Justo*

Elektrizitate eta Elektronika Saila, Euskal Herriko Unibertsitatea (UPV/EHU), Leioa

LABURPENA: Lan honetan sare neuronalen bidez euskaraz hitz egiten ikasten duen elkarrizketa sistema automatiko bat aurkezten dugu. Horretarako, Turingen testaren ideia era konputazionalean inplementatzen duten sare neuronal sortzaile aurkariak erabili ditugu. Normalean erabiltzen diren ingelesezko corpusak baino bi magnitude ordena txikiagoa den euskarazko corpus batekin halako sareak doitzea badagoela frogatzen dugu. Amaitzeko, euskararen morfologia kontuan hartzen duen aurreprozesamendua erabiltzea komenigarria dela erakusten dugu. Sare neuronaletan oinarrituta dagoen euskarazko lehen elkarrizketa sistema aurkezten dugu.

HITZ GAKOAK: elkarrizketa sistema automatikoak, sare neuronalak, sare neuronal sortzaile aurkariak, euskara.

AbstrACt: *This work presents a neural dialogue system capable of learning Basque. To this end, we build upon generative adversarial networks which implement the idea of the Turing test. We demonstrate that training such a dialogue system with corpora two orders of magnitude smaller than usual English corpora is feasible. Finally, we also found that preprocessing the Basque language according to its morphology helps training these neural models. To the best of our knowledge, this is the first attempt to develop a neural dialogue system in Basque.*

KEYWORDS: dialogue systems, deep learning, generative adversarial networks, Basque language.

* ***Harremanetan jartzeko / Corresponding author:*** Asier López Zorrilla. Speech Interactive Research Group, Elektrizitate eta Elektronika Saila, Zientzia eta Teknologia Fakultatea, Euskal Herriko Unibertsitatea (UPV/EHU), Sarriena Auzoa, zg, 48940, Leioa, Bizkaia, Euskal Herria. – asier.lopezz@ehu.eus – https://orcid.org/0000-0003-1739-1397.

Nola aipatu / How to cite: Eseberri, Itziar; López Zorrilla, Asier; De Velasco Vázquez, Mikel; Justo, Raquel (2020). «Euskarazko elkarrizketa sistema automatikoa sare neuronalen bidez»; *Ekaia*, 37, 2020, 327-341. (https://doi.org/10.1387/ekaia.20987).

Jasoa: 08 uztaila, 2019; Onartua: 29 urria, 2019.

ISSN 0214-9001 - eISSN 2444-3255 / © 2020 UPV/EHU

1. SARRERA

Elkarrizketa sistema automatikoek pertsona eta makinen arteko komunikazioa eta interakzioa ahalbidetzen dute, lengoaia naturalaren bidez. Adibide moduan azken urteotan hedatu diren laguntzaile birtualak aipa ditzakegu, hala nola Siri, Cortana, Google Assistant edo Alexa. Horiek normalean elkarrizketa sistema helburuduntzat hartzen dira, haien lana erabiltzailearen aginduak burutzea baita; esate baterako, dei bat egitea edo Interneten biharko eguraldiaren iragarpena bilatzea. Horietaz gain, busen ordutegiak eta lineak kontsultatzeko [1] eta jatetxeetan edo hoteletan erreserbak egiteko [2] balio duten sistemak helburudunak dira ere.

Beste alde batetik, aldez aurretik definitutako helbururik edo gairik ez duten elkarrizketa sistemak ere badira: eremu irekikoak. Sistema horietan erabiltzaileak eta makinak ez diote elkarri hitz egiten helburu espezifiko batekin; interakzioa bera naturala eta zentzuduna izatea da helburua. Horretarako, sistemak esaldi ahal bezain logiko, koherente eta informatzaileenekin erantzun behar dio erabiltzaileak esaten duenari. Beste modu batean esanda, sistemak era gizatiarrean hitz egin behar du. Lan honetan elkarrizketa sistema mota horietan zentratuko gara.

Era gizatiarrean hitz egitearen ideiarekin lotuta, Alan Turing matematikariak 1950. urtean bere test famatua aurkeztu zuen: Turingen testa [3]. Testaren ideia nagusia honakoa da: sistema automatiko bat kalitatezkoa edo adimenduna dela esateko, sistema horrek eta pertsona batek bereizezinak izan behar dute haiekin hitz egiteko orduan. Sistema batek halako propietatea betetzen duen egiaztatzeko, Turingek hainbat epaile zenbait makinarekin hitz egiten jartzea proposatu zuen, makina batzuen atzean sistema automatikoak eta besteen atzean pertsonak daudelarik. Egoera horretan epaileek proportzio handi[1] batean usteko balute sistema automatikoa pertsona bat dela, orduan sistema hori erabat adimenduna dela esan liteke.

Denbora pasatu ahala, Turingen testa gainditzearen ideiak gero eta ikerketa gehiago bultzatu zituen adimen artifizialaren arloan. Adibidez, 1966. urtean ELIZA programa [4] aurkeztu zuten MIT-eko ikertzaileek. Programaren funtsa hitz gakoak detektatzean eta horien arabera aurredefinitutako esaldi bat aukeratzean datza. Algoritmo hori sinplea izan arren, hainbat epailek pertsonatzat hartzea lortu zuen.

Hurrengo hamarkadetan ikerketek aurrera jarraitu zuten arren, benetan Turingen testa gainditzeko gai zen sistemarik ez zen lortu. 2011. urtean

[1] Eztabaida handia dago sistema batek Turingen testa gainditzeko behar duen portzentajearen inguruan. Erreferentzia gisa, 2011. urtean Indian Institute of Technology Guwahati institutuan ospatutako Turingen test batean, epaileek %63,3n pertsonak pertsona moduan sailkatu zituzten.

gauzak aldatu ziren, Turingen test batean inoiz lortutako emaitzarik onenak Cleverbot sistemak[2] lortu zituenean; berarekin hitz egin zuten 1.334 epaileetatik % 59,3-ak pertsonatzat hartu zuen. ELIZA-k ez bezala, Cleverbot-ek ez ditu aurredefinitutako esaldiak erabiltzen. Horren ordez urteetan zehar pertsonekin edukitako elkarrizketak erabiltzen ditu erantzuterako orduan. Hitz gutxitan esanda, esaldi bati erantzuteko Cleverbot-ek esaldi hori edo antzeko bat esan duenean zein erantzun jaso duen bilatzen du, eta erantzun horretatk abiatuta sortzen du bere erantzuna. Ideia hori interesagarria bada ere, konputazionalki nahiko garestia da, denbora zein memoriaren ikuspegitik, datu-base oso handi batean bilaketak egitea baitakar. Are gehiago, datu-basea zenbat eta handiagoa izan, orduan eta denbora eta memoria gehiago beharko du halako sistema batek erantzun bat sortzeko.

Eragozpen horiek saihesteko, baita adimen artifizialaren beste arloetan izan duten emaitzengatik ere, azken urteetan sare neuronalak elkarrizketa sistemak eraikitzeko teknologia nagusia bilakatu dira. Sare neuronalak datuetatik eredu konputazional konplexuak lortzeko balio duten paradigma konputazional bat dira, bereziki eraginkorra datuen kantitatea oso handia denean. Ulertzekoa da, beraz, arloko autore gehienek ingelesez dauden datu-baseekin lan egitea, horiek izanda baitira handienak, eta, hortaz, sare neuronalek hobeto funtzionatuko dutelako. Baina zer gertatzen da baliabide gutxiagoko hizkuntzekin? Ba al dago sare neuronaletan oinarrituriko elkarrizketa sistema automatikoak eraikitzerik euskaraz?

Lan honetan erakusten dugu baietz, badagoela. Normalean erabiltzen diren datu-baseak baino bi magnitude ordena txikiagoko datu-baseak erabiliz modu koherente eta zentzudunean euskaraz hitz egiten duen elkarrizketa sistema automatikoa aurkezten dugu.

2. ARLOKO EGOERA ETA IKERKETAREN HELBURUAK

Sare neuronalen bidezko eremu irekiko elkarrizketa sistemak itzulpen automatikorako erabiltzen diren sareetan oinarritzen dira, hots, sekuentziatik sekuentziarako sareetan [5, 6] (*Sequence to sequence networks* ingelesez). Sare neuronal horiek luzera arbitrarioko bektore segida bat har dezakete sarrera moduan, eta era berean beste luzera arbitrario bateko segida bat sortu. Hala, transdukzio problemak ebazten saiatzeko baliagarriak dira. Itzulpen automatikoaren kasuan, sarreran hizkuntza batean idatzitako esaldia hartuko du sareak, eta irteeran esaldi hori beste hizkuntza batean sortu. Bestalde, elkarrizketa sistemak eraikitzerako orduan, sarrera erabiltzaileak esandako hitzen segida izango da, eta irteera sistemaren erantzunari dagozkion hitzen sekuentzia.

[2] https://www.cleverbot.com/, azken bisita 2019ko uztailaren 1ean.

https://doi.org/10.1387/ekaia.20987

Sare horiek entrenatzeko edo haien parametroak doitzeko, ikasketa metodo gainbegiratuak erabili ohi dira, aipatutako sarrera-irteera bikoteez osaturiko corpusen bat erabiliz. Adibidez, lan honetan filmen azpitituluak erabiliko ditugu corpusa eratzeko: sarrera bakoitza aktore batek esandako esaldi bat izango da, eta dagokion irteera beste aktore batek emandako erantzuna.

Metodologia hori erabiliz emaitza interesgarriak lortu ahal diren arren, askotan horrela entrenatutako sareek informaziorik gabeko erantzun orokorrak sortzeko joera dute, hala nola «*I don't know* » edo «*I'm sorry* » [7, 8]. [9] lanean adierazten den moduan, ikasketa metodo gainbegiratuek irteera bakarra esleitzen diote sarrera bakoitzari, baina horrek ez ditu elkarrizketen propietateak behar bezala jasotzen. Izatez, hitz egiten dugunean, norbaitek esan duenari erantzuteko hamaika esaldi erabili ahalko genituzke, guztiak onargarriak. Horrela, esaldi askoren erantzuna izan daitezkeen esaldi generikoak probabilitate handiarekin sortzen ditu sareak.

Arazo hori konpontzeko, ikasketa gainbegiratuaren ordez sare sortzaile aurkariak (*Generative adversarial networks* ingelesez) [10] erabiliko ditugu lan honetan. Sare sortzaile aurkariek Turingen testaren ideia era konputazionalean aplikatzea ahalbidetzen dute. Kasu honetan, erantzunak sortzen dituen sareari (sare sortzailea hemendik aurrera) ez zaio adieraziko zer irteera dagokion sarrera bakoitzari. Horren ordez, beste sare batek, sare diskriminatzaileak, ebaluatuko ditu sare sortzaileak emandako erantzunak, zein punturaino diren gizatiarrak esanez, Turingen testaren epaile batek egingo lukeen modu berean. Sare sortzailearen helburua sare diskriminatzaileak berari emandako ebaluazioa ahal bezainbeste hobetzea izango da. Sare diskriminatzailearena, aldiz, pertsonek sortutako eta sare sortzaileak sortutako esaldien artean bereiztea izango da. Modu horretan, bi sareak iteratiboki entrenatuko dira; sortzailea saiatuko da diskriminatzaileak hura pertsonatzat har dezan, diskriminatzaileak sortzailearen eta pertsonen artean bereizten ikasten duen bitartean.

Halako optimizazio prozesua egitea, dena den, ez da sinplea, sareak entrenatzeko normalean erabiltzen diren gradienteetan oinarritutako optimizazio metodoak ez baitira zuzenean aplikagarriak. Xehetasunetan sartu gabe, sare diskriminatzailearen irteera ez da diferentziagarria sare sortzailearen parametroekiko, sortzaileak sortutako hitzak diskretuak dira eta [11]. Errefortzu bidezko ikasketa erabili daiteke gradienteetan oinarritutako metodoen ordez [12, 13], baina horrek entrenamenduaren konbergentzia zaildu dezake [14]. Beste aukera bat *straight-through Gumbel-softmax* [15, 16] zenbateslearen bidez gradientearen hurbilketa bat egitea da, [17] eta [18] laneetan erakusten duten moduan. Azkenik, lan honetako autoreek guztiz diferentziagarria den sare sortzaile aurkari bat aurkeztu berri dute [19], hitzen errepresentazio bektorial hurbilduak erabiltzen dituena, ondoren azalduko dugun moduan.

Testuinguru horretan, lan honen ekarpenak hiru dira: alde batetik, [19] lanean proposatutako sare sortzaile aurkaria balioztatzen dugu, ingelesez ez ezik euskaraz ere eraginkorra dela frogatuz; bigarrenik, modu koherente eta zentzudunean hitz egiten duen sare neuronaletan oinarritutako elkarrizketa sistema automatikoa euskaraz eraikitzea badagoela frogatzen dugu; eta, amaitzeko, lematizazio prozesu baten bidez corpusaren tamaina txikiagoagatik sortutako desabantailak nola leundu daitezkeen erakusten dugu.

3. atalean, proposatutako elkarrizketa sistema osatzen duten sareen egituraren deskribapena emango dugu. Hurrena, 4. atalean sare horien parametroak doitzeko egingo saiakuntzak eta lortutako emaitzak erakutsiko ditugu, eta, 7. atalean, ondorio batzuk eta etorkizunerako planteatzen den norabidea aipatuko ditugu.

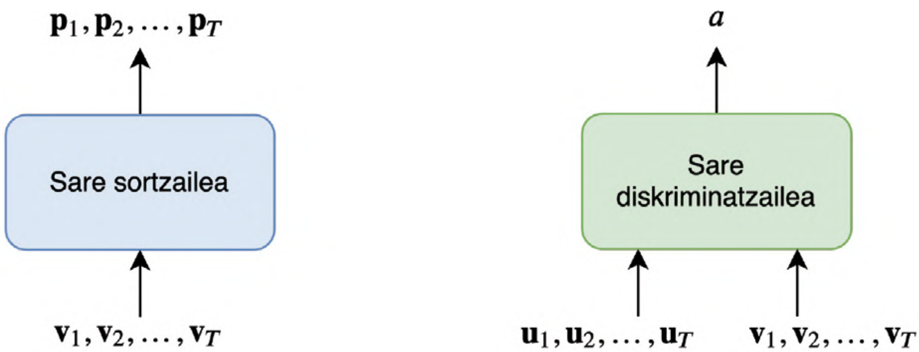
3. SARE SORTZAILE ETA DISKRIMINATZAILEA

Esan bezala, elkarrizketa sistema eraikitzeko sare sortzaile aurkariak erabili ditugu. Beraz, bi sare entrenatu behar dira elkarrizketa sistema lortzeko: sare sortzailea, sarrera mezu baten aurrean erantzuna sortzen duena; eta sare diskriminatzailea, sarrera mezu baten aurrean emandako erantzun bat ebaluatzen duena.

Sare sortzailea sekuentziatik sekuentziarako sare bat da, *long short-term memory* edo LSTM [20] kodetzaile eta deskodetzaile errekurrente independenteekin [5] eta arreta modulu batekin [21, 22]. Sare horrek T luzera arbitrarioko hitzen errepresentazio bektorialen [23] segida bat hartuko du sarrera moduan: $\mathbf{v} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T$. Sarrera hori prozesatu ostean, irteera moduan beste τ luzera arbitrarioko segida bat bueltatuko du, elementu bakoitza sareak sor ditzakeen hitz guztien arteko probabilitate-banaketa izanik: $\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_\tau$.

Bestalde, sare diskriminatzailea beste bi kodetzaile errekurrentez osaturik dago, guztiz konektatutako geruza batzuez jarraituak, [24] lanean azaltzen den antzera. Kodetzaile bakoitzak esaldi bat hartzen du sarrera moduan. Batek $\mathbf{v}$, erabiltzailearen mezua, prozesatuko du, eta besteak $\mathbf{u}$, erantzuna. Sistemaren irteera 0 eta 1-en arteko zenbaki erreal bat, a, izango da, erabiltzailearen mezuari emandako erantzuna gizatiarra zer punturaino den adierazten duena. Irteera zenbat eta baxuagoa, orduan eta gizatiarragoa izango da erantzuna, sarearen irizpidearen arabera. Bi sarrerak, berriz ere, hitzen errepresentazio bektorialen moduan hartuko ditu sareak.

Bi sareen sarrerak eta irteerak 1. irudian erakusten ditugu. Gainontzeko xehetasunak [19] lanean aurki daitezke.



1. irudia. Sare sortzailearen eta diskriminatzailearen sarrerak eta irteerak.

4. IKASKETA ALGORITMOA

Bi sareak entrenatzeko, hiru optimizazio prozesu era iteratiboan egingo ditugu. Lehen aipatu dugun moduan, alde batetik, sare sortzailea entrenatuko dugu diskriminatzaileak hura pertsonatzat har dezan, hau da, diskriminatzailearen irteera minimizatzeko. Bigarrenik, diskriminatzailea entrenatuko dugu sare sortzaileak sortutako erantzunak eta corpusetik hartutako erantzunak desberdintzeko. Amaitzeko, [12] lanean proposatzen den legez, sare sortzailearen parametroak ikasketa metodo gainbegiratuen bidez doituko ditugu ere noizean behin, prozedura orokorraren konbergentzia bermatzeko.

Optimizazio prozesu horiek definitzeko, horietako bakoitzean gradienteetan oinarritutako optimizazio metodoekin minimizatuko ditugun galera-funtzioak zehaztuko ditugu.

4.1. Sare sortzailearen parametroen egiantz handieneko zenbatespena

Sare sortzailearen parametroak ikasketa gainbegiratuaren bidez doituko ditugu egiantz handieneko zenbatespen baten bidez. Hau da, corpuseko sarrera-irteera bikote bakoitzarentzat, sareak sarrera prozesatzean irteera desiratua sortzeko duen probabilitatea maximizatuko dugu. 1. ekuazioan agertzen den galera-funtzioa erabiliko dugu.

$$L_{EH} = \frac{1}{|C|} \sum_{\mathbf{v}, s \in C} \frac{1}{|s|} \sum_{t=1}^{|s|} -\log \mathbf{p}_t[s_t] \tag{1}$$

non C $\mathbf{v}$ sarrerek eta s irteera desiratuek osatutako corpusa den, s_t irteera desiratuaren t-garren hitzari dagokion indizea den, eta $\mathbf{p}_t$ $[s_t]$ sareak t-garren denbora unean *st* hitzari esleitutako probabilitatea den. Sarearen irteera, **p**, **v** sarreraren menpe dago noski, baina mendekotasun hori ez dugu esplizituki adierazi notazioa ez korapilatzeko.

4.2. Sare diskriminatzailearen galera-funtzioa

Sare diskriminatzailearen entrenamendua egiteko lehenik eta behin corpus berri bat sortu beharko dugu, C_D, C corpusetik abiatuz eta sare sortzailea erabiliz. Diskriminatzaileak pertsonek emandako eta sare sortzaileak sortutako erantzunak desberdintzen ikasi behar duenez, bi eratako laginak behar ditu bereb artean diskriminatu ahal izateko. Horretarako, bi motatako hirukoteez osatuko dugu C_D corpusa. Hirukote bakoitza erabiltzaileak bidalitako mezu batez, erantzun batez eta 0 edo 1 izan daitekeen etiketa batez osatuta egongo da. Lehenengo motako hirukoteek gizakiek emandako erantzunak edukiko dituzte, eta, beraz, etiketa 0 izango da. Hirukote horiek lortzeko C corpuseko bikoteak erabili ditugu zuzenean. Bestalde, bigarren motako hirukoteek sare sortzaileak sortutako erantzunak edukiko ditu, eta, ondorioz, etiketaren balioa 1 izango da. Hirukote horiek eratzeko, C corpusetik hartu dira erabiltzailearen mezuak, gero horiek sare sortzaileari pasatu sarrera moduan, eta sarearen irteera erabili erantzun moduan. C_D eraiki ondoren, entropia gurutzatuko galera-funtzioa erabili dugu sare diskriminatzailearen parametroak doitzeko:

$$L_D = \frac{1}{|C_D|} \sum_{\mathbf{v},\mathbf{u},l \in C_D} -\left[l \cdot \log a + (1-l) \cdot \log(1-a) \right], \tag{2}$$

non **v** erabiltzailearen mezuaren hitzen errepresentazio bektorialen segida den, **u** erantzunarena, l erantzuna pertsona batena edo sare sortzailearena den adierazten duen eskalarra, eta a diskriminatzailearen irteera. Berriro ere, a-k **v**-rekiko eta **u**-rekiko duen mendekotasuna ez dugu esplizituki adierazi.

4.3. Sare sortzailearen galera-funtzio aurkaria

Azkenik, sare sortzailea diskriminatzailearen irteera minimizatzeko galera-funtzioa definitzea erraza da, diskriminatzailearen irteera bera baita, 3. ekuazioan ageri den bezala.

$$L_S = \frac{1}{|C_S|} \sum_{\mathbf{v} \in C_S} a, \tag{3}$$

non C_S corpusa C corpusean dauden sarrera mezuez osatuta dagoen, **v** horietako bakoitza izanik. a diskriminatzailearen irteera da.

https://doi.org/10.1387/ekaia.20987

3. ekuazioko galera-funtzioa gradienteetan oinarritutako optimizazio metodoekin minimizatu ahal izateko, a sare sortzailearen parametroekiko diferentziagarria izan behar du. Sare sortzaileak $\mathbf{v}$ sarrera $\mathbf{p}$ irteeran era guztiz diferentzialean transformatzen du. Era berean, sare diskriminatzaileak bere bi sarrerak, $\mathbf{v}$ eta $\mathbf{u}$, era guztiz diferentzialean transformatzen ditu a irteeran. Hortaz, diferentziagarritasuna ez galtzeko, $\mathbf{p}$ $\mathbf{u}$-n transformatu behar da transformazio diferentziagarri baten bidez. $\mathbf{p}$-ko elementu bakoitza, hots, $\mathbf{p}_t$, sareak esan ditzakeen hitz guztien arteko probabilitate-banaketa bat da. Normalean $\mathbf{p}_t$-ko maximoaren argumentua hartuko genuke sareak t-garren denbora unean esan duen hitzat. Baina argmax operazioa ez da deribagarria.

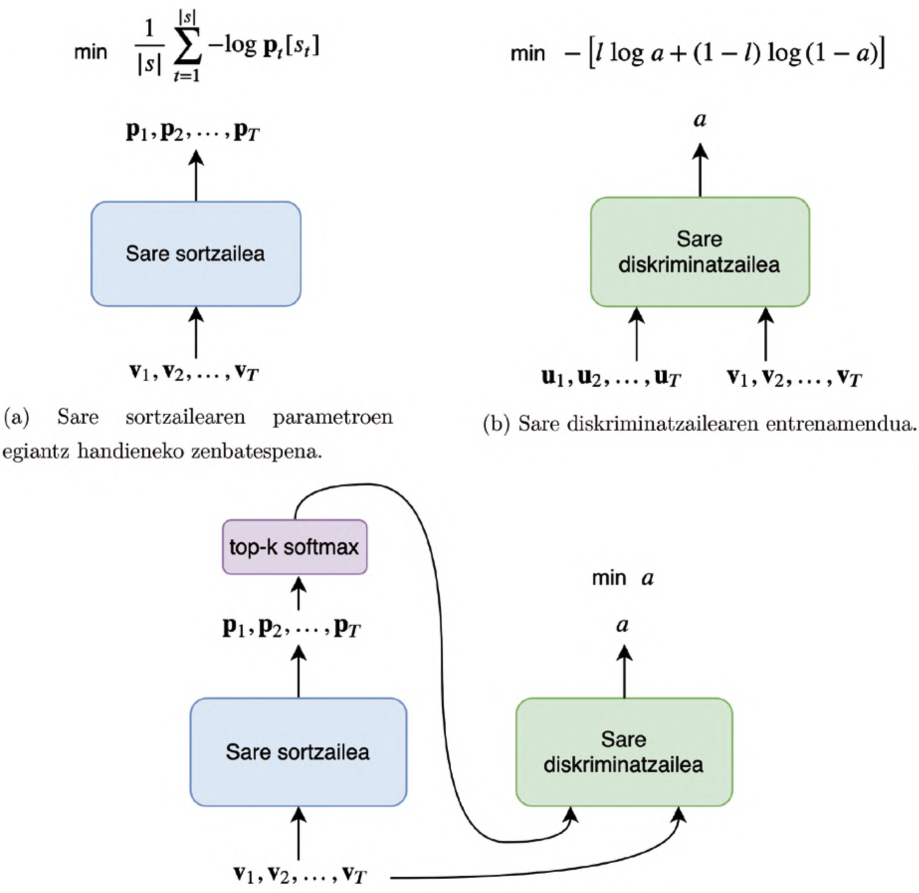
Arazo horri irtenbidea emateko, [19] laneko prozedura berdina erabiltzen dugu lan honetan. $\mathbf{p}_t$-ri dagokion errepresentazio bektoriala, $\mathbf{u}_t$, lortzeko, $\mathbf{p}_t$-ko k elementurik handienak hartzen ditugu, *top-k* operazio baten bidez. Horrela, elementu horien $\tilde{\mathbf{p}}_t$ balioak eta $\mathbf{k}_t$ indizeak lortzen ditugu. Jarraian, $\tilde{\mathbf{p}}_t$ normalizatzen dugu *softmax* normalizazio batekin, eta $\tilde{\mathbf{p}}_t t$ lortu. Azkenik, $\mathbf{u}_t$ kalkula dezakegu $\mathbf{k}_t$ indizeei dagozkien hitzen errepresentazio bektorialen arteko batazbesteko aritmetiko haztatua eginez, pisuak $\hat{\mathbf{p}}_t$ izanik.

4.4. **Goi mailako optimizazio algoritmoaren deskribapena**

Erabiliko diren hiru galera-funtzioak deskribatu ondoren, horiek iteratiboki minimizatzeko prozedura zehaztuko dugu. Hasteko, sare sortzailearen parametroak ez ditugu ausaz hasieratuko. Horren ordez, hainbat iteraziotan zehar doituko ditugu hasieran, 1. ekuazioko egiantz handieneko galera-funtzioa minimizatuz. Behin sare sortzaileak kalitate onargarriko esaldiak sortuz gero, C_D corpusa bere erantzunekin eta C-ko pertsonen erantzunekin hasieratuko dugu, eta sare diskriminatzailea lehenengo aldiz entrenatuko dugu.

Ondoren, algoritmoaren begizta nagusia hasten da. Horretan, sare sortzailea eta diskriminatzailea iteratiboki entrenatzen dira. Sare sortzailea entrenatzeko galera-funtzio aurkaria (3. ekuazioa) eta egiantz handieneko galera-funtzioak (1. ekuazioa) txandakatzen dira. Prozedura osoan zehar, sare sortzailearen entrenamendu prozesu bakoitza amaitu ostean, hainbat sarrera ausaz aukeratzen dira, C-tik eta sare sortzaileak sortutako erantzunak C_D-ra gehitzen dira, eta diskriminatzailea entrenatzen da hainbat iteraziotan zehar (2. ekuazioa). Prozeduraren konbergentzia bermatzeko, diskriminatzailea entrenatzerakoan probabilitate handiagoarekin hartzen dira C_D-n sartutako erantzun berriagoak.

2. irudian algoritmoan egindako hiru optimizazio problemen adierazpen grafikoa erakusten dugu. Kasu bakoitzean optimizatzen den galera funtzioa agertzen da, sarrera-irteera bikote bakarrarentzat.



(a) Sare sortzailearen parametroen egiantz handieneko zenbatespena.

(b) Sare diskriminatzailearen entrenamendua.

(d) Sare sortzailearen entrenamendu aurkaria.

2. irudia. Entrenamendu algoritmoan egiten diren optimizazio prozesuen laburpena.

5. EUSKARAREN AURREPROZESAMENDUA ETA LEMATIZAZIOA

Orain, euskaraz dagoen corpusa nola aurreprozesatu dugun deskribatuko dugu. Esan dugunez, euskarazko corpus batekin entrenatuko ditugu sareak. Ingelesa ez bezala, euskara hizkuntza eranskaria da egitura morfologikoaren aldetik. Hau da, euskarak monema independenteak elkartuz sortzen ditu hitzak. Horrela, askotan euskaraz hitz batekin esan daitekeena ingelesez hainbat hitz erabiliz adierazi behar da. Adibidez, ingelesezko «*to the cinema*» euskaraz «zinemara» itzuliko litzateke, edo «*because of the*

https://doi.org/10.1387/ekaia.20987

baby» «haurrarengatik». Sareen ikuspegitik hitz bakoitza token independente bat denez, sareak ez ditu ikusten euskaraz gertatzen diren hitzen arteko erlazioak, eta horrek euskararen prozesamendu automatikoa zailtzen du. Hasiera batean, behintzat, sarearentzat «haurrarengatik» eta «haurraren» hitzak «haurrarengatik» eta «daitezke» bezain ezberdinak dira.

Honek hitzen errepresentazio numerikoa zailtzen du bi sareen sarreran, baita sare sortzailearen irteeran ere. Sareen sarreretan, hitzen egituran arreta jartzen duten errepresentazio bektorialak erabiliko ditugu hitzen arteko erlazio horiek sortzeko, *Fastext* [25] hain zuzen ere. Dena den, irteeran ezin da arazoa horrela konpondu, sare sortzaileak hitzen arteko probabilitate-banaketa bat sortzen duelako. Horri irteera ematen saiatzeko, hitzen lexemak kasu marketatik eta postposizioetatik banatzea proposatzen dugu. Zehazki, izen, izenordain, adjektibo eta determinanteak bananduko ditugu lan honetan.

Hitzen lexema eta kategoria gramatikala topatzeko, [26] lanean aurkeztutako kode irekiko lematizatzailea erabiliko dugu. Izen, izenordain, adjektibo edo determinante baten lexema eta postposizioa banatuko diren ala ez erabakitzeko, baldintza sinple bat erabiliko da. Halako hitz baten bukaera postposizio baten berdina balitz, orduan hitzak lexematan eta postposiziotan bananduko dugu. Adibidez, «zeruko» hitza «zeru» lexeman eta «-ko» postposizioan banatuko genuke. Hogeita bi postposizio eta kasu marka hartu genituen kontuan: «-ri», «-ei», «-rekin», «-ekin», «-ren», «-en», «-n», «-tik», «-dik», «-rik», «-ra», «-tara», «-rengana», «-engana», «-rantz», «-raino», «-z», «-rako», «-ko», «-entzat», «-tzat» eta «-gatik».

Lematizazioaz gain, izen propioak <izen> tokenera bihurtuko ditugu, normalean pertsonen izenak baitira, eta, beraz, funtzio berdina dutelako esaldietan. Era berean, zenbakiak <zenbaki> tokenera bihurtuko dira [26] laneko lematizatzailea erabili genuen bi ataza hauetarako ere.

6. **SAIAKUNTZAK ETA EMAITZAK**

Orain arte azaldutako sareak, ikasketa algoritmoa eta euskararen aurreprozesamendua balioztatzeko, OpenSubtitles [27] corpusaren euskarazko bertsioarekin entrenatu dugu deskribatutako elkarrizketa sistema. Corpus horretatik milioi bat sarrera-irteera bikote atera ditugu, ingelesezko bertsioan baino 420 aldiz gutxiago. Corpusa 5. atalean azaldutako metodologiarekin aurreprozesatu dugu; ondorioz hitz desberdinen kopurua berrehun milatik ehun milara jaitsi da. Normalean egiten den bezala, hitz horietako azpimultzo bat baino ez dugu kontuan hartuko saiakuntzetarako: maiztasun handieneko 15.000 hitzak. Gainontzekoak corpusetik kendu dira. Aurreprozesamenduaren efektua erakusteko, corpus aurreprozesatua zein aurreprozesatu gabearekin entrenatu ditugu sareak.

Kasu bietan, dena den, hiper-parametro berdinak erabiliko ditugu sareen arkitekturan eta baita ikasketa algoritmoan. Hiper-parametro horietatik inportanteenak jarraian aipatzen ditugu. Sare errekurrente guztiak, hau da, sare sortzailearen kodetzailea, deskodetzailea, eta sare diskriminatzailearen bi kodetzaileak, bi LSTM geruzaz osatuta daude. Sare sortzailearen geruzek 1.028 zelda dituzte, eta diskriminatzailearenak 128. Adam optimizazio metodoa [28] erabiliko dugu 4. ataleko hiru galera-funtzioak minimizatzeko, 512 tamainako *batch*-ak erabiliz. Hitzen errepresentazio bektorialak *Fastext* metodologiarekin hasieratuko dira, eta entrenamenduan zehar optimizatuko dira. Sare sortzailea 50.000 iteraziotan zehar entrenatuko dugu, ikasketa begizta hasi baino lehen. Hori 500 aldiz errepikatu dugu ondoren. Iterazio bakoitzean sare sortzailea zein diskriminatzailea 40 iterazioetan zehar entrenatuko da.

[29] lanean adierazten den moduan, ebaluazio automatikoak ez dira komenigarriak elkarrizketa sistemen kalitatea neurtzeko, normalean ez baitago korrelazio nabaririk horien eta pertsonek egindako ebaluazioen artean. Hortaz, entrenatutako sistemen funtzionamendua erakusteko, zenbait sarrerako mezuen aurrean emandako erantzuna erakusten dugu emaitza modura. Erreferentzia gisa, metodologia berdinarekin baina OpenSubtitles corpusaren ingelesezko bertsioarekin entrenatutako sareak emandako erantzunak ere erakusten ditugu[3]. Erantzun guztiak 1. taulan ageri dira.

7. ONDORIOAK ETA ETORKIZUNERAKO PLANTEATZEN DEN NORABIDEA

1. taulan ikusi daitekeen moduan, sare neuronal sortzaile aurkarien bidez euskaraz era nahiko koherente eta zentzudunean hitz egiten duen elkarrizketa sistema automatikoa lortu dugu. Ingelesarekin konparatuz euskaraz dauden baliabideen tamaina askoz txikiagoa izan arren, sare neuronalen bidezko metodologiak erabiltzea badagoela frogatu dugu. Horretarako, euskararen morfologia kontuan hartzea inportantea dela ere erakutsi dugu. Izen, izenordain, adjektibo edo determinanteak lexematan eta postposiziotan banatzea komenigarria da, sareak hala era eraginkorrago batean prozesatzen baitu lengoaia. 1. taulari begira, aurreprozesu horrekin sareak esaldi konplexuagoak sortzeko joera duela esan dezakegu.

Gainera, lan honekin proposatu berri dugun [19] eta testuarekin era guztiz diferentzialean lan egin dezakeen sare sortzaile aurkarien arkitektura baliozkotzen dugu, eta elkarrizketa sistema automatikoak euskaraz eraikitzeko aproposa dela egiaztatu.

---

[3] Ingelesez corpusaren tamaina handiagoa denez, sareak ere handiagoak dira, eta iterazio gehiagotan zehar entrenatu dugu. Zehaztasunak [19] erreferentzian ematen dira.

https://doi.org/10.1387/ekaia.20987

1. taula. Hainbat mezuren aurrean elkarrizketa sistemek emandako erantzunak. **A** ingelesez entrenatutako sistema da, **B** euskaraz aurreprozesamenduarekin entrenatutakoa eta **D** aurreprozesamendurik gabekoa

Sarrera mezua	Sistemaren erantzuna
A: hi !	hello , <person>.
B: kaixo !	kaixo , <izen> naiz , <izen>-ren semea naiz .
D: kaixo !	kaixo , kaixo.
A: I love you	you know , I think maybe we could talk ...
B: maite zaitut	nik ere maite zaitut .
D: maite zaitut	maite zaitut
A: who is she ?	she 's my girlfriend , <person> , and ...
B: nor da neska hori ?	<izen> da neska .
D: nor da neska hori ?	nor da neska ?
A: what's the color of the sky ?	it 's a beautiful blue sky
B: zein da zeruko kolorea ?	zeru urdin iluna da .
D: zein da zeruko kolorea ?	zein da zure kolore faboritoa ?
A: do n't worry	I do n't want to hear about it.
B: ez kezkatu	ez dut ulertzen zergatik egin duen hau zutzat, <izen> ?
D: ez kezkatu	ez dut ahaztuko
A: could you come with us there ?	I 'll come with you .
B: toki horitara lagunduko diguzu ?	bai , horixe uste dut nik ere .
D: toki horretara lagunduko diguzu ?	jakina baietz
A: I forgot it .	I did n't know you were coming .
B: ahaztu egin zait .	<izen> .
D: ahaztu egin zait .	zer ?
A: can you see anything ?	I see you in the mirror .
B: zerbait ikusten ?	ez , ez
D: zerbait ikusten ?	ez .
A: it was a plane crash .	I was in the hospital
B: hegazkin istripu bat izan zen .	<izen> larri zaurituta zegoen eta bere etxera bueltatu zen
D: hegazkin istripu bat izan zen .	sute bat izan zen .
A: did he wake up again ?	no he did not, but he did n't sleep .
B: jaiki al da berriz ?	zer esan nahi du horrek ?
D: jaiki al da berriz ?	astelehenean atera behar duzu .
A: yesterday I saw you in the mountains	and now I see you
B: atzo menditik ikusi zintudan	ikusi zintudan <zenbaki> aldin .
D: atzo menditik ikusi zintudan	zer ?
A: see you	see you later , <person> .
B: agur	agur , aita .
D: agur	agur , ene erregea

Dena den, lan honetan aurkeztutako metodologia eta ideiak asko garatu behar dira benetan pertsona baten moduan euskaraz hitz egiten duen sistema lortzeko. Izatez, ingelesez ere oraindik urrun gaude halako sistemak sortzetik. Oraingoz baliabide handiagoko eta txikiagoko lengoaiekin sortutako sistemak parekatzea da gure hurrengo helburua. 1. taulan ageri den moduan, ingelesez entrenatutako sare sortzaile aurkariak era zentzudunagoan eta gizatiarragoan hitz egiten du euskarazko sistemarekin konparatuta.

Diferentzia horiek murrizteko, ezagutzaren transferentzia (*transfer learning* ingelesez) egiteko teknikak erabiltzeko asmoa daukagu. Ezagutzaren transferentziaren ideia nagusia da corpus handiagoekin baina eginkizun ezberdin baterako entrenatutako ereduak eredu berriak sortzeko erabiltzea. Kasu honetan, beraz, ingelesez sortutako sarea euskarazko sistema hobetzeko erabiltzea izango da gure helburua.

Amaitzeko, etorkizunean [30] lanean proposatutako eta *byte pair encoding* edo BPE izeneko aurreprozesamenduarekin saiakuntzak egingo ditugu, eta guk proposatutako lematizazioarekin konparatu. BPE-a tokenizatzaile edo hitz-banatzaile estatistiko eta automatikoa da, guk proposatutako banaketak ez ezik, printzipioz beste banaketa zentzudun batzuk ere egiteko gai dena.

8. **BIBLIOGRAFIA**

[1] OLASO J.M. eta TORRES M.I. 2017. «User experience evaluation of a conversational bus information system in spanish». *8th IEEE International Conference on Cognitive Infocommunications.*

[2] BORDES, A. eta WESTON, J. 2016. «Learning end-to-end goal-oriented dialog». *CoRR abs/1605.07683.*

[3] TURING A.M. 1950. «Computing machinery and intelligence». *Mind*, LIX, 433-460.

[4] WEIZENBAUM, J. 1966. «ELIZA - a computer program for the study of natural language communication between man and machine». *Communications of the ACM*, 9, 36-45.

[5] SUTSKEVER I., VINYALS O. eta LE Q.V. 2014. «Sequence to sequence learning with neural networks». *Advances in neural information processing systems*, 3104-3112.

[6] CHO K., MERRIE¨ NBOER B., C¸ AGLAR G., BAHDANAU D., BOUGARES F., SCHWENK H., eta BENGIO Y. 2014. «Learning phrase representations using RNN encoder-decoder for statistical machine translation». *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP*, 1724-1734.

[7] SORDONI A., GALLEY M., AULI M., BROCKETT C., JI Y., MITCHELL M., NIE J., GAO J. eta DOLLAN B. 2015. «A neural network approach to

context-sensitive generation of conversational responses». *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 196-205.
[8] SERBAN I.V., SORDONI A., BENGIO Y., COURVILLE A. eta PINEAU J. 2016. «Building end-to-end dialogue systems using generative hierarchical neural network models». *Thirtieth AAAI Conference on Artificial Intelligence*.
[9] TUAN Y. eta LEE H. 2019. «Improving conditional sequence generative adversarial networks by stepwise evaluation». *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
[10] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A. eta BENGIO Y. 2014. «Generative adversarial nets». *Advances in neural information processing systems*, 2672-2680.
[11] YU L., ZHANG W., WANG J. eta YU Y. 2017. «SeqGAN: Sequence generative adversarial nets with policy gradient». *AAAI*, 2852-2858.
[12] LI J., MONROE W., SHI T., JEAN S., RITTER A. eta JURAFSKY D. 2017. «Adversarial learning for neural dialogue generation». *arXiv preprint arXiv:1710.06547*
[13] HORI T., WANG W., KOJI Y., HORI C., HARSHAM B. eta HERSHERY J.R. 2019. «Adversarial training and decoding strategies for end-to-end neural conversation models». *Computer Speech* & *Language*, 54, 122-139.
[14] SUTTON R.S. eta BARTO A.G. 1998. «Introduction to reinforcement learning». *MIT press Cambridge*.
[15] BENGIO Y., LÉONARD N. eta COURVILLE A. 2013. «Estimating or propagating gradients through stochastic neurons for conditional computation». *arXiv preprint arXiv:1308.3432*.
[16] JANG E., GU S. eta POOLE B. 2016. «Categorical reparameterization with gumbel- softmax». *arXiv preprint arXiv:1611.01144*.
[17] LU J., KANNA A., YANG J., PARIKH D. eta BATRA D. 2017. «Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model». *Advances in Neural Information Processing Systems*, 314-324.
[18] SHETTY R., ROHRBACH M., HENDRICKS L.A., FRITZ M. eta SCHIELE B. 2017. «Speaking the same language: Matching machine to human captions by adversarial training». *Proceedings of the IEEE International Conference on Computer Vision*.
[19] LÓPEZ ZORRILLA A., DEVELASCO VÁZQUEZ M. eta TORRES M.I. 2019. «A differentiable generative adversarial network for open domain dialogue». *Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS).*
[20] HOCHREITER S. eta SCHMIDHUBER J. 1997. «Long short-term memory». *Neural computation*, 9, 1735-1780.

[21] BAHDANAU D., CHO K. eta BENGIO Y. 2016. «Neural machine translation by jointly learning to align and translate». *CoRR abs/1409.0473*.

[22] LUONG M., PHAM H. eta MANNING C.D. 2015. «Effective approaches to attention- based neural machine translation». *arXiv preprint arXiv:1508.04025*.

[23] MIKOLOV T., CHEN K., CORRADO G.S. eta DEAN J. 2013. «Efficient estimation of word representations in vector space». *CoRR abs/1301.3781*.

[24] KANNAN A. eta VINYALS O. 2017. «Adversarial evaluation of dialogue models». *arXiv preprint arXiv:1701.08198*.

[25] BOJANOWSKI P., GRAVE E., JOULIN A. eta MIKOLOV T. 2016. «Enriching word vectors with subword information». *arXiv preprint arXiv:1607.04606*.

[26] RODRIGO A., BERMUDEZ J. eta RIGAU G. 2014. «Ixa pipeline: Efficient and ready to use multilingual NLP tools». *LERC*, 3823-3828.

[27] LISON P. eta TIEDEMANN J. 2016. «Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles». *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

[28] KINGMA D.P. eta BA J. 2014. «Adam: A method for stochastic optimization». *arXiv preprint arXiv:1412.6980*.

[29] LIU C., LOWE R., SERBAN I., NOSEWORTHY M., CHARLIN L. eta PINEAU J. 2016. «How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation». *EMNLP*.

[30] SENNRICH, R., HADDOW, B., BIRCH, A. 2016. «Neural Machine Translation of Rare Words with Subword Units». *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1715-1725.

Publication 10

Contrasting the Emotions identified in Spanish TV debates and in Human-Machine Interactions

IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Contrasting the Emotions identified in Spanish TV debates and in Human-Machine Interactions

Mikel deVelasco, Raquel Justo, Leila Ben Letaifa, M. Inés Torres

Speech Interactve Research Group, Universidad del País Vasco UPV/EHU
{mikel.develasco, raquel.justo, leila.benletaifa, manes.torres}@ehu.es

Abstract

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. To this end we developed a crowd annotation procedure to label the speech signal in terms of both, emotional categories and Valence-Arousal-Dominance models. The analysis of these data showed interesting findings that allowed to profile both the speakers and the task. Then, Convolutional Neural Networks were used for the automatic classification of the emotional samples in both tasks. Experimental results drew up a different human behavior in both tasks and outlined different speaker profiles.

Index Terms: emotions recognition from speech, perception, communication, human-machine interaction, crowd annotation, speech processing.

1. Introduction

Speech signal includes information about the personal characteristics of the speaker, the content of the message delivered or the language used to code it, among others [1]. The analysis of the speech also allows to estimate, to some extent, the current emotional status of the speaker [2, 3, 4], even the basal mood, or the probability to be suffering a particular mental disease [5]. However, speech may also be influenced by several other variables, such as the habits of the speaker, his personality, culture or the particular task being performed [6, 7]

This work is aimed to contrast the similarities and differences for the emotions identified in two very different scenarios: human-to-human interaction on Spanish TV debates and human-machine interaction with a virtual agent in Spanish. Thus, we focus on spontaneous emotions appearing in each task that show significant differences to the six basic emotions [8] that have been many times simulated by professional actors [9, 10, 11] and recorded in the lab [12]. In fact, spontaneous emotions have been hypothesized to be extremely task dependent [2, 3, 7, 4, 6]. Further to this, emotions cannot be unambiguously identified. As a consequence, not even expert labelling procedure can lead to a ground truth for learning. As an alternative, crowd annotation implementing perception experiments has also been proposed as a way to establish the ground truth [13]. However, human perception of emotions does not usually show a high agreement. As a consequence, a certain ambiguity and uncertainty always remains, which adds an stochastic component to the emotion identification problem.

In order to verify whether actually the task plays a significant role when dealing with emotion detection, a preliminary comparison of the emotional content in two very different Spanish tasks was carried out in this research work. To this end, we chose the following set of features to be analysed: agreement in crowd annotation, perceived emotions and significance in the particular task, distribution of categories in both tasks, distribution of dimensional axes of emotions, namely Valence, Arousal and Dominance (VAD), and the representation of the categories into the 3D VAD model. An additional contribution is the comparative analysis of the results in terms of categories of the automatic classification of the samples based on Convolutional Neural Networks (CNN).

Section 2 describes the two tasks addressed as well as the annotation procedure and its outcomes. Then Section 3 develops the analysis of emotional content of the corpora and Section 4 describes the preliminary classification experiments carried out. Finally, Section 5 summarizes the concluding remarks and future work.

2. Perception of emotions

2.1. Description of the tasks

TV Debates Firstly, a data-set that gathers real human-human conversations extracted from TV debates, specifically the Spanish TV program “La 6 Noche”, was selected. In this weekly broadcasted show, hot news of the week are addressed by using social and political debate panels that were led by two moderators. There is a very wide range of talk-show guests (politicians, journalists, etc.) who analyse, from their perspective, social topics. Given that the topics under discussion are usually controversial it is expected to have emotionally rich interactions. However, the participants are used to speak in public so they do not lose control of the situation and even if they might overreact sometimes, it is a real scenario, when emotions are subtle. The spontaneity in this situation makes a great difference from scenarios with acted emotions as shown in [2]. The selected programs were broadcasted during the electoral campaign of the Spanish general elections in December 2015.

Elder interaction with simulated virtual agent Empathic is a European Research & Innovation Project [1] [14, 7] that implements personalized virtual coaching interactions to promote healthy and independent aging. As a part of the project, a series of spontaneous conversations between elderly and a Wizard of OZ (WOZ) have been recorded in three languages: Spanish, French and Norwegian. WOZ’s technique allows users to believe that they are communicating with a human (and not a machine) in order to make their reaction more natural [7]. The conversations are related to four main topics: leisure, nutrition, physical activity and social and family relationships [14, 7]. In this work we focused on the Spanish dialogues that were recorded by 79 speakers resulting in 7 hours and 15 minutes of audio extracted from the recordings [3].

[1] www.empathic-project.eu

10.21437/IberSPEECH.2021-11

2.2. Crowd perception

TV Debates and Spanish Virtual agent interaction data were labeled in terms of emotions using the crowd annotation technique. To begin with, we automatically extracted segments of audio that we estimated to match a clause. A clause can be defined as “a sequence of words grouped together on semantic or functional basis” [15]. Thus, we can hypothesize that the emotional status does not change inside a clause. This procedures allowed to get 4118 chunks from the TV Debate corpus and 2000 from the Virtual agent corpus. Then, all these segments were crowd annotated by native speakers. To this end, both categorical and VAD model of emotions were considered. For the categorical model we first consider the categories proposed in [16] and then we reduce and adapt the list of each of the tasks. For TV Debates task we selected a list of ten labels to be considered by annotators. Then we added three questions to annotate the perception of each of the axes of the dimensional model, namely Valence, Arousal and Dominance

Three of them are related to the arousal: Excited, Slightly excited and Neutral.Valence is annotated as Positive or Slightly positive or Neutral or Slightly negative or Negative in TV Debates. For Virtual Agent task, valence is assigned one of only three labels that are Positive, Neutral and Negative for valence. The dominance labels are: Rather dominant / controlling the situation, Rather intimidated / defensive, and Neither dominant nor intimidated.. The whole questionnaire is reported in [2]. For Virtual Agent task we also selected list of ten categories adapted to the task that differs from the previous one. As an example *Sad* was only included in this task whereas *Annoyed* was only proposed to TV Debates annotators.

Annotators agreement Each audio segment was annotated by 5 different annotators. Table 1 shows the statistics of agreement per audio chunk for the categorical model. This table shows that for about 70% of the data and in both tasks, the agreement is 3/5 or 2/5. This confirms the ambiguity and subjectivity of the task. Moreover the Krippendorff’s $alpha$ coefficient was also low for both tasks resulting in 0.11 and 0.13 values respectively. This coefficient reflects the agreement degree but is very dependent on the number of labels, which was high and sometimes difficult to be perceived.

In the rest of the document, we do not consider samples with agreement below 0.6, which means we have used the 64.13% of the corpus for the TV debates task and the 66.20% of the Virtual Agent task.

Table 1: *Statistics of the agreement per audio chunk*

	TV Debates		Virtual Agent	
Agr	No. audios	% audios	No. audios	%. audios
5/5	197	4.72%	149	7.45%
4/5	799	19.40%	421	21.05%
3/5	1645	39.95%	754	37.7%
2/5	1431	34.75%	636	31.8%
1/5	46	1.18%	40	2%
Tot. audios	4118		2000	

Annotation labels The defined sets of labels were then reduced by merging overlapping categories that we selected for the tag pairs with high level of confusion among them in the annotation procedure. Then, a minimum agreement of 0.6 (3/5) was requested for each sample as a well as a minimum number of samples. Table 2 shows the resulting list of categories considered for each task along with the percentage of samples. This Table shows that different categories appear in each corpus. Some of them could be equivalent, such as *Calm/Indiferent* and *Calm/relaxed* but *annoyed/tense* does not appear in Virtual Agent task whereas puzzled is not in the list for TV debates.

Table 2 also shows that both data-sets are imbalanced, being the *Calm* category the majority class with around 75% of the samples. This reflects the spontaneous nature of the data. There are more positive emotions in the Virtual Agent annotations and more negative emotions in TV Debates. This difference comes from the tasks characteristics. During political debates, people try to convince or even impose their opinions on other interlocutors. However, during the coaching sessions, people speak with a machine. They are quiet and paying attention to the answers to their expectations.

For the dimensional model we got a set of scale values for each axe. For for Arousal we proposed Neutral, Slightly excited and Excited in both databases. For Dominance we proposed Rather intimidated / defensive, Neither dominant or intimidated, Rather dominant / controlling the situation fro both databases. For Valence we got Negative, Slightly Negative, Neither negative or positive, Slightly Positive and Positive for TV Debates whereas we reduced the scale to Rather Negative, Neither negative or positive, Rather Positive for the Virtual Agent task. These dimensions are considered in Section 3

Table 2: *Categories more frequent in the corpora*

TV Debates		Virtual Agent	
Category	% audios	Category	% audios
Calm/Indiferent	73.64	Calm/Relaxed.	78.32
Annoyed/Tense	14.32	Happy/Pleased	8.76
Enthusiast	4.72	Interested	5.66
Satisfied	3.23	Puzzled	2.95
Worried	2.12		
Interested.	1.57		
Others	0.40	Others	4.31

3. Analysis of emotions

Figure 1 shows the probability density function of each variable (Valence, Arousal, Dominance) of VAD model that has been obtained by a Gaussian kernel density estimator (upper row). Figure 1 also shows different 2D projections of sample distribution in the 3D space (row below), representing each scenario in a different colour. When regarding Arousal, Virtual Agent seems to work in a very neutral scenario where excitement is almost absent. In TV debates, although neutrality is also predominant, some excitement is perceived, due to the debate nature of the conversations. Valence distribution shows a clear deviation towards positive values when considering Virtual Agent scenario, a sign of the good acceptance of the system among the users, whereas in TV debates neutrality is predominant with only a slight nuance towards positiveness. On the contrary Dominance is shifted towards Dominant values, in TV debates, but keeps

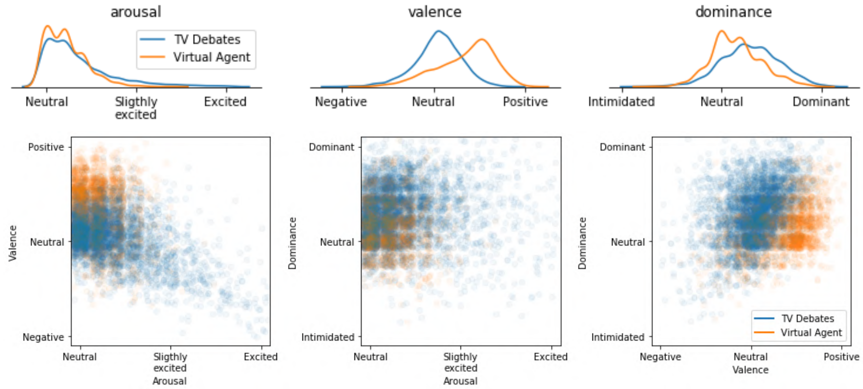


Figure 1: *VAD representation.*

neutral when users interact with the Virtual Agent. These results correlate well with the kind of audios we are dealing with in the two scenarios. In TV debates, people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally they appear to be neutral when communicating their opinions (valence tends to be neutral or slightly positive). In the Virtual Agent scenario the users are volunteers, with a good predisposition, and thus they seem to be pleased with the system (Positive Valence values). They are relaxed talking to the agent (levels of excitement tend to neutrality) and although they do not have to convince anyone they know well what they are talking about and are not intimidated (dominance values are around neutrality with a slight shift to the right).

The categorical model is also considered in this work and each category is represented in the 3D VAD space for comparison purposes. Specifically, the average of the Valence, Arousal and Dominance values of all the audios labeled within a specific category was computed and the resulting value was represented as a point in the 3D space. Figure 2 shows 2D projection of the resulting representation. If we focus on TV Debates, it can be noticed that *Interested* and *Worried*, the least representative categories, according to Table 2, are very close to the category with the highest number of samples, *Calm/Indifferent*, in all the 2D projections (purple, orange and deep blue points), so they were merged in an only one category. The same happens with Enthusiastic and Satisfied (light blue and green points). When considering Virtual Agent scenario although the category *Calm/Relaxed* is the most relevant one with more than the 75% of the samples we decided to keep the remaining categories because the fusion is not as clear as in the previous case, as shown in Figure 2. Thus the final set of categories used for the classification experiments reported in this work (Sectiion **??**) is the following one for TV Debates: 1) *Annoyed/Tense*, 2) *Enthusiastic + Satisfied*, 3) *Calm/Indifferent + Interested + Worried* and for as the Virtual Agent the list is: 1) *Calm/Relaxed*, 2) *Happy/Pleased*, 3) *Interested*, 4) *Puzzled*.

As shown above, there are some categories that are not in both sets due to the nature of the different tasks, like Annoyed/Tense that is only In TV Debates or Puzzled that only appears in the interaction with the Virtual Agent. Moreover, Figure 2 shows that there is not any point in Virtual Agent scenario around the location of the red point (Annoyed/Tense) of TV Debates (higher excitement levels and negative values of Valence), which is in fact quite separated from the other categories. The same happens with *Puzzled* represented by the brown point (low levels of Valence and Dominance) that has not any representation in TV Debates and it is a bit separated from the other categories in Virtual Agent scenario. This correlates well with the idea that people interacting with the Virtual Agent are not in general annoyed or tense, while this is a quite common feeling in a debate. Furthermore, speakers in the debates do not usually show that they are in an unexpected situation, since it can be interpreted as a weak point, while it is quite easy to imagine it in the interaction with a machine. There are also categories, like *Calm* that has a similar location in both scenarios but with higher values of Valence for Virtual Agent interactions. That is, the users interacting with the Virtual Agent perceived as calm tend to be more positive than the ones in TV Debates. The same happens with *Enthusiastic + Satisfied* from TV Debates and *Happy/Pleased* from Virtual Agent, that although they are very close in their location in both scenarios (with a very similar meaning) *Happy/Pleased* seems to have more positive Valence values than *Enthusiastic + Satisfied*, but a bit lower Dominance and Arousal values.

4. Experiments and results

To complete the work, some classification problems were carried out in both tasks described in Section 2.1. For TV Debates, 4118 chunks were selected distributed in the 3 classes mentioned avobe (*Annoyed/Tense*, *Enthusiastic + Satisfied*, and *Calm/Indifferent + Interested + Worried*) and for the Virtual Agent, 2000 samples were selected divided into 4 classes (*Calm/Relaxed*, *Happy/Pleased*, *Interested*, and *Puzzled*).

One of the challenges of both data-sets is the different length of each audio sample. Some kind of Neural Networks are specifically well suited to deal with this problem and given that deep learning is the state of the art in many AI areas, including emotion recognition, a Convolutional Neural Network architecture was designed for this work. Let us note that in [17] a neural network architecture provided promising results when comparing ot to classical Support Vector Machines, for a regression problem over the task related to TV debates.

The number of samples in both data-sets are also a challenge. It makes nonsense to try to identify the emotions from

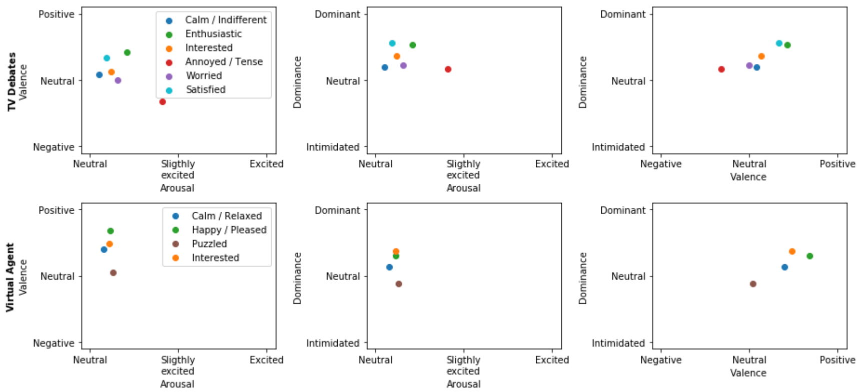


Figure 2: *Categories in dimensional representation.*

raw-audio. Different works suggest that there is not a standard audio feature-set that works well for all emotion recognition corpora [18, 19, 20]. In this context, we decided to use the audio Mel-frequency spectrogram as the classifier's input. It is known that the spectrogram encodes almost all audio information and should be possible to identify from that.

Figure 3 shows the architecture of the network used in this work. It takes the mel-spectrogram input and reduce both mel-frequency and time dimensions using 2D convolutions and max-poolings (red boxes). This sub-network reduces time dimension but creates richer audio representation. Then, the network takes the new representation and try to classify each time step. After classifying all time steps, the network averages it in order to provide an output for the input audio.

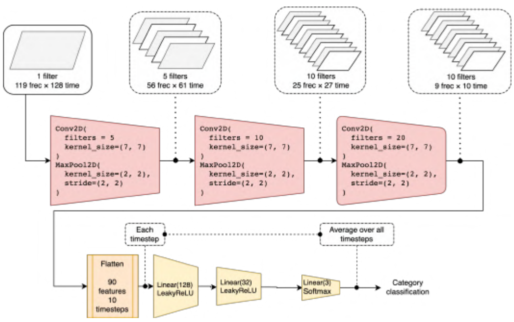


Figure 3: *Architecture of the Network used*

In the training process, several decisions were chosen. On the one hand, the network will only see a sub-part of the full audio. Thus, the training process is easier if all the batches work with the same input length, which can be considered as a dropout mechanism. On the other hand, an repetition over-sampling method was chosen, where all the non-majority class samples were provided 5 times. It helps the network to avoid the exclusive prediction of the majority class. Adam optimizer is used with a learning rate of 0.001 and 150 epochs were training on all database. These experiments were carried out over a 10-fold cross-validation procedure.

Classification results are given in Table 3. Most promising results come from TV Debates, in fact, the model guesses 72% of the test samples, and achieves a F1 Score of 0.59, that can be considered a good result taking into account the ambiguity and subjectivity of the task.

As expected, the category *Calm/Indifferent* + *Interested* + *Worried* got better results since it is the majority class with a F1 Score of 0.82.In contrast, *Annoyed/Tense* and *Enthusiastic* + *Satisfied* perform a little bit worse, with a 0.56 and 0.43 in F1 Score.

Table 3: *Evaluation of the classification results for the categorical model*

TV Debates				Virtual Agent			
Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
72%	0.56	0.66	0.59	74%	0.32	0.27	0.27

Nevertheless, Virtual Agent experiments obtained lower results. Table 3 show a very high accuracy (74% of the samples) along with low values of F1, precision and recall values. The majority class, i.e.*Calm*, achieved an F1 score of 0.88 whereas all minority classes remain under 0.32 fro F1. This is mainly due to the huge imbalance of this data-set along with the very reduced number of samples.

5. Conclusions

This work provides a comparison of the emotional content in two different Spanish corpora dealing with very different tasks. The emotional labels, associated to spontaneous emotions, were achieved by means of perception experiments using crowd annotation. The agreement among the annotators was considered to build the ground truth. The analysis carried out shows the main differences associated to each task, in terms of both, the emotional category distribution and the level of Valence, Arousal and Dominance and brings out the relevance of the task when addressing an emotion recognition problems. This analysis also highlights that the perception experiments carried out were able to outline a different speaker profile for each of the tasks. Thus, crowd annotation seems to be valid approach for emotions. Finally, some preliminary classification experiments were also conducted showing very promising results for TV Debate task whereas the Virtual Agent task needs more samples and a more sophisticated oversampling method. Future work includes a deeper and interrelated analysis of the data as well getting a higher number of annotated samples for the Virtual Agent classification task.

6. Acknowledgements

The research presented in this paper is conducted as part of the AMIC and EMPATHIC projects project that have received funding from the Spanish Minister of Science under grant TIN2017-85854-C4-3-R and from the European Union's Horizon 2020 research and innovation program under grant agreement No 769872. First author has also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

7. References

[1] A. López-Zorrilla, N. Dugan, M. Torres, C. Glackin, G. Chollet, and N. Cannings, "Some asr experiments using deep neural networks on spanish databases," in *IberSpeech*, Lisbon, 2016, pp. 149–158.

[2] M. deVelasco, R. Justo, A. López-Zorrilla, and M. Torres, "Can spontaneous emotions be detected from speech on tv political debates?" in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*, Naples, 2019.

[3] L. B. Letaifa, M. I. Torres, and R. Justo, "Adding dimensional features for emotion recognition on speech," in *IEEE International Conference on Advanced Technologies for Signal and Image Processing*, Tunisia, 2020.

[4] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech*, 2018.

[5] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," 2020.

[6] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech Language*, vol. 53, pp. 156 – 180, 2019.

[7] R. Justo, L. B. Letaifa, C. Palmero, E. G. Fraile, A. Johansen, A. Vazquez, G. Cordasco, S. Schlogl, B. F. Ruanova, M. Silva, S. Escalera, M. D. Velasco, J. T. Laranga, A. Esposito, M. Kornes, and M. I. Torres, "Analysis of the interaction between elderly people and a simulated virtual coach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 6125–6140, 2020.

[8] P. A. Davidson. R. J., Ekman, *Nature of emotion: Fundamental questions*, ser. Oxford University Press, P. E. . R. J. Davidson, Ed. New York: Springer, 1994.

[9] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions, IJSE*, pp. 68–99, 2010.

[10] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[11] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: How does an automated system compare to naive human coders?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2274–2278.

[12] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition."

[13] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English," in *Proc. Interspeech 2019*, 2019, pp. 316–320. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1413

[14] M. I. Torres, J. M. Olaso, C. Montenegro, R. Santana, A. Vázquez, R. Justo, J. A. Lozano, S. Schlögl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. P. Cantariño, O. Deroo, O. Gordeeva, J. Tenorio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto, "The empathic project: Mid-term achievements," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 629–638.

[15] A. Esposito, V. Stejskal, and Z. Smékal, "Cognitive role of speech pauses and algorithmic considerations for their processing," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, pp. 1073–1088, 2008.

[16] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, pp. E7900 – E7909, 2017.

[17] M. de Velasco, R. Justo, J. Antón, M. Carrilero, and M. I. Torres, "Emotion detection from speech and text," in *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, J. Luque, A. Bonafonte, F. A. Pujol, and A. J. S. Teixeira, Eds. ISCA, 2018, pp. 68–71. [Online]. Available: https://doi.org/10.21437/IberSPEECH.2018-15

[18] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 698–704.

[19] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," 2017.

[20] S. Parthasarathy and I. Tashev, "Convolutional neural network techniques for speech emotion recognition," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 121–125.

Publication 11

A Spanish Corpus for Talking to the Elderly

A Spanish Corpus for Talking to the Elderly

Raquel Justo, Leila Ben Letaifa, Javier Mikel Olaso, Asier López-Zorrilla, Mikel Develasco, Alain Vázquez, and M. Inés Torres

Abstract In this work, a Spanish corpus that was developed, within the EMPATHIC project (http://www.empathic-project.eu/) framework, is presented. It was designed for building a dialogue system capable of talking to elderly people and promoting healthy habits, through a coaching model. The corpus, that comprises audio, video an text channels, was acquired by using a Wizard of Oz strategy. It was annotated in terms of different labels according to the different models that are needed in a dialogue system, including an emotion based annotation that will be used to generate empathetic system reactions. The annotation at different levels along with the employed procedure are described and analysed.

1 Introduction

Although the use of conversational systems in our daily life seemed to be science fiction not much time ago. Nowadays they are pretty integrated in our homes (Alexa

R. Justo ( ) · L. B. Letaifa · J. M. Olaso · A. López-Zorrilla · M. Develasco · A. Vázquez · M. I. Torres
Universidad del Pais Vasco UPV/EHU, Sarriena s/n, 48940 Leioa, Spain
e-mail: raquel.justo@ehu.eus

L. B. Letaifa
e-mail: leila.benletaifa@ehu.eus

J. M. Olaso
e-mail: javiermikel.olaso@ehu.eus

A. López-Zorrilla
e-mail: asier.lopezz@ehu.eus

M. Develasco
e-mail: mikel.develasco@ehu.eus

A. Vázquez
e-mail: alain.vazquez@ehu.eus

M. I. Torres
e-mail: manes.torres@ehu.eus

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_13

speaker by Amazon), jobs (Cortana or Siri to manage our agenda) or even in our leisure (Siri or Samsung's Bixby for smartphones). They are becoming useful in more and more different domains ranging from game environments to educational contexts. Some of them can pass the Turing test (e.g., Eugene Goostman[1]). Thus, we can say that the way in which people interact with computers is shifting to the use of natural language.

There are many different systems in the literature built for different purposes and that make use of different technologies [7, 21, 23, 26]. However, one of the most extended categorization of conversational systems is the one that distinguishes among "chatbots" and "dialogue systems" [9, 12, 18]. Although the frontiers among those categories are not always clear, focusing on the differences related to the goal, chatbots are aimed at generating appropriate, relevant, meaningful and human-like utterances and there is not an specific goal to be achieved during the interaction like in the case of dialogue systems. Dialogue systems are often developed for a specific domain, whereas simulated conversational systems [chatbots] are aimed at open domain conversation [13].

In this work we deal with a dialogue system developed within the EMPATHIC project [11, 24, 25] framework. The goal of this project is to design and validate new interaction paradigms for personalized Virtual Coaches to promote healthy and independent aging. Thus, a dialogue system that can talk to the elderly, understand them, empathise with them and promote healthy habits is being developed. This kind of dialogue systems need different modules like automatic speech recognizer, natural language understanding module, dialogue manager, natural language generator, etc. Moreover, a module that tries to detect the emotion of the speaker is also being developed in order to provide a system response that can be empathetic with regard to the user emotional status. The methodologies employed to develop these modules are mainly based on machine learning and data driven approaches. When using these approaches, data are needed to be able to train robust models. Moreover, the data have to be closely related to the specific task, environment, channel, etc. Thus, it is very difficult to get valuable resources when specific tasks, like the one presented in this framework are considered. Furthermore the lack of resources is even more noticeable when we consider other languages (apart from English) like Spanish.

The main contribution of this paper is the description of a Spanish corpus devoted to train different models that will be employed in a dialogue system that tries to talk to the elderly people and promotes healthy habits being aware of the affective component. The corpus was annotated in terms of different labels that will be used by the different modules. The annotation procedures, that will be described in the following sections, were selected to allow the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. Section 2 provides a description of the dialogues that comprise the corpus and the way in which they were acquired. In Sect. 3 the annotation procedure developed to build the modules related to dialogue generation are described (natural language understanding, dialog manager and

[1] http://www.reading.ac.uk/news-archive/press-releases/pr583836.html.

natural language generation). Then in Sect. 4 the annotation carried out to detect emotions in different channels (audio, video and text) is detailed.

2 Dialogues in the EMPATHIC Framework

In order to develop a dialogue system, like the one described above, a data acquisition procedure has to be designed first. In this process we used a Wizard of Oz (WoZ) platform [19, 20] for the acquisition of the database. The WoZ constitutes a prototyping method that uses a human operator (the so-called wizard) to simulate non- or only partly- existing system functions. It was used to make users think that they are interacting with a real automatic dialogue system. In this way, the data acquisition procedure considers human-machine conversations that were carried out in an environment as most realistic as possible.

The dialogues in the EMPATHIC project are leaded according to a coaching model, a GROW coaching model in this case, that tries to get the desired goals related to healthy habits. A GROW coaching dialogue consists on four main phases: Goals or objectives, Reality, Options and Will or action plan. During the first phase, the dialogue aims at establishing explicit objectives that the user wants to achieve, e.g. reduce the amount of salt. During the next phase, taking into account the user's personal context, the dialogue tries to detect potential obstacles that prevent fulfilling the previously established objectives. For the next phase, the goal is to analyse the options the user has in order to face the obstacles and achieve the objectives. In the last phase, the dialogue tries to specify an action plan for the user to carry out in order to advance towards the objectives. The final goal for the EMPATHIC virtual coach is to deal with four different domains: leisure, nutrition, physical activity and social and family relationships [17]. However, in the initial phase described in this paper, not all the scenarios were used; two scenarios were integrated in this platform. A first introductory scenario, which in turn was used to strengthen the user in the interaction with the platform. And a second one to simulate a GROW session on nutrition. These scenarios were designed using the documentation provided by a professional coach. Although different acquisition procedures were carried out in the project for different languages: Spanish, French and Norwegian, in this work we focus on the Spanish dataset.

Making use of the aforementioned WoZ platform, 79 native Spanish users selected among the target population (healthy elderly above 65) interacted with the system. The majority of them participated in the two predefined scenarios, but in some cases, due to different reasons, only one of these sessions was carried out. Thus, 142 dialogues were collected. These include around 4,500 user turns and the same amount of machine turns.

The acquired conversational sessions between elderly people and the simulated virtual coach were recorded in order to have an audio-visual database. Each session takes about 10 min so the total recordings correspond to about 23 h of video. The audio part represents about 30% of the database.

3 Resources for Building the Dialogue

Once the acquisition procedure was finished the data were annotated in order to build the different models involved in the conversational process.

3.1 speech to text Annotation

One of the first annotation needed for training robust models to be used in a dialogue system is the transcription of the speech. This is essential for the Automatic Speech Recognizer for instance. Thus after the acquisition procedure, the dialogues were manually transcribed. The vocabulary size resulted to be 5,543 for the user turns and 2,941 for the virtual coach's turns. As for the running words, the corpus contains 72,350 in the user turns and 30,389 in the he virtual coach's turns.

The transcriptions of the acquired dialogues were further annotated in order to facilitate the modeling of the dialogues. The following two sections explain how the turns of both the users and the virtual coach were labeled. The two annotation tasks were carried out by 9 annotators, who were instructed about the structure of the labels, the GROW coaching model, and about the context of the project. Each annotator labeled roughly the same number of dialogues.

3.2 semantic Annotation

The taxonomy of the labels used to represent each of the users' utterances was designed so as to be usable for the dialogue agent that is being deployed in the EMPATHIC project. Several works have addressed the question of defining dialogue act taxonomies [3, 22]. Among them, the DIT++ taxonomy [2] and the more recent ISO 24617-2 standard [4, 15], which is intended to be a development of the previous one, can be considered the general methodological framework of the taxonomy defined in this section. It is a dialogue-act taxonomy aimed to represent the user utterances in a particular human-machine communication framework, which develops a coaching model aimed at keeping a healthy and independent life of elderly. Thus, the taxonomy allows the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. To fulfill its needs, we employed three different types of labels: the topic, the intent and the name entities. The topic label identifies the general context of utterance, such as nutrition, leisure or family; and also some subtopics. The intent label determines the communicative intention of the user, e.g. greetings, agreement, opinion and so on. Additionally, it also includes information about which stage of the GROW model the user is talking about. Finally, the name entities are tuples containing small segments of the utterance and their category.

Table 1 Most frequent topic, intent and entity labels in the corpus

Topic	Intent	Entities
sport & leisure - travelling	*generic - agreement*	*actions*
sport & leisure - hobbies - type	*GROW - habit - present*	*quantities*
nutrition - regularity - ordered	*generic - opinion - positive*	*places*
sport & leisure - motivation	*generic - disagreement*	*amount of time*
nutrition - quantity	*generic - greetings*	*frequencies*
sport & leisure - music	*GROW - plan*	*hobbies*

They can be very useful for understanding the user but also for enriching the natural language generator. We have included, for example, people's names, places, and books.

The topic and intent labels are hierarchical, i.e., each utterance is labeled with multiple tags that can be ordered from more general to more specific. To make the annotation more consistent, each turn was split into several subsentences if there were more than one topic or intents in that turn. In total, 56 different labels were used for the topic representation, 34 for the communicative intent and 22 types of entities were identified. The complete list of labels is provided in detail in [14]. Since it is too large, Table 1 shows the most frequent labels for the topic and intent, and the most frequent entities.

3.3 Dialogue Act Annotation

Dialogue Act (DA) annotation is the equivalent task to the semantic annotation for the turns of the virtual coach. In this case, the outputs of the coach are labelled considering five criteria: DA, polarity, gender of the user and coach and possible appearance of entities in the responses of the coach. Such annotation is highly related to the Natural Language Generation (NLG), one of the modules included in the dialogue system developed in the EMPATHIC project. The NLG is in charge of generating the responses of the virtual coach to the users through a unit of information which contains a set of labels. The inverse process is made in the annotation: one set of labels is assigned to each turn of the virtual coach contained in the data.

The data was extracted from two different sources: the WoZ sessions and a set of handmade dialogues prepared by a professional coach. In both cases, only the turns of the coach were relevant to build this part of the data. Indeed, each turn can be split in different utterances, where an utterance is considered each element which can be labelled with a different DA. In total, the number of utterances is 8173 where 5985 are from the real session with users and 2188 from the handcrafted conversations.

All these utterances were labelled in terms of the five aforementioned criteria. The DA, which is built for one principal label and sublabel in the case of EMPATHIC, describes the communicative function and the semantic of the coach's sentences. There are 10 different values for the principal label and more than 100 for the sublabel. However, the DAs do not allow all the possible combinations, as each label only can be joined with a reduced group of sublabels. The polarity defines the emotional state of the coach, which can be selected between positive and neutral. The possible values for the genders are male, female and not identifiable, since what is annotated is if the gender of the two participants can be known through the coach turn alone (without any context). Finally, the detection of entities followed the same procedure carried out in the semantic annotation.

In the DAs, we identified three different blocks with the following distribution: the GROW block (19.6%), the Introduction one (24.6%) and General one (55.8%). The first block contains eight of the ten principal labels. These labels are the eight typical questions used in the GROW model. The other blocks, each one only contains one principal label. The Introduction label is used to annotate usual turns in a first session with the user (ask for the name, self-introduction, information of the project, ...). Finally, the General one is used to label all the expression which can be part of any conversation (thanking, greetings, agreement, ...). In terms of the polarity, the positive utterances (63.0%) were almost two times the neutral ones (37.0%). For both user and coach gender, they were not identifiable in almost the 99% of the data. Finally, the most frequent entities in the data were actions, dates and food.

4 Resources for Empathizing with the Elderly

Within the EMPATHIC project framework, the idea of empathising was very important. Thus, we wanted not only to understand what elderly is requesting to the system, but also to know their emotional status when interacting with it. Therefore, an annotation in terms of emotion was carried out by Spanish native annotators. The representation of emotional status is not straightforward and different models can be used according to Affective Computing literature [1, 5, 6, 16]. In this work we employed both a categorical model and a three-dimensional VAD (Valence, Arousal and Dominance) model in order to be able to compare both criteria.

Both data modalities, audio and video, were considered. In order to avoid interference between modalities, only audio (i.e. no images) was provided to the speech annotators and only video (without sound) was used by the video annotators.
In this section, we describe and analyse each modality annotation. For more information about the annotation procedure, refer to [8] and [10].
Finally, at the same time as the semantic annotation was carried out, the polarity of the transcribed utterances was also labeled by the same annotators.

Table 2 Audio annotated segments

	Calm	Sad	Happy	Puzzled	Tense
Annotation 1	7017	17	260	347	12
Annotation 2	7794	19	291	297	24
Annotation 3	7655	21	244	360	20

Table 3 Video annotated segments

	Sad	Annoyed	Surprised	Happy	Pensive	Other	Neutral
Annotation1	0	0	12	234	2032	0	2278
Annotation2	0	1	44	151	2059	3	2258

4.1 Audio Annotation

Only the audio part of the conversations between the virtual coach and elder people (which duration is about 7 h) is concerned by the audio annotation process. A manual labeling procedure from scratch was carried out by three native people. The perceived emotion was labelled in terms of categorical labels and the three-dimensional VAD model labels. The labels assigned to the dimensional VAD model were:

- Valence: Positive, Neither positive nor negative, Negative
- Arousal: Excited, Slightly excited, Neutral
- Dominance: Dominant, Neither dominant nor intimidated, Defensive

The categorical labels were Calm, Sad, Happy, Puzzled and Tense. For each emotion label, the number of segments labeled by each annotator is reported in Table 2. “Calm” is the most frequent label. “Happy” and “Puzzled” are less present but “Sad” and “Tense” are quite absent.

Dealing with the duration of emotion labels, “Calm” occurs in 94% of the audio database size which correspond to more than 6 h. “Happy” and “Puzzled” labels are present in only 4% of the database with respective durations of 9 and 8 min. The negative emotions “Sad” and “Tense” have a total duration lower than one second. This could indicate that the dialog system is user friendly.

4.2 Video Annotation

For the video annotation, all the database is involved and the recordings were labeled by two native people. Six video emotion categories were selected: Sad, Annoyed/Angry, Surprised, Happy/Amused, Pensive and Other. The label Other is assigned to segments containing different emotions that the sub-mentioned ones or including simultaneous emotions. The non annotated parts are considered neutral. For each emotion label, the number of segments labeled by each annotator is reported in Table 3.

With respectively 0, 1 and 3 occurrences, “Sad”, “Annoyed” and “Other” are almost absent. “Pensive” and “Neutral” represent the most frequent labels. Indeed, as more than 14 h are not labeled, the content of the database is mainly neutral. The participants are annotated “Pensive” within a duration of about 2 h. Finally they are sometimes happy or amused (during 5–10 min).

4.3 text Annotation

Emotions were not only labeled from audio and video (Sects. 4.1 and 4.2) but also from text, that is from the manual transcriptions achieved in Sect. 3.1. It was carried out along with the semantic annotation (Sect. 3.2) providing an emotional annotation for each transcribed utterance.

Although the audio and video has richer annotations, the text annotation includes a very significant one related to polarity labels on a scale of three values: negative, neutral and positive. This might be very useful to be combined with the audio annotation in terms of the VAD model. Specifically, when combining of Valence (audio) and Polarity (text) labels we can get the same annotation for different channels.

Looking at the annotated set it can be concluded that Neutral is the most common polarity, representing the 66.24% of the corpus, then a positive behaviour can be analysed, with a 27.21% of the corpus, and finally, negative polarity is almost absent (with 6.55% of occurrences).

5 Concluding Remarks

In this work a Spanish corpus devoted to the development of a dialogue system, oriented to promoting healthy habits among elderly is presented. The corpus was annotated in terms of different labels in order to obtain robust models for generating coherent system responses according to a coaching model. Moreover, an emotion-based annotation is also provided in order to detect emotional status of the speakers and provide a response adapted to it. The procedure carried out to obtain the annotations along with the obtained results is described.

References

1. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59. https://doi.org/10.1016/0005-7916(94)90063-9. http://www.sciencedirect.com/science/article/pii/0005791694900639
2. Bunt H (2009) The DIT++ taxonomy for functional dialogue markup. In: AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, pp 13–24

3. Bunt H, Alexandersson J, Choe JW, Fang AC, Hasida K, Petukhova V, Popescu-Belis A, Traum DR (2012) ISO 24617-2: a semantically-based standard for dialogue annotation. In: LREC, pp 430–437
4. Bunt H, Petukhova V, Malchanau A, Wijnhoven K, Fang A (2016) The dialogbank. In: Chair NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France
5. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1(1):18–37. https://doi.org/10.1109/T-AFFC.2010.1
6. Calvo RA, Mac Kim S (2013) Emotions in text: dimensional and categorical models. Comput Intell 29(3):527–543. https://doi.org/10.1111/j.1467-8640.2012.00456.x
7. Gao J, Galley M, Li L (2019) Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots
8. Justo R, Letaifa LB, Palmero C, Gonzalez-Fraile E, Johansen A, Vazquez A, Cordasco G, Schlogl S, Fernandez-Ruanova B, Silva M, Escalera S, Velasco MD, Tenorio-Laranga J, Esposito A, Kornes M, Torres M (2020) Analysis of the interaction between elderly people and a simulated virtual coach. Amb Intell Humanized Comput
9. Klüwer T (2011) From Chatbots to Dialog Systems. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. https://doi.org/10.4018/978-1-60960-617-6.ch001
10. Ben Letaifa L, Raquel JTMI (2020) Adding dimensional features for emotion recognition on speech. In: International Conference on advanced technologies for signal and image processing, Tunisia, pp 109–114
11. López Zorrilla A, Velasco Vázquez MD, Irastorza J, Olaso Fernández JM, Justo Blanco R, Torres Barañano MI (2018) Empathic: empathic, expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. Procesamiento del Lenguaje Natural
12. Masche J, Le NT (2018) A review of technologies for conversational systems. In: Advanced Computational Methods for Knowledge Engineering. Springer, Cham, pp 212–225. https://doi.org/10.1007/978-3-319-61911-8_19
13. Mctear M (2004) Spoken Dialogue Technology - Toward the Conversational User Interface. Springer, Heidelberg
14. Montenegro C, López Zorrilla A, Mikel Olaso J, Santana R, Justo R, Lozano JA, Torres MI (2019) A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. Multimodal Technol Interact 3(3):52
15. Petukhova V, Bunt H (2012) The coding and annotation of multimodal dialogue acts. In: LREC, pp 430–437
16. Russell J (2003) Core affect and the psychological construction of emotion. Psychol Rev 110:145–172. https://doi.org/10.1037/0033-295X.110.1.145
17. Sayas S (2018) Dialogues on Leisure and Free Time, Dialogues on Physical Exercise, Dialogues on Nutrition. Technical Report DP1, DP2, DP3, Empathic Project; Internal Documents: Tampere, Finland
18. Scerri D, Dingli A (2013) Dialog systems and their inputs. In: Stephanidis C (ed) HCI International 2013 - Posters' Extended Abstracts. Springer, Heidelberg, pp 601–605. https://doi.org/10.1007/978-3-642-39476-8_121
19. Schlögl S, Doherty G, Karamanis N, Luz S (2010) Webwoz: a wizard of Oz prototyping framework. In: Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing System (EICS 2010), pp 109–114. https://doi.org/10.1145/1822018.1822035
20. Schlögl S, Milhorat P, Chollet G, Boudy J (2014) Designing language technology applications: a wizard of Oz driven prototyping framework. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, pp 85–88. https://doi.org/10.3115/v1/E14-2022

21. Serban I, Sordoni A, Bengio Y, Courville AC, Pineau J (2015) Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI
22. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M (2000) Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput Linguist 26(3):339–373
23. Su PH, Gašić M, Young S (2018) Reward estimation for dialogue policy optimisation. Comput Speech Lang 51:24–43. https://doi.org/10.1016/j.csl.2018.02.003
24. Torres MI, Olaso JM, Glackin N, Justo R, Chollet G (2019) A spoken dialogue system for the empathic virtual coach. In: D'Haro LF, Banchs RE, Li H (eds) 9th International Workshop on Spoken Dialogue System Technology, Singapore, pp 259–265
25. Torres MI, Olaso JM, Montenegro C, Santana R, Vázquez A, Justo R, Lozano JA, Schlögl S, Chollet G, Dugan N, Irvine M, Glackin N, Pickard C, Esposito A, Cordasco G, Troncone A, Petrovska-Delacretaz D, Mtibaa A, Hmani MA, Korsnes MS, Martinussen LJ, Escalera S, Cantariño CP, Deroo, O, Gordeeva O, Tenorio-Laranga J, Gonzalez-Fraile E, Fernandez-Ruanova B, Gonzalez-Pinto A (2019) The empathic project: mid-term achievements. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '19. ACM, New York, pp 629–638. https://doi.org/10.1145/3316782.3322764
26. Williams JD, Asadi K, Zweig G (2017) Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: ACL

Publication 12

A Content and Knowledge Management System Supporting Emotion Detection from Speech

A Content and Knowledge Management System Supporting Emotion Detection from Speech

Binh Vu, Mikel deVelasco, Paul Mc Kevitt, Raymond Bond, Robin Turkington, Frederick Booth, Maurice Mulvenna, Michael Fuchs, and Matthias Hemmje

Abstract Emotion recognition has recently attracted much attention in both industrial and academic research as it can be applied in many areas from education to national security. In healthcare, emotion detection has a key role as emotional state is an indicator of depression and mental disease. Much research in this area focuses on extracting emotion related features from images of the human face. Nevertheless, there are many other sources that can identify a person's emotion. In the context of MENHIR, an EU-funded R&D project that applies Affective Computing to support people in their mental health, a new emotion-recognition system based on speech is being developed. However, this system requires comprehensive data-management

---

B. Vu ( )
FernUniversität in Hagen, Hagen, Germany
e-mail: binh.vu@fernuni-hagen.de

M. deVelasco
Universidad del Pais Vasco UPV/EHU, Leioa, Spain
e-mail: mikel.develasco@ehu.eus

P. Mc Kevitt
Ulster University, Derry/Londonderry, Northern Ireland
e-mail: p.mckevitt@ulster.ac.uk

R. Bond · R. Turkington · F. Booth · M. Mulvenna
Ulster University, Newtownabbey, Northern Ireland
e-mail: rb.bond@ulster.ac.uk

R. Turkington
e-mail: turkington-r@ulster.ac.uk

F. Booth
e-mail: booth-f@ulster.ac.uk

M. Mulvenna
e-mail: md.mulvenna@ulster.ac.uk

M. Fuchs · M. Hemmje
GLOBIT GmbH, Barsbüttel, Germany
e-mail: m.fuchs@globit.com

M. Hemmje
e-mail: matthias.hemmje@globit.com

L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_28

support in order to manage its input data and analysis results. As a result, a cloud-based, high-performance, scalable, and accessible ecosystem for supporting speech-based emotion detection is currently developed and discussed here.

1 Introduction and Motivation

Affective Computing is an emerging inter-disciplinary field developing technology that attempts to detect, analyse, process, and respond to important psychological traits such as emotions, feelings, or behaviours with the goal of improving human-computer interaction [1]. Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is a 4-year project funded by the European Union (EU), that applies Affective Computing to enhance and advance future healthcare processes and systems, especially in providing assistance to people with dementia, medical professionals, and caregivers [2]. By gathering activity and related sensor data to infer the emotional state of the patient as a knowledge stream of emotional signals, SenseCare can provide a basis for enhanced care and can alert medics, professional carer, and family members to situations where intervention is required [3, 4].

One of the systems developed in SenseCare is a machine-learning-based emotion detection platform [5] which provides an early insight into the emotional state of an observed person. SenseCare can process a live video stream or a pre-recorded video which enables analysis to be completed on the fly or at a later stage. Similar to SenseCare, the MENtal Health monitoring through InteRactive conversations (MENHIR) is a EU-funded project that aims to support and improve the mental wellbeing of people by applying Affective Computing, especially conversational technologies, such as emotion recognition in speech, automatic conversation management (chatbots), and other multidisciplinary topics [6]. According to the World Health Organization (WHO), mental, neurological, and substance use disorders make up 10% of the global, and 30% of non-fatal, disease burden. The global economy loses about US$ 1 trillion per year in productivity due to depression and anxiety [7].

In MENHIR, new research assists people with improving their current state of emotion and provides a long-term overview of their state over time. A machine-learning-based emotion detection platform has been developed. Unlike SenseCare, where human emotions are extracted from a live video stream or a pre-recorded video, the MENHIR emotion detection platform identifies emotions from speech. The system relies on short-term features such as pitch, vocal tract features such as formants, prosodic features such as pitch loudness, as well as speaking rate to perform effectively. Furthermore, recurrent neural networks are applied to predict emotion in real-time using a 3D emotional model. This paper discusses the challenges of emotion detection based on speech and its corresponding transcription in the MENHIR project. Furthermore, it provides a solution to overcome these challenges. The architecture of the proposed system and its constituent components are described. Finally, we conclude and discuss future work.

2 Problem Statement

One of the goals of the MENHIR project is to further extend the results of earlier research work, expanding the set of identified depressive speech acoustic features and automating their detection so that depressed and anxious speech can be accurately distinguished from healthy speech [8]. To enable this, challenging scenarios need to be considered and overcome as discussed here.

After a series of human-to-human counselling conversations are recorded in a laboratory setting, a corpus of audio data of conversations is formed. Along with the audio files, their metadata, which consists of documents describing the conversations and spreadsheets describing the conversation results, are also provided for advanced annotation and analysis. All these data need to be stored in a high-performance repository where other analysis systems can connect to and download them when needed. Furthermore, multimedia objects usually take up a lot of storage space. This means the data repository also needs to be scalable to fulfil users' demands in the future.

In MENHIR, not only multimedia objects but also other kinds of scientific content, knowledge, and their metadata need to be imported, stored, and managed. Sharing and exchanging research results powers collaborative and co-creative networking among project participants. Therefore, a solution is needed to support the ingestion of scientific publications from different sources. Here, the imported content can be managed and transformed into learning materials. Similar to multimedia objects, scientific data content also needs high-performance, scalable, and fault-tolerant storage. Furthermore, a content management system will enable users to edit, share, and publish their content.

There are a number of collaborative services producing analysis results and generating observed subject and patient conversational behaviour, such as, authentication, authorization, speech analysis data services, big data speech analysis, collaboration and coordination services, psychological/affective analytics, reporting/result sharing and reproducibility services [8]. It is crucial to have an integration architecture for all the mental health services and applications employed in MENHIR. This architecture will provide a common platform for these systems to communicate in a predefined flow, where input data is received and results are stored.

For research results to make an impact, they need to be easily found and used. Meanwhile, related publications, datasets, and analysis results are distributed in different locations. Therefore, one needs to find a means to automatically gather and combine all these resources into scientific asset packages. Otherwise, users can only find fragments of related information. It will prevent them from having a complete overview of the research topic and discovering important relationships between factors. Organizing related information and data into scientific asset packages is a powerful method of systemizing results produced by conversational technologies.

Finally, classification helps to narrow the choices among content, information, and knowledge resources. By dividing the material into reduced subsets, classification can make content, information, and knowledge resource retrieval and access faster and more accurate [9]. In MENHIR, a considerable volume of subject data will be analysed by an emotion detection server and will be made available for use by, for example, chatbots. Furthermore, the analysis results and related scientific publications will also be generated and managed in MENHIR. Without organizing the content created into suitable categories, researchers will not have capacity for insight on the key data produced in MENHIR, discover connections between data or whether something is missing. Therefore, a system that allows the content, knowledge, analysis results, and datasets to be classified is critical for the success of MENHIR.

3 System Design

Based on these challenges, we have developed a system design to support MENHIR in the task of conversational technologies research and development. In this section, a cloud-based Content and Knowledge Management Ecosystem (KM-EP) for audio files and metadata persistence, human emotion detection, as well as asset packaging, classification, and management is introduced and described. Figure 1 illustrates the architecture of the system, which comprises the MENHIR KM-EP and supporting systems.

The MENHIR Content and Knowledge Management Ecosystem (KM-EP) provides a platform for managing scientific as well as educational content and knowledge resources. Furthermore, the KM-EP will act as a framework for researchers to deploy their work without spending time reimplementing basic functionalities, such as, e.g. user management and task scheduling. In Fig. 1, four components of the MENHIR KM-EP, which are related and crucial for the tasks of audio data persistence, emotion detection, as well as asset packaging, classification, and management, are shown. The components are Media Archive (MA), Digital Library (DL), Taxonomy Manager (TM), and Asset Manager (AM).

The first component that is needed for the MENHIR KM-EP is the Media Archive (MA). The MA manages all multimedia objects in the KM-EP. The MA enables users to create, persist, manage, and classify different types of multimedia objects, such as, e.g. video, audio, images, presentation slides, along with their metadata. In MENHIR, audio files and their metadata need to be imported and stored together in the system. The initial audio files are, e.g., recordings of interviews, which are conducted in order to form a corpus of conversational audio data. This corpus will be used to validate the operation of the Emotion Detection Server. Their metadata consists, e.g., of documents describing the interviews and spreadsheets describing the interview results. Furthermore, audio files of conversations and interviews can also be uploaded and linked to user accounts automatically or manually. The KM-EP provides an interface where users can upload these files into the system and populate basic metadata information related to them, such as, e.g. title, description, authors, and creation date. The uploaded files are stored in a cloud storage service, which is fault-tolerant, flexible, scalable, and has high performance. This will enable users to have fast and stable access to the files worldwide. Furthermore, with the support of

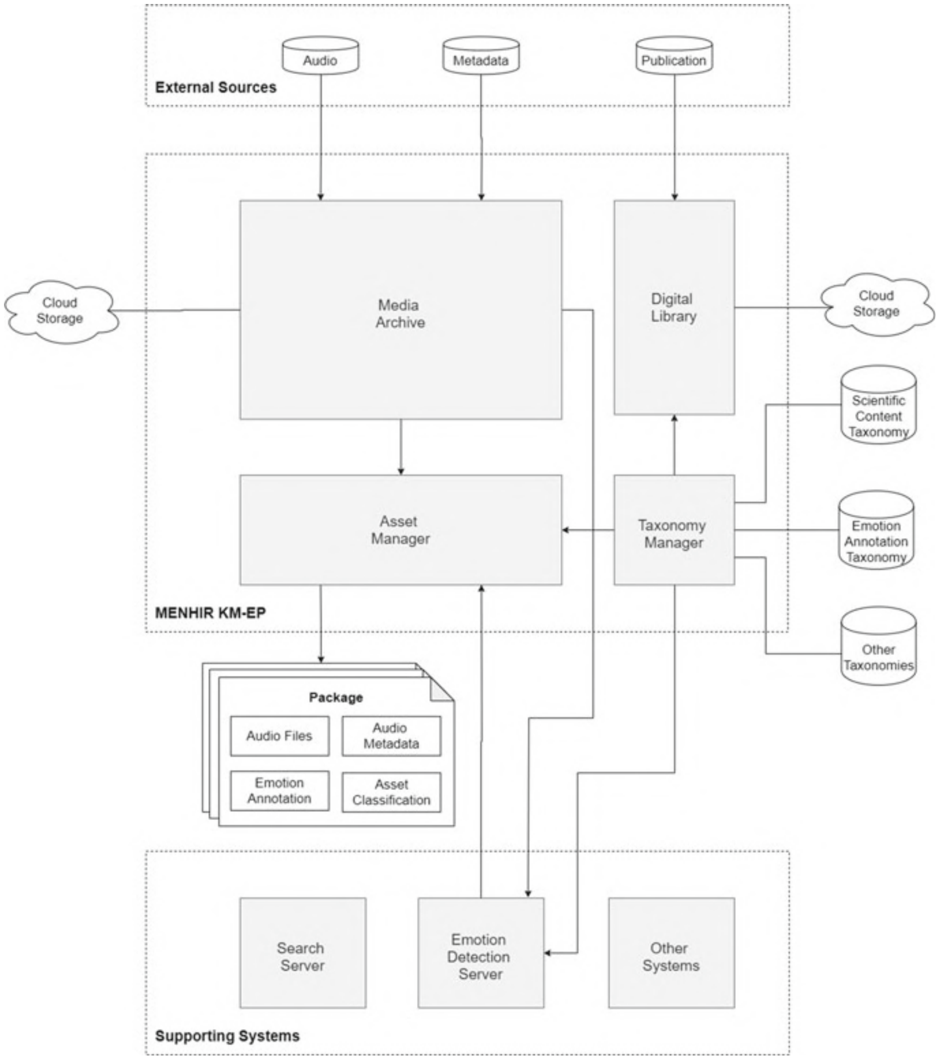


Fig. 1 Architecture of MENHIR content and knowledge management ecosystem (KM-EP)

the TM component, multimedia objects can be classified into different categories. Classification enables objects to be searched and accessed easily and quickly by users.

The next component of the KM-EP is the Digital Library (DL). The DL enables users to import publications into the KM-EP, persist, and manage them. Using a Mediator-Wrapper Architecture, publications from different sources, such as, e.g. Mendeley [10], SlideShare [11], and in different formats, such as BibTex [12] and OAI-PMH [13], can be queried, uploaded, and integrated into the DL [14]. Similar to the MA, after importing or creating a new publication using the DL, users have the option to fill in its metadata, such as, e.g. title, abstract, conference, publisher, and upload the document. The uploaded files will also be stored in cloud storage to maintain their availability and scalability. By indexing file metadata and classifying publications into existing categories, they can be searched by users based on these criteria.

The Taxonomy Manager (TM) component supports the construction, collaboration, management, and evolution of taxonomies. The TM enables users to develop and manage their own taxonomies. With the support of its version control system, users can manage the changes of their taxonomies. Every modification is tracked and can be reverted. A complete history of changes helps users to compare different versions of a taxonomy. Furthermore, the branching feature enables users to create multiple versions of a taxonomy.

Multimedia objects, publications, and assets of the MENHIR KM-EP can be classified with support of the TM. As a result, users can search and browse contents quickly and easily. Classification also enables navigation inside the KM-EP. A persistent identifier introduced for each term in a taxonomy enables taxonomy evolution without affecting existing classifications. A rating system is implemented based on crowd voting to support the evaluation of taxonomies in the KM-EP. With the rating system, authors can improve the accessibility of their taxonomies, and users can also choose quickly more relevant taxonomies. A caching system enables thousands of taxonomies and terms to be retrieved and constructed in just a few milliseconds.

In MENHIR, the TM can not only be used to collect, classify, and provide access to audio materials from initial emotion analysis and results but can also support the emotion detection platform by providing an emotion annotation taxonomy. The machine learning platform can use this taxonomy to label its training and validation set. This creates a standard emotion classification that can be used for classifying results produced later by the platform. This process would be more costly without the classification, annotation, and access support of the TM in the MENHIR KM-EP supporting scientific research in the domain of Affective Computing.

The Asset Manager (AM) component is where related data, metadata, analysis results, and classification are gathered and combined into packages. In order to do this, a cronjob is developed and scheduled to run regularly after a given period of time. This cronjob has 3 tasks, which are: (1) searching for new audio files and their metadata and adding them into a new asset, (2) sending the new audio files and their metadata to the emotion detection server for analysis, and (3) receiving and adding analysis results into its package. This guarantees that new data will always be processed after it is uploaded to the MENHIR KM-EP.

After the cronjob has been started, the daemon searches for audio files along with their metadata in the MA. For each audio record found, the daemon will check if it belongs to a package in the AM or not. If it exists in a package, the emotion detection process, along with other processes, has been already completed for this audio record and the daemon continues to work with other audio records. Otherwise, the daemon needs to gather necessary data, such as uploaded files, documents describing the counseling interview where the audio file was recorded, and the spreadsheets describing the interview results. These files will be downloaded from the current cloud storage service to a temporary location in the local server. Next, the daemon

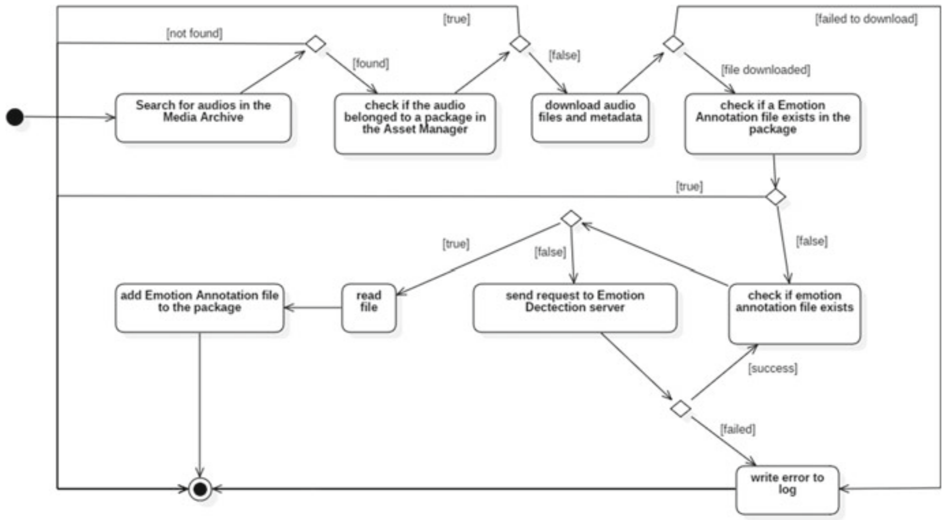


Fig. 2 Activity diagram of the asset manager (AM) cronjob

checks if an emotion annotation was produced for the audio record. If it has been produced, the daemon will go to the next step. Otherwise, it will search for the annotation file produced by the Emotion Detection server. If this file does not exist, the daemon will send a request to the server and let it process the downloaded audio file along with its metadata. An annotation file will then be produced by the server. The daemon reads the file and adds it to the new package. Errors that occur whilst the daemon is running will be written to the log file in order to enable the system administrator to debug them later. Figure 2 describes the activity flow of the cronjob.

After an emotion annotation file is added, users can use the Emotion Audio Player (EAP), which is an important feature of the AM, to play the audio files and discover the current emotional state of the subject in the audio. The emotion in the annotation file will be indexed based on its timestamp. When the audio playback reaches a timestamp, the player will display the emotion associated with it. Furthermore, with this annotation file, emotions of the subject in the audio file can be visualized with various visualization techniques. This enables users to have an overview of the current emotional state of the subject and provides an opportunity to explore hidden information behind human emotion.

Finally, the AM enables users to edit, delete and classify their packages through interacting with a user interface. Not only audio records and respective analysis results inside a package can be classified using the scientific content, emotion annotation, and other types of taxonomies, but the package itself can be classified into different categories using the TM. This classification will be stored in the package as well as indexed in a search server. With the AM, scientific content and respective analysis results can be managed in a central repository. This will reduce dramatically the effort to deploy, maintain, search, and reuse scientific data.

Supporting systems such as Emotion Detection Server and Search Server provide standalone, high-performance services that the MENHIR KM-EP can take advantage of. They provide interfaces, so the KM-EP can send requests and later receive results. In the context of this paper, we focus on the Emotion Detection Server, which is being developed in MENHIR.

The Emotion Detection Server detects human emotion from speech signals extracted from the audio files and their transcriptions. The files will be downloaded from the MA and sent to the processing server by the introduced cronjob from the AM component. The audio samples are processed by the server and their results are exported to annotation files and stored in the local server for the KM-EP to access and use. Automatic recognition of spontaneous emotions from speech is complex [15]. To overcome its challenges, two procedures have been conducted. The first one is the annotation task, that involves the segmentation of the audio samples in order to label them with emotions, and the second one is building a model that is able to distinguish between different emotional states.

In relation to the annotation, transcriptions are used to identify the spoken turns, and those turns have been split automatically into segments of between 2 and 5 5 s, because it is known that there is no emotion change in this time window. Subsequently, each segment is labelled by both professional and crowd annotators following the same questionnaire. The questionnaire includes both categorical and dimensional annotation (valence, arousal, and dominance). Using these annotations, we have experimented with the creation of a model capable of identifying the mood of the speaker through application of neural network algorithms. This model infers the subject's emotional state using both audio features (such as e.g. pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs)) or the spectrogram. With this model, an emotion detection server can be developed to provide the MENHIR KM-EP with emotion annotations from both acoustic signals and their corresponding transcription in real-time.

A high-performance search server is needed to index content objects in the MA and the DL, so they can be searched quickly by users. Furthermore, indexing classifications enables faceted search, which is a way to add specific, relevant options to the results pages, so that when users search for content, they can see where in the catalogue they've ended up [16]. With the faceted feature, users can have an overview of the classification of contents in real-time and quickly find results by selecting only relevant categories. Furthermore, faceted search enables navigation using taxonomy hierarchies, which are created and managed using the TM [17].

Besides a search server, other systems, such as, e.g. cache server, queuing system, are also important for the MENHIR KM-EP. Caching improves performance of the system by pre-processing and storing frequently used data in memory, each time it is required, it can be retrieved from there without requiring reconstruction resources. A queuing system enables the KM-EP to process data in an organized manner. Processing all data at once requires considerable computing power and resources. Therefore, organizing data into a queue and processing them accordingly would allow the resource to be distributed evenly and reduce stress on system components.

4 Conclusion and Future Work

The MENHIR project provides rapid intervention, appropriate feedback and overview on the state of development of subject mood and anxiety levels over time, by monitoring moods, behaviour, and symptoms of subjects in real time. The objective of the work reported in this paper is to develop an integration platform to support the ingestion and management of audio files and their metadata, results on human emotion detection from speech, and scientific asset packaging, classification, and management.

Here, we have described the challenges involved in the development and integration of such a platform. The content and knowledge management ecosystem (KM-EP) proposed here is a cloud-based, high-performance, scalable, and easy to use solution. By relying on its Media Archive and Digital Library, the KM-EP is able to ingest, modify, share, and preserve scientific publications and multimedia objects, such as audio files and their metadata. The Taxonomy Manager enables users to classify content and knowledge, which leads to better quality and faster exploration. Finally, the Asset Manager combines related scientific publications, multimedia objects, datasets, and analysis results into packages. With the Asset Manager, all the related data, information, and knowledge can be gathered and managed in one central repository, which is easier to maintain and reuse. The MENHIR KM-EP will provide a useful foundation for the development of conversational systems in mental health promotion and assistance.

The current emotion detection server uses a model, which needs to be trained offline by AI experts. This model also needs to be re-trained frequently with updated corpora to enhanced its accuracy. The MENHIR KM-EP can be extended in the future to use the uploaded audio records in the MA to form a new corpus. Then, the new model can be trained based on the new data corpus and replace the former model automatically. By doing this, the cost of developing an advanced emotion detection model can be reduced.

Acknowledgements This publication has been produced in the context of the MENHIR project. This project has received funding from the European Union's H2020 Programme under grant agreement No. 823907. However, this paper reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

1. Healy M, Donovan R, Walsh P, Zheng H (2018) A machine learning emotion detection platform to support affective well being. In: IEEE international conference on bioinformatics and biomedicine
2. Sensor enabled affective computing for enhancing medical care, 19 April 2017. https://cordis.europa.eu/project/rcn/199563/factsheet/en. Accessed 27 Aug 2019

3. Engel F, Bond R, Keary A, Mulvenna M, Walsh P, Zheng H, Wang H, Kowohl U, Hemmje M (2016) SenseCare: towards an experimental platform for home-based, visualisation of emotional states of people with dementia. In: Advanced visual interfaces. Supporting big data applications
4. Healy M, Walsh P (2017) Detecting demeanor for healthcare with machine learning. In: IEEE international conference on bioinformatics and biomedicine
5. Donovan R, Healy M, Zheng H, Engel F, Vu B, Fuchs M, Walsh P, Hemmje M, Kevitt PM (2018) SenseCare: using automatic emotional analysis to provide effective tools for supporting wellbeing. In: IEEE international conference on bioinformatics and biomedicine
6. MENHIR. https://menhir-project.eu/. Accessed 20 Jan 2020
7. Mental health. WHO, 2 Oct 2019. https://www.who.int/news-room/facts-in-pictures/detail/mental-health. Accessed 20 Jan 2020
8. Consortium M (2018) MENHIR proposal. European Commission
9. Vu B, Mertens J, Gaisbachgrabner K, Fuchs M, Hemmje M (2018) Supporting taxonomy management and evolution in a web-based knowledge management system. In: HCI 2018, Belfast, UK
10. Mendeley. https://www.mendeley.com/. Accessed 27 Jan 2020
11. SlideShare. https://de.slideshare.net. Accessed 27 Jan 2020
12. Your BibTeX resource (2016). BibTeX. http://www.bibtex.org/. Accessed 28 Oct 2019
13. Protocol for Metadata Harvesting. Open Archives Initiative. https://www.openarchives.org/pmh/. Accessed 28 Oct 2019
14. Vu B, Wu Y, Afli H, Kevitt PM, Walsh P, Engel F, Fuchs M, Hemmje M (2019) A metagenomic content and knowledge management ecosystem platform. In: IEEE international conference on bioinformatics and biomedicine, San Diego, USA
15. de Vázquez M, Justo R, López Zorrilla A, Inés Torres M (2019) Can spontaneous emotions be detected from speech on TV political debates? In: IEEE international conference on cognitive infocommunications
16. What is faceted search and navigation? (2016) Loop54. https://www.loop54.com/knowledge-base/what-is-faceted-search-navigation. Accessed Apr 2019
17. Vu B, Donovan R, Healy M, Kevitt PM, Walsh P, Engel F, Fuchs M, Hemmje M (2019) Using an affective computing taxonomy management system to support data management in personality traits. In: IEEE international conference on bioinformatics and biomedicine, San Diego, USA

Publication 13

A Differentiable Generative Adversarial Network for Open Domain Dialogue

A Differentiable Generative Adversarial Network for Open Domain Dialogue

Asier López Zorrilla, Mikel deVelasco Vázquez, and M. Inés Torres

Abstract This work presents a novel methodology to train open domain neural dialogue systems within the framework of Generative Adversarial Networks with gradient based optimization methods. We avoid the non-differentiability related to text-generating networks approximating the word vector corresponding to each generated token via a *top-k softmax*. We show that a weighted average of the word vectors of the most probable tokens computed from the probabilities resulting of the top-k softmax leads to a good approximation of the word vector of the generated token. Finally we demonstrate through a human evaluation process that training a neural dialogue system via adversarial learning with this method successfully discourages it from producing generic responses. Instead it tends to produce more informative and variate ones.

Keywords Dialogue systems · Generative adversarial networks · Open domain dialogue

1 Introduction

Open domain dialogue systems or chatbots are systems deployed to interact with humans offering coherent responses according to the dialogue history. Unlike task-oriented dialogue systems, there is no specific goal to be achieved during the inter-

A. López Zorrilla ( ) · M. deVelasco Vázquez · M. I. Torres
Faculty of Science and Technology, University of the Basque Country UPV/EHU, Sarriena s/n, 48940 Leioa, Spain
e-mail: asier.lopezz@ehu.eus

M. deVelasco Vázquez
e-mail: mikel.develasco@ehu.eus

M. I. Torres
e-mail: manes.torres@ehu.eus

E. Marchi et al. (eds.), *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Lecture Notes in Electrical Engineering 714,
https://doi.org/10.1007/978-981-15-9323-9_24

action by the system. The only goal is to generate appropriate, relevant, meaningful and human-like utterances.

This area of research has gained an increasing amount of interest from the community since the advent of sequence-to-sequence neural network models [22]. These neural networks are capable of processing and generating sequences of data of arbitrary length, which makes them very suitable for this research [21, 24]. The task of open domain dialogue generation can easily be cast as a sequence transduction problem, where the input is the sequence of words corresponding to the last user's utterance, and the output are the words of the system's response. It is also possible to condition the output of the network to a broader dialogue context or other knowledge sources in order to increase the coherence of the responses [6, 19], but in this work we will not research in that direction.

These neural models are usually learnt from corpora composed of input utterance-response pairs, via supervised learning. Movies subtitles, Twitter or online forums can be used as the source of these data. In this framework, the neural network is trained to minimize a distance between the generated response and the desired one. Even though interesting performances can be obtained with this procedure, it frequently yields models that tend to generate dull and safe responses which appear frequently in the corpus, such as *I don't know* or *I'm sorry*.

We build upon Generative Adversarial Networks (GANs) [7] to overcome this problem and to increase the overall variety in the responses of the neural dialogue model, as these have shown promising results in many data generation tasks. While in supervised learning a unique desired output is assigned to each input in the corpus, GANs allow many correct outputs, which makes much more sense in dialogue, and models better the one-to-many property of input-output pairs [23]. The learning methodology for GANs involves training two neural networks, a generator and a discriminator, in an adversarial fashion. The generator tries to learn a data distribution while the discriminator learns whether a given sample corresponds to the training data or has been generated by the generator. In the context of dialogue systems, the generator would be the sequence-to-sequence model and the discriminator would act as a Turing Test.

GANs were first successful in image generation tasks. More recently text-related problems, such as machine translation [25], text generation [26, 27] or image captioning [20] have also been tackled within this framework. GANs have also been applied in the research of dialogue systems, yet only on a few occasions. References [5, 11] experiment with training discriminators that could measure the quality of the utterances generated by chatbots. On the other hand [9, 14] go a step further and train neural dialogue systems via adversarial learning, but with the drawback that they make use of reinforcement learning instead of gradient-based optimization methods. This is due to text being represented as a sequence of discrete tokens, which breaks the differentiability of the discriminator's output with respect to the generator's parameters, as explained in Sect. 3.

In this context, the contributions of our work are twofold. First, we present a novel methodology to avoid this non-differentiability: the *top-k softmax*. Since the top-k softmax allows to plug-in the output of the generator into the discriminator in a differentiable manner, our approach is simpler and easier to implement than other dialogue systems trained in the GAN framework. Second, we demonstrate that training a neural dialogue system via adversarial learning with this method successfully discourages it from producing generic responses, and that it often leads to more informative responses too.

The rest of the paper is organized as follows. In Sect. 2 we specify the chosen architecture for the sequence-to-sequence dialogue model and the baseline training procedure. In Sect. 3 we describe the proposed GAN for dialogue generation based on the top-k softmax and compare it to alternative approaches to deal with the differentiability problem. In Sect. 4 we give all the details about our experimental setup and hyper-parameter choice. Section 5 shows the results of two experiments to validate our proposal. We conclude with the final remarks in Sect. 6.

2 Sequence-to-Sequence Dialogue Model Architecture

The chosen architecture for the dialogue model is a standard sequence-to-sequence network with attention [1]. Given an input sequence of length T of discrete integer tokens $x = x_1, x_2, ..., x_T$, the corresponding sequence of vectorial word representations $\mathbf{v} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T$ can be obtained via the word vector matrix $\mathbf{W}$, just by taking the corresponding row $\mathbf{v}_i = \mathbf{W}[x_i]$ per each token x_i. The size of $\mathbf{W}$ is $V \times D$, where V is the vocabulary size and D the dimension of each word vector. The encoder takes this sequence of vectors and produces another sequence of vectors of the same length $\mathbf{h} = \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T = \text{encoder}(\mathbf{v})$. In our work the encoder is a deep bidirectional Long Short Term Memory (LSTM) Recurrent Neural Network (RNN).

To proceed with the generation of the output sequence $y = y_1, y_2, ..., y_\tau$, a global attention mechanism is applied as in [17]. At the time step t of the generation, the decoder is fed with the discrete integer token generated at previous time step, y_{t-1}. Then the corresponding word vector $\mathbf{W}[y_{t-1}]$ is input to the decoder's RNN and this outputs $\mathbf{o}_t$. Of course, due to the architecture of RNNs, $\mathbf{o}_t$ is conditioned, though implicitly, not only to y_{t-1} but also to all the previously generated tokens. In our experiments this neural network is also a deep LSTM. $\mathbf{o}_t$ is then transformed to $\tilde{\mathbf{o}}_t$ via a multilayer perceptron (MLP) that takes as input o_t and also $\mathbf{c}_t$, the context-vector produced by the attention mechanism at time step t. $\mathbf{c}_t$ is a weighted average of the encoder's output vectors:

$$\mathbf{c}_t = \sum_{j=1}^{T} a_{jt}\mathbf{h}_j \ , \tag{1}$$

where a_{jt} is the score between $\mathbf{h}_j$ and $\mathbf{o}_t$, i.e., how much attention should be put on the output of the encoder at the encoding time step j on the time step t of the decoding phase. a_{jt} is a softmax-normalized scalar output of another MLP, that takes

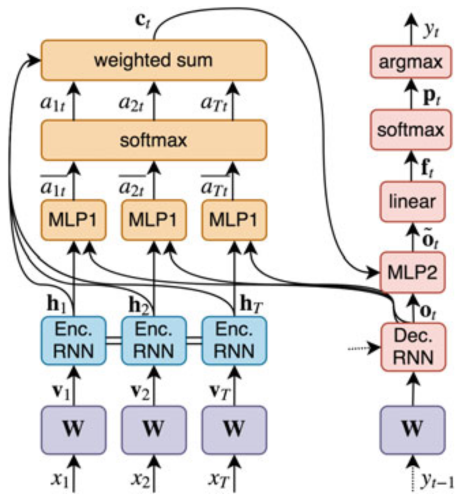


Fig. 1 A diagram of the chosen sequence-to-sequence network: blue transformations refer to the encoder, orange to the attention mechanism, purple to the word matrix (shared between the encoder and decoder), and red to the decoder. For simplicity, only the time step t of the decoding is shown

as input $\mathbf{h}_j$ and $\mathbf{o}_t$, and outputs $\overline{a_{jt}}$. With the softmax normalization we ensure that all the scores at time step t are positive and sum one:

$$a_{jt} = \frac{\exp\left(\overline{a_{jt}}\right)}{\sum_{j'=1}^{T} \exp(\overline{a_{j't}})} \tag{2}$$

Finally, $\tilde{\mathbf{o}}_t$ is linearly projected to a vector of dimension V: $\mathbf{f}_t = \text{linear}(\tilde{\mathbf{o}}_t)$. This vector represents an unnormalized probability distribution over all the possible words in the vocabulary. A softmax normalization is then applied to $\mathbf{f}_t$ to get $\mathbf{p}_t = \text{softmax}(\mathbf{f}_t)$, the normalized version of $\mathbf{f}_t$. The output token at time step t, y_t, can be sampled from $\mathbf{p}_t$ taking the argument of the maxima:

$$y_t = \arg\max_i \left(\mathbf{p}_t[i]\right) \tag{3}$$

Generation stops at time τ, when y_τ corresponds to the end-of-sequence token. The architecture of the network is summarized in Fig. 1.

maximum likelihood estimation via supervised learning As aforementioned, this neural network can be trained from a corpus composed of input-output sequence pairs via supervised learning. A maximum likelihood estimation (MLE) of the parameters of the network can be carried out by minimizing the word level cross entropy loss L_{MLE}:

$$L_{MLE} = \frac{1}{|\mathcal{C}|} \sum_{x,s \in \mathcal{C}} \frac{1}{|s|} \sum_{t=1}^{|s|} -\log \mathbf{p}_t[s_t] \ , \tag{4}$$

where $\mathcal{C}$ is a corpus composed of pairs of inputs x and desired outputs s, s_t each of the words in s, and $\mathbf{p}_t[s_t]$ the output of the network in the t-th time step corresponding to the token s_t. We omit the output's dependence on x to keep the notation simple.

During training we employ the teacher forcing strategy, i.e., in the t-th step of the decoding we feed the ground true token s_{t-1} to the decoder's RNN instead of the prediction y_{t-1}. We experimented with other sampling techniques such as scheduled sampling [2], but we found no improvement.

3 Sequence Generative Adversarial Network Training

In the context of dialogue systems, the generator network in the GAN is the sequence-to-sequence dialogue model, which produces a response y to the input utterance x. The discriminator is another network that acts like a Turing Test: it takes an input utterance x and a response r as inputs, and outputs a scalar between 0 and 1 representing the network's confidence level on r being produced by a chatbot. Namely, the lower the output of the discriminator is, the more human-like r is according to the discriminator's criteria.

The procedure to train the dialogue system in this framework involves iteratively updating the generator and the discriminator. The generator is trained to fool the discriminator and make it think that its responses are human-like, and in contrast the discriminator is trained to distinguish between human and bot responses.

Let us now define the losses to be minimized in this two optimization procedures. Given a batch of input utterances, responses and labels indicating whether each response has been generated by a bot or a human, the discriminator's parameters will be updated to minimize the next cross-entropy loss:

$$L_D = \frac{1}{|\mathcal{B}_D|} \sum_{x,r,l \in \mathcal{B}_D} -\left[l \cdot \log a + (1-l) \cdot \log(1-a)\right] , \tag{5}$$

where $\mathcal{B}_D$ is a batch composed of tuples of input utterances x, responses r and boolean labels l, and a the output of the network given x and r.

The objective for the generator is just to minimize the output of the discriminator when the latter is fed with a batch of input utterances and the responses of the generator to those same input utterances:

$$L_G = \frac{1}{|\mathcal{B}_G|} \sum_{x \in \mathcal{B}_G} a , \tag{6}$$

where $\mathcal{B}_G$ is a batch composed of input utterances x. a is the output of the discriminator given x and y, where y is the output of the generator given x.

the differentiability problem We have already described the architecture of the generator in Sect. 2. On the other hand, the discriminator is a composed of two deep bidirectional LSTM-RNNs, for x and r respectively, followed by some fully-connected layers. Before being processed by the RNNs, both x and r integer sequences are converted to word vector sequences via the same word vector matrix $\mathbf{W}$, as explained in Sect. 2.

Being these the network architectures, it is not possible to differentiate L_G (Eq. 6) with respect to the parameters of the generator. The problem arises with the argmax operation in the sequence of transformations that converts $\mathbf{f}_t$ into $\mathbf{u}_t$:

$$\mathbf{f}_t \underset{\longleftarrow}{\xrightarrow{\text{softmax}}} \mathbf{p}_t \underset{\not\longleftarrow}{\xrightarrow{\text{argmax}}} y_t \underset{\longleftarrow}{\xrightarrow{\mathbf{W}[y_t]}} \mathbf{u}_t \ , \tag{7}$$

where $\mathbf{f}_t$ is the unnormalized probability distribution over all the possible words in the vocabulary in the step t of the generation, $\mathbf{p}_t$ the softmax-normalized version of $\mathbf{f}_t$, y_t the argument of the maxima of $\mathbf{p}_t$, and $\mathbf{u}_t$ is the word vector corresponding to the token y_t. Green arrows indicate that the operation is differentiable, whereas red arrows that it is not.

the top-k softmax We propose a novel alternative computation path that approximates $\mathbf{u}_t$ in a fully differentiable manner, allowing the generator to be trained with very convenient gradient-based methods. The idea behind this path is to generate a word vector $\tilde{\mathbf{u}}_t$, hopefully similar to $\mathbf{u}_t$, as a weighted average over the word vectors corresponding to the k most probable words according to $\mathbf{f}_t$. $k \geq 2$ is an integer parameter of the transformation. In short, the differentiable computation path is as follows:

$$\mathbf{f}_t \underset{\longleftarrow}{\xrightarrow{\text{top-k}}} \mathbf{k}_t, \tilde{\mathbf{f}}_t \underset{\longleftarrow}{\xrightarrow{\text{softmax}}} \mathbf{k}_t, \tilde{\mathbf{p}}_t \underset{\longleftarrow}{\xrightarrow{\sum_i \tilde{\mathbf{p}}_t[i] \cdot \mathbf{W}[\mathbf{k}_t[i]]}} \tilde{\mathbf{u}}_t \tag{8}$$

The first operation in Eq. 8 performs a selection of the *top-k* elements in $\mathbf{f}_t$. It outputs $\mathbf{k}_t$ and $\tilde{\mathbf{f}}_t$. $\mathbf{k}_t$ are the indices corresponding to the k elements in $\mathbf{f}_t$ with the highest values, and $\tilde{\mathbf{f}}_t$ are those values. In other words, $\mathbf{k}_t$ represents the most probable words, and $\tilde{\mathbf{f}}_t$ their unnormalized probabilities. The second operation is just a softmax normalization of these k probabilities. It converts $\tilde{\mathbf{f}}_t$ into $\tilde{\mathbf{p}}_t$. See Fig. 2 for a graphical example. Finally, the approximated word vector that will be fed to the discriminator's RNN is computed as the weighted average of the word vectors corresponding to tokens $\mathbf{k}_t$, where the weights are the probabilities $\tilde{\mathbf{p}}_t$:

$$\tilde{\mathbf{u}}_t = \sum_{i=1}^{k} \tilde{\mathbf{p}}_t[i] \cdot \mathbf{W}[\mathbf{k}_t[i]] \tag{9}$$

Note that in the whole process the differentiability has not been broken. Therefore, and in contrary to the previous computation path (Eq. 7), the partial derivatives of $\tilde{\mathbf{u}}_t$ with respect to $\tilde{\mathbf{f}}_t$ exist and are non zero. In Sect. 5 we show that $\tilde{\mathbf{u}}_t$ is a good approximation of $\mathbf{u}_t$ when k is small. In fact, $\mathbf{u}_t$ is the nearest neighbor of $\tilde{\mathbf{u}}_t$ the 98% of the times with $k = 2$.

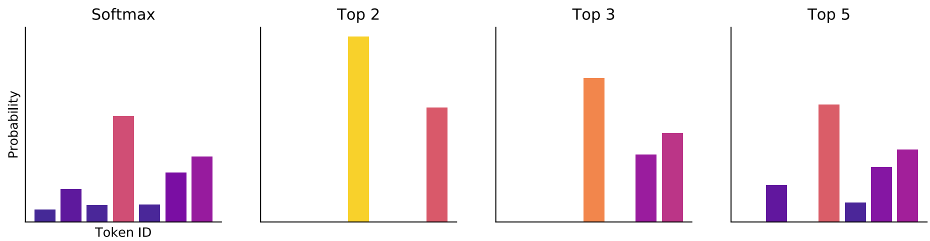


Fig. 2 On the left, a graphical example of a softmax normalization of a $\mathbf{f}_t$ distribution. The rest of the plots show the top-k softmax normalizations of $\mathbf{f}_t$ for different values of k

related approaches Before continuing with our proposal for training the GAN, let us briefly compare the top-k softmax with alternative approaches to deal with the non-differentiable argmax operation. Apart from the aforementioned reinforcement learning-related methodologies based on [27], we are only aware of works [13] that in one way or another tackle this problem with the concrete or Gumbel-softmax distribution [10, 18]. This is a continuous relaxation of discrete random variables. In short, it transforms a probability distribution into a relaxed one-hot vector corresponding to a randomly taken sample from that distribution. That relaxed vector is different from the result of the top-k softmax in two important aspects. First, it is non-deterministic, which could be interesting but also unnecessary for our application. Second, all its elements are non-zero, which means that approximating a word vector as a weighted average according to those probabilities would imply mixing all the word vectors in the vocabulary, which seems again inadequate for our application.

A discrete version of this transformation is the Straight-Through Gumbel-softmax estimator [3, 10], which was used by [16, 20]. It serves to approximate the gradients of a one hot vector sampled according to a probability distribution. Thus it avoids the problem of averaging over all the word vectors, but it is still non-deterministic. Moreover, the operation is still non-differentiable. Even though this method provides an estimation of the gradients in this scenario, but using it could be risky because it might cause discrepancies between the forward and backward passes, as stated in the original work [10].

training procedure The top-k softmax allows L_D to be differentiable with respect to the parameters of the generator. Thus gradient-based optimization methods can be applied to train both the generator and the discriminator. Let us now specify the general training loop and the pretraining strategies applied in this work.

Prior to the training of the dialogue system, we pretrain the word vector matrix in the same corpus that will be used later. Following the work of [9, 14], we also pretrain the generator using the MLE criteria, and the discriminator with the responses generated by the pretrained generator and with responses from the corpus. In order to stabilize the rest of the training process and to avoid the catastrophic forgetting phenomenon of the discriminator, each time we sample a response of the generator to a given input, we add it to a corpus of generator's turns $\mathcal{C}_D$.

Now we enter the main training loop, where the generator and the discriminator will be trained adversarially. This loop will be run for many iterations. We start it training the generator to minimize the output of the discriminator according to Eq. 6 during a number of iterations. Then we increase the corpus $\mathcal{C}_D$ with the current state of the generator, and train the discriminator during another number of iterations. More recent input-response pairs are taken with a higher probability than the older ones from $\mathcal{C}_D$ when training the discriminator.

We finally repeat this process of training the generator, adding samples to $\mathcal{C}_D$ and training the discriminator, but this time training the corpus with the MLE criteria. This approach is also taken in [9, 14], and it aims at stabilizing the training process. In order to further stabilize it, we reduce the learning rate of the training optimizer throughout the global iterations.

This whole procedure is summarized in the Algorithm 1.

Algorithm 1 An adversarial training strategy for neural dialogue models

Require: Generator G, Discriminator D, Corpus $\mathcal{C}$, training hyper-parameters.
Pretrain word vector matrix **W** on $\mathcal{C}$.
Pretrain G minimizing L_{MLE} (Eq. 4).
Initialize $\mathcal{C}_D$ with G's responses y to some inputs x.
Pretrain D minimizing L_D (Eq. 5).
for the number of total iterations, and with a decaying learning rate **do**
    Update G minimizing L_G on inputs x in $\mathcal{C}$ (Eq. 6).
    Add (x, y) pairs to $\mathcal{C}_D$ using G.
    Update D minimizing L_D.
    Update G minimizing L_{MLE} on $\mathcal{C}$.
    Add (x, y) pairs to $\mathcal{C}_D$ using G.
    Update D minimizing L_D.

4 Experimental Setup

All the experiments in this work were carried out with the OpenSubtitles2018 corpus [15], which is composed of around 400M utterances from movie subtitles. As proposed in [24], since the turns are not clearly indicated, we treat each utterance as the desired output for the previous one.

As for the text preprocessing, we removed some symbols and converted all the names, numbers and places to tags *<person>*, *<number>* and *<place>*, respectively. This was done with the Spacy entity recognizer [8]. Finally we defined the vocabulary with most 30000 frequent words, and deleted every other token from the corpus. We pretrained 300 dimensional word vectors of those tokens on the corpus, with FastText [4]. These are then optimized again throughout the training process.

Let us now give details about the architecture of the sequence-to-sequence generator. The deep bidirectional RNN encoder is made of two LSTM networks (one per direction) of 4 layers, 512 cells each. On the other hand, the decoder's LSTM has 4 layers of 1028 cells. The MLP that converts $\mathbf{o}_t$ and $\mathbf{c}_t$ into $\tilde{\mathbf{o}}_t$ (see Sect. 2 for more details) has one *leaky*-ReLU layer. The size of $\tilde{\mathbf{o}}_t$ is 500. The MLP that computes the attention score has two layers. The first one is a 250-sized hyperbolic tangent layer, and the second is a linear output layer that computes the scalar score.

Regarding the discriminator, its two deep bidirectional encoders share the same architecture: two LSTM networks of two layers, 128 cells each. This vector is then fed to a MLP of two layers: a *leaky*-ReLU layer of size 100 followed by a single sigmoidal unit. The chosen value for the k parameter of the top-k softmax was 2.

The most promising hyper-parameters we have found for the training procedure are summarized next. First of all, we used the Adam optimizer [12] with batch size of 256 throughout all the optimization processes. We pretrained the generator during 50000 training iterations with a fixed learning rate of 0.001. We sampled 125000 responses from that generator and then pretrained the discriminator during 1000 iterations, with the same learning rate. All the batches fed to the discriminator were balanced: there was a human example per each generator's example. Human and generator's example were uncorrelated; they did not share the input.

The main iteration loop was run 200 times. The initial learning rate was 0.001 with a decaying factor of 0.995 when training the discriminator and the generator with the MLE criteria. It was ten times smaller when training the generator to minimize the output of the discriminator. Every MLE step was run during 50 iterations, and every step of minimizing the discriminator's output was run during 35 iterations. After each of these steps, 5000 input-response pairs were sampled from the generator, and the discriminator was trained during 40 iterations.

It is worth mentioning that we did not vary the architectural hyper-parameters much during our experiments. They are similar to many other sequence-to-sequence networks in the literature. On the other hand, selecting good and stable training hyper-parameters is challenging. This requires a deeper and more specific research that we leave for future work.

5 Experiments

We now present an experimental validation of the proposed differentiable sequence generative adversarial network for dialogue generation in two series of experiments. First we validate our differentiable GAN architecture measuring the quality of the word vectors obtained with after the top-k softmax computation path presented in Eq. 8. Additionally, we compare a neural dialogue trained with this computation path and with the adversarial learning procedure summarized in the Algorithm 1 with a standard MLE model.

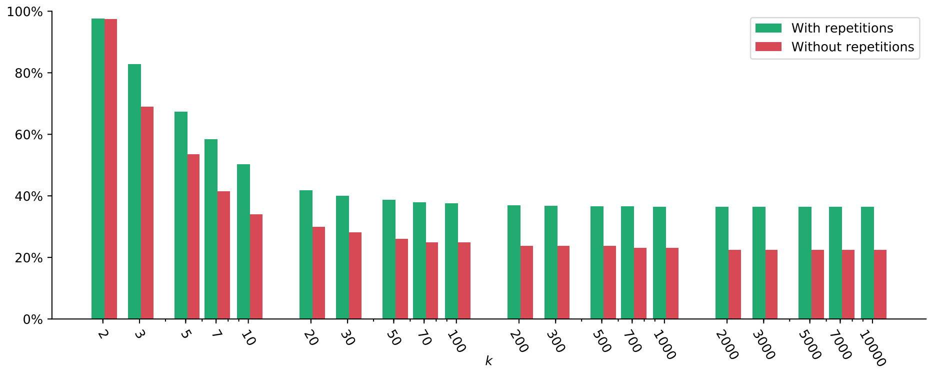


Fig. 3 Number of times that the actual word vector is the nearest neighbor of an approximated word vector produced by the top-k softmax, for different values of *k*

Table 1 Some responses to the same inputs in four different stages of the training procedure. *A* is the response of the pretrained MLE model. *B* is a response of an intermediate model just after a MLE step. *C* is the response of the same model than *B*, but just after a step of minimizing the output of the discriminator. *D* is the output of the final model

Input utterance	Response
Hi!	**A**: I'm <person> <person>.
	B: hi.
	C: hi, <person>.
	D: hello, <person>.
I love you.	**A**: I love you, too.
	B: I love you
	C: I love you too, <person>
	D: you know, I think maybe we could talk...
Who is she?	**A**: I don' know.
	B: she's my friend
	C: you'll never regret again.
	D: she's my girlfriend, <person>, and...
What's the color of the sky?	**A**: it's a blue light.
	B: the blue blue.
	C: it's the red of the moon.
	D: it's a beautiful blue sky
See you	**A**: <person> <person>
	B: I'm sorry, I'm sorry
	C: I'm sorry, I didn't mean to interrupt you
	D: see you later, <person>

Approximated word vectors We fed 1000 random inputs from the corpus to the dialogue system, and computed which was the closest word vector to each approximated one according to the euclidean distance, for different values of k. With $k = 2$, the closest word vector was the correct one the 98% of the times if we consider all the produced tokens, and the 97% if we do not consider repetitions. This two percentages decrease to 83%/69% respectively with $k = 3$, and to 74%/60% with $k = 4$. Figure 3 shows this statistic for more values of k. We therefore conclude that the proposed method to make the output of the discriminator differentiable with respect to the generator's parameters is appropriate, at least with $k = 2$.

Comparison between the mle baseline and the gAn Let us show a preliminary comparison between the pretrained MLE dialogue model with the final system after the adversarial learning. We asked 10 human evaluators to interact freely with the two systems during some few minutes, which resulted in dialogues of 25 turns on average. Then they were asked to decide which of them was better in terms of (1) the variety of the responses, (2) coherence and (3) informativeness. 7 out of the 10 evaluators opined that the final system was more variate and informative, and there was a draw in terms of coherence.

This can also be seen in Table 1. It shows responses to the same inputs in different stages of the training procedure. Not only are the baseline and final models compared in the table, but it also lets us gain an insight into the short-term effect of each of generator's minimizing the output of the discriminator. It tends to complex and enrich the model's responses, sometimes at the cost of losing some coherence.

6 Conclusion

We have presented a novel methodology to allow text generating models be trained in the GAN framework with gradient based optimization methods, the top-k softmax, and we have validated it in the open domain dialogue generation task. We have shown that good approximations of the word vector corresponding to each of the tokens generated by the dialogue system can be obtained with the top-k softmax. Moreover, we have demonstrated through a human evaluation process that a dialogue model trained in these conditions produces more variate and informative responses than the baseline MLE model, while being as coherent as it. Ultimately, the intersection between dialogue systems and GANs is a very promising area of research. We expect many more ideas from the two fields will be combined, and that many more applications of the GANs in the dialogue research will arise.

Acknowledgements This work has been partially funded by the Basque Government under grant PRE_2017_1_0357, by the University of the Basque Country UPV/EHU under grant PIF17/310, and by the H2020 RIA EMPATHIC (Grant N: 769872).

References

1. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473
2. Bengio S, Vinyals O, Jaitly N, Shazeer N (2015) Scheduled sampling for sequence prediction with recurrent neural networks. In: Advances in neural information processing systems, pp 1171–1179
3. Bengio Y, Léonard N, Courville A (2013) Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:13083432
4. Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching word vectors with subword information. arXiv preprint arXiv:160704606
5. Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R, Bengio S (2015) Generating sentences from a continuous space. arXiv preprint arXiv:151106349
6. Ghazvininejad M, Brockett C, Chang MW, Dolan B, Gao J, Yih W, Galley M (2017) A knowledge-grounded neural conversation model. arXiv preprint arXiv:170201932
7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
8. Honnibal M, Montani I (2017) Spacy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear
9. Hori T, Wang W, Koji Y, Hori C, Harsham B, Hershey JR (2019) Adversarial training and decoding strategies for end-to-end neural conversation models. Comput Speech Lang 54:122–139
10. Jang E, Gu S, Poole B (2016) Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:161101144
11. Kannan A, Vinyals O (2017) Adversarial evaluation of dialogue models. arXiv preprint arXiv:170108198
12. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980
13. Kusner MJ, Hernández-Lobato JM (2016) Gans for sequences of discrete elements with the gumbel-softmax distribution. arXiv preprint arXiv:161104051
14. Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D (2017) Adversarial learning for neural dialogue generation. arXiv preprint arXiv:170106547
15. Lison P, Tiedemann J (2016) Opensubtitles2016: extracting large parallel corpora from movie and tv subtitles. European language resources association
16. Lu J, Kannan A, Yang J, Parikh D, Batra D (2017) Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model. In: Advances in neural information processing systems, pp 314–324
17. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:150804025
18. Maddison CJ, Mnih A, Teh YW (2016) The concrete distribution: a continuous relaxation of discrete random variables. arXiv preprint arXiv:161100712
19. Serban IV, Sordoni A, Bengio Y, Courville AC, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. AAAI 16:3776–3784
20. Shetty R, Rohrbach M, Hendricks LA, Fritz M, Schiele B (2017) Speaking the same language: matching machine to human captions by adversarial training. In: Proceedings of the IEEE international conference on computer vision (ICCV)
21. Sordoni A, Galley M, Auli M, Brockett C, Ji Y, Mitchell M, Nie JY, Gao J, Dolan B (2015) A neural network approach to context-sensitive generation of conversation responses. arXiv preprint arXiv:150606714
22. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112
23. Tuan YL, Lee HY (2019) Improving conditional sequence generative adversarial networks by stepwise evaluation. IEEE/ACM Trans Audio, Speech, Lang Process

24. Vinyals O, Le Q (2015) A neural conversational model. arXiv preprint arXiv:150605869
25. Wu L, Xia Y, Zhao L, Tian F, Qin T, Lai J, Liu TY (2017) Adversarial neural machine translation. arXiv preprint arXiv:170406933
26. Xu J, Ren X, Lin J, Sun X (2018) Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 3940–3949
27. Yu L, Zhang W, Wang J, Yu Y (2017) Seqgan: sequence generative adversarial nets with policy gradient. In: AAAI, pp 2852–2858

Publication 14

The EMPATHIC Virtual Coach: a demo

The EMPATHIC Virtual Coach: a demo

Javier M. Olaso
Alain Vázquez
Leila Ben Letaifa
Mikel deVelasco
Universidad del País Vasco, Spain

Aymen Mtibaa
Mohamed Amine Hmani
Dijana Petrovska–Delacrétaz
Gérard Chollet
Institut Mines Télécom, France

César Montenegro
Asier López Zorrilla
Raquel Justo
Roberto Santana
Universidad del País Vasco, Spain

Jofre Tenorio-Laranga
Eduardo González-Fraile
Begoña Fernández-Ruanova
BIOEF; Osakidetza, Osatek, Spain

Gennaro Cordasco
Anna Esposito
Università degli Studi della Campania, Luigi Vinvitelli, Italy

Kristin Beck Gjellesvik
Anna Torp Johansen
Maria Stylianou Kornes
Oslo Universitetssykehus, Norway

Colin Pickard
Cornelius Glackin
Gary Cahalane
Intelligent Voice Ltd., UK

Pau Buch
Cristina Palmero
Sergio Escalera
Universitat de Barcelona, Spain

Olga Gordeeva
Olivier Deroo
Acapela Group; Belgium

Anaïs Fernández
Daria Kyslitska
e_Seniors Association, France

Jose Antonio Lozano
María Inés Torres
Universidad del País Vasco, Spain

Stephan Schlögl
Management C. Innsbruck, Austria

ABSTRACT

The main objective of the EMPATHIC project has been the design and development of a virtual coach to engage the healthy-senior user and to enhance well-being through awareness of personal status. The EMPATHIC approach addresses this objective through multimodal interactions supported by the GROW coaching model. The paper summarizes the main components of the EMPATHIC Virtual Coach (EMPATHIC-VC) and introduces a demonstration of the coaching sessions in selected scenarios.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Interactive systems and tools*;

KEYWORDS

interactive virtual coaching, empathic communication, elderly care

ACM Reference Format:

Javier M. Olaso, Alain Vázquez, Leila Ben Letaifa, Mikel deVelasco, Aymen Mtibaa, Mohamed Amine Hmani, Dijana Petrovska–Delacrétaz, Gérard Chollet, César Montenegro, Asier López Zorrilla, Raquel Justo, Roberto Santana, Jofre Tenorio-Laranga, Eduardo González-Fraile, Begoña Fernández-Ruanova, Gennaro Cordasco, Anna Esposito, Kristin Beck Gjellesvik, Anna Torp Johansen, Maria Stylianou Kornes, Colin Pickard, Cornelius Glackin, Gary Cahalane, Pau Buch, Cristina Palmero, Sergio Escalera, Olga Gordeeva, Olivier Deroo, Anaïs Fernández, Daria Kyslitska, Jose Antonio Lozano, María Inés Torres, and Stephan Schlögl. 2021. The EMPATHIC Virtual Coach: a demo. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3462244.3481574

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10.
https://doi.org/10.1145/3462244.3481574

1 INTRODUCTION

In this paper we introduce a demonstration of the virtual coach session conducted by the EMPATHIC-VC. The main objective of the EMPATHIC project has been the design of a virtual coach to engage the healthy-senior user and reach pre-set benefits to enhance well-being through awareness of personal status, by improving diet and nutritional habits and by developing more physical and social activities [9]. The EMPATHIC approach addresses this objective through conversational interactions. Figure 1 shows a picture of the interaction of an end-user and the VC. For this purpose:

- end-users have been involved from the beginning of the project to meet personalised needs and requirements, derived by the coach [1, 6];
- EMPATHIC-VC incorporates non-intrusive, privacy-preserving, empathic, and expressive interaction technologies [2, 3];
- the coach efficiency and effectiveness have been validated across 3 distinct European societies (Norway, Spain, and France), with 280 subjects who were involved from the start.

During the interactions the dialog manager (DM) develops coaching sessions guided by a GROW model [9] that is summarized in Section 2. The EMPATHIC-VC also includes multimodal human sensing modules aimed at getting the current emotional status of the end-user that will also have impact in the emotional behaviour of the VC resulting in an EMPATHIC communication model. Section 3 summarizes the VC architecture as well as the included modules.

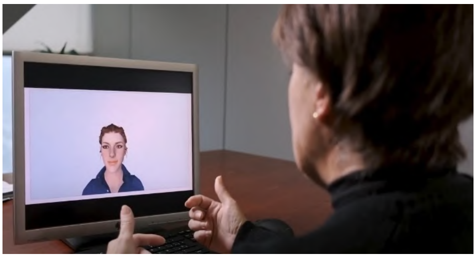


Figure 1: A virtual coaching session

The description of the Video demonstrator is shown in Section 4 and final conclusions in Section 5.

2 THE GROW MODEL OF COACHING

A core task of the project has been the incorporation of a coaching methodology to the EMPATHIC-VC through a well-being personalized coaching plan, which is driven by four basic requirements:

- Processes driven by questions;
- User defines and accepts the goals;
- Understand the User habits and routines related to the defined goal;
- EMPATHIC-VC is not an expert, but is able to provide expert's information;

Coaching is about raising awareness and responsibility. Through active listening and appropriate questions, this methodology aims to promote a positive change in the person coached. In that sense, the GROW model is a simple way of structuring an effective coaching conversation. The key to using GROW successfully is first to spend sufficient time exploring the goals until the coached sets one which is both inspirational and stretching, and then to move through the sequence, including revisiting the goal if needed. Figure 2 shows the main steps of the methodology [9].

Figure 2: Steps of the GROW model: identify the user Goal, assess their Reality by understanding the internal obstacles, analyse Options and alternatives and set up an action plan (Will). Visual appearance of one of the female virtual agent

3 SYSTEM OVERVIEW

Figure 3 shows the VC architecture that consists of:

- WebRTC used on the client side for:
  - Capturing voice and video from the participant's input devices;
  - Displaying of the virtual coach animation, speech and subtitles.

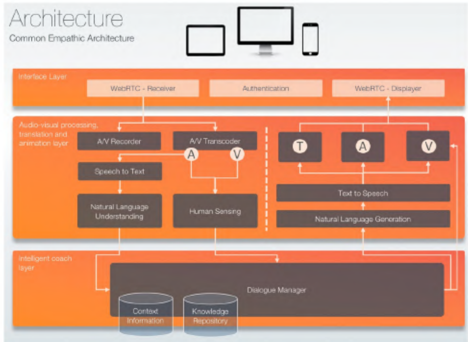


Figure 3: EMPATHIC specified architecture. "T", "A" and "V" stand for "Text", "Audio" and "Video"

- An audio/video recorder capturing all recordings in its native format and storing it so that it can be retrieved later if necessary.
- An audio/video transcoder transforming the various audio and video formats into a single format for further processing downstream.
- Biometric authentication based on face analysis [2].
- "Sensing" [3] the participant as follows:
  - User emotions are automatically detected from the audio signal. The output is provided in two emotional spaces: the categorical space (calm, happy, puzzled,..) and the dimensional one (valence, arousal and dominance) [4]
  - Emotion Recognition from participant's faces in the categorical space.
  - The multimodal fusion module is in charge of taking emotional status predictions from the other two independent emotional modules (face and speech), in order to fuse them into a singular emotional state score.
- Speech to text conversion with timestamps and confidence.
- Natural language understanding translating words into a sequence of semantic units. It provides the intents, the topic and the name entities from the analysis of the text transcription of the participant [5].
- The Dialogue manager controlling the flow of the conversation according to the strategy based on the GROW model and the scenario (nutrition, physical activity and Leisure) implemented. It provides an advanced management structure with distributed software agents reasoning about their state and the next action and enforces a clear separation between the domain-dependent and the domain-independent aspects of the dialogue control logic [8].
- The Intelligent Coach designed to use data from different sources: a) conversation data to provide complex information to the DM; b) previous coaching sessions to decide the coaching topic to be proposed; c) external information from the user's location and profile.

- Natural language generation converting conceptual representations, such as dialogue acts, into natural language. It is based on the GROW coaching model and has been developed specifically for the EMPATHIC project (called GROWsetta).
- Text to speech conversion. The EMPATHIC-VC includes male and female voices for each of the languages and cultures and implements neutral and positive moods.
- Five different virtual agents designed to increase the user acceptance [7, 9]. They have been enriched with animations and emotions to improve the interactivity. The agent runs in the user's browser with WebGL.
- The Media Server (see Figure 4) integrating three main components: a) the EMPATHIC media server application which is responsible for the setup of the pipelines and for communication with other modules; b) the control of the Kurento media server; c) the Kurento media server[1].

Figure 4 shows the Media Server implementation.

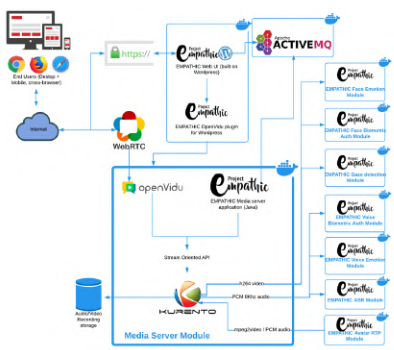


Figure 4: EMPATHIC Media Server

4 DESCRIPTION OF THE DEMO

The demonstration video shows an interaction with the EMPATHIC-VC in French (Spanish and Norwegian versions are also available). The video expounds a demo of a coaching session in a nutrition scenario as follows:

- The user chooses the agent with whom he/she wants to interact.
- The facial biometric module creates a template during the first encounter with the user, enabling further authentication.
- The participant uses a mouse click to start (and end) speaking. As an alternative the participant can also write the turn instead of speaking.
- Agent subtitles are provided to assist elderly with potential hearing impairments.
- The voice of the agent changes according to the identified emotional status of the user.
- The interaction starts with a discussion about the weather, to engage the user and show that the system is able to retrieve external information.
- Then the EMPATHIC-VC develops a coaching session on nutrition following the GROW model.
- The user ends and restarts the session to take a break. This shows that the user can resume the conversation at a different day or time. The agent can pick up where it left off, after verifying his/her identity with the facial biometric module.
- The EMPATHIC-VC demonstrates the ability to get some potential objectives and get an action plan from the user.

Note: The video recording has not been edited. It shows the development of the sessions as it happened.

5 CONCLUSIONS

This paper summarizes the main components of the EMPATHIC Virtual Coach, which are based on a multimodal interaction between elderly people and an agent who implements the GROW coaching approach. The demonstration video shows an interaction between a French user and the VC developing a coaching session in a nutrition scenario. A Spanish and a Norwegian video demo can also be provided. Additional technological products of the projects are a simulated virtual coach based on a Wizard-of-Oz platform [3] and an annotated corpus of dialogues in four languages (Spanish, French, Norwegian and English), which will be distributed for research purposes.

ACKNOWLEDGMENTS

This work has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 76987.

REFERENCES

[1] Anna Esposito, Terry Amorese, Nelson Mauro Maldonato, Alessandro Vinciarelli, Maria Ines Torres, Sergio Escalera, and Gennaro Cordasco. 2020. Seniors' ability to decode differently aged facial emotional expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 716–722. https://doi.org/10.1109/FG47880.2020.00077

[2] Mohamed Amine Hmani, Aymen Mtibaa, and Dijana Petrovska-Delacrétaz. 2021. Joining Forces of Voice and Facial Biometrics: a Case Study in the Scope of NIST SRE'19. In *Voice Biometrics: Technology, Trust and Security*, Carmen Garcia-Mateo and Gérard Chollet (Eds.). IET to appear, Chapter 9.

[3] Raquel Justo, Leila Ben Letaifa, Cristina Palmero, Eduardo Gonzalez-Fraile, Alain Torp Johansen, Annaand Vázquez, Gennaro Cordasco, Stephan Schlögl, Begoña Fernández-Ruanova, Micaela Silva, Sergio Escalera, Mikel deVelasco, Joffre Tenorio-Laranga, Anna Esposito, Maria Korsnes, and M. Inés Torres. 2020. Analysis of the Interaction between Elderly People and a Simulated Virtual Coach. *Journal of Ambient Intelligence and Humanized Computing* 11, 12 (2020), 6125–6140. https://doi.org/10.1007/s12652-020-01983-3

[4] Leila Ben Letaifa and M. Inés Torres. 2021. Perceptual Borderline for Balancing Multi-Class Spontaneous Emotional Data. *IEEE Access* 9 (2021), 55939–55954.

[5] César Montenegro, Asier López Zorrilla, Javier Mikel Olaso, Roberto Santana, Raquel Justo, Jose A. Lozano, and María Inés Torres. 2019. A Dialogue-Act Taxonomy for a Virtual Coach Designed to Improve the Life of Elderly. *Multimodal Technologies and Interaction* 3, 3 (2019), 1–19. https://doi.org/10.3390/mti3030052

[6] Jofre Tenorio-Laranga, Begoña Fernandez Ruanova, M. Inés Torres Barañano, Raquel Justo, Alfredo Alday, and Josu Xabier Llano Hernaiz. 2019. Designing a virtual coach: Involvement of end-users from early design to prototype. *International Journal of Integrated Care* 19 (2019), 1–8. Issue S1. https://doi.org/10.5334/ijic.s3207

[7] Markus Thaler, Stephan Schlögl, and Aleksander Groth. 2020. Agent vs. Avatar: Comparing Embodied Conversational Agents Concerning Characteristics of the Uncanny Valley. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209539

[1]Online: https://www.kurento.org/ [accessed: August 12th 2021]

[8] M. Inés Torres, Javier Mikel Olaso, Neil Glackin, Raquel Justo, and Gérard Chollet. 2019. A Spoken Dialogue System for the EMPATHIC Virtual Coach. In *9th International Workshop on Spoken Dialogue System Technology (Lecture Notes in Electrical Engineering, Vol. 579)*, Luis Fernando D'Haro, Rafael E. Banchs, and Haizhou Li (Eds.). Springer Singapore, Singapore, 259–265.

[9] M. I. Torres, J. M. Olaso, C. Montenegro, R. Santana, A. Vázquez, R. Justo, J. A. Lozano, S. Schlögl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. Palmero Cantariño, O. Deroo, O. Gordeeva, J. Tenorio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto. 2019. The EMPATHIC Project: Mid-Term Achievements. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '19)*. Association for Computing Machinery, New York, NY, USA, 629–638. https://doi.org/10.1145/3316782.3322764

Publication 15

Automatic Analysis of Emotions from the Voices/Speech in Spanish TV Debates

Automatic Analysis of Emotions from the Voices/Speech in Spanish TV Debates

Mikel deVelasco, Raquel Justo, Asier López Zorrilla, M. Inés Torres

Speech Interactive Research Group, Universidad del País Vasco UPV/EHU
Barrio Sarriena s/n, 48940 Leioa, Spain
{mikel.develasco, raquel.justo, asier.lopezz, manes.torres}@ehu.eus

Abstract: The goal of this work is to automatically analyze the emotional status of speakers, in human-human interactions, carried out in TV debates, where controversial topics are often presented. Human observers provide their perception about the emotional status associated to the interventions of the participants. An analysis of the resulting annotation was carried out by using different models for representing the emotions. The obtained labeled corpus was used to build an automatic system capable of detecting the emotional status associated to each acoustic signal, making use of the deep learning paradigm. The use of a corpus, where the real emotions that appear in a Spanish TV debate (with subtleties and often closer to neutrality than acted ones), are represented is crucial for learning models properly. In fact, although the level of accuracy depends on the problem complexity and the model employed for representing the emotional status, F1 scores of 0.7 were attained.

Keywords: emotion detection; human-human interaction; speech; behavioral analysis

1 Introduction

Affective computing has become a very interesting research area for scientific community, due to, inter alia, its potential capability to change the way in which human-machine, or even human-human, interaction is carried out. It is related to the idea of cognitive processes working together with ICT applications in order to take benefit of each other and go beyond their isolated capabilities [1] [2].

One of the goals of affective computing is to use the linguistic analysis of human-human interactions to detect the emotional status of human beings interacting together. In this work, we focus on the development of a system that can detect emotions from the speech extracted from a video recording. Speech can be defined as human vocal communication using language and it is inseparably intertwined with the emotional status during the cognitive process in human communication.

Furthermore, it seems to be a good indicator of depression [3], very related to the emotional status, or even Parkinson's disease [4].

Many works in the literature, that deal with emotional status identification from video, consider a reduced set of acted emotions [5] [6] [7]. Specifically, the basic set of emotions defined by Eckman [8] are usually employed when dealing with facial expression. In this way, a considerable amount of labeled data can be obtained in order to train machine learning algorithms with a limited effort. Moreover, corpora can be reused and the results obtained with different models can be easily compared to each other. However, the emotions that can be found in real scenarios are pretty different. In fact, the surface realizations of the underlying spontaneous emotions are different to those associated to acted emotions [9] [10], which complicates the direct application of the results of the investigations carried out with acted emotions as well as the use of acted data for training purposes. Furthermore, the set of emotions that appear in each specific real scenario is very task dependent and, thus, also the related automatic detection is. For example, the goal may just be to recognize anger through a simple anger/no anger classification in call centers [11] or to identify annoyance activation levels [12] [13] in customer assistance calls.

In this work the emotion detection is tackled in a specific scenario, where human-human interaction is carried out within the framework of a TV show. It is worth noting that TV shows are broadcasted for the general public and its semi-institutional character framed with specific roles, such as moderator and guest, affect the set of emotions that are expressed as well as their intensity.

On the other hand, research works on emotions have established that ordinary communication involves a variety of complex feeling states that cannot be characterized by a reduced set of categories, which does not cover the wide range of affect states. Therefore a number of researchers [14] [15] propose a dimensional representation [16] where each affect state is represented by a point in a two-dimensional space, namely valence and arousal, which some authors extend to three by also considering dominance.

An additional point to take into account when regarding spontaneous emotions is the labelling procedure, given that the current emotion of a speaker cannot be unequivocally established. In fact, the emotional label assigned by a speaker to his own utterance might differ to the one assigned by a listener to the same utterance, being the first one closer to the current emotion [17]. However the speaker self-annotation is not usually a realistic approach. As a consequence, the annotation of utterances in terms of spontaneous emotions is generally carried out through perception experiments, which are based on the particular judgement of every single annotator. Therefore, the disagreement among annotators as well as the distance between the emotion expressed and the emotion perceived can be significant. In contrast, if emotions are expressed by professional actors, or just elicited, then the annotation procedure is not required [18]. Thus, the generated emotion is always labelled by the intent of the actor. Finally, it is relevant to note, that the emotion perception and representation is very dependent on sociocultural aspects [19] [20] [21] [22]. Thus, another drawback in the labelling procedure might be the sociocultural differences among the annotator and the speaker that could lead to a low quality annotated corpus.

The previous framework shows spontaneous emotions generated and perceived to be very dependent of a variety of factors that make every data analysis and every automatic recognition task challenging and difficult for comparison. In this context, the main contributions of this work can be summarized as follows.

- An emotional analysis of the human behavior, from the perspective of external observers that listen to the acoustic signals, by making use of two different models for representing the emotional status.
- An emotionally labeled corpus where spontaneous emotions given in the scenario of interest, instead of acted ones, can be found. Let us note that machine learning algorithms need corpora where the intensity and the set of emotions match the involved task, in order to successfully learn the representation of these subtle emotions.
- An automatic system capable of successfully carrying out emotional status detection for the specific task we are dealing with, that was built using the deep learning paradigm along with the aforementioned corpus.

This work is organized as follows. Section 2 provides the description of the data of interest, the specific task and corpus, the different models employed to represent the emotional status and the annotation procedure. In Section 3, the analysis of the data is carried out with regard to the two models employed for emotional status representation and the relations among them. Section 4 provides a brief description of the employed feature sets and Section 5 summarizes the regression and classification experiments carried out with the corresponding neural network architectures. Section 6 discusses the Experimental Results and finally, Section 7 provides Conclusions for this work.

2 Describing the Data

In this section the data used in this work are presented: the task is described, the models employed for representing emotions are defined and the data annotation procedure is detailed.

2.1 Task and Corpus

In this work the data were extracted from *La Sexta Noche* Spanish TV program. In this weekly broadcasted show, hot news of the week are addressed by using social and political debate panels, led by two moderators. There is a very wide range of talk-show, guests (politicians, journalists, etc.) who analyze, from their perspective, social topics using Spanish language. Their interventions are mixed with edited videos and research reports. People in the set can give their opinion about the topics on the table and also people following the program at home using social networks. Given that the topics under discussion are usually controversial it is expected to have emotionally rich interactions. However, the participants are used to speak in public so they do not lose control of the situation and even if they might overreact sometimes, it is a real scenario, where emotions are subtle. This makes a great difference from scenarios with acted emotions as shown in [23]. Thus, it is very important to have a corpus consisting of real data related to the task we are dealing with in order to be able to train robust models which will represent emotional status.

In order to build the corpus, La sexta Noche programs broadcasted during the electoral campaign of the Spanish general elections in December 2015 were selected. This corpus was developed by a consortium of Spanish Universities under the umbrella of AMIC, “Affective multimedia analytics with inclusive and natural communication” project [24][1].

Acoustic signals were extracted from the TV shows videos and then segmented into clauses. A clause can be defined as “a sequence of words grouped together on semantic or functional basis” [25] and it can be considered that the emotional status does not change inside a clause. Therefore, in this work the clause is used as the working unit. An algorithm that considered silences and pauses, as well as the text transcriptions, was designed to identify the utterances compatible with clauses (Algorithm 1). It provided audio chunks from two to five seconds long, assuming that they match with the aforementioned clauses. Using this algorithm acoustic signals extracted from the TV programs were segmented into chunks. This procedure provided a set of 5500 audio chunks that were used as our data set. These chunks can correspond to any section of the program (including advertisements) in which people are speaking, either moderator of the show, guests, audience or all of them. However, most of the chunks correspond to the guests or/and the moderator. Later, within the labelling procedure (see questionnaire in Section 2.2), the annotators are asked to indicate whether the audio is correct, there is a high overlapping between the speakers, it corresponds to and advertisement or whether it has other issues. Thus, 1382 audios that did not correspond to “correct audios” were removed and only the remaining 4118 were used. Regarding the speaker features, the gender distribution in this set was 30% females and 70% male, with a total number of 238 different speakers and the age of them ranges from 35-65.

[1] ATRESMEDIA, producer and owner of the copyright of LaSextaNoche program’s contents, provided the consortium with the rights to use the audio files only for research purposes.

Algorithm 1: Segmentation algorithm

```
Function AudioSegmentation(audio, text_transcription):
    all_chunks ← ∅;
    for user_turn in text_transcription do
        audio_chunk ← get_audio_from(user_turn);
        chunks ← SplitByLowestEnergy(audio_chunk);
        all_chunks ← all_chunks + chunks;
    end
    return all_chunks ;
End Function

Function SplitByLowestEnergy(audio_chunk):
    chunks ← ∅;
    if audio_chunk > 5s then
        lowest_energy_point ← find_lowest_energy(audio_chunk);
        part1, part2 ← split_by(audio_chunk, alowest_energy_point);
        chunks ← chunks + SplitByLowestEnergy(part1);
        chunks ← chunks + SplitByLowestEnergy(part2);
    else
        chunks ← audio_chunk;
    end
    return chunks ;
End Function
```

2.2 Emotional Status from Acoustic Signals

The representation of the emotional status can be carried out using different models according to the Affective Computing literature. One popular approach involves the use of a categorical representation, in which emotions consist of discrete labels, such as boredom, frustration, anger, etc. [26] [27]. An alternative approach emphasizes the importance of the fundamental dimensions of valence and arousal in understanding emotional experience [28]. They are postulated as universal primitives in [28] and the feeling at any point on this two-dimensional space is called core affect. Other researchers have found “dominance”, a third dimension, important to represent emotional phenomena [29], particularly in social situations. For this work we used the set of categories of interest based on the selection provided in [30]. Then, it was adapted to the specific features of the task. For instance, *Sad* was not included since it is not expected to appear in political debates. With regard to the dimensional model the three dimensions were considered Valence, Arousal and Dominance (VAD).

The data set was annotated in terms of emotions to achieve a labeled corpus. The intrinsically subjectivity of the task makes it difficult to get a ground truth for the emotional status associated with an audio chunk using either categorical or dimensional model. One way to deal with this problem is to carry out expert annotations. However, according to some works, like the one presented in [31], the idea of a single correct truth is antiquated in determined contexts and needs to be disrupted. They propose to use crowd truth, that is based on the intuition that human interpretation is subjective, and that measuring annotations on the same objects of interpretation across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations. In this work crowd annotations, using a crowdsourcing platform [32] was carried out to get emotional labels for both, VAD and categorical models. The idea is to divide the work in micro-tasks that are carried out by a large number of annotators, that are not trained and do not speak to each other. This makes it possible to have an annotation task completed by a wide variety of different annotators, in cases where the diversity means a plus [33]. In this work, each audio chunk was annotated by 5 different annotators that were asked to fill the following questionnaire for each audio-clip.

How do you perceive the speaker?

- Excited
- Slightly excited
- Neutral

His/her mood is:

- Positive
- Slightly positive
- Neutral
- Slightly negative
- Negative

How do you perceive the speaker in relation to the situation which he/she is in?

- Rather dominant / controlling the situation
- Rather intimidated / defensive
- Neither dominant nor intimidated

Select the emotion that you think describes better the speaker's mood:

- Embarrassed
- Bored/Tired
- Disconc./Surp.
- Angry
- Interested
- Satisfied/Pleased
- Worried
- Enthusiastic
- Annoyed/Tense
- Calm/Indifferent

Quality of the audio:

- Correct
- Overlapping of several speakers, that do not identify the main speaker
- Advertisement
- Other

The chunks were given to the external observers randomly, thus the audios labelled by a specific annotator might not be from the same speaker nor even from the same TV show. However, it was guaranteed that all the annotators mother tongue was Spanish (like the speakers' one) and their cultural environment matched with the speakers' one as well (all coming from Spain). Table 1 shows the specific features of the 126 annotators set. Note that although most participants have only secondary studies a high percentage of them (about 80%) are University students.

Table 1
Different features of the Crowd Annotators Set: Sex, Education level (Undergraduate (U), Graduated (G)), Age and University Student (Yes/No)

Sex		**Education**		**Age**			**Student**	
M	F	U	G	20-30	30-40	> 40	Yes	No
65	61	103	23	80	27	19	82	44

3 Analyzing the Data

This section provides an analysis of the annotated data that will help to better understand what is being perceived in human interactions (by other human annotators) with regard to their emotional status.

3.1 Data Distribution in Categories and Dimensions

First of all, an analysis in terms of categories was carried out (Fourth question in the questionnaire). Let us note that 5 annotators provided possible different labels for each category, so that a unifying criterion was needed to associate a category to each audio chunk. In this work a majority voting criterion was employed, that is, an agreement >= 60% was required to assign a specific category to a sample. In this way it was guaranteed that at least 3 from the 5 annotators provided the same specific label to an audio chunk and otherwise the annotation was not valid. For instance, annotations in which 2/5 provided label1, 2/5 provided label2 and 1/5 provided label3 were discarded. According to this criterion the obtained distribution of samples is given in Figure 1.

As Figure 1 shows some categories were only selected in few occasions. This might be due to some categories being frequently mixed up with other ones, so they rarely reached the required threshold (see Angry, Bored/Tired, Disconcerted/Surprised or even Interested). We decided to keep only the classes with at least 2% of the samples not to have a so highly unbalanced dataset, so we finally considered the set of the following 5 classes: *Calm*, *Annoyed*, *Enthusiastic*, *Satisfied* and *Worried*. The sample distribution in categories can be explained focusing on the specific task. As mentioned before, most of the audio chunks are related to politicians, journalists, etc. talking about a controversial topic.

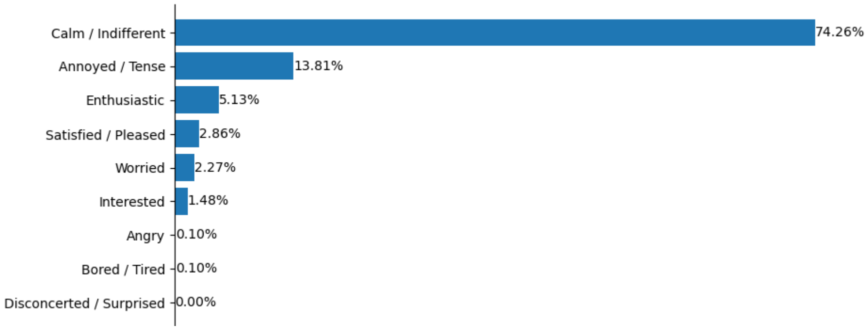


Figure 1
Distribution of samples into different categories

In these situations, speakers do not usually show themselves *Embarrassed* or *Disconcerted*. It is not either their role to show *Boredom/Tiredness* and regarding *Anger*, a subtler category like *Annoyed* seems to match better with annotators' perception.

Then, an analysis according to the dimensional representation of the emotional status was carried out. In this case, each sample was annotated with 3 different labels representing Arousal, Valence and Dominance (First 3 questions in the questionnaire). Let us note that for each dimension different levels representing a discrete scale were provided. Then, a numerical value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited:1, Slightly excited: 0.5, Neutral: 0. Then the average value considering the 5 annotations was computed to represent each annotated sample in a 3D space.

Figure 2, shows the probability density function of each variable (Valence, Arousal, Dominance) estimated by using a Gaussian kernel density estimator. The vertical line markers will be described in Section 6.2.

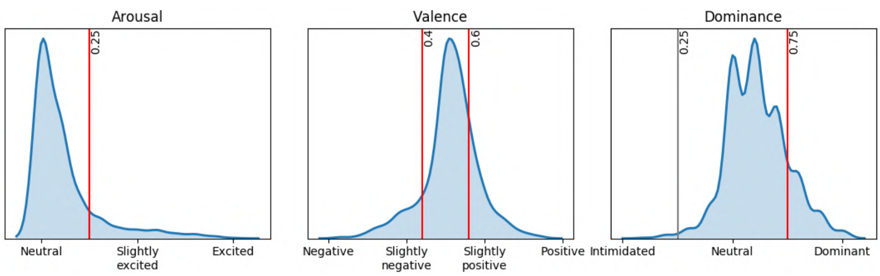


Figure 2
Probability Density Function of each VAD dimension

The results show that, in most cases, Arousal values tend to be among Neutral and Slightly Excited with more tendency to Neutrality. Most Valence values seem to be also quite Neutral although a slight nuance of positivity can be observed.

Dominance values instead, are clearly shifted towards Dominant, in fact most values are distributed among Neutral and Dominant while Intimidated almost never appeared. These results correlate well with the kind of audios we are dealing with, in which people express themselves without getting angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally, they appear to be neutral with regarding their opinions (valence tends to be neutral or slightly positive).

3.2 Relations among Categorical and Dimensional Models

In order to assign a label to an audio chunk using the categorical model the aforementioned agreement threshold (60%) was selected. However, this specific value matches with patterns of 3-2/2-3 annotations (3 annotations c_i category and 2 annotations c_j category or 2 annotations c_i and 3 annotations c_j). In these cases, according to the agreement criterion, the sample is given to the category associated with 3 votes, but this decision is questionable. Thus, a confusion matrix was built with these samples (Figure 3), showing that *Annoyed* and *Worried* were mixed up frequently and the same happens for *Enthusiastic* and *Satisfied*. Thus, it was decided to finally mix those categories, leading to a final set of three classes: *Calm, Annoyed/Worried, Enthusiastic/Satisfied*.

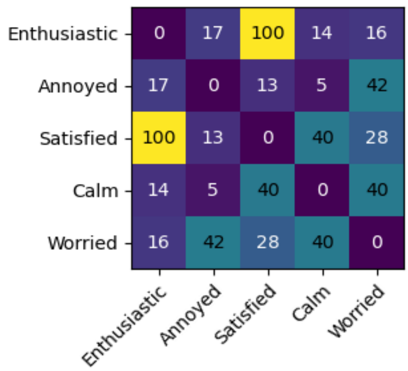


Figure 3
Confusion between annotations

Figure 4 shows different 2D projections of sample distribution in the 3D space representing each of the 3 resulting classes in a different color. Thus, the location of each category in the 3D space, according to the specific data and annotation procedure, can be explored.

It can be concluded, according to Figures 4, that when regarding Valence, samples labeled as Calm are almost perfectly centered at Neutral, although two peaks can be differentiated due to the discrete levels offered to the annotators in the questionnaire.

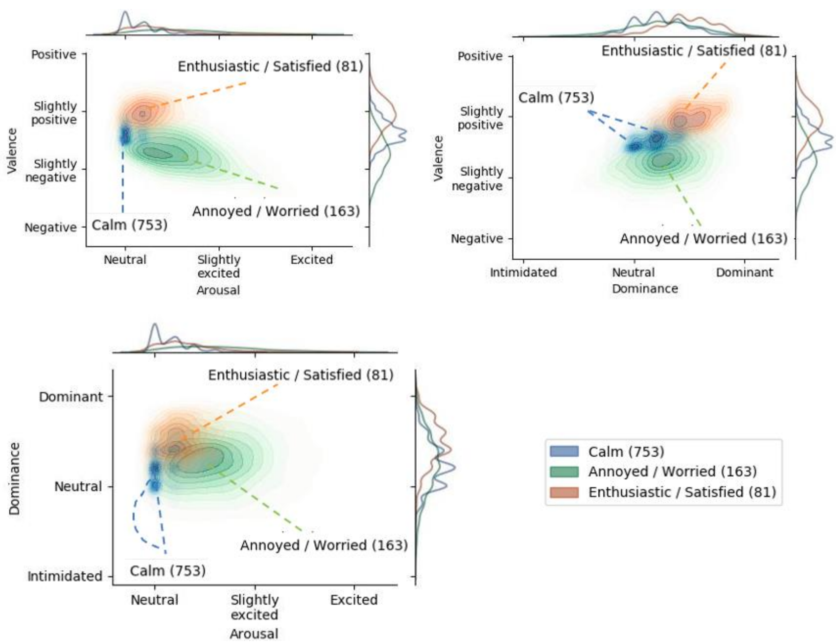


Figure 4
2D projections of 3D spaces

A Gaussian centered in Slightly Positive, is achieved for Enthusiastic/Satisfied and another one in Slightly Negative, for Annoyed/Worried. These results seem to be very coherent and validate the annotation procedure carried out in terms of both categorical and dimensional model. For Arousal, Calm is almost totally Neutral, although a second lower peak can be seen at the right. Enthusiastic/Satisfied, although more active it is also close to Neutral and the sparsest category with regard to Arousal is Annoyed/Worried, that seems to be closer to Slightly Excited than the other ones. This can be explained as mentioned before due to the specific data where speakers' role is to stay calm. Only when people are Annoyed the excitement seems to be a bit higher. Finally, with regard to Dominance, Calm is the most Neutral class but its higher peak is shift to Dominant values. Annoyed/Worried is also located between Neutral and Dominant and Enthusiastic/Satisfied is the most Dominant category. It is very interesting the tendency of samples towards dominant values that reveals the specific nature of the data, where speakers (politicians, journalists, etc.) try to be always dominating the situation. This tendency is very different from the results obtained for other tasks where Dominance is mainly Neutral [34].

4 Feature Extraction

There is no agreement in the state of the art about which features are the most relevant for emotion recognition from speech. Some authors rely on a small set of acoustic features [35] [36], whereas others have found that using the raw audio signal as input leads to good results [37]. Therefore, 3 sets of features were selected to be compared across all of the experiments.

4.1 Baseline Set

On the one hand, the first feature set we experimented with, was derived from a feature set that seemed to be useful in a previous work, where acoustic features were also employed as the input of a classification problem [36]. This baseline set is formed by 16 audio features: pitch, energy, entropy of energy and 13 MFCCs. The pitch was extracted with Praat [38] while the others were achieved by using pyAudioAnalysis [39]. To obtain all these features a step size of 10ms and a window size of 25 ms were used.

4.2 LLDs-GeMAPS

The GeMAPS feature set is a recommended minimalistic set of acoustic parameters described in [35] which was built for Voice Research and Affective Computing. These features were selected trying to fit with 3 different criteria:

1) The potential of an acoustic parameter to index physiological changes in voice production during affective processes.
2) The frequency and success with which the parameter has been used in the past literature.
3) Its theoretical significance.

This set is made up of 62 features that describe each full audio, regardless of its length. However, as this work makes use of convolutional neural networks, it has been decided to use the 18 Low Level Descriptors (LLDs) on which GeMAPS is based for all its final features.

These LLDs include information about prosodic, excitation, vocal tract, temporal and spectral descriptors. Briefly, they can be grouped as:

- Frequency related parameters (pitch, jitter, frequency of formats 1, 2 and 3, and bandwidth of the formant 1)
- Energy/Amplitude related parameters (shimmer, loudness and Harmonic-to-Noise Ratio)
- Spectral parameters (Alpha Ratio, Hammarberg Index, Spectral Slope 0-500 Hz and 500-1500 Hz, relative energy of formants 1, 2 and 3, Harmonic differences H1–H2 and H1-H3)

4.3 Spectrogram

In addition to the aforementioned sets of features, we also attempt to use more general and lossless acoustic features. To this end we implemented a much richer input: a mel-frequency spectrogram. Besides being richer, it does not require any feature engineering; it just represents the audio almost losslessly. The mel-frequency spectrogram was extracted using 128 FFT components, with a step of 2.66 ms and a window size of 42.66 ms. We first computed the squared Short-time Fourier transform of the audio wave, then filter it through a Mel filter bank, and finally take its logarithm (i.e. convert the power spectrogram to decibel units for an easier processing). Librosa [40] was used throughout this process.

5 Automatic Detection of Emotions

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition and also emotion detection from both speech and image [41] [42] [43]. In this section we describe the deep learning architecture designed in this work to solve emotion detection problems.

Tackling the problem of emotion recognition from audio requires dealing with variable length inputs, because each audio in the corpus has a different length. Thus, the neural network that is going to be used needs to take this into consideration. In the literature we can find different approaches to dealing with variable length inputs. One of the simplest and more used one is to compute the mean and standard deviation of each feature, and then use a classifier that takes as input that fixed length vector [23] [35]. Another approach could be to use a time step level classifier to try to classify the feature vector corresponding to each time step (and maybe some context), and then output the mean of all the low level classifications to get the final prediction. This approach is often used, for example, in image processing [44].

These two approaches though, share the same disadvantage: none of them is able to take into account the long term dependencies that may exist in the input. Therefore, we propose a network architecture which is divided into two different sections or subnetworks: an embedding network, and a classifier or regressor. Our approach is similar to [45]. The embedding network is responsible for getting an embedded and fixed-length representation of the input audio. The classification or regression network takes as input this embedding representation and classifies it in one of the defined classes or makes the desired regression.

5.1 Embedding Network

The embedding network's architecture is capable of working with different audio lengths and it always outputs a fixed output size length. Depending on the selected feature set, a slightly different network has been implemented. The embedding network for acoustic features of 4.1 and 4.2 is built with 2 small 1D convolutional layers (Figure 5). These two convolutional layers aim to extract some patterns on each feature as well as to reduce time dimension.

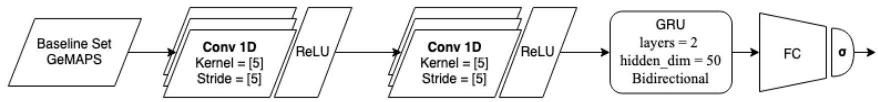


Figure 5
Embedding network for acoustic features

The spectrogram embedding network (Figure 6) is composed of 2 small 2D convolutional layers. These layers try to find some patterns and reduce dimensionality in both time and mel-spectrum dimensions at the same time.

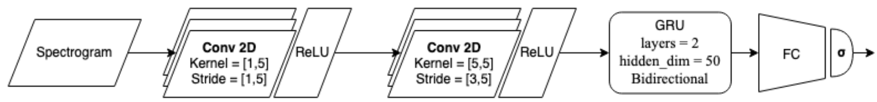


Figure 6
Embedding network for the spectrogram

But both architectures end up with a bidirectional-2 layer-GRU module to handle different input lengths and a fully connected layer to get the high-dimensional embedding vector.

As shown above, the main difference lies in the convolutional layers. The spectrogram is a two-dimensional feature matrix being time and mel-frequency its dimensions. Therefore, they can be processed with 2D convolutional layers. Baseline or LLDs-GeMAPS sets are groups of one-dimensional features. Thus, they will only be convolved across the time dimension.

5.2 Classification & Regression Networks

The classification and regression networks are two simple multilayer perceptrons, identical in terms of structure (Figure 7). Both are composed of two fully-connected layers, the first one is 15 with a ReLU activation function and a second layer is the output layer with the dimension equal to the number of outputs. The architecture is simple because we assume that the embedding should be already related enough to the output at this point of the network. The number of outputs is different for each classification problem (3 for categories, 2 for arousal, 3 for valence and 2 for dominance, as we will discuss in the following Section) and is set to 1 in regression problems. In classification problems a softmax function is set as activation function and a sigmoid function in regression problems.

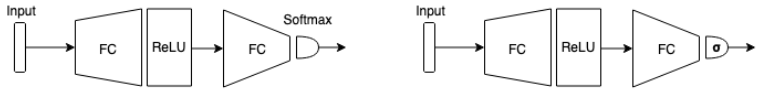


Figure 7
Classification (left) and regression (right) networks

6 Experimental Results

We analyzed the performance of the proposed regression and classification systems for each feature set. All experiments were performed with a 10-fold cross validation system in order to achieve a stronger statistical result. In all the experiments, we trained the network with the Adam optimizer with an early stopping strategy. In Section 5.1, we present the experiments related to the categorical model, and in Section 5.2, those related to the VAD model.

6.1 Categorical Model

As discussed in Section 3, the acquired corpus was first filtered so the agreement was at least 60% in all the samples. Then, some of the classes were dismissed due to the very few samples corresponding to them, and others were grouped because they were mixed up frequently by the annotators. Finally, we ended up with these three different classes: *Calm* (753 samples), *Annoyed/Worried* (163 samples) and *Enthusiastic/Satisfied* (81 samples).

Since the classes are very unbalanced, we observed that oversampling the samples of the minority classes in the training set led to a better performance. The oversampling ratio was 4 for *Annoyed/Worried and* 9 for *Enthusiastic/Satisfied*.

Thus, Table 1 shows the macro-average F1 score achieved with each feature set and the network presented in Section 5.2 for classification. In order to check whether the best model is significantly better than the others, we also computed a Wilconxon signed-ranks test [46] for each pair of classifiers over the cross validation results. It tests the null hypothesis that two related paired samples (the results of the cross validation) come from the same distribution. In particular, it tests whether the distribution of the differences of each cross validation iteration is symmetric about zero. In this table (and in all the tables throughout the work) values in bold indicate that the p-values for the two comparisons of the best model with the rest are lower than 0.10 (if an asterisk is used *) or 0.05 (if two asterisk are used **). No bold values are shown in a row if there is a comparison with a p-value greater than 0.10.

According to these results, the spectrogram is the most suitable input for the model. It works better than the Basic Set of features, and much better than LLDs-GeMAPS. The reason for this could be that the spectrogram represents the audio much less loosely than the other feature sets.

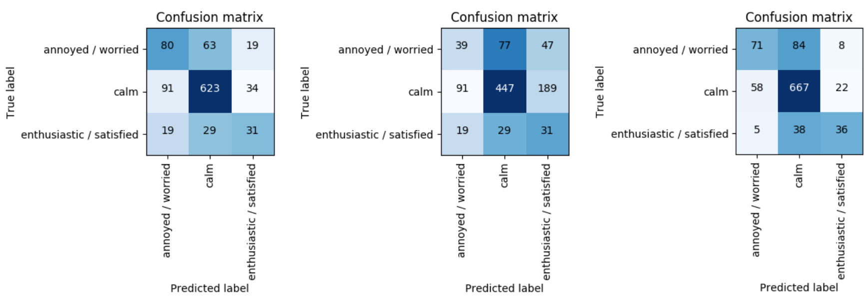
Table 2
Average F1 score after the 10-fold cross validation

	Baseline Set	LLDs-GeMAPS	Spectrogram
F1 score	0.56	0.38	**0.61***
Precision	0.55	0.4	0.64
Recall	0.57	0.42	0.59

It does not assume which features of the audio are the most relevant to tackle the emotion recognition problem, and it lets the network learn them implicitly. It also contains more temporal information, because the step size used to extract it is four times smaller compared to the other feature sets. Other works, such as [47] or [48] have also shown the potential performance benefits of using less losslessly or even raw inputs as opposed to common sets of acoustic features. Additionally, the neural networks used to process the three types inputs might well be another reason why the spectrogram is working better than the other features. We are using CNNs, which are very suitable for processing image-like inputs, like the spectrogram. Thus, we believe that the other sets of inputs (Baseline Sets or LLDs-GeMAPS) could benefit from the use of other kind of classifiers. However, in order to enable the use of classical classification methods, fixed length feature sets should be used. To do so, some functions should be applied to get rid of the temporal dimension of the input, such as the ones used in GeMAPS.

Focusing on Table 2, 0.61 is the best F1 score achieved. It has been achieved with the spectrogram and it outperforms both Baseline (0.56) and LLDs-GeMAPS (0.38) features set by 0.05 and 0.23 respectively. Figure 8 shows that the models perform quite well in the majority class (*Calm*) but some confusion was encountered when predicting other classes (*Annoyed/Worried* and *Enthusiastic/Satisfied*).

Figure 8
From left to right, test confusion matrix for Baseline, LLDs-GeMAPS and Spectrogram sets. The color scale is normalized by true label.



In any case, these results are quite promising taking into account that we are dealing with very difficult audios of spontaneous speech and a very ambiguous task like emotion detection that is not obvious even for a human. Note that similar results about 50-60% accuracy are reported in other works published in Emotion Recognition in the Wild Challenge [49] when regarding emotion detection from audio.

6.2 VAD Model

The VAD prediction problem was tackled in two different approaches:

- Building a regressor with each of three real dimensions of the model
- Discretizing those dimensions and trying to learn a categorical classifier to predict each of the discretized classes.

The discrete levels for the classification problem were selected according to the distributions of the annotated data of Figure 2 with the selected frontiers (red lines). For Arousal, given that there are few samples labeled as Excited or Slightly Excited, only two different levels were differentiated: Neutral (Samples with Arousal values < 0.25 for training purposes) and Excited (Samples with Arousal values ≥ 0.25). In the case of Valence, although many samples are labeled as Neutral, three different regions can be differentiated: Negative (Valence values ≤ 0.4), Neutral: ($0.4 <$ Valence values < 0.6) and Positive (Valence values ≥ 0.6). Finally, for Dominance two different values were selected: Neutral ($0.25 \leq$ Dominance value < 0.75) and Dominant (Dominance value ≥ 0.75).

Finally, the resulting number of samples in the categories is given below:

- Valence: Negative (473 samples), Neutral (2439 samples) and Positive (1191 samples).
- Arousal: Neutral (3057 samples) and Excited (1046 samples).
- Dominance: Dominant (1075 samples) and Neutral (3004 samples).

Due to the obtained unbalanced values in the categories, we also decided to apply an oversampling procedure during the training processes of the regressor and classifier. For Valence we chose an oversampling ratio of 3 for Negative and 2 for Positive. The chosen balance ratio for Arousal was 3 for Excited. Finally, when regarding Dominance, an oversampling ratio of 3 was also selected for the Dominant category.

6.2.1 Regression

During the first series of experiments, we tried to fit each dimension of the VAD model. To this end, the selected optimization objective was the batch-level coefficient of determination (R^2). We also experimented with other loss functions such as the MSE error, but got worse results. Table 2 and Figure 10 show the performance of the regressors after the 10-fold cross validation procedure.

Table 3
Mean and standard deviation of R^2 score after the 10-fold cross validation for the VAD model

R^2 Score (Test)	Baseline Set	LLDs-GeMAPS	Spectrogram
Arousal	0.20 ± 0.06	0.1 ± 0.1	**0.3 ± 0.1 ****
Valence	0.08 ± 0.04	0.02 ± 0.02	**0.11 ± 0.06 ****
Dominance	0.02 ± 0.04	0.03 ± 0.03	0.06 ± 0.04

Table 2 shows that low R^2 score values were obtained and the spectrogram features still have better performance across all tasks. Both basic and LLDs-GeMAPS sets of features seem to have similar performance in these regression problems.

Trying to have a deeper understanding of what actually is happening in these models, we displayed the resulting density plot for each set of features and dimension in the test partitions of the cross validation. Figure 9 reveals that models end with a narrowed range of outputs. The models tend to output very similar values, which result in high peaks, while the real distribution is more uniform.

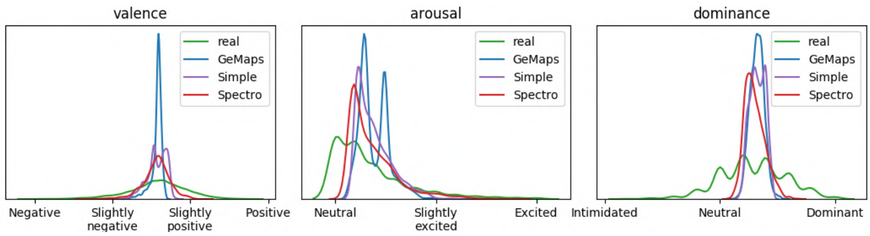


Figure 9
Density plots of the output of each model

In the same vein, Figure 10 shows that even the best models cannot perform well in the regression task; the real value vs. predicted value plots are still far from being diagonal straight lines.

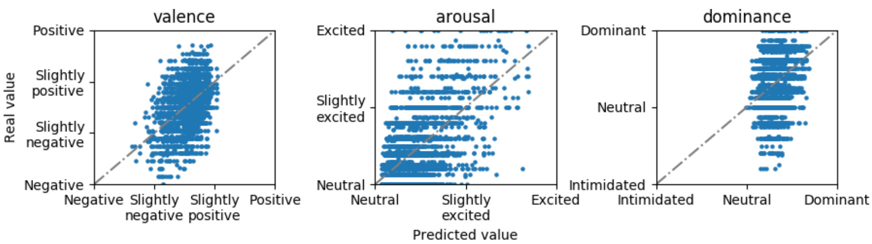


Figure 10
Test samples for the best regression model (Spectrogram)

These results have not been as good as expected. The R^2 score for the three tasks are not higher than 0.3 in the best case. It means that there is a low correlation between the real and predicted values. The spectro seems to be the best input for all the dimensions with an R^2 score of 0.3 in Arousal, 0.11 in Valence and 0.06 in Dominance. Figures 9 and 10 allow us to understand why the scores are that low; all models tend to output values in a narrow range. The Arousal model (the one with higher R^2 score) is the only one that predicts over almost the full range of values, but without a high accuracy.

6.2.2 Classification with the Discretized Classes

In this section we focus on a less ambitious scenario for the VAD emotional model. The aim of this second series of experiments was to classify the audios into the dimensional discrete classes. In the same way as with the other experiments, the three sets of features were compared and the average F1 score reported (Table 4).

Table 4
Mean and standard deviation of F1 score after the 10-fold cross validation for the VAD model

F1 Score (Test)	Baseline Set	LLDs-GeMAPS	Spectrogram
Arousal	0.66 ± 0.03	0.63 ± 0.02	0.70 ± 0.04
Valence	0.44 ± 0.02	0.41 ± 0.02	**0.52 ± 0.03***
Dominance	0.56 ± 0.02	0.58 ± 0.02	**0.60 ± 0.02***

Likewise, Figure 11 shows the confusion matrices obtained in the three classification tasks with the best model (Spectrogram).

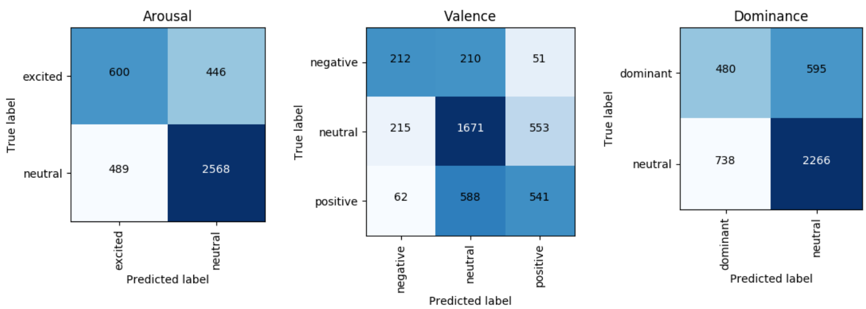


Figure 11
Test confusion matrix for Spectrogram set on each dimension.
The color scale is normalized by the true label.

The results in this case are much more promising, Baseline and LLDs-GeMAPS set of features achieves similar results, but spectro still have a better performance in these 3 tasks. Some results outperform the F1 value achieved for categorical models but we consider that similar performance are achieved compared to the categorical results, taking into account that Arousal and Dominance have 2 output values instead of 3. That is why Valence has lower scores than others.

These results demonstrate the benefits of the VAD model, because it is not needed a previous study or reflection about the specific emotions of the task and without any refinement about the category set (analyzing the annotation results or the confusion matrices), the obtained results are very similar (sometimes even better) to those achieved with the categorical model. Moreover, the VAD model is much more general and a specific label/category can be associated to different points or regions of the 3D VAD model [50] [51].

Concluding Remarks

This work provides a deep analysis of the emotional information gathered from the acoustic signal associated to a debate TV show. This emotional information is related to the perception of people who listen to the acoustic signals, thus, it can be seen as human perception of human-human interactions. The analysis was carried out by using different models for representing the emotional status which were also analyzed and compared to each other. In the different analysis it can be concluded that emotions in this specific real scenario (TV shows) are subtle, with a strong tendency to neutrality. However, the specific features of this task show a significant and non-obvious bias to dominance of the speakers.

The aforementioned information led to the achievement of an emotionally labeled corpus, made up of real and non-acted emotions, that was employed to build an automatic system capable of detecting the emotional status of an acoustic signal in the presented scenario. Different experiments were carried out using the different models for representing emotions and also different deep learning paradigms. The obtained results show that having a corpus of real interactions, that matches with the task under consideration, where emotions are not acted is crucial for getting good results in such environment where emotions are subtle (real life). Moreover, although the more ambitious regression paradigm provides poor results, when the problem is discretized and transformed into a classification one, very promising results can be achieved. This suggests that a higher number of external observers answering to the same VAD questionnaire, but with a higher number of responses for each dimension (a scale closer to a continuous scenario) might improve the regression results in this task.

Acknowledgements

This work has been partially supported by the Spanish Government under grant TIN2017-85854-C4-3-R (MCIU/AEI/FEDER, UE), the Basque Government under grant and from the European Union’s Horizon 2020 research and innovation program under grant agreement no 769872.

References

[1] Baranyi P, Csapó A, Sallai G. Cognitive Infocommunications (CogInfoCom). Springer International; 2015

[2] Baranyi P. Special Issue on Cognitive Infocommunications Preface. Acta Polytechnica Hungarica. 2018;15(5):7-10

[3] Gábor K, Vicsi K. Comparison of read and spontaneous speech in case of Automatic Detection of Depression; 2017

[4] Sztahó D, Tulics MG, Vicsi K, Valálik I. Automatic estimation of severity of Parkinson's disease based on speech rhythm related features. - 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2017

[5] Kim JC, Clements MA. Multimodal Affect Classification at Various Temporal Lengths. IEEE Transactions on Affective Computing. 2015;6(4):371-84

[6] Eskimez SE, Imade K, Yang N, Sturge-Apple M, Duan Z, Heinzelman W:. Emotion classification: How does an automated system compare to Naive human coders? - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016

[7] 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(9):1062-87

[8] Ekman P. Basic Emotions. Handbook of Cognition and Emotion. 1999:45-60

[9] Chakraborty R, Pandharipande M, Kopparapu SK. Analyzing Emotion in Spontaneous Speech; 2017

[10] Schuller B, Weninger F, Zhang Y, Ringeval F, Batliner A, Steidl S, et al. Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge. Comput Speech Lang. 2019;53:156-80

[11] Pappas D, Androutsopoulos I, Papageorgiou H. Anger detection in call center dialogues. - 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2015

[12] Irastorza J, Torres MI. Analyzing the expression of annoyance during phone calls to complaint services. - 2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); 2016

[13] Irastorza J, Inés Torres M. Tracking the Expression of Annoyance in Call Centers. In: Klempous R, Nikodem J, Baranyi PZ, editors. Cognitive Infocommunications, Theory and Applications. Cham: Springer International Publishing; 2019, pp. 131-51

[14] Gunes H, Pantic M. Automatic, Dimensional and Continuous Emotion Recognition. International Journal of Synthetic Emotions (IJSE). 2010;1(1):68-99

[15] 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(9):1062-87

[16] Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161-78

[17] Chakraboty R, Pandharipande M, Kopparapu SK. Analyzing Emotions in Spontaneous Speech. Springer Nature; 2017

[18] Bänziger T, Mortillaro M FAU - Scherer, Klaus, R., Scherer KR. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion (Washington, D.C.) JID - 101125678

[19] Mesquita B, Leu J. The cultural psychology of emotion. In: New York, NY, US: The Guilford Press; 2007, pp. 734-59

[20] Averill JR. Emotion and anxiety: Sociocultural, biological, and psychological determinants. In: Oxford, England: Lawrence Erlbaum; 1976, p. x, 362

[21] Vea T. The learning of emotion in/as sociocultural practice: The case of animal rights activism. null. 2020;29(3):311-46

[22] Riviello MT, Esposito A, Vicsi K. A Cross-Cultural Study on the Perception of Emotions: How Hungarian Subjects Evaluate American and Italian Emotional Expressions. Cognitive Behavioural; Systems; Berlin, Heidelberg: Springer Berlin Heidelberg; 2012

[23] de Velasco M, Justo R, López-Zorrila A, Torres MI. Can Spontaneous Emotions be Detected from Speech on TV Political Debates? 10th IEEE International Conference on Cognitive Infocommunications. 2019:289-94

[24] Ortega A, Lleida E, San-Segundo R, Ferreiros J, Hurtado L, Sanchís E, et al. AMIC: Affective multimedia analytics with inclusive and natural communication. Procesamiento del Lenguaje Natural. 2018;61:147-50

[25] Esposito A, Marinaro M, Palombo G. Children speech pauses as markers of different discourse structures and utterance information content. International. Conference: From Sound to Sense; June 10-13, 2004; MIT Cambridge USA

[26] Calvo RA, D'Mello S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing. 2010;1(1):18-37

[27] Calvo RA, Kim SM. EMOTIONS IN TEXT: DIMENSIONAL AND CATEGORICAL MODELS. Comput Intell. 2013;29(3):527-43

[28] Russell J. Core Affect and the Psychological Construction of Emotion. Psychological review. 2003;110(1):145–17

[29] Bradley MM, Lang PJ. Measuring emotion: The self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry. 1994;25(1):49-59

[30] Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc Natl Acad Sci USA. 2017:201702247

[31] Aroyo L, Welty C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AIMag. 2015;36(1):15-24

[32] Justo R, Alcaide JM, Torres MI. Crowdzientzia: Crowdsourcing for research and development; 2016

[33] Justo R, Torres MI, Alcaide JM. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. Pattern Recognition and Image; Analysis; Cham: Springer International Publishing; 2017

[34] Justo R, Ben Letaifa L, Palmero C, Gonzalez-Fraile E, Torp Johansen A, Vázquez A, et al. Analysis of the interaction between elderly people and a simulated virtual coach. Journal of Ambient Intelligence and Humanized Computing. 2020;11(12):6125-40

[35] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. 2016;7(2):190-202

[36] López-Zorrilla A, de Velasco M, Cenceschi S. Corrective focus detection in italian speech using neural networks. Acta Polytechnica Hungarica 2018;15(5):109-27

[37] Tzirakis P, Chen J, Zafeiriou S, Schuller B. End-to-end multimodal affect recognition in real-world environments. Information Fusion. 2021;68:46-53

[38] Boersma P, Weenink D. Praat: doing phonetics by computer [Computer program] 2018;6.0.37

[39] Giannakopoulos T. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLOS ONE. 2015;10(12):e0144610

[40] McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: Audio and music signal analysis in python. 2015

[41] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44

[42] Amer MR, Siddiquie B, Richey C, Divakaran A. Emotion detection in speech using deep networks. - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014

[43] Bertero D, Fung P. A first look into a Convolutional Neural Network for speech emotion detection. - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017

[44] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and

[45] Yoon S, Byun S, Jung K. Multimodal Speech Emotion Recognition Using Audio and Text; 2018

[46] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-3

[47] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016

[48] Avci U. Speech Emotion Recognition Using Spectrogram Patterns as Features. Speech and; Computer; Cham: Springer International Publishing; 2020

[49] Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, Shaun Canavan. Recognizing Emotion in the Wild using Multimodal Data. 2020; Association for Computing Machinery (ACM); 2020

[50] Russell JA. Pancultural aspects of the human conceptual organization of emotions. J Pers Soc Psychol. 1983;45(6):1281-8

[51] Scherer KR. What are emotions? And how can they be measured? Social Science Information. 2005;44(4):695-729

Publication 16

Automatic Identification of Emotional Information in Spanish TV Debates and Human–Machine Interactions

applied sciences

MDPI

Article

Automatic Identification of Emotional Information in Spanish TV Debates and Human–Machine Interactions

Mikel de Velasco , Raquel Justo * and María Inés Torres

Iniversidad del País Vasco UPV/EHU, Department of Electrical and Electronics, Faculty of Science and Technology, 48940 Leioa, Spain; mikel.develasco@ehu.eus (M.d.V.); manes.torres@ehu.eus (M.I.T.)
* Correspondence: raquel.justo@ehu.eus

Abstract: Automatic emotion detection is a very attractive field of research that can help build more natural human–machine interaction systems. However, several issues arise when real scenarios are considered, such as the tendency toward neutrality, which makes it difficult to obtain balanced datasets, or the lack of standards for the annotation of emotional categories. Moreover, the intrinsic subjectivity of emotional information increases the difficulty of obtaining valuable data to train machine learning-based algorithms. In this work, two different real scenarios were tackled: human–human interactions in TV debates and human–machine interactions with a virtual agent. For comparison purposes, an analysis of the emotional information was conducted in both. Thus, a profiling of the speakers associated with each task was carried out. Furthermore, different classification experiments show that deep learning approaches can be useful for detecting speakers' emotional information, mainly for arousal, valence, and dominance levels, reaching a 0.7 *F*1-score.

Keywords: speech processing; emotion detection; machine learning; behavioral analysis; human–machine and human–human interaction

Citation: de Velasco, M.; Justo, R.; Inés Torres, M. Automatic Identification of Emotional Information in Spanish TV Debates and Human–Machine Interactions. *Appl. Sci.* **2022**, *12*, 1902. https://doi.org/10.3390/app12041902

Academic Editors: Francesc Alías

Received: 30 December 2021
Accepted: 4 February 2022
Published: 11 February 2022

1. Introduction

Emotion expression and perception is a very important issue in human interactions and is one of the bases upon which the communication between humans is established. Therefore, the automatic detection of emotions by a computer has become a very attractive topic due to its impact on the effort towards more natural and empathic human–machine interaction systems. Emotions can be expressed in different ways, including facial expression, speech, gestures, etc. In this work, we focus on speech and its ability to provide diverse information.

In addition to the message communicated, speech signals can provide information related to different aspects of the speaker. In fact, speech signals can give insights into the emotional state of the speaker or even their baseline mood, as shown in many studies about this issue [1,2]. The probability of suffering from a disease, such as depression, Alzheimer's disease [3–5], or even COVID-19 [6], can also be extracted from speech. However, speech may also be influenced by several other variables, such as the speaker's habits, personality, culture, or specific objective [7,8].

Human–human interactions take place in specific contexts where, to some extent, people know each other. However, current artificial agents have little capacity to imitate a real user, resulting in shallow interactions [9]. In fact, users find it hard to interact with agents with rudimentary visual and speech capacities [10]. The literature suggests that human behavior in human–human interactions is guided by the other human's behavior and is, thus, reactionary behavior [11]. However, comparisons between these two scenarios have almost only been carried out at the interaction and dialogue levels [9,11]. The emotional exchange in both scenarios is completely different due the rudimentary emotional capacity of the agent, which results in very subtle emotions. This work aims to contrast the similarities and differences for emotions identified in two very different scenarios: human–human interactions on Spanish TV debates (TV Debates) and human–machine interactions with a virtual agent developed by the H2020 EMPATHIC project (http://www.empathic-project.eu, accessed on 3 February 2022) (Empathic VA), also in Spanish. Thus, we profile the task in each scenario or, more specifically, the speakers involved in each task. Although they are quite different, they both share the spontaneity of speech, as well as the spontaneity of the expression of emotions in real scenarios [12].

Disfluencies or spontaneous speech events, such as filled pauses or speech repairs, enrich spontaneous communication [13] with paralinguistic information that depends on the context, on the speaker profile, and on emotional state. In recent years, research on spontaneously expressed emotions in everyday tasks has gained interest in the scientific community [14,15]. However, this research has typically been conducted on emotions simulated by professional actors in artificially designed databases such as EMODB [16] or IEMOCAP [17]. The six basic emotions defined by Eckman [18] (anger, surprise, disgust, enjoyment, fear, and sadness) can be represented by facial expressions that typically characterize these emotions and thus can be used in the automatic identification of emotions on a face [19]. However, spontaneous emotions are more varied and complex. Furthermore, emotions expressed during acting or during a real-life scenario show significant differences [20]. In fact, only a small set of complex and compound emotions [21] can be found in real scenarios [2,15,22], and this subset is strongly dependent on the situation. Therefore, a set of categories including the emotions that arise in each specific task has to be defined. To this end, some perception experiments have to be conducted to specify the set of emotions of interest. However, this process is expensive; time consuming; and sometimes, not viable. Alternatively, and assuming that ordinary communication involves a variety of complex feelings that cannot be characterized by a reduced set of categories, a number of researchers [23,24] proposed a dimensional representation [25] of the emotional space. Thus, each affective state is represented by a point in a two-dimensional space, namely valence and arousal, which space some authors extend to three dimensions by also considering dominance (also known as the VAD model). This work employs both approaches to analyze emotional information.

Additionally, spontaneous emotions cannot be unambiguously perceived, not even by experts. In fact, the emotional label assigned by a speaker to their own utterances might differ from those assigned by a listener, with the former being, of course, more accurate [26]. In this work, we draw from some works dealing with the annotation of a virtual agent [22,27] that provide insights into the problems associated with this kind of annotation. The intrinsic subjectivity of this task makes obtaining a ground truth for emotional states associated with an audio signal using either the categorical or the dimensional model difficult. According to some work, such as the one presented in [28], this subjectivity cannot be properly gathered when experts label emotions; therefore, a more useful representation based on the interpretation of emotions across a crowd should be used. In this work, crowd annotations, using a crowdsourcing platform [29], was carried out to obtain emotional labels for both the VAD and categorical models. This methodology led to two corpora for each task: (a) TV debates labeled in terms of discrete categories, (b) TV debates labeled in terms of the VAD model, (c) empathic VAs labeled in terms of discrete categories, and (d) empathic VAs labeled in terms of the VAD model.

In the context of interactions, annotations are usually carried at the turn or dialogue levels [9]. However, the debate on the minimum temporal length of the audio for which the emotions can be extracted reliably remain open. This length has usually been set in tuning experiments for a particular situation [26]. In contrast, in this work, we propose utterances compatible with clauses as segments to be annotated and develop an algorithm to obtain them from speech signal.

Once a labeled corpus is designed, a machine learning-based system can be built to carry out automatic emotion detection. One of the first steps in creating such a system is to identify which acoustic features are the most suitable for detecting emotions. In recent years, promoted by challenges such as the INTERSPEECH Computational Paralinguistic

Challenge [30], several attempts have been made to obtain such a set, such as the minimalist set of GeMAPS speech features proposed in [31]. However, several studies [32,33] suggested that no universal acoustic features that extract emotional content and work well in all contexts exist. Low-level descriptors (LLD) [33,34] based on characteristics related to prosody (pitch, formants, energy, jitter, and shimmer) or to the spectrum (centroid, flux, and entropy), and its functionals (mean, std, quartiles 1–3, delta, etc.) have been widely used. Alternatively, some authors avoided LLD features and let a neural network extract the emotional features in the first layers using other speech representations, such as a spectrogram [35–37] or a raw audio signal [38]. Moreover, the rise in the self-supervised learning paradigm and the recently proposed transformer architecture [39], have led to novel speech representations, such as wav2vec [40,41] or HuBERT [42]. These representations were extracted from raw audio and can be used to feed a neural network. In this work, we primarily design and build a deep neural network architecture fed with a spectrogram. Furthermore, we also provide some preliminary experiments for which the network is fed with the wav2vec model to obtain preliminary insights into such an approach to working with the tasks tackled in this work.

Within this framework in which the perception, modeling, and detection of emotions constitute a challenge, the main contributions of this work can be summarized as follows:

- An in depth analysis of the emotions arising in two different scenarios as a way of profiling the speakers associated with a task using both the categorical and the VAD model to represent the emotional state.
- Two Spanish corpora are emotionally labeled by the crowd, where spontaneous emotions can be found instead of acted ones.
- An emotion-detection system based on deep learning is specifically designed to the tasks considered. In this framework, this paper discusses the issues derived from the detection of realistic emotions in Spanish tasks as an attempt to progress research on emotion detection.
- The preliminary experiments aimed to evaluate the convenience of the recent wav2vec representation of speech for the automatic detection of spontaneous emotions in Spanish Language.

This paper is structured as follows: Section 2 describes the tasks and the associated corpora tackled in this work (Section 2.1) and provides insights into the annotation procedure (Section 2.2) as well as insights into the design of the automatic detection system including the neural network architecture (Section 2.3). In Section 3, the results obtained in terms of both an analysis of emotions (Section 3.1) and the classification performance (Section 3.2) are given. Finally, Section 4 provides a discussion of the results.

2. Materials and Methods

2.1. Task and Corpus

This section describes the two tasks tackled in this work.

2.1.1. TV Debates

First, a set of real human–human conversations was gathered from TV debates. Specifically, the Spanish TV program "La Sexta Noche" was selected. In this weekly broadcast show, news about hot topics from the week are addressed by social and political debate panels led by two moderators. A very wide range of talk-show guests (politicians, journalists, etc.) analyze social topics from their perspectives. Given that the topics under discussion are usually controversial, emotionally rich interactions can be expected. However, the participants are used to speaking in public so they do not lose control of the situation. Thus, even if they might overreact sometimes, this is a real scenario, where emotions are subtle. The spontaneity in this situation is vastly different from scenarios with acted emotions, as shown in [15]. The selected programs were broadcast during the electoral campaign of the Spanish general elections in December 2015. Table 1 shows a small excerpt of a dialogue taken from the TV Debates corpus.

Table 1. Small excerpt extracted from the TV Debates corpus. This is an emotionally rich example of a discussion between two talk-show guests debating politics. The same excerpt is shown in Spanish (the original language) above and in English below.

	Spanish
Speaker 1:	Yo entiendo que de España y de datos y de hechos no quieras hablar, pero resulta…
Speaker 2:	Claro que puedo hablar…
Speaker 1:	Que acaban de imputar también al quinto tesorero en la historia de tu partido.
Speaker 2:	Y dale.
Speaker 1:	De cinco…
	English
Speaker 1:	I understand that you do not want to talk about Spain, about neither data nor facts, but it turns out…
Speaker 2:	Of course I can talk…
Speaker 1:	That they have just imputed the fifth treasurer in the history of your party as well.
Speaker 2:	And hit it.
Speaker 1:	Five out of five…

To start building the corpus, the whole audio signal was separated into shorter segments or chunks useful for crowd annotation. The segments have to be short enough to avoid variations in emotional information but long enough to allow for their identification. Thus, the audio signal was divided into clauses. A clause was defined as "a sequence of words grouped together on semantic or functional basis" [43], and it can be hypothesized that the emotional state does not change inside a clause. An algorithm that considered silences and pauses as well as the text transcriptions was designed to identify the utterances compatible with the clauses [2]. This procedure provided a set of 4118 audio chunks from two- to five-seconds long that comprises our working corpus. Regarding the speaker's features, the gender distribution in this set was 30% females and 70% male, with a total of 238 different speakers within the age range from 35 to 65.

This corpus was developed by a consortium of Spanish Universities under the umbrella of AMIC, "Affective Multimedia Analytics with Inclusive and Natural Communication" project [44].

2.1.2. Empathic VA

The Empathic VA is a European Research and Innovation project [45,46] that implements personalized virtual coaching interactions to promote healthy and independent aging. As a part of the project, a series of spontaneous conversations between people who are older and a Wizard of OZ (WoZ) were recorded in three languages: Spanish, French, and Norwegian. The WoZ technique allows users to believe that they are communicating with a machine (and not a human), which makes their reaction more similar to one they have when interacting with a fully automatic virtual agent [8,45]. The conversations are related to four main topics: leisure, nutrition, physical activity, and social and family relationships [8,45]. Table 2 presents an example of a conversation about nutrition between a user and the WoZ, where some emotional content can be found. In this work, we focused on Spanish dialogues that were recorded by 31 speakers with an average age of 71.6 years. The gender distribution was 45% female and 55% male. After applying the algorithm to extract chunks that matched clauses, a corpus of 4525 audios from two to five seconds long was achieved.

Table 2. Small excerpt taken from the Empathic VA corpus. This sample is a segment of a conversation between a user and the WoZ. It shows that the WoZ generally asks questions and tries to obtain extensive answers from the user. The excerpt is shown in Spanish (original language) above and then in English (translation) below.

	Spanish
User:	Para cenar suelo tomar o huevos con. . . o pescado con algo de también de verdura, y luego suelo tomar o tomo algunas frutas, suelo tener fruta. Y luego pues no tomo nada más.
System:	Sueles merendar?
User:	A veces sí, depende. . . A veces sí meriendo y suelo merendar pues un. . . un café y algún dulce o algún bocadillito de jamón también suelo tomar. Depende de la ocasión, si estoy en casa o no.
System:	Entiendo, ¿Crees que comes bien?
User:	Creo que sí.
System:	Dime Ana, ¿Hay algo que te gustaría cambiar en tu forma de comer?
User:	Pues nada porque como a gusto, no. . . no quiero otra cosa.
	English
User:	For dinner, I usually have either eggs with . . . or fish with some vegetables too, and then I usually take or I take some fruits, I usually have fruit. And then, well, I do not take anything else.
System:	Do you usually have a snack?
User:	Sometimes, yes, it depends. . . Sometimes, yes, I take a snack and I use to have a snack; let us say a. . . a coffee and I also used to take some sweets or some small ham sandwich. It depends on the occasion, if I am at home or not.
System:	I understand. Do you think you eat well?
User:	I think so.
System:	Tell me Ana, Is there anything you would like to change in the way you eat?
User:	Well, nothing because I eat at ease; no, . . . I do not want anything else.

2.2. Annotation Procedure

The TV Debates and Spanish Empathic VA datasets were labeled by emotion to achieve two useful and very valuable corpora to model emotions in Spanish.

Emotions are traditionally represented by two models: a categorical representation, in which emotions consist of discrete labels, such as happiness, anger, etc. [47,48], or an alternative approach that emphasizes the importance of the fundamental dimensions of valence and arousal in understanding emotional experience [49]. They are postulated as universal primitives in [49], and a feeling at any point on this two-dimensional space is called a core affect. A representation of the core affect is shown in Figure 1, where an emotion such as *sad* is represented with a very low value of arousal and a neutral valence slightly shifted to the negative side. Other researchers have found a third dimension, dominance, to be important in representing emotional phenomena [50], particularly in social situations. In this work, we use both representations, the categorical one and the dimensional one. A set of categories of interest based on the selection provided in [51] was first considered. Then, the set was adapted to the specific features of each of the tasks presented above. For instance, *sad* was not included in the TV Debates set, since sad emotions were not expected to appear in political debates. For the Empathic VA, a study was conducted to identify the emotions that were perceived by the users. In addition, the dimensional VAD model was also considered for both datasets.

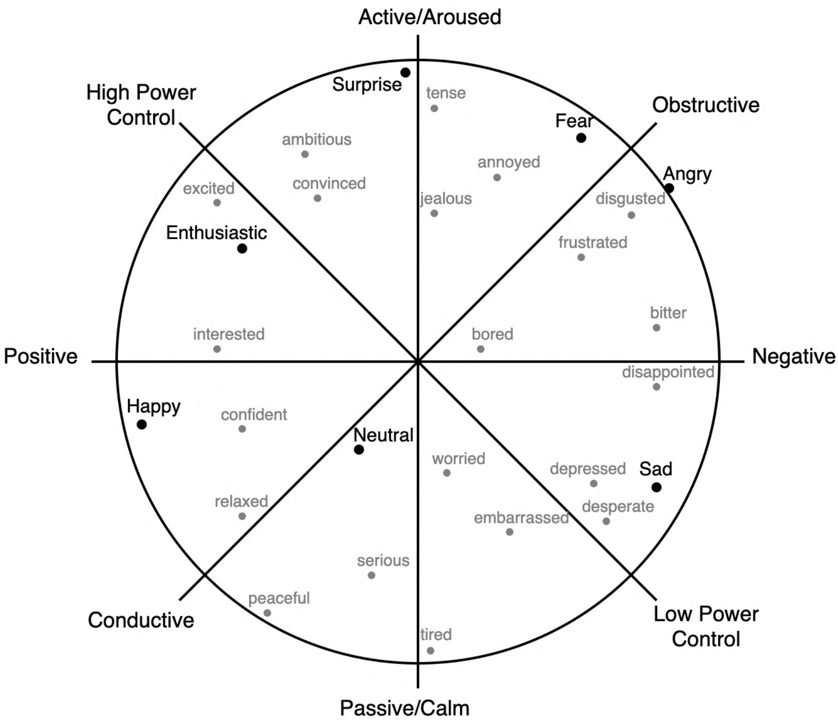


Figure 1. Illustration of Scherer's circumplex [52], which shows categories represented by the dimensions arousal and valence.

The intrinsic subjectivity of the tasks makes obtaining a ground truth for the emotional status associated with an audio chunk difficult. One way to deal with this problem is to use the crowd truth [28], which is based on the intuition that human interpretation is subjective. Thus, measuring annotations on the same objects of interpretation across a crowd provides a useful representation of their subjectivity and the range of reasonable interpretations. In this work, crowd annotations was carried out through a crowdsourcing platform [29] to obtain emotional labels for both the VAD and categorical models. To this end, the annotation work was divided in micro-tasks that were performed by a large number of untrained annotators who did not speak to each other. This division in tasks made the annotations diverse, which is a plus for our dataset [53] and was made possible by the wide variety of different annotators. In this work, each audio chunk was annotated by five different annotators who were asked to fill in the following questionnaire for each audio clip.

Some categories have two or three names as a result of the preliminary task adaptation carried out over the categories selected from [51].

- How do you perceive the speaker?
  - Excited
  - Slightly Excited
  - Neutral
- His/her mood is
  - Rather Positive
  - Neither Positive nor Negative
  - Rather Negative
- How do you perceive the speaker in relation to the situation which he/she is in?
  - Rather dominant/controlling the situation
  - Rather intimidated/defensive
  - Neither dominant nor intimidated
- Select the emotion that you think best describes the speaker's mood:

  (TV Debates)
  - Calm/Indifferent
  - Annoyed/Tense
  - Puzzled
  - Angry
  - Interested
  - Satisfied/Pleased
  - Worried
  - Enthusiastic
  - Embarrassed
  - Bored/Tired

  (Empathic VA)
  - Calm/Bored/Tired
  - Sad
  - Happy/Amused
  - Puzzled
  - Annoyed/Tense

Annotators Agreement

Given that each audio chunk was labeled by five different annotators, an analysis of the agreement among the annotators was carried out. Table 3 gathers the statistics of agreement per audio chunk for the categorical model. This table shows that, for about 70% of the data in the TV Debates dataset, the agreement was 2/5 or lower. For the Empathic VA dataset, a higher agreement was achieved due to the lower number of categories that were selected. However, almost 50% of the samples still showed an agreement of lower than 4/5. This confirms the ambiguity and subjectivity of the task. Moreover, Krippendorff's $alpha$ coefficient [54] was also low for both tasks, resulting in values of 0.11 and 0.2, respectively. This coefficient reflects the degree of agreement but is very dependent on the number of labels.

Table 3. Statistics of the agreement per audio chunk for each corpus. Column Agr. Level indicates the condition, i.e., the minimum inter-annotator agreement, and the next two columns (No. Audios and % audios), indicate how many samples and what percentage of them fulfilled the agreement condition.

	TV Debates		Empathic VA	
Agr Level	**No. Audios**	**% Audios**	**No. Audios**	**% Audios**
⩾1/5	4118	100.00%	4525	100.00%
⩾2/5	3035	73.70%	4522	99.93%
⩾3/5	1266	30.74%	4023	88.91%
⩾4/5	392	9.52%	2519	55.67%
⩾5/5	82	1.99%	1086	24.00%

In the rest of the document, we do not consider samples with an agreement below 3/5 for the categorical model, which means that we used 30.74% of the annotated audio files of the TV Debates dataset and 88.81% of the Empathic VA for the experiments with emotional categories. Then, the majority voting method was used to establish the ground truth for these sets.

The answers to the questionnaires related to the VAD model were transformed into real values, ranging from 0 to 1, by applying the rules of Table 4 to each response. Then, these values were averaged per sample over all five annotators to obtain a real value in the 3D space. In this case, carrying out majority voting and thus obtaining a minimum agreement level were not required. The average was computed due to the vast diversity derived from the subjectivity of this task, which was reflected in the different answers provided by the diverse labels generated by the annotators. The size of the resulting labeled corpus (100% of the audio clips shown in Table 3) was bigger than the corpus labeled in terms of the categorical model.

Table 4. Transformation of the answers to the VAD questions into continuous values in the range [0, 1]. Later, the means of the transformed values of the five annotators were computed to obtain continuous values for the dimensional model.

Arousal	Valence	Dominance	Value
neutral	rather negative	rather intimidated/defensive	0.0
slightly excited	neither positive nor negative	neither intimidated/defensive	0.5
excited	rather positive	rather dominant/controlling the situation	1.0

2.3. Classification Framework

The automatic detection of emotions was carried out within the machine learning paradigm using the aforementioned corpora for training and test purposes. To this end, the usual pipeline includes a first procedural stage to extract features from a speech signal that feeds a classifier in a second stage. The feature set can make a difference in the resulting performance. However, no standard audio feature set seems to work well for all emotion recognition corpora [32,33,55]. The audio Mel-frequency spectrogram was considered in this work given that it demonstrated an efficient method to encode the information extracted from audio clips, as shown in [56,57]. Thus, each audio chunk in the aforementioned corpora was transformed into its corresponding spectrogram using the librosa toolkit [58]. This decision led us to the pipeline described in Figure 2.

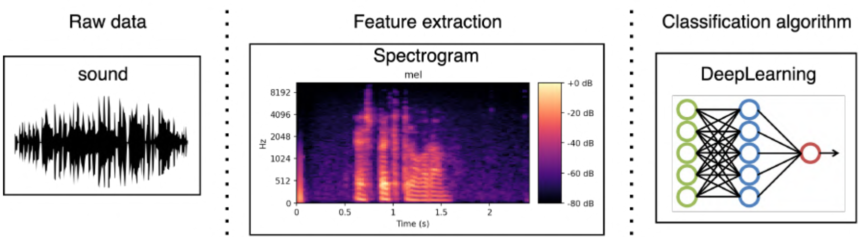


Figure 2. Pipeline of a basic procedure in audio classification problems. First, raw data are identified, i.e., the wav audio itself. Then, the characteristics are extracted with a tool, the spectrogram, through librosa [58], in this case. Finally, the classification problem is carried out, in this case, using neural networks. Color in Deep Learning diagram means input, intermediate and output layers.

Furthermore, one of the challenges to be addressed in both datasets was the difference in the audio sample lengths. Recurrent neural networks (RNNs), such as LSTMs, are specifically well suited to dealing with this problem [59,60] in the framework of neural networks (NN). However, convolutional architectures can outperform recurrent networks on tasks related to sequence modeling, as shown in [61]. Moreover, the training of convolutional neural networks (CNNs) is a simpler process that neither requires as many computational resources as RNNs nor suffers from a vanishing gradient [62]. Nevertheless, a common approach allowing for the use of CNNs is to pad all samples in such a way that all of them have the same input length [63,64], which also allows the network to learn which parts are relevant for the task.

The network architecture proposed in this work (Figure 3) takes the padded mel-spectrogram input and reduces both the mel-frequency and the time dimensions using 2D convolutions and max-poolings. This sub-network reduces the time dimension but creates a richer audio representation. Then, the network takes the new representation and tries to classify each time step using a multi-layer perceptron of three layers. After classifying all of the time steps, the network averages it to provide a single output for the input audio. The same architecture is employed for classification in terms of the categorical and VAD models. In the latter, the annotation values were discretized as described in Section 3.2.

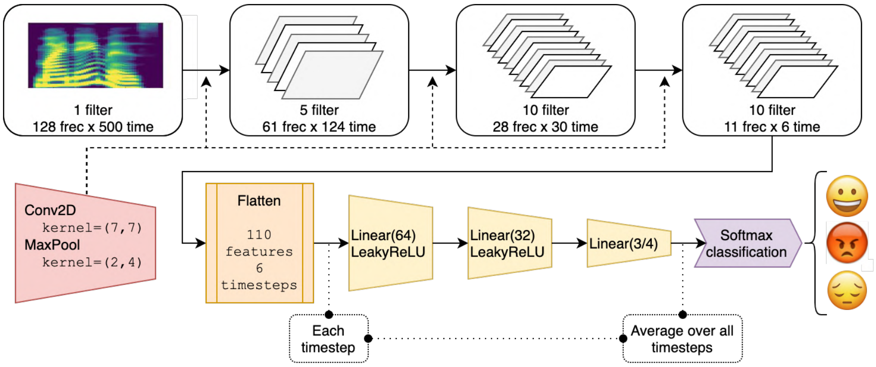


Figure 3. Neural network architecture used for classification when the spectrogram represents the speech signal. First, a succession of convolutional and maxpooling layers reduce the dimensions, obtaining a small time dimension with 110 features each (10 filters times 11 frequencies). Then, some logits are obtained for each of the features of the time dimension over three linear layers. Finally, the mean of all of the logits is computed to classify the sample.

In the training process, several decisions were made. On the one hand, all samples were padded to obtain the same time dimension, as mentioned above. Thus, the training process is easier when all of the batches have the same input length. On the other hand, an ADAM optimizer with stochastic weight averaging (SWA) [65] procedure was used as the optimization method. SWA can be used with a learning rate schedule that encourages exploration of the flat region of solutions. To this end, a cyclical learning rate schedule was used (see Figure 4). First, 60,000 batch updates were performed with a constant learning rate of 10^{-4}. Second, a decaying schedule with a learning rate of 10^{-5} over 1000 batch updates was applied. Finally, cyclical learning rates were adopted over five cycles, with a decaying learning rate schedule from 10^{-3} to 10^{-5}. The models for averaging were collected at the end of each cycle, corresponding to the lowest values of the learning rate.

The imbalance in classes of the training corpora can negatively influence the performance of the machine learning algorithms [22]. In some cases, this imbalance can even lead to completely ignoring the minority class, which is often the class with which we are more interested. An approach to dealing with this challenge is the use of oversampling/undersampling methodologies to duplicate/delete samples from the minority/majority class, respectively. In this work, a repetition oversampling method was chosen, where all of the non-majority class samples were duplicated. This procedure helped the network alleviate the problem of exclusively predicting the majority class. Finally, the experiments were carried out over a 10-fold cross-validation procedure.

Figure 4. Learning rate schedule for SWA updates. Each SWA update is performed when the learning rate is at the minimum ($u1$, $u2$, . . . , $u6$).

In addition to the architecture mentioned above, a preliminary work that deals with a novel methodology based on pretrained networks was also considered. The wav2vec 2.0 [41] speech representation was used, which is a pretrained end2end network for speech feature extraction (https://github.com/pytorch/fairseq/blob/main/examples/wav2vec/README.md (accessed on 3 February 2022)). Specifically, xlsr_53 was considered, a multilingual model that was trained on the MLS, CommonVoice, and BABEL databases. MLS [66] is a multilingual dataset derived from audiobooks. The Common Voice corpus [67] is a massive multilingual collection of transcribed speech built using crowdsourcing for both data collection and data validation. Crowd contributors record their voice by reading sentences displayed on the screen. The goal of the BABEL project [68] is to produce a multi-language database of speech for five of the most widely differing Eastern European languages. We note that, in these datasets, the amount of Spanish speech is not significant. In fact, BABEL does not include it at all. Moreover, some parts do not include European Spanish but, rather, American Spanish, which makes a great difference. Furthermore, the datasets include non-spontaneous speech, and as a consequence, emotional content is not expected. The wav2vec 2.0 representation has been recently proposed for speech emotion recognition in English, for which specific pretrained networks can be found [69].

The pipeline used for these preliminary experiments is similar to the previous one. Only the feature extraction module differs and is now implemented by the pretrained network that transforms the speech signal into sequences of vectors. This pipeline is shown in Figure 5.

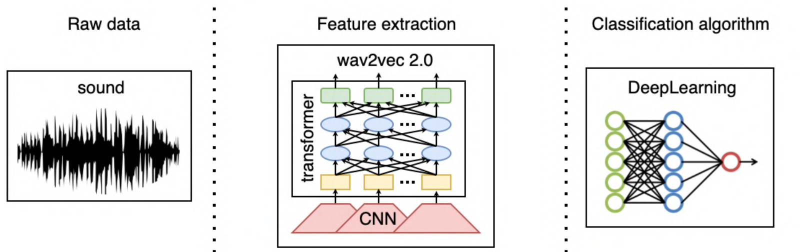


Figure 5. Pipeline for emotion detection from audio signals using wav2vec 2.0 [41]. First, raw data must be identified, i.e., the wav audio itself. Then, the characteristics are extracted with the pretrained wav2vec model, and finally, the classification problem is carried out, in this case, using neural networks. Colors in Deep Neural Networks means different type layers.

In the wav2vec architecture, the output of the last layer of the pretrained wav2vec 2.0 model was chosen to feed the network as audio representations. This representation has a dimension of 1024 features plus the time dimension (250 time samples for 5 s). The network architecture implemented for the wav2vec 2.0 input reduces the time dimension over several 1D convolutions and max-poolings and then takes the new representation and tries to classify each time step using a multi-layer perceptron of three layers. Finally, once all time steps are classified, the network averages the logits in the same way as the network when using the spectrogram, as the input in Figure 3 shows.

The training process used in the wav2vec network architecture was the same as the one used with the spectrogram network architecture, explained above.

3. Analysis of Emotions and Classification Results

For this section, we conducted an analysis of the emotions perceived by the annotators in the different tasks, and then, different series of classification experiments were carried out.

3.1. Analysis of Emotions

First, the categorical model annotation was analyzed. Table 5 shows the list of categories for each task along with the percentage of samples in each category in descending order. A minimum agreement of 0.6 (3/5) was requested to consider a sample to be valid, as mentioned above. Moreover, a minimum number of samples (1% of the total) was required in each class. These requirements led to a reduction in the valid samples, resulting in a set of 1266 samples for the TV Debates dataset and 4023 for the Empathic VA dataset when considering the categorical model. This table shows that different categories are predominant in each of the corpora. Some of them could be considered equivalent, such as calm/indifferent and calm/bored/tired, which are the most frequent categories in both sets. However, annoyed/tense, for instance, is the second most frequent class in the TV Debates dataset but was almost last in the Empathic VA dataset. In the same way, puzzled is almost absent (included in others) in the list for the TV Debates dataset.

Table 5 also shows that both datasets are imbalanced, with the calm category being the majority class, with more than 70% of the samples. This reflects the spontaneous nature of the data, showing that, most of the time, people do not show extreme emotions. Moreover, more positive emotions, such as happy/amused, appear in the Empathic VA annotations and more negative emotions, such as annoyed/tense, appear in the TV Debates set. This difference comes from the specific nature of the tasks. During political debates (human–human interactions), people try to convince or even impose their opinions on other interlocutors. However, during coaching sessions (human–machine interaction), speakers are quiet and pay attention to the virtual agent while preparing their next exchange.

Table 5. Frequency of the different categories in the corpora. Both the TV Debates and Empathic VA datasets are unbalanced. The majority class is the neutral emotion (calm/indifferent and calm/bored/tired), with more than 70% of the samples.

TV Debates		**Empathic VA**	
Category	**% Audios**	**Category**	**% Audios**
calm/indifferent	73.64	calm/bored/tired	79.47
annoyed/tense	14.32	happy/amused	13.55
enthusiastic	4.72	puzzled	3.11
satisfied/pleased	3.23	annoyed/tense	2.83
worried	2.12	sad	1.04
interested.	1.57		
others	0.40		

As mentioned above, all of the samples were considered for the VAD model. Figure 6 shows the probability density function of each variable (valence, arousal, and dominance) that was obtained by a Gaussian kernel density estimator (upper row). Figure 6 also shows different 2D projections of the sample distribution in the 3D space (row below), representing each scenario in a different color. When regarding arousal, the Empathic VA dataset works in a very neutral scenario, where excitement is almost absent. In TV debates, although neutrality is also predominant, some excitement is perceived due to the nature of debates. The distribution of valence also shows a clear deviation towards positive values for the Empathic VA scenario, which is an indicator of the good acceptance of the system among users, whereas in TV debates, neutrality is predominant, with only a slight nuance towards positiveness. On the contrary, dominance is shifted towards dominant values in TV debates but remains neutral when users interact with the Empathic VA case. These results correlate well with the types of audio we deal with in the two scenarios. In the TV debates, people express themselves without becoming angry (low levels of excitement) but in a very assertive way (quite high dominance levels). Additionally, they appear to be neutral when communicating their opinions (valence tends to be neutral or slightly positive). In the Empathic VA scenario, the users are volunteers with a good predisposition and seem to be pleased with the system (positive valence values). They are relaxed while talking to the agent (levels of excitement tend toward neutrality) and are not intimidated (dominance values are around neutrality, with a slight shift to the right). The differences between human–human and human–machine interactions are also noticeable in the specific tasks we are dealing with. Human–human communication appears to be more intense and emotional, with higher arousal and dominance values. During communication with a machine, on the contrary, people are not confident and they tend to be expectant, which might be translated into low values of arousal/dominance and higher values of valence.

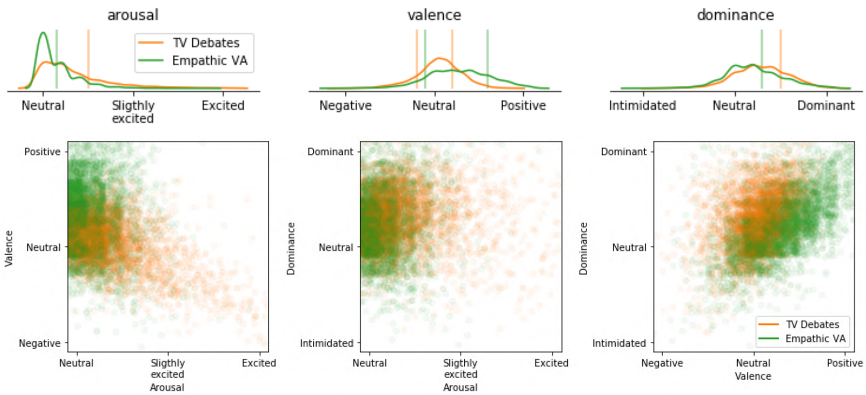


Figure 6. Representation of the VAD dimensional model. In the first row, each of the dimensions is displayed independently, letting us compare each corpus. The vertical lines are the cuts that have been used for the discretization (see Section 3.2). In the second row, a representation of the same dimensions but taken two-by-two is displayed, helping to provide a better understanding of the corpora.

The two models, categorical and VAD, were also considered together. Each category was represented in the 3D VAD space for comparison purposes. Specifically, the average of the valence, arousal, and dominance values of all of the audios labeled within a specific category was computed, and the resulting value was represented as a point in the 3D space. Figure 7 shows a 2D projection of the resulting representation. If we focus on the TV Debates dataset, we notice that interested and worried, the least representative categories, according to Table 5, are very close to the category with the highest number of samples, calm/indifferent, in all of the 2D projections (the purple, orange, and deep-blue points), so they were merged into only one category. The same happens with enthusiastic and satisfied (light-blue and green points). Overall, three different categories were finally considered for the TV Debates dataset. With regard to the Empathic VA scenario, puzzled and sad were merged into a single category because they are extremely close in all three projections, as shown in Figure 7. Thus, the final set of categories used for the classification experiments reported in this work are shown in Table 6.

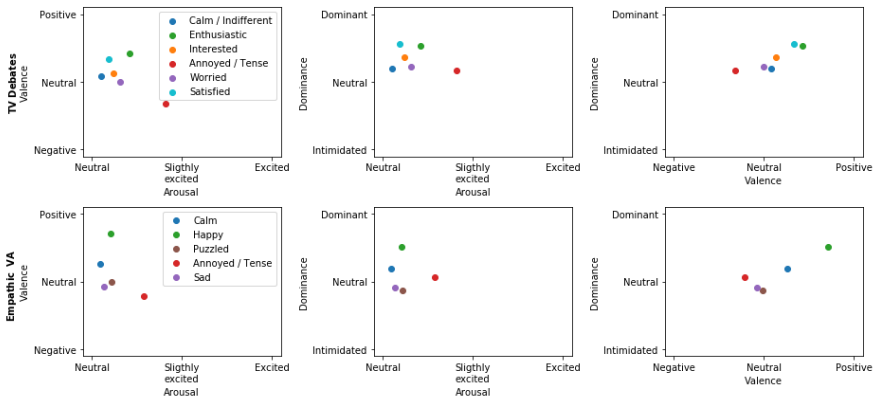


Figure 7. Representation of the mean value of each of the categories in the dimensional model. The first row shows the TV Debates representations, and the second row shows the Empathic VA representations. Some emotions are very close to one another in all dimensions, which is why we considered merging them into a single category.

The distribution of categories is quite different for both sets due to the nature of the scenarios (see Table 6). For instance, annoyed/tense, although present in both tasks, has a very different significance in the TV Debates dataset (almost 15%) and in the Empathic VA dataset (less than 3%). Puzzled was not considered in the TV Debates dataset due to the low number of samples labeled with that emotion, but it entails 3% of the samples in the Empathic VA dataset. Moreover, the final category, puzzled + sad, represented by the union of brown and purple points (low levels of valence and dominance) is not represented in the TV Debates dataset and is slightly separated from the other categories in the Empathic VA scenario. Moreover, Figure 7 shows that the location of annoyed/tense (red point, which is in fact quite separated from the other categories) is closer to neutrality in the Empathic VA scenario (lower excitement levels and lower negative values of valence) than in the TV Debates dataset, meaning that this negative feeling is softer when interacting with the machine and within this specific scenario.

Table 6. Composition of the final categories of each corpus with the number of samples that each category contains.

TV Debates			Empathic VA		
calm/indifferent + interested + worried	CALM	983	calm/bored/tired	CALM	3197
enthusiastic + satisfied	ENT	101	Happy/Amused	HAPPY	545
annoyed/tense	ANN	182	annoyed/tense	ANN	114
			puzzled + sad	PUZZ	167

This correlates well with the idea that people interacting with the Empathic VA scenario are not angry with the system, and if they experience any anger, their feelings are more related to annoyance, which is quite common during debates. Furthermore, speakers in debates do not usually show that they are in an unexpected situation (puzzled), since this emotion can be interpreted as weakness, while it is often shown in interactions with machines. In fact, puzzled was detected in the Empathic VA scenario. Categories such as calm also had a similar location in both scenarios, but with higher values of valence for the Empathic VA interactions; what the annotators perceived as calm tend to be more positive in the Empathic VA scenario than in the TV Debates scenario. The same occurs with enthusiastic + satisfied from the TV Debates scenario and with happy/pleased from the Empathic VA scenario, which although are very close in location in both scenarios (with very similar meanings), happy/pleased seems to have more positive valence values than enthusiastic + satisfied but a bit lower dominance and arousal values.

3.2. Classification Results

Some classification experiments were carried out for the tasks described in Section 2.1. In the TV Debates dataset, 1266 chunks were selected and distributed into the three categories mentioned above (CALM, ANN, and ENT), and for the Empathic VA, 4023 samples were selected and divided into four categories (CALM, HAPPY, ANN, and PUZZ).

When using the dimensional model, previous studies showed that trying to predict a specific value in 3D space (as a regression problem) leads to very poor results [2,15] due to the scarcity of data and the tendency toward neutrality. To solve this problem, a discretization of each dimension was carried out and the regression problem was converted into a classification one. The discrete levels were selected according to the distributions of the annotated data in Figure 6, with orange lines as selected frontiers for the TV Debates dataset and green lines selected for the Empathic VA dataset.

According to the top row displayed in Figure 6, arousal can be approximated by a log-normal distribution with a longer tail towards higher values of excitement. Thus, we decided to discern between only two values: neutral and excited. The thresholds (0.25 for TV Debates and 0.075 for Empathic VA) were selected to keep the classes as balanced as possible without distorting the limits imposed by the density function form.

In the case of valence, three categories were kept because of their similarity to a Gaussian distribution. The decisions related to these thresholds also avoided the imbalance problem. In the TV Debates set, since many of the samples are neutral, the values outside the limits [0.4, 0.6] were considered negative or positive samples respectively. The Empathic VA corpus was slightly more positive, and as a consequence, the limits were shifted towards the more positive values 0.45 and 0.8, respectively.

Finally, the dominance distribution was similar to a Gaussian distribution. However, it shifted towards dominant values; intimidated samples were almost absent. Consequently, only two categories were considered: dominant and neutral. The cutoff limit between neutral samples and dominant ones was set to 0.75 for the TV Debates dataset and to 0.65 for the Empathic VA dataset, which was the less dominant corpus.

Once the aforementioned discretization was applied, the distribution of samples in the different classes remained, as Table 7 shows.

Table 7. Final categories for each dimension of the VAD model with the number of samples they contain in each of the corpora.

		TV Debates	Empathic VA
arousal	neutral	3068	2498
	excited	1050	2027
valence	negative	682	520
	neutral	2239	2811
	positive	1197	1194
dominance	neutral	3039	2946
	dominant	1079	1579

The classification results for the TV Debates and Empathic VA datasets are given in Table 8. The experiments were carried out by considering the categorical and VAD models in an independent way. In both series, the spectrogram represented the speech signal. Different evaluation metrics were given to provide better insight into the capabilities of the neural network in predicting: the accuracy (ACC), precision (P), recall (R), and $F1$-score ($F1$). Since we dealt with a multi-class classification problem, weighted and macro averages were considered. Macro $F1$ is the average of the $F1$-scores for all classes; thus, it penalizes imbalanced datasets, which was the case in this work. It was computed as shown in Equation (1):

$$F1 = \frac{\sum_{i=1}^{N_c} F1^i}{N_c} \tag{1}$$

where N_c is the number of classes and $F1^i$ is the $F1$-score computed assuming that the i-th class is the positive one and that the negative one is composed by the remaining classes.

In weighted $F1$ ($F1_W$) (Equation (2)) instead, the $F1$-scores were calculated for each label, and then, their average is weighted with the number of true instances for each label.

$$F1_W = \frac{\sum_{i=1}^{N_c} n_{C_i} F1^i}{n} \tag{2}$$

where n_{C_i} is the number of samples in C_i class and n is the total number of samples in the test set.

Note that, hereafter, macro averages are denoted as P, R, and $F1$, whereas weighted averages are denoted as P_W, R_W, and $F1_W$.

In the results associated with the TV Debates experiments, a macro $F1$-score of 0.56 was achieved in the categorical model. Interestingly, all of the evaluation metrics (P, R, and $F1$) provided results in the same range and were quite compensated for. Weighted $F1$ ($F1_W$) provided better results (about 0.65) than macro $F1$ due to the imbalance that could be appreciated in the dataset (the minority class comprises only 8% of the corpus, as seen in Table 6). If we focus on the specific categories, the best results were achieved for the most frequent one (CALM), but the $F1$ scores for the rest were still acceptable. Focusing on the VAD model, we notice that arousal provideds a much better $F1$-score, reaching 0.7; the $F1$-score dropped again for valence (0.47) and, then, increased a bit for dominance (0.58). Let us note that three different labels were provided for valence, which made the classification task more difficult, while only two were provided for arousal and dominance. Dominance was the most difficult dimension to perceive for the annotators and the most ambiguous one. Nevertheless, in this dataset, dominance had a significant presence and was efficiently perceived and classified.

The experiments in the Empathic VA task resulted in lower performances. The categorical model provided a much lower $F1$-score when compared with those obtained in the TV Debates dataset, which may be due to the imbalance being even higher in this dataset. The minority class comprised 2.8% of the whole corpus, which was lower than that found in the TV Debates dataset (8%), as shown in Table 6. In fact, looking at the independent categories, the evaluation metrics were very low for less-frequent classes, such as PUZZ or ANN. Moreover, in this set, the number of labels was higher (four instead of three), which also leads, in general, to more confusion and lower performance. The VAD model followed the same tendency observed in the TV Debates dataset, with the highest performance for arousal and lower values for valence and dominance. However, in this case, the results achieved in the previous corpus were not reached, either, because it was a more neutral dataset and little emotional information could be learned.

Table 8. Classification results on the spectrogram input for each of the problems (categorical and each dimension of the VAD) and corpus (TV Debates and Empathic VA). Each problem has a row per category with the precision, recall, and *F*1-score metrics and a row of means (Overall) that shows the macro and weighted average measures of the metrics for all categories.

		TV Debates				Empathic VA				
		Acc	**P/P$_W$**	**R/R$_W$**	***F*1/*F*1$_W$**	**Acc**	**P/P$_W$**	**R/R$_W$**	***F*1/*F*1$_W$**	
Cat.	overall	0.65	0.56/0.65	0.57/0.65	0.56/0.65	0.73	0.34/0.64	0.27/0.73	0.26/0.66	overall
	CALM		0.75	0.74	0.75		0.76	0.94	0.84	CALM
	ENT		0.42	0.45	0.43		0.26	0.08	0.13	HAPPY
	ANN		0.51	0.50	0.51		0.20	0.01	0.03	ANN
							0.14	0.03	0.05	PUZZ
Aro.	overall	0.76	0.69/0.77	0.71/0.76	0.70/0.77	0.58	0.58/0.58	0.56/0.58	0.54/0.55	overall
	neutral		0.86	0.82	0.84		0.59	0.81	0.68	neutral
	excited		0.53	0.60	0.56		0.57	0.30	0.40	excited
Val.	overall	0.51	0.48/0.52	0.47/0.51	0.47/0.52	0.55	0.41/0.52	0.39/0.55	0.38/0.52	overall
	negative		0.41	0.36	0.38		0.21	0.21	0.21	negative
	neutral		0.60	0.59	0.59		0.63	0.77	0.69	neutral
	positive		0.42	0.47	0.44		0.40	0.18	0.25	positive
Dom.	overall	0.63	0.58/0.69	0.60/0.63	0.58/0.65	0.63	0.59/0.62	0.58/0.63	0.59/0.63	overall
	neutral		0.80	0.66	0.72		0.71	0.74	0.73	neutral
	dominant		0.36	0.54	0.43		0.47	0.42	0.45	dominant

Preliminary Classification Results Using wav2vec Model

Some preliminary experiments were also carried out using the wav2vec model, as shown in Table 9. The performance achieved was significantly lower. Minority categories, such as HAPPY and PUZZ, were almost never recognized. However, the same tendency observed with the spectrogram was also perceived, here: the VAD model performed better than the categorical one, where arousal was the best recognized dimension and weighted averages provided better results due to the imbalanced nature of these scenarios. Thus, the results achieved were promising, considering the pretrained nature of the model and the specific datasets employed in the training process. These datasets were based on speech that is quite far from the conversational nature of the scenarios we deal with in this work. Their contents were mostly neutral, and Spanish was scarcely included. A fine-tuning process would be needed, in this case, to adapt the model to specific features of the task and language. However, the aforementioned corpora might not be large enough to robustly perform such an adaptation, which, currently, makes the use of pretrained representations of the speech signal to model emotions in Spanish really difficult.

Table 9. Classification results on wav2vec input for each of the problems (categorical and each dimension of the VAD) and corpus (TV Debates and Empathic VA). Each problem shows the accuracy, and macro and weighted average measures of the metrics (precision, recall, and *F*1-score) for all categories.

	TV Debates				Empathic VA			
	Acc	**P/P$_W$**	**R/R$_W$**	***F*1/*F*1$_W$**	**Acc**	**P/P$_W$**	**R/R$_W$**	***F*1/*F*1$_W$**
categorical	0.63	0.44/0.53	0.34/0.63	0.27/0.50	0.79	0.20/0.63	0.25/0.79	0.22/0.70
arousal	0.75	0.64/0.71	0.56/0.75	0.55/0.70	0.53	0.51/0.51	0.51/0.53	0.49/0.51
valence	0.38	0.29/0.39	0.35/0.38	0.27/0.33	0.62	0.33/0.48	0.33/0.62	0.26/0.48
dominance	0.74	0.37/0.54	0.50/0.74	0.42/0.63	0.63	0.55/0.59	0.53/0.63	0.52/0.59

4. Discussion

4.1. Analysis of Emotions

The perceived emotions provide very valuable information to profile the specific features of the speakers in a scenario. The analyses carried out showed, for instance, the predominant neutrality in scenarios where spontaneous speech was considered. As mentioned above, most of the time, conversational speech does not show extreme emotions and tends to be calm. However, some differences are found when analyzing different scenarios such as the ones tackled in this work. Quite noticeable was that human–human interactions in a TV Debate format showed higher levels of excitement and dominance and more negativeness, whereas human–machine interactions in the Empathic VA task showed more positive feelings and lower values of excitement and dominance. These observations could be easily reflected using the VAD model. However, when trying to translate it to categories, finding appropriate ones is difficult without a previous perception study. This fact makes the VAD model very appropriate when dealing with a new real task that is not an artificial database specially designed for carrying out machine learning studies (with all five basic emotions equally distributed).

When focusing on a speaker, their emotions can also help profile them. Looking at their dominance, for instance, provides good insight into the kind of person they are when taking part in a conversation (speaking in public to convince others vs. talking in a relaxed environment an interest). Valence can also provide information about the speaker's interest during a conversation.

The experiments conducted also showed that the location of a specific category in the 3D VAD space could vary depending on the scenario considered. As shown above, the category CALM in the Empathic VA dataset was more positive than in TV Debates, showing the ambiguity in the definition of the categories and the relevance of the VAD model, which might consider more general definitions.

4.2. Classification Results

The classification experiments carried out show that the system performance was significantly better in the TV Debates scenario than in Empathic VA one when using the categorical model, which may be due to the specific composition of the tasks. In fact, even though the TV Debates scenario is a heavily imbalanced task, the percentage of minority classes was higher than in the Empathic VA one. This deviation towards only one category is very difficult to tackle, even using oversampling methods. In future work, the use of a data-augmentation technique, such as SMOTE algorithm [22,70] may be useful. Moreover, the analysis of emotions in the Empathic VA dataset revealed that the dataset is a very neutral corpus (much more than the TV Debates one), which complicates the detection of emotional information. This fact is a major challenge when designing emotionally conscious human–machine interaction systems. However, the differences among each value of the VAD dimensions (excited/neutral for arousal, negative/neutral/positive for valence, or dominant/neutral for dominance) were not very significant. Thus, we can hypothesize that the VAD model might helped extract more precise and valuable information when considering spontaneous emotions that tended toward neutrality. Finally, the preliminary experiments conducted with the wav2vec model showed that these kinds of representations, although very powerful, would require a tuning and adaptation process specific for the task and language under consideration.

5. Conclusions

This work analyzed the emotional features found in two very different spontaneous speech scenarios: human interactions during TV debates and human–machine conversations with a virtual agent. In both scenarios, a very reduced set of emotions was perceived by a large number of annotators. Moreover, the emotional information had a high tendency to be neutral, with the rest of the emotions being of a clear minority. This fact raised a difficult pattern recognition problem, which was the imbalance in the data. Overall, the automatic identification of spontaneous speech and related emotional content is still a difficult problem to address. However, this work also showed that human interactions could be more emotional and, thus, easier to tackle than human–machine interactions.

Thus, the design of human–machine conversational systems, aimed at integrating a user's emotional state, are still challenging tasks.

In this framework, the VAD model was demonstrated to be more adequate in representing emotional information. The dimensional VAD space, in fact, could be better managed than categories in terms of annotation and automatic identification. The classification experiments carried out in this work showed that deep learning approaches are useful for detecting speakers' emotional information, reaching a 0.7 $F1$-score for arousal. The preliminary experiments with the novel wav2vec2.0 representation of speech signals seem to be promising. However, this representation needs large sets of spontaneous emotional speech in the target language, i.e., Spanish, which are not currently found.

Author Contributions: Conceptualization, M.d.V., R.J. and M.I.T.; methodology, M.d.V. and R.J.; software, M.d.V.; validation, M.d.V. and R.J.; formal analysis, M.d.V., R.J. and M.I.T.; investigation, M.d.V. and R.J.; resources, M.d.V., R.J. and M.I.T.; data curation, M.d.V., R.J. and M.I.T.; writing—original draft preparation, M.d.V. and R.J.; writing—review and editing, M.d.V., R.J. and M.I.T.; visualization, M.d.V. and R.J.; supervision, R.J. and M.I.T.; project administration, R.J. and M.I.T.; funding acquisition, R.J. and M.I.T. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this paper was conducted as part of the AMIC and EMPATHIC projects, which received funding from the Spanish Minister of Science under grants TIN2017-85854-C4-3-R and PDC2021-120846-C43 and from the European Union's Horizon 2020 research and innovation program under grant agreement No. 769872. The first author also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD Comisión Europea

Institutional Review Board Statement: The experiments recorded conversations between seniors and a WoZ (the Empathic VA corpus) in Basque Country, Spain. The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee for Research involving Human Beings of the University of the Basque Country and the Basque Ethical Committee for the Clinical Research (Comité de ética de la investigación clínica (CEIC) de Euskadi).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study resulting in the Empathic VA corpus.

Data Availability Statement: The Empathic VA corpus is distributed for research purposes by the European Language Resources Association (ELRA http://www.elra.info/en/about/ (accessed on 3 February 2022)) at a very low price for academic and research institutions, as well as for small companies. The TV Debates corpus will be made available upon request only for research purposes.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lalitha, S.; Tripathi, S.; Gupta, D. Enhanced speech emotion detection using deep neural networks. *Int. J. Speech Technol.* **2019**, *22*, 497–510. [CrossRef]
2. deVelasco, M.; Justo, R.; López-Zorrilla, A.; Torres, M.I. Automatic analysis of emotions from speech in Spanish TV debates. *Acta Polytech. Hung.* 2022, in press.
3. Kiss, G.; Vicsi, K. Comparison of read and spontaneous speech in case of automatic detection of depression. In Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, 11–14 September 2017; pp. 213–218. [CrossRef]
4. He, L.; Cao, C. Automated depression analysis using convolutional neural networks from speech. *J. Biomed. Inform.* **2018**, *83*, 103–111. [CrossRef] [PubMed]
5. Balagopalan, A.; Eyre, B.; Rudzicz, F.; Novikova, J. To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv* **2020**, arXiv:2008.01551.
6. Han, J.; Qian, K.; Song, M.; Yang, Z.; Ren, Z.; Liu, S.; Liu, J.; Zheng, H.; Ji, W.; Koike, T.; et al. An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. *arXiv* **2020**, arXiv:2005.00096.

7. Schuller, B.; Weninger, F.; Zhang, Y.; Ringeval, F.; Batliner, A.; Steidl, S.; Eyben, F.; Marchi, E.; Vinciarelli, A.; Scherer, K.R.; et al. Affective and Behavioural Computing: Lessons Learnt from the First Computational Paralinguistics Challenge. *Comput. Speech Lang.* **2019**, *53*, 156–180. [CrossRef]
8. Justo, R.; Letaifa, L.B.; Palmero, C.; Fraile, E.G.; Johansen, A.; Vazquez, A.; Cordasco, G.; Schlogl, S.; Ruanova, B.F.; Silva, M.; et al. Analysis of the Interaction between Elderly People and a Simulated Virtual Coach. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 6125–6140. [CrossRef]
9. Vinciarelli, A.; Esposito, A.; André, E.; Bonin, F.; Chetouani, M.; Cohn, J.F.; Cristani, M.; Fuhrmann, F.; Gilmartin, E.; Hammal, Z.; et al. Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. *Cogn. Comput.* **2015**, *7*, 397–413. [CrossRef]
10. Chiba, Y.; Nose, T.; Ito, A. Analysis of efficient multimodal features for estimating user's willingness to talk: Comparison of human-machine and human-human dialog. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 428–431.
11. Aïsha, S.; Elisabeth, P.; Ouriel, G.; Bruno, B. Predictive Mechanisms Are Not Involved the Same Way during Human-Human vs. Human-Machine Interactions: A Review. *Front. Neurorobot.* **2017**, *11*, 52.
12. deVelasco, M.; Justo, R.; Letaifa, L.B.; Torres, M. Contrasting the emotions identified in spanish tv debates and in human-machine interactions. In Proceedings of the IberSPEECH, Valladolid, Spain, 24–25 March 2021.
13. Rodríguez, L.J.; Torres, M.I. Spontaneous Speech Events in Two Speech Databases of Human-Computer and Human-Human Dialogs in Spanish. *Lang. Speech* **2006**, *49*, 333–366. [CrossRef]
14. Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; Pantic, M. AVEC 2012: The continuous audio/visual emotion challenge. In Proceedings of the 14th ACM International conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; pp. 449–456.
15. deVelasco, M.; Justo, R.; López-Zorrilla, A.; Torres, M. Can Spontaneous Emotions be Detected from Speech on TV Political Debates? In Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications, Naples, Italy, 23–25 October 2019.
16. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005.
17. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
18. Davidson, R.J.; Ekman, P.A. *Nature of Emotion: Fundamental Questions*; Oxford University Press: New York, NY, USA; Springer: New York, NY, USA, 1994.
19. Nasri, M.A.; Hmani, M.A.; Mtibaa, A.; Petrovska-Delacrétaz, D.; Slima, M.B.; Hamida, A.B. Face Emotion Recognition From Static Image Based on Convolution Neural Networks. In Proceedings of the 5th International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020, Sousse, Tunisia, 2–5 September 2020; pp. 1–6. [CrossRef]
20. Vogt, T.; Andre, E. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; pp. 474–477. [CrossRef]
21. Scherer, K.R. *Approaches To Emotion. Chapter: On the Nature and Function of Emotion: A Component Process Approach*; Scherer, K.R., Ekman, P., Eds.; Taylor and Francis Group: New York, NY, USA, 1984.
22. Letaifa, L.B.; Torres, M.I. Perceptual Borderline for Balancing Multi-Class Spontaneous Emotional Data. *IEEE Access* **2021**, *9*, 55939–55954. [CrossRef]
23. Gunes, H.; Pantic, M. Automatic, Dimensional and Continuous Emotion Recognition. *Int. J. Synth. Emot.* **2010**, *1*, 68–99. [CrossRef]
24. 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. [CrossRef]
25. Russell, J. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
26. Chakraborty, R.; Pandharipande, M.; Kopparapu, S.K. *Analyzing Emotion in Spontaneous Speech*; Springer: Berlin/Heidelberg, Germany, 2017.
27. Greco, C.; Buono, C.; Buch-Cardona, P.; Cordasco, G.; Escalera, S.; Esposito, A.; Fernandez, A.; Kyslitska, D.; Kornes, M.S.; Palmero, C.; et al. Emotional Features of Interactions with Empathic Agents. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2168–2176. [CrossRef]
28. Aroyo, L.; Welty, C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.* **2015**, *36*, 15–24. [CrossRef]
29. Justo, R.; Alcaide, J.M.; Torresn, M.I. CrowdScience: Crowdsourcing for research and development. In Proceedings of the IberSpeech 2016, Lisbon, Portugal, 23–25 November 2016; pp. 403–410.
30. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013.

31. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
32. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *arXiv* **2017**, arXiv:1706.00612.
33. Parthasarathy, S.; Tashev, I. Convolutional Neural Network Techniques for Speech Emotion Recognition. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 121–125. [CrossRef]
34. Huang, K.; Wu, C.; Hong, Q.; Su, M.; Chen, Y. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5866–5870. [CrossRef]
35. Marazakis, M.; Papadakis, D.; Nikolaou, C.; Constanta, P. System-level infrastructure issues for controlled interactions among autonomous participants in electronic commerce processes. In Proceedings of the Tenth International Workshop on Database and Expert Systems Applications. DEXA 99, Florence, Italy, 3 September 1999; pp. 613–617. [CrossRef]
36. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-Based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484. [CrossRef]
37. Ocquaye, E.N.N.; Mao, Q.; Xue, Y.; Song, H. Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *Int. J. Intell. Syst.* **2021**, *36*, 53–71. [CrossRef]
38. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 13 September 2018; pp. 5089–5093. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *arXiv* **2017**, arXiv:1706.03762.
40. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
41. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:2006.11477.
42. Hsu, W.; Bolte, B.; Tsai, Y.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447.
43. Esposito, A.; Stejskal, V.; Smékal, Z. Cognitive Role of Speech Pauses and Algorithmic Considerations for their Processing. *Int. J. Pattern Recognit. Artif. Intell.* **2008**, *22*, 1073–1088. [CrossRef]
44. Ortega, A.; Lleida, E.; Segundo, R.S.; Ferreiros, J.; Hurtado, L.F.; Arnal, E.S.; Torres, M.I.; Justo, R. AMIC: Affective multimedia analytics with inclusive and natural communication. *Proces. Leng. Natural* **2018**, *61*, 147–150.
45. Torres, M.I.; Olaso, J.M.; Montenegro, C.; Santana, R.; Vázquez, A.; Justo, R.; Lozano, J.A.; Schlögl, S.; Chollet, G.; Dugan, N.; et al. The EMPATHIC Project: Mid-Term Achievements. In Proceedings of the PETRA '19: 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece, 5–7 June 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 629–638.
46. Brinkschulte, L.; Mariacher, N.; Schlögl, S.; Torres, M.I.; Justo, R.; Olaso, J.M.; Esposito, A.; Cordasco, G.; Chollet, G.; Glackin, C.; et al. The EMPATHIC Project: Building an Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly. *arXiv* **2021**, arXiv:2104.13836.
47. Calvo, R.; D'Mello, S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Trans. Affect. Comput.* **2010**, *1*, 18–37. [CrossRef]
48. Calvo, R.; Kim, S. Emotions in text: Dimensional and categorical models. *Comput. Intell.* **2012**, *29*, 527–543. [CrossRef]
49. Russell, J.A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **2003**, *110*, 145–172. [CrossRef]
50. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
51. Cowen, A.S.; Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E7900–E7909. [CrossRef]
52. Scherer, K.R. What are emotions? And how can they be measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [CrossRef]
53. Justo, R.; Torres, M.; Alcaide, J. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. In *Iberian Conference on Pattern Recognition and Image Analysis*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 58–68. [CrossRef]
54. Wester, F.; Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; Communications 2005; Sage: Thousand Oaks, CA, USA, 2004; pp. 124–126.
55. Tian, L.; Moore, J.D.; Lai, C. Emotion recognition in spontaneous and acted dialogues. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 698–704.

56. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
57. Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. *Speech Emotion Recognition Using Spectrogram & Phoneme Embedding*. In Proceedings of the Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 6 September 2018; pp. 3688–3692.
58. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
59. Tao, F.; Liu, G. Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 13 September 2018; pp. 2906–2910. [CrossRef]
60. Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; Tarokh, V. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6474–6478. [CrossRef]
61. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
62. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness-Knowl.-Based Syst.* **1998**, *6*, 107–116. [CrossRef]
63. Jin, Z.; Finkelstein, A.; Mysore, G.J.; Lu, J. FFTNet: A real-time speaker-dependent neural vocoder. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 13 September 2018; pp. 2251–2255.
64. Akiyama, O.; Sato, J. Multitask learning and semisupervised learning with noisy data for audio tagging. In Proceedings of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2019), New York, NY, USA, 25–26 October 2019.
65. Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; Wilson, A.G. Averaging weights leads to wider optima and better generalization. *arXiv* **2018**, arXiv:1803.05407.
66. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech* **2020**, *2020*, 2757–2761. [CrossRef]
67. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
68. Cui, J.; Cui, X.; Ramabhadran, B.; Kim, J.; Kingsbury, B.; Mamou, J.; Mangu, L.; Picheny, M.; Sainath, T.N.; Sethy, A. Developing speech recognition systems for corpus indexing under the IARPA Babel program. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6753–6757.
69. Luna-Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.M.; Fernández-Martínez, F. A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. *Appl. Sci.* **2022**, *12*, 327. [CrossRef]
70. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

Publication 17

Mental Health Monitoring from Speech and Language

Mental Health Monitoring from Speech and Language

Irune Zubiaga, Ignacio Menchaca, Mikel de Velasco, Raquel Justo

Universidad del Pais Vasco UPV/EHU. Sarriena s/n. 48940 Leioa. Spain.

irune.zubiaga@ehu.eus, mikel.develasco@ehu.eus, raquel.justo@ehu.eus

Abstract

Concern for mental health has increased in the last years due to its impact in people life quality and its consequential effect on healthcare systems. Automatic systems that can help in the diagnosis, symptom monitoring, alarm generation etc. are an emerging technology that has provided several challenges to the scientific community. The goal of this work is to design a system capable of distinguishing between healthy and depressed and/or anxious subjects, in a realistic environment, using their speech. The system is based on efficient representations of acoustic signals and text representations extracted within the self-supervised paradigm. Considering the good results achieved by using acoustic signals, another set of experiments was carried out in order to detect the specific illness. An analysis of the emotional information and its impact in the presented task is also tackled as an additional contribution.

Index Terms: acoustic signal, textual information, mental health monitoring

1. Introduction

Mental Health is an essential component of overall health. However, depression and anxiety are still common disorders. As an example, in Europe the overall prevalence of current depressive disorder is between 5-10%, according to [1], with potentially large differences between countries and time. This kind of disorders are a major cause of disability, increasing the risk of premature mortality, decreasing life quality, and creating a substantial burden on health systems [2]. Many people may be described as “living with” a mental illness, and managing their own symptoms. However, they are often unsure of the thresholds for treatment, how to control their symptoms, which are the best coping strategies or the available resources.

AI based systems may facilitate watchful waiting and symptom monitoring by initiating contact and symptom checking at various times of the day and night. Thus, they can play a relevant role in the detection of the illness and in the patients care. These automatic systems can take advantage of different information sources like voice, gait, EEG, facial expressions, etc. [3, 4]. Speech can be an easily accessible, non-invasive marker, whose features can significantly change due to slight psychological or physiological changes [5]. Therefore, it can be considered a key marker for detecting depression and suicide risk [5, 6]. Different features like acoustic parameters extracted from fundamental frequency [7] or vocal-source-based features (jitter, shimmer, etc.) [5] have been used as successful cues for predicting depression. These features, among others, have been exploited by different machine learning algorithms. Spectral and prosodic features along with their statistics extracted using openSMILE toolkit [8] have been used to train support vector machines and random forest models for depression detection [9]. More recently deep Convolutional Neural Networks (CNNs) have been used to extract acoustic embeddings [10] and detect depression from speech [11, 12].

In this work, we focus on an alternative and more efficient way of representing the audio. The rise in the self-supervised learning paradigm and the recently proposed transformer architecture [13], have led to novel speech representations, such as wav2vec [14], HuBERT [15] or the most recent UniSpeech-Sat [16]. These models are inspired by deep Transformer-based text generation models[17] such as GPT[18, 19, 20] and BERT[21] that are able to extract features from simple text without additional annotation. Audio encoder models have also been used to extract speech representations. In this work, the HuBERT representation was selected to feed a simple neural network that can be trained with small amounts of data. This way, we can tackle an anxiety and depression detection task in a realistic environment where getting a large training set is time consuming and expensive. The HuBERT representation was also compared to the spectral and prosodic features achieved with openSMILE. Moreover, the transcriptions of the utterances were also considered as an information source. A BERT based representation of the text corresponding to the audio transcriptions, was also employed with the same aim. This way, an audio based system and a text based system were built to perform the anxiety and/or depression detection task. We also carried out an audio centered analysis in order to measure the impact of emotions, represented as a 2 dimensional model (Valence and Arousal), in the prediction of depression and/or anxiety.

The manuscript is organized as follows, Section 2 describes the task and corpus tackled in this work. Section 3 deals with the representation of the audio signal and the textual transcriptions. Section 4 and Section 5 detail the different sets of experiments that were carried out and Section 6 sums up the main conclusions and future work.

2. Mental health monitoring task

This work deals with the data acquired within the framework of the H2020-MSCA-RISE project MENHIR [22] (Mental health monitoring through interactive conversations). In this project 60 conversations between a counsellor and a participant were recorded to form a corpus. Participants were divided in two groups: **AMH** and **Control**. AMH (32 members) consists of users of the Action Mental Health (AMH) foundation[1], diagnosed with some kind of mental illness, depression and anxiety being the most common ones. In contrast the Control group (28 members) is formed of people who have never been diagnosed with any kind of mental illness.

The interviews consist of three main sections; In the first section the counsellor asks the patient 5 questions that lead to non-emotional conversation. The second part consists of fourteen affirmations from the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [23]. The participants have 5 possible answers that go from ”None of the Time” to ”All of the Time” to indicate how often they feel the way that these affirmations

[1] https://www.amh.org.uk/

10.21437/SMM.2022-3

Table 1: *Distribution of Anxiety/Depression in AMH and Control.*

	No. speakers	No. interventions
Depression	3	276
Anxiety	2	140
Both	16	824
Control	20	614

express. In this part, in addition to the answer, sometimes the interviewees added some dialogue of their own to explain further the answer they gave. In this section the counsellor perception of the speaker emotional status was also annotated according to the following questionnaire:

How do you perceive the client?

- Excited/activated/agitated
- Slightly Excited/activated/agitated
- Calm

His/her mood is:

- Positive
- Slightly positive
- Slightly negative
- Negative

Finally, in the last part, the participants were asked to read a text passage of the popular tale ”The Boy and the Wolf”.

2.1. Speech corpus

The second section of the interviews, with its corresponding emotional information, was used For generating the speech corpus. This section consists of 14 questions that the potential patients (speakers) respond to. The audio sessions were split according to each turn or intervention of the speaker. This way, the corpus was formed by the audio files corresponding to those interventions for which the counsellor annotated their perception of Valence and Arousal levels on the participants speech. Said interventions were extracted from 41 interview recordings; 21 regarding AMH and 20 regarding Control. The total corpus consists of 1854 audio files (interventions) and has a length of 4 hr 15 min and 45 sec. The number of speakers and interventions associated to each mental illness is given in Table 1.

2.2. Textual corpus

The textual corpus consists of transcriptions of the aforementioned interviews. These transcriptions include both dialogues written in a literal way and annotations regarding paralinguistic or acoustic information; *noises (music, footsteps etc.)*, *pauses*, *laughs* and such. We ignored annotations regarding noise and will refer to the rest of annotations as **paralinguistic tokens** from now on. For our task, we only considered the interviewee's conversational information and paralinguistic tokens. The transcription of the reading phase was also removed since there is no distinguishing semantic information associated with it. All the text associated to the remaining dialogues was gathered as a corpus consisting of 6741 sentences; 4743 sentences regarding AMH and 1998 regarding Control. Starting from this, we created four textual data sets to work with.

To create the first two data sets a cleaning process was carried out and thus we will refer to them as clean data. First, the typos in the text were corrected as far as possible. Then, the answers to the questions from the WEMWBE scale (fixed response like “None of the time”, etc.) and the yes/no answers

Table 2: *Paralinguistic tokens in CD and their meanings.*

Token	Meaning
{INTERJECTIONS}	uses an interjection
{CUT}	stops speaking mid-word
{inhaling}	inhales
{laugh}	laughs
<pause>WORDS</pause>	says something between pauses
{STALLING}	makes a sound denoting they are thinking
<laugh>WORDS</laugh>	says something while laughing
{breathing}	takes a deep breath
{tsk}	makes a flicking sound with their mouth
{STUTTER}	stutters when saying a word
{PUZZLEMENT}	makes a sound denoting puzzlement
{cough}	coughs
<breathing>WORDS</breathing>	says something while taking a deep breath
<pause>empty</pause>	makes a long pause
pause at start	marks if the sentence starts with a pause

Table 3: *Paralinguistic tokens in D and their meanings.*

Token	Meaning
(inhaling)	inhales
(laugh)	laughs
(breathing)	takes a deep breath
(tsk)	makes a flicking sound with their mouth
(cough)	coughs

were removed. In the next step, paralinguistic features that were gathered as plain text were represented using a unique token from the ones given in Table 2 (ex. em, ew, jeeze, oh... were represented as {*INTERJECTIONS*}). This process is explained in more detail in [24]. This way, two data sets were created; one considering clean plain text, which we will refer to as **CD**, and another one considering clean plain text and paralinguistic tokens from Table 2, which we will refer to as **CD+T**. These two data sets consist of 3955 sentences; 2810 sentences from the AMH group and 1145 from the Control group. The other two data sets were formed with a more natural approach; we did not do any kind of text cleaning, typo correction or word grouping. We used the literal transcriptions of the dialogues and the paralinguistic tokens were written between parentheses to process them as plain text. Their representations are shorted in Table 3. Two new data sets were created; one considering plain text, which we will refer to as **D**, and another one considering plain text and paralinguistic tokens from Table 3, which we will refer to as **D+T**.These data sets consists of 4872 sentences; 3308 regarding AMH and 1564 regarding Control.

3. Data processing

3.1. Acoustic Features

Affective processes can change Arousal and tension influencing both voice and speech production. These changes can be estimated with different parameters of the acoustic waveform. In this research, we used GeMAPS and Hubert to extract features of the vocal expressions, from acoustic waveforms, that provide this kind of information.

The GeMAPS (Genova Minimalistic Parameter Set) [25] feature set is divided into two main blocks.; the first block is composed of some prosodic, excitation and vocal tract descriptors, and the second one of some dynamic and cepstral descriptors. The first descriptors are classified depending on their relation to different physical characteristics like frequency (pitch or jitter) energy (shimmer, loudness or harmonics-to-noise ratio) and spectrum (alpha ratio, hammaberg index, spectral slope or harmonic differences). The second ones are parameters such as mel-frequency cepstral coefficients (MFCC) or spectral flux. Furthermore, features such as arithmetic mean, coefficients of variation, 20/50/80 percentiles and other temporal parameters have been added to the set, making a total of 88 different features. The OpenSmile Python library [8] was used to achieve this set of 88 features from an audio file in wav format.

Alternatively, Hubert was chosen with the pre-trained parameters *hubert-large-ll60k*, which have been fitted with the Libri-Light [26] corpus at 16k hertz. Such models trained only on raw audio have proven to be able to extract very representative features that have been widely used for different tasks. The model extracts a total of 1024 features every 20ms which have then been reduced by averaging to 1024 features.

3.2. Text representation

For text representation we worked with BERT (Bidirectional Encoder Representations from Transformers) [21]. BERT is based on a deep Transformer encoder network [17]. This kind of network can process long texts efficiently by using self-attention. BERT is bidirectional, which means that it uses the whole input text to understand the meaning of each word. In its base size (the one presented in [21]), BERT is composed of 12 successive transformer layers, each having 12 attention heads and has a total number of 110 million parameters. The BERT Encoder block calculates a 768-long vector representing an embedding of each input token.

4. Prediction of a mental disorder

Regarding the detection of mental illness, two different experiments were conducted. The first one tried to differentiate AMH and Control group members (Section 4.1).The second one will try to identify which participants suffer from anxiety and which ones from depression within the AMH members (Section 4.2).

For this purpose, 8 folds of the data set have been built in order to create a Cross-Validation in which no intervention from the interviews that form the test subset appears in the train partition. This way more robust and reliable results were achieved.

4.1. Prediction of interventions related to Healthy and Ill subjects

The task to discriminate healthy and ill subjects was carried out using both acoustic and textual information separately.

As for the acoustic information, the speech corpus described in Section 2.1 was considered. The GeMAPS and HuBERT feature sets were used to feed a Deep Neural Network and a random oversampling method was used in the training set to balance the data. In both cases the network consisted of two hidden layers, the first one with a ReLu activation function and an output layer with the softmax activation function. Due to the different feature dimension in each set the hidden layer was 32 dimensional when using GeMAPS and 128 dimensional when using HuBERT. Adam optimizer was used with a learning rate of 1e-4. The batch size was set to 32 and the cross-entropy loss function was used. The training was done over 250 epochs.

When regarding textual information, BERT was used to classify mentally ill and healthy people. To carry out this task, we used a model that consisted of a BERT main layer and a classification head [27]. Three different pre-trained models were used as a checkpoint for the BERT main layer and for the tokenizer: *bert-base-cased* and *bert-base-uncased* witch were presented in [21] and *bert-base-uncased-emotion* [28] which is a pre-trained *bert-base-uncased* model fine-tuned with the *emotion* data set [29]. Then, the whole model was fine-tuned using the four variations of the textual corpus described in Section 2.2. The chosen optimizer was AdamW and a linear scheduler was used. The learning rate was set to 5e-5, the batch size to 16 and the fine-tuning was done in 3 epochs. These parameters were chosen following the recommendations in [21]. In the case of **CD+T** the paralinguistic tokens were added to the tokenizer as *additional special tokens* to try and learn a representation for them. In contrast, when working with **D+T**, we processed paralinguistic tokens as if they were plain text. The obtained results for this task are shorted in Table 4.

Table 4: *Macro F1-scores of the results obtained in the audio based and text based approaches to the AMH and Control group discrimination problem.*

	GeMAPS	HuBERT
Acoustic DNN	0.88	**0.94**

	CD	CD+T	D	D+T
bert-base-cased	0.61	0.56	0.66	0.60
bert-base-uncased	0.60	0.55	0.67	0.56
bert-base-uncased-emotion	0.60	0.60	0.65	**0.68**

As seen in Table 4, when working with text the best result is obtained when using *bert-base-uncased-emotion*, which can be a cause of this model being fine-tuned on emotional data. The general lack of accuracy in BERT's predictions is probably a consequence of the majority of sentences having a length of one or two words. In addition, most long sentences were difficult to classify even by humans (ex."two rounds of granary toast with a banana") and the data set was noisy, which has shown to significantly degrade BERT's performance when fine-tuning it for tasks such as sentiment analysis [30]. The results are better when working with non clean data. This might be because BERT uses features to provide context information that are being removed by cleaning words and representing paralinguistic features with tokens (ex. replacing *wh-what?* with {*STUTTER*}). It can also be observed that when using *bert-base-cased* and *bert-base-uncased* taking paralinguistic tokens into account makes the results worse. In the case of **CD+T**, this can be because there was not a pre-learned representation for these tokens and the performed training is not enough as to learn one. In the case of **D+T** this might be because they contain no semantic information. In contrast, when working *bert-base-uncased-emotion*, **D+T** has a better outcome than **D**. This might mean that there is some kind of emotional information in the tokens from Table 3. In the case of **CD** and **CD+T** it appears that while a representation for tokens in Table 2 still seems hard to learn, their presence does not worsen the results as seen when using other model checkpoints. It is worth noting that in this case the transcriptions were done by humans and if a full automatic system is required, an ASR for text detection would have to be introduced, which would lead to slightly worse results.

When working with acoustic signals the system performance was significantly better. This shows that the acoustic model is able to identify some features that are not detectable in text and that help differentiating one group from the other. The GeMAPS features turned out to be worse than those extracted with the Hubert model. This highlights the power of the acoustic embeddings achieved making use of semi-supervised learning. Besides, the number of features extracted with GeMAPS (88) is smaller than those extracted with HuBERT (1024).

Table 5: *Averaged F1-Score for the Cross-Validation on anxiety and depression detection problems.*

	GeMAPS	HuBERT
Anxiety	0.64	**0.71**
Depression	0.53	**0.70**

4.2. Prediction of Depression and Anxiety Interventions

The results achieved with speech signal encouraged us to try to discriminate between depression and anxiety. These sets of experiments were only performed on the AMH group.

With this aim, a system was designed to predict whether an intervention corresponds to a patient with or without depression. Alternatively, an additional network was designed to predict whether it corresponds to a patient with or without anxiety.

Both experiments replicate the procedure mentioned in Section 4.1 but with an adapted output to the new classification problem. The results are shown in Table 5. The results are not as good as those in Section 4.1 because the task is now more challenging as a consequence of the two classes being more similar between them. All samples belong to the AMH group, which means that even if the subject does not have the target illness, it does suffer from another one. Additionally, the class imbalance problem is more relevant in this case. However, the results with HuBERT features are still promising.

5. Detection of Emotional Information

As a first step to include affective information, an analysis of the counselors emotional annotations was carried out. Figure 1 shows the histogram for the interventions annotated with different values of Valence and Arousal from the AMH and the Control group. As expected, lower values of pleasure were perceived in AMH group, meaning that their status is generally more negative than the status of the members of Control group, whose values are more positive. Something similar can be observed in the Arousal levels, where AMH members show lower levels of excitement. However, the difference is not as significant in this case, revealing that Valence might be a more informative feature for the detection of anxiety and depression.

When focusing on specific illnesses and differences between them, the analysis shown in Figure 2 was carried out. This figure shows the percentage of interventions from each group (Depression, Anxiety, Both, None) that were annotated-with an emotional label for Valence and Arousal. When regarding Valence, patients with depression are the most negative ones and those with no illnesses the most positive. Patients with anxiety who are not also diagnosed with depression are in between, suggesting that they might be differentiated from depressed ones using Valence as a cue for illness detection. On the other hand, the differences in Arousal levels are not as meaningful when comparing depression and anxiety.

Figure 1: *Percentage of interventions per group.*

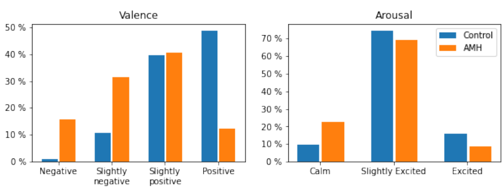


Figure 2: *Percentage of interventions per illness.*

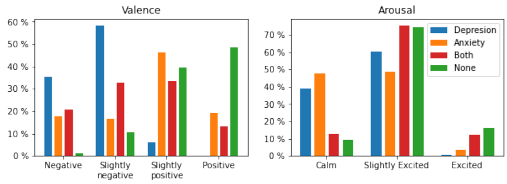


Table 6: *F1-Score on Valence and Arousal detection problems.*

	GeMAPS	HuBERT
Valence	0.35	0.46
Arousal	0.41	**0.57**

Finally, an experiment was conducted in order to predict Valence and Arousal associated to each intervention. Thus, a classifier was designed to predict among the 4 different Valence categories and the 3 statuses of Arousal. Once again the DNN architecture and training paradigm explained in Section 2 were implemented for the classification, but considering a random 90%-10% train-test split instead of the designed 8-folds Cross-Validation, due to a lack of emotion distribution across the folds. The achieved results are given in Table 6.

The task of identifying emotions is even more complicated since it involves a larger number of classes and the emotional annotation is a perception of the interviewer, what makes the task very subjective. Even so, we still draw the same conclusions as in Sections 4.1 and 4.2. Looking at the achieved results and by looking at Figures 1 and 2 it might be interesting to focus on Valence for future work, since it seems to provide more relevant information related to depression and anxiety than Arousal despite having worse results. Moreover, simplifying the Valence information into two different classes (positive and negative) might lead to more accurate results.

6. Conclusions

This manuscript provides a system capable of detecting depression and/or anxiety from speech signal uttered by potential patients in an interview. Our experiments show that acoustic features based on HuBERT transformer significantly outperform the classical GeMAPS extended set. Thus, an additional experiment was carried out in order to distinguish between anxiety and depression. Although the achieved results are not as impressive, HuBERT features still provide promising results. The text associated to transcriptions is also taken into account to build an alternative system that, although provides worse results, can be considered as an alternative information source. Finally, an analysis of emotional information associated to the interventions was conducted to study its potential use for future work showing that Valence might be an interesting marker.

7. Acknowledgements

This work was partially funded by the European Commission, grant number 823907 and the Spanish Ministry of Science under grant TIN2017-85854-C4-3-R.

8. References

[1] J. Arias-de la Torre, G. Vilagut, A. Ronaldson, A. Serrano-Blanco, V. Martín, M. Peters, J. M. Valderas, A. Dregan, and J. Alonso, "Prevalence and variability of current depressive disorder in 27 european countries: a population-based study," *The Lancet Public Health*, vol. 6, no. 10, pp. e729–e738, 2021.

[2] G. Archer, D. Kuh, M. Hotopf, M. Stafford, and M. Richards, "Association Between Lifetime Affective Symptoms and Premature Mortality," *JAMA Psychiatry*, vol. 77, no. 8, pp. 806–813, 08 2020.

[3] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, "A case-based reasoning model for depression based on three-electrode EEG data," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 383–392, 2020.

[4] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 09 2013.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[6] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2020.

[7] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: features and normalization," in *Proc. Interspeech 2011*, 2011, pp. 2997–3000.

[8] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor." in *ACM Multimedia*, A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462.

[9] M. F. Valstar, B. Schuller, K. Smith, T. R. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *AVEC '14*, 2014.

[10] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[11] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 12 2019.

[12] S. H. Dumpala, S. Rempel, K. Dikaios, M. Sajjadian, R. Uher, and S. Oore, "Estimating severity of depression from acoustic features and embeddings of natural speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7278–7282.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[22] Z. Callejas, K. B. Akhlaki, M. Noguera, M. I. Torres, and R. Justo, "MENHIR: mental health monitoring through interactive conversations," *Proces. del Leng. Natural*, vol. 63, pp. 139–142, 2019.

[23] R. Tennant, L. Hiller, R. Fishwick, S. Platt, S. Joseph, S. Weich, J. Parkinson, J. Secker, and S. Stewart-Brown, "The warwick-edinburgh mental well-being scale (wemwbs): Development and uk validation," *Health and Quality of Life Outcomes*, vol. 5, no. 1, 2007.

[24] I. Zubiaga and R. Justo, "Multimodal feature evaluation and fusion for emotional well-being monitorization," in *Pattern Recognition and Image Analysis: 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, May 4–6, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 242–254.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[26] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[27] "Bertforsequenceclassification," accessed: 2022-06-15. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

[28] "bhadresh-savani/bert-base-uncased-emotion," accessed: 2022-06-15. [Online]. Available: https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion

[29] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3687–3697.

[30] A. Kumar, P. Makhija, and A. Gupta, "Noisy text data: Achilles' heel of bert," 2020.

Publication 18

Speech emotion recognition in Spanish TV Debates

Speech emotion recognition in Spanish TV Debates

Irune Zubiaga, Raquel Justo, Mikel De Velasco, M. Inés Torres

Universidad del Pais Vasco UPV/EHU. Sarriena s/n. 48940 Leioa. Spain.

irune.zubiaga@ehu.eus, raquel.justo@ehu.eus, mikel.develasco@ehu.eus, manes.torres@ehu.eus

Abstract

Emotion recognition from speech is an active field of study that can help build more natural human–machine interaction systems. Even though the advancement of deep learning technology has brought improvements in this task, it is still a very challenging field. For instance, when considering real life scenarios, things such as tendency toward neutrality or the ambiguous definition of emotion can make labeling a difficult task causing the data-set to be severally imbalanced and not very representative.

In this work we considered a real life scenario to carry out a series of emotion classification experiments. Specifically, we worked with a labeled corpus consisting of a set of audios from Spanish TV debates and their respective transcriptions. First, an analysis of the emotional information within the corpus was conducted. Then different data representations were analyzed as to choose the best one for our task; Spectrograms and UniSpeech-SAT were used for audio representation and DistilBERT for text representation. As a final step, Multimodal Machine Learning was used with the aim of improving the obtained classification results by combining acoustic and textual information.

Index Terms: Acoustic Signal, Textual Information, Multimodal Machine Learning, Emotion Recognition

1. Introduction

The automatic detection of emotion from speech and language has gained popularity in recent years due to its capability to promote natural human-machine interaction, better comprehension of human interventions, etc. In order to be useful, the emotion detection systems need to work properly in real life scenarios, where emotions are not very extreme and only subtle expressions can be appreciated. Most of existing systems and approaches deal with emotions simulated by professional actors, leading to poor performances when trying to extrapolate to more realistic tasks.

Emotional responses result in changes in facial expression, in vocal expression, speaking style, in the way the language is used as well as in changes in physiological signals, such as the electroencephalographic signals (EEG) or galvanic skin responses, among others (GSR) [1]. The information provided by each signal can contribute to the selection of different features, which can be complementary. In this work, we focus on speech and language as information sources that can help in the automatic identification of emotions. Moreover, we will also explore whether the two sources can contribute together to a better system performance.

The six basic emotions defined by Eckman [2] (anger, surprise, disgust, enjoyment, fear, and sadness) can be represented by facial expressions that typically characterize these emotions [3]. However, spontaneous emotions that can be perceived from speech or language, are more varied and complex. Only a small set of complex and compound emotions [4] can be found in real scenarios [5, 6], and this subset is strongly dependent on the task. Therefore, a set of categories including the emotions that arise in each specific task has to be defined, according to perception experiments. However, this process is expensive and time consuming. In this work, we deal with a real life scenario; speech gathered from TV debates was considered to train an automatic emotion detection system.

For supervised learning, researchers need a ground truth to be used as a reference for automatic emotion identification. Usually, human annotators establish their own perception of the emotional data as the ground truth. So, in addition to being expensive and time consuming, these perceptual experiments also add subjectivity and complexity to the already complex and, to some extent, subjective emotional constructions, mainly in speech processing. In this work we carried out an annotation procedure based on crowdsourcing, that tries to gather the diversity from a bigger set of annotators [7].

As an alternative to working with categorical emotions, a number of researchers [8, 9] proposed a dimensional representation [10] of the emotional space. Thus, each affective state is represented by a point in a two-dimensional space, namely Valence and Arousal. This two dimensional model has been replaced by a three dimensional model, according to some authors work [11], including Dominance as a third dimension, to represent the complete range of human responses. This work employs both approaches to analyze emotional information.

The contribution of this work lies on the idea of using transformer-based representations for acoustic and textual information in a multimodal environment, in order to detect emotional information perceived in a real scenario. The achieved results show that multimodality is mainly helpful when considering Valence dimension and the categorical emotional information.

The manuscript is organized as follows: Section 2 deals with the specific task and corpus and the employed modelization of emotions. Section 3 describes the different features and methodologies employed to carry out the experiments and Section 4 summarizes the achieved results. Finally, Section 5 underlines extracted conclusions and future work.

2. Task and Corpus

In this work a set of human–human conversations was gathered from TV debates. Specifically, the Spanish TV program “La Sexta Noche” was used. In this weekly broadcast show, news about hot topics from the week are addressed by social and political debate panels led by two moderators. A very wide range of talk-show guests (politicians, journalists, etc.) analyze social topics from their perspectives. Given that the topics under discussion are usually controversial, emotionally rich interactions can be expected. However, the participants are used to speaking in public so they do not lose control of the situation. Thus,

 10.21437/IberSPEECH.2022-38

even if they might overreact sometimes, this is a real scenario, where emotions are subtle. The spontaneity in this situation is vastly different from scenarios with acted emotions, as shown in [15]. The selected programs were broadcast during the electoral campaign of the Spanish general elections in December 2015. Table 1 shows a small excerpt of a dialogue taken from the TV Debate corpus.

Table 1: *Emotionally rich excerpt from the corpus in which three talk-show guests debate about politics. The excerpt is shown in Spanish (the original language) and in English.*

	Spanish
Speaker 1:	Sí, sí, efectivamente, efectivamente, cuatro que optan a ganar estas elecciones
Speaker 2:	Por eso
Speaker 1:	Pero hay muchos más partidos
Speaker 3:	Van a ganar, yo creo que un tanto a dos
Speaker 1:	Bueno, están en un pañuelo
	English
Speaker 1:	Yes, yes, indeed, indeed, four who opt to win these elections
Speaker 2:	That is why
Speaker 1:	But there are many more parties
Speaker 2:	They are going to win, I think that one to two
Speaker 1:	Well, they are too close to call

The whole audio signal associated to an specific show, was separated according to the interventions of the speakers. This way, an audio file was achieved for each speaker intervention. The example of Table 1 would correspond to 5 different audio files, associated to Speaker 1, Speaker 2, Speaker 1, Speaker 3, Speaker 1.

In contrast with previous research [12] in which audio segments between 2 and 5 seconds were considered, in this work we used the full audio of each speaker intervention without slicing it since we considered this could be a more representative unit for emotional recognition.The audio files in which speakers could not be told apart and the ones that were not related to the debates (music, ads, etc.) were removed from the corpus.

The diarization and transcription were carried out manually within the framework of the Affective Multimedia Analytics with Inclusive and Natural Communication (AMIC) project [13]. The labeling was done using crowd annotation by 5 annotators. This procedure provided a set of 2964 audio files from 2 to 20 seconds long. Their respective transcriptions were also gathered. Said transcriptions have a length between 1 and 86 words, with a mean sentence length of 33 words and a mode of 37.

Regarding speaker features, the gender distribution was 24.8% female and 75.2% male, with a total of 88 speakers.

Table 2: *Number of audio files for each VAD category.*

			Audio n°
V	Negative	$(v \leqslant 0.4)$	669
	Neutral	$(0.4 < v < 0.6)$	1597
	Positive	$(v \geqslant 0.6)$	698
A	Neutral	$(a \leqslant 0.15)$	2113
	Excited	$(a > 0.15)$	851
D	Intimidated	$(d \leqslant 0.75)$	1533
	Dominant	$(d > 0.75)$	1431

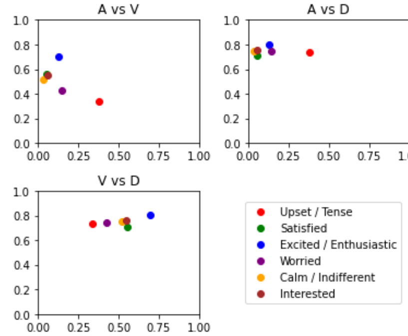


Figure 1: *Representation of the mean value of each emotion in the dimensional model.*

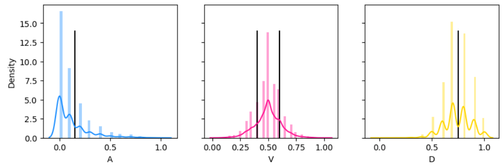


Figure 2: *Distribution of VAD values in the data-set*

.

2.1. Data-set for the VAD model

The Valence-Arousal-Dominance (VAD) model, also known as Pleasure-Arousal-Dominance (PAD), is a three-dimensional model that was introduced by Mehrabian and Russell in [14]. Mehrabian and Russell propose three independent dimensions for emotional representation; Valence (Pleasure), which ranges from displeasure to pleasure and expresses the pleasant or unpleasant feeling about something, Arousal, that ranges from nonarousal to arousal and represents the level of affective activation , and Dominance, which shows the level of control or influence on events and surroundings and goes from submissiveness to dominance.

To label Valence, Arousal and Dominance all 5 annotators were asked to answer the following set of questions for each intervention:

¿How do you perceive the speaker?

- Excited (1)
- Slightly excited (0.5)
- Neutral (0)

His emotional state is:

- Positive (1)
- Slightly positive (0.75)
- Neutral (0.5)
- Slightly negative (0.25)
- Negative (0)

¿How do you perceive the speaker in relation to the situation they are in?

- Rather dominant / Controlling the situation (1)

Table 3: *Number of audio files for each emotion.*

Emotion	**Audio nº**
Upset / Tense	361
Satisfied	221
Excited / Enthusiastic	27
Suprised	2
Worried	92
Calm / Indiferent	643
Bored	0
Interested	179
Total	1525

- Neither dominates the situation nor is intimidated (0.5)
- Rather cowed / Defensive (0)

These qualitative answers were encoded with the values in parentheses. Then, the mean of each set of labels (each set consisting of the answers of the 5 annotators for the intervention) was computed in order to have a single label for each intervention (ex. Labels of intervention 1 = positive, negative, neutral, neutral, negative = 1, 0, 0.5, 0.5, 0 = 0.4).

Our first approach was to carry out a set of regression experiments but the task was too complex and the obtained results were not satisfying. As a consequence, we decided to simplify the task by discretizing the data. We used Figure 2 as a guideline to choose the threshold values for each class, which are represented in the figure by black vertical lines. This way we were left with the classes shown in Table 2. We have **2964** samples for this task.

2.2. Data-set for the Categorical Emotion Model

To label categorical emotions annotators were to choose an emotion from Table 3 that, in their opinion, better suited the speakers state. Because of the perception of emotion being very ambiguous, for the emotion recognition task we only selected samples in which %60 of the annotators agreed in an emotional label with the goal of reducing noise in the data-set. The number of audio samples which belong to each emotion after applying this filter is shown in Table 3. As can be seen, there is not enough data regarding some of the classes for the model to learn a representation. Figure 1 presents the distribution of the emotions in our task within the dimensional emotional space, spanned by Valence, Arousal and Dominance. As seen there, when representing emotions in the VAD space some of them are difficult to tell apart.

Taking into account these facts we chose to try to discriminate between three different emotions: **Calm**, **Upset/Tense** and **Worried**. Even though *Excited/Enthusiastic* seems quite distinguishable from other emotions we did not work with it as a consequence of having very little data from this class (27 samples). The rest of samples were dismissed since merging them with the Calm class (the one they are closer to in the VAD space) would make the class imbalance even bigger than it already is (1:4:7). This way, we are left with **1096** samples for the emotion recognition task.

3. Experimental Setup

Both acoustic based and text based systems were built and trained with the aforementioned training corpus. In all of the cases 10 fold cross validation was used for validation.

3.1. Acoustic information

To analyze acoustic data Mel Spectrograms and the UniSpeech-SAT model were used.

3.1.1. Mel Spectrograms

The Mel Spectrogram is a spectrogram where the frequencies are converted to the Mel Scale [15], this being a perceptual scale of pitches judged by listeners to be equal in distance from one another. Mel Spectrograms have been proved to be a good audio representation for several tasks including emotion recognition [16].

Our first approach to carry out the emotion classification task was to use Mel Spectrogram representations of each audio file as an input to a Deep Convolutional Neural Network (DCNN). Zero padding was used on the spectrograms for them to have the same length, the achieved shape being 128x625. The network consisted of three convolutional layers with 3, 5 and 10 filters and three linear layers with 70, 40 and n neurons and ReLu as the activation function, n being the number of classes in each task. The model was trained for 300 epochs with Adam as the optimizer a learning rate of 1e-4 and a batch size of 16.

3.1.2. UniSpeech-SAT

Another outlook to deal with this task was to use speech representation models such as Wav2vec, Hubert, WavLM and UniSpeech-SAT. The best outcome was achieved when using the UniSpeech-SAT model architecture, specifically, when working with the *microsoft/unispeech-sat-large* [17] pre-trained model. This being the case, we will only focus in the results that were obtained with this setting. The Universal Speech Representation Learning with Speaker Aware Pre-Training model (UniSpeech-Sat) [18] performs specially well on speaker verification, speaker identification, and speaker diarization tasks. UniSpeech-SAT has been pre-trained on 16kHz sampled speech audio with utterance and speaker contrastive loss. The model is pre-trained on 94k hours of public English audio data; 60K hours of Libri-Light [19], 10K hours of GigaSpeech [20] and 24K hours of VoxPopuli [21].

To carry out the classification experiments, we froze the UniSpeech-SAT model and added a 1024 dimensional and a n dimensional linear layer to the last hidden layer, n being the number of classes we want to predict in each case. After that, the model was trained for 80 epochs using Adam as an optimizer and with a batch size of 8 and learning rate of 5e-5.

3.2. Textual information

To work with textual information the DistilBert [22] model was used. DistilBERT is a small, fast, cheap and light Transformer model. By leveraging knowledge distillation during the pre-training phase, the reduction of BERTs size by 40 % is achieved along with the model running 60% faster while preserving 97% of its language understanding capabilities.

We used the *CenIA/distillbert-base-spanish-uncased* pre-trained model. Said model is the *distilbert-base-uncased* model trained in *The Large Spanish Corpus* [23], which is a compilation of 15 unlabelled Spanish corpora spanning Wikipedia to European parliament notes.

Then the model was fine-tuned for two epochs using Adam as an optimizer with a batch size of 8 and a learning rate of 3e-5.

distilbert-base-uncased consists of 6 layers of transformers block with a hidden size of 768 and 12 self-attention heads and has a total of 66M trainable parameters.

3.3. Multimodal Machine Learning

Multimodal machine learning (MMML) is a multi-disciplinary research field that addresses some of the original goals of artificial intelligence by building models that can process and relate information from multiple modalities, including linguistic, acoustic and visual information. This approach outperforms single modal AI in many real-world problems [24] [25].

In this research we created a model that takes textual and acoustic information as an input with the aim of improving emotion classification results.The model architecture is shown in Figure 3. In this model we concatenated the logits from the audio model described in Section 3.1.2 and the text model described in Section 3.2 and used them as an input to a Neural Network that consisted of a n*2 dimensional and a n dimensional linear layer, n being the number of classes in each task.

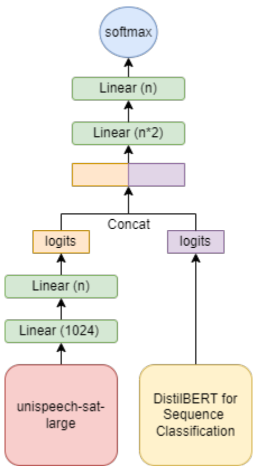


Figure 3: *MMML model architecture.*

4. Results

The obtained results are shown in Table 4.

The first observation when focusing on audio models is that UniSpeech-SAT outperforms the DCNN with Mel Spectrograms as an input. For this reason, from now on we will only compare UniSpeech-SAT (we will call it the audio model from now on) and the text based model.

The audio model outperforms the text based model when focusing on the categorical emotion and arousal while having very similar results in valence and dominance classification.

Looking more closely into each emotional feature, in the categorical emotion classification task the best results are the ones obtained with the audio model. Even in that case, the results are not very good. One of the facts that may cause this is that while the model has a high classification performance when regarding *Calm* and *Upset/Tense*, it performs very poorly when focusing on *Worried*. For example,in the case of the audio model, the obtained F-scores for *Calm* and *Upset/Tense* are respectively **0.78** and **0.81** while the F-score for *Worried* is as low as **0.19**. This makes sense considering that *Worried* is the minority class with only 92 samples vs the 643 and 361 that *Calm* and *Upset/Tense* have, which makes it hard to learn a representation for this class. However when using MML there is an improvement in the classification.

When considering Valence, results are similar when working with text and audio, being, in both cases, pretty low. It is the feature of the VAD with the lowest F-score values, which might be related to having to predict three labels instead of two (as is the case for Arousal and Dominance) since this reduces the quantity of training data for each class. Valence has also shown to generally perform worse in audio centered researches than arousal [26]. However, when using MMML we can see a very slight improvement. The results in [27] show that even the best-performing HuBERT representation under-performs on Valence prediction compared to a multimodal model that also incorporates text representation. The results in [26] also show that, while Valence is hard to detect in audio, text based features do add to the accuracy of prediction of Valence for speech stimuli. Our results might be in line with these observations and it might be interesting to further look into it.

Arousal is the feature that has the highest prediction accuracy. This can also be seen in other researches considering VAD [26]. This result is achieved when working with audio, which is in line with the results in [28] that show that Valence is better estimated using semantic features while Arousal is better estimated using acoustic features.

Regarding Dominance the achieved results are very similar when working with all the tested models. This might be a consequence of not having enough data as to learn a representation.

Table 4: *Classification results F-score.*

	E	V	A	D
Spectrograms	0.49	0.36	0.63	**0.57**
UniSpeech-SAT	0.59	0.46	**0.73**	**0.57**
DistilBERT	0.52	0.46	0.59	0.56
MMML	**0.61**	**0.47**	0.70	0.56

5. Conclusions and Future Work

We can remark the value of MMML models for categorical emotion recognition in our task. It would also be interesting to further analyze their use for Valence classification.

We observed that Valence is the best recognized dimension when using textual information, while Arousal has better outcomes when working with audio. This might make sense considering that positive or negative feelings about something are easier to detect in semantics than Arousal or Dominance [29] which might be features that are more related to acoustics.

For future work it would be interesting to explore other classification architectures and label more data to improve the results and make it possible to learn representations for more classes. For example, it would be interesting to have more data of the *Enthusiastic* class, since, as seen in Figure 1 and in [12], it is quite distinguishable from other emotions in our corpus.

6. Acknowledgements

The research presented in this paper was conducted as part of the AMIC PdC project, which received funding from the Spanish Minister of Science under grants TIN2017-85854-C4-3-R, PID2021-126061OB-C42 and PDC2021-120846-C43 and it was also partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 823907 (MENHIR).

7. References

[1] A. Raheel, M. Majid, M. Alnowami, and S. M. Anwar, “Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia,” *Sensors*, vol. 20, no. 14, p. 4037, 2020.

[2] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2004.

[3] M. A. Nasri, M. A. Hmani, A. Mtibaa, D. Petrovska-Delacrétaz, M. B. Slima, and A. B. Hamida, “Face emotion recognition from static image based on convolution neural networks,” in *5th International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020, Sousse, Tunisia, September 2-5, 2020*. IEEE, 2020, pp. 1–6.

[4] K. R. Scherer, *Approaches To Emotion. Chapter: On the nature and function of emotion: A component process approach*. K. R. Scherer & P. Ekman. Taylor and Francis Group, 1984.

[5] M. deVelasco, R. Justo, A. López-Zorrilla, and M. Torres, “Can spontaneous emotions be detected from speech on tv political debates?” in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*, Naples, 2019.

[6] M. deVelasco, R. Justo, A. López-Zorrilla, and M. I. Torres, “Automatic analysis of emotions from speech in spanish tv debates,” *Acta Polytechnica Hungarica (In Press)*, vol. 19, pp. 149–171, 2022.

[7] R. Justo, M. I. Torres, and J. M. Alcaide, “Measuring the quality of annotations for a subjective crowdsourcing task,” in *Pattern Recognition and Image Analysis*, L. A. Alexandre, J. Salvador Sánchez, and J. M. F. Rodrigues, Eds. Springer International Publishing, 2017, pp. 58–68.

[8] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” *International Journal of Synthetic Emotions, IJSE*, pp. 68–99, 2010.

[9] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[10] J. Russell, “A circumplex model of affect,” *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] I. Bakker, T. Van der Voordt, J. Boon, and P. Vink, “Pleasure, arousal, dominance: Mehrabian and russell revisited,” *Current Psychology*, vol. 33, pp. 405–421, 10 2014.

[12] M. de Velasco, R. Justo, and M. Inés Torres, “Automatic identification of emotional information in spanish tv debates and human-machine interactions,” *Applied Sciences*, vol. 12, no. 4, 2022.

[13] A. Ortega, E. Lleida, R. S. Segundo, J. Ferreiros, L. F. Hurtado, E. S. Arnal, M. I. Torres, and R. Justo, “Amic: Affective multimedia analytics with inclusive and natural communication.” *Proces. del Leng. Natural*, vol. 61, pp. 147–150, 2018.

[14] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of Research in Personality*, vol. 11, pp. 273–294, 1977.

[15] S. S. Stevens, J. E. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[16] K. Venkataramanan and H. R. Rajamohan, “Emotion recognition from speech,” *arXiv preprint arXiv:1912.10458*, 2019.

[17] “Unispeechsatlarge,” accessed: 2022-08-10. [Online]. Available: https://huggingface.co/microsoft/unispeechsatlarge

[18] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022.

[19] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-light: A benchmark for asr with limited or no supervision,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[20] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, “Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio,” *Interspeech 2021*, Aug 2021.

[21] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, “Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *ArXiv*, vol. abs/1910.01108, 2019.

[23] “Datasets: large_spanish_corpus,” accessed: 2022-08-10. [Online]. Available: https://huggingface.co/datasets/large_spanish_corpus

[24] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, “Multimodal deep learning models for early detection of alzheimer’s disease stage,” *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[25] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, “Emonets: Multimodal deep learning approaches for emotion recognition in video,” *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[26] M. Asgari, G. Kiss, J. van Santen, I. Shafran, and X. Song, “Automatic measurement of affective valence and arousal in speech,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 965–969.

[27] S. Srinivasan, Z. Huang, and K. Kirchhoff, “Representation learning through cross-modal conditional teacher-student training for speech emotion recognition,” 2021.

[28] S. G. Karadoğan and J. Larsen, “Combining semantic and acoustic features for valence and arousal recognition in speech,” in *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, 2012, pp. 1–6.

[29] Z. Yao, X. ru Zhu, and W. Luo, “Valence makes a stronger contribution than arousal to affective priming,” *PeerJ*, vol. 7, 2019.

Publication 19

Analysis of deep learning-based decision making in an emotional spontaneous speech task

applied sciences

MDPI

Article

Analysis of Deep Learning-Based Decision-Making in an Emotional Spontaneous Speech Task

Mikel de Velasco , Raquel Justo *, Asier López Zorrilla and María Inés Torres

Department of Electricity and Electronics, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain; mikel.develasco@ehu.eus (M.d.V.); asier.lopezz@ehu.eus (A.L.Z.); manes.torres@ehu.eus (M.I.T.)

* Correspondence: raquel.justo@ehu.eus

Abstract: In this work, we present an approach to understand the computational methods and decision-making involved in the identification of emotions in spontaneous speech. The selected task consists of Spanish TV debates, which entail a high level of complexity as well as additional subjectivity in the human perception-based annotation procedure. A simple convolutional neural model is proposed, and its behaviour is analysed to explain its decision-making. The proposed model slightly outperforms commonly used CNN architectures such as VGG16, while being much lighter. Internal layer-by-layer transformations of the input spectrogram are visualised and analysed. Finally, a class model visualisation is proposed as a simple interpretation approach whose usefulness is assessed in the work.

Keywords: emotion detection; speech processing; explainable artificial intelligence; machine learning

Citation: de Velasco, M.; Justo, R.; López Zorrilla, A.; Torres, M.I. Analysis of Deep Learning-Based Decision-Making in an Emotional Spontaneous Speech Task. *Appl. Sci.* **2023**, *13*, 980. https://doi.org/10.3390/app13020980

Academic Editor: Douglas O'Shaughnessy

Received: 30 November 2022
Revised: 3 January 2023
Accepted: 8 January 2023
Published: 11 January 2023

1. Introduction

Emotion theories agree that an emotional episode consists of several components, such as the stimulus, motivation for action, central and peripheral physiological responses, behaviour (e.g., facial and vocal expressions, among others) and subjective experiences or feelings [1]. In addition, the behavioural and physiological expression of emotions and the stimulus quality depend on the person and on the specific scenario [2].

Affective computing often uses a categorical model based on a set of predefined emotional labels that are roughly supported by the basic emotions defined by the affect program theory [3], which might cover the whole emotional space. Each basic emotion encompasses a wide subset of emotions that can be understood as blends or elaborations of the basic ones. An alternative theory [1] discriminates emotions on the basis of combinations of continuous variables aiming to characterise the contents of feelings [4]. Typical variables are Valence and Arousal, which define a 2D space for representations, even though Dominance has also been proposed, resulting in a 3D model usually called a VAD model.

Furthermore, emotional responses result in changes in gaze, facial and vocal expressions, speaking style, in the way the language is used as well as in changes in physiological signals, such as the electroencephalographic signals or galvanic skin responses, among others [5]. The information provided by each signal has distinctive features, which can be complementary. This might result in a variety of approaches and systems for emotion recognition with different goals and application tasks [6]. However, universal facial expressions [7], also considered as short-term stereotypical responses, are the more extensively analysed emotional expressions.

Speech signals encode speaking styles, paralinguistic features, the usage of language, message contents to be transmitted, environmental sounds, etc., which contain varied information about the speaker profile, intent, current emotional status, and even information about some mental diseases [8].

In contrast to this complexity, computational researchers of emotions need an exact ground truth to be used for supervised learning, decoding and evaluating computational models of emotions. Usually, human annotators establish their own perception of the emotional data as the ground truth and reference for the automatic identification of emotions. These perceptual experiments add subjectivity and complexity to the already complex and, to some extent, subjective emotional constructions, mainly in speech processing.

In the next step, researchers submit the data to black boxes, i.e., to complex architectures of neural networks, whose behaviour is not fully understood but that might perform well in terms of usual scores. Therefore, the key to successful or unsuccessful classification rates remains unknown. In other words, we are unaware of what the computational model identifies as emotional cues.

Over the last few years, some techniques have been proposed to explain the internal behaviour of complex computational models, resulting in what is called XAI, i.e., eXplainable Artificial Intelligence. Some of them propose simple models that can represent the aforementioned external behaviour. However, the most amount of effort has been put into the image analysis domain because the action of the network on the original images can be visually represented and, thus, it can be more easily understood. In contrast, XAI methodologies have been scarcely used in voice processing.

The aim of our work is to contribute to the understanding of computational methods and decision-making involved in the identification of emotions in spontaneous speech. To this end, we selected a task consisting of TV debates in which spontaneous emotions can be investigated. To this end, we follow the transformation of the input data, from layer to layer, until the classification is carried out in the output layer of the network. If the whole architecture becomes too deep, this process is hard and it becomes difficult to extract valuable conclusions. Thus, we propose a CNN-based deep architecture capable of providing good results but simple enough to be able to follow and interpret the decisions taken.

The main contributions of the work can be summarized as follows:

- We develop a multitask architecture to simultaneously classify discrete categories and VAD dimensions, in the aforementioned realistic task. This requires a previous annotation of the corpus in terms of both categorical and VAD models through human perception experiments, which define the ground truth. The proposed model is also compared to a more complex state-of-the-art image processing network such as VGG-16 [9], resulting in a better performance (even if slightly) for the target task.
- In an attempt to explain the decisions of our automatic system, we analyse the evolution of the categorical representations of our model layer-by-layer. Thus, we analyse the evolution of the data until they become predictions, i.e., from input spectrograms to the results.
- As a final contribution, we use the spectrogram to parameterise the voice signal to process it as an image. This allows us to obtain a visual class model [10] (deep dream) that can be used to visualise the patterns learnt by the proposed network. This technique is widely used when dealing with images, but as far as we know, it has never been applied to speech.

The paper is organised as follows: Section 2 reports some related works. Section 3 describes the methodology selected to develop an automatic recogniser of emotions from spontaneous speech. This section includes the description of the task and corpus, the neural network model proposed for the joint classification of categories and emotional dimensions, and also a comparison of classification results obtained with our network and a pretrained VGG-16 net. Then, Section 4 deals with the interpretation of the model behaviour. It first presents a joint analysis of the results in terms of both categories and dimensions. Then, the evolution of the model across the layers is visualised and examined. Finally, the proposed simple interpretation model, i.e., the class visualisation model is introduced and assessed. Finally, Section 5 reports the main conclusions of this work.

2. Related Work

Affective computing has become more relevant due to its impact on person–computer interactions [11,12]. This has translated into significant progress in all its modalities [13–15]: face [7,16], gestures [17], text [18–21], audio [22] and others. Some investigations do not only focus on a single modality but also multi-modal approaches [23].

With regard to the detection of emotion from facial cues, most studies deal with the categorical model for emotional state representation. Within this framework, the most employed set of emotions is the one proposed by Ekman [3], which is widely accepted under the name of "The Big Six" [24]. Ekman's proposal consists of six basic and universal emotions: surprise, disgust, sadness, anger, fear, and happiness. However, in other works dealing with speech, emotions are also represented using a dimensional model [25–27]. Dimensional theories postulate that the vast array of emotions cannot be simplified to a basic set but can be mapped to a continuous low-dimensional spatial representation. Most of the works in this context propose a two-dimensional model, comprised of valence (whether the emotional state is positive/pleasant or otherwise negative/aversive) and arousal (intensity or level of arousal) [28,29]. Dominance is a third dimension also included in some works that encodes the level of control (leading to feelings of power/dominance or weakness/submission) [30]. The aforementioned two models (categorical and dimensional) show a close relationship according to the Core Affect theory [4], where each categorical emotion is represented in a point/area of the dimensional model. An example of this is the illustration of Sherer's circumplex [31], which makes use of the arousal/valence two-dimensional model to represent categorical emotions. In this work, the two models were considered to represent the emotional status of the speakers because they can complement each other.

In order to build an emotion detection system from scratch, annotated data are needed, assuming the supervised machine learning paradigm. Finding corpora where real emotions appear is a really difficult task. Thus, most of the research in this field relies on data sets where emotions have been acted or forced [26], as occurs with the EMODB [32] or IEMOCAP [33] corpora. However, in recent years, there has been an attempt to put emphasis on creating corpora with spontaneous emotions such as AVEC2012 [34], EmoL6N [35] or DBATES [36]. However, this is a challenging task because, on the one hand, the perception of emotions is not as intense as in the corpora with acted emotions [35] and, on the other hand, the annotation procedure is very subjective, leading to low inter-annotator agreements [37–39]. This work deals with a task in which spontaneous emotions are involved. This entails an additional challenge for the system that has to deal with speech chunks with subjective and subtle emotional representations.

Another important issue to be addressed is how to identify the most suitable features, i.e., speech representations, for detecting emotions. In recent years, there have been several attempts to build a set of features suitable for the identification of emotions in the speech signal [40,41]. Several works are based on Low-Level Descriptors [42–46], whose characteristics are related to prosody (pitch, formants, energy, jitter and shimmer), the spectrum (centroid, flux, entropy) and their functionals (mean, std, quartiles 1–3, delta, etc.). In this context, Ref. [47] proposes the GeMAPS set of speech features that has been considered as a standard. However, and thanks to challenges such as INTERSPEECH [48], other sets have also been proposed (ComParE) to become a reference in this area. However, none of these sets has actually proven to be superior to the rest in a global environment. Several works [43,46,49] suggest that there are no universal acoustic features that extract the emotional content and work well in all contexts. In this direction, some works propose working with the spectrogram [45,46,50–54] since it contains almost all the information about a speech signal. More recently, self-learning based approaches [55,56] have shown to find good representations of the speech signal. Indeed, self-learning has been applied in a variety of applications of speech processing with successful results in solving different tasks [56–59]. In fact, our task has already been addressed using such speech representations [2]. However, pretrained (and not fine-tuned) models were needed to obtain good results, and thus the analysis of the decisions made by such models cannot be easily conducted.

The understanding of which patterns are detected by deep neural networks, or how they work, can help to design new architectures or new learning paradigms that can make a difference. The introduction of such advances has been decisive to achieve the current automatic systems' performance. One of the clearest examples appears when the field of computer vision introduces convolutional [60] and pooling [61] networks, which marked the beginning of a new age. In 1997, the recurrent networks based on LSTM cells were proposed [62], which are known for their ability to process long sequences that were first used in NLP or speech processing. However, attention networks [63,64] have been the ones that have made progress in the NLP field. In emotion recognition, although some work has been conducted [2,65,66], promising results have not yet been achieved. Different works based on CNNs [67], LSTMs [68,69] and attention mechanisms [67,69] have reached accuracy values (or F1 scores) of around 0.7 with the most commonly employed acted data sets.

Moreover, this kind of deep neural architecture is sometimes so complex that even experts are hardly able to interpret it [70]. As a consequence, understanding the behaviour of a model to make predictions is becoming as important as its accuracy. In fact, interpretability is nowadays a key to improving the performance of complex neural architectures. Several methodologies have recently been proposed to explain the importance of particular features for decision-making. These methodologies are sometimes integrated into the models, but they also very often consist of postprocessing analysis and models [71]. On the other hand, explanation and interpretation are context-, domain- and task-dependent concepts. In this way, XAI targets are the end-users who depend on the decisions taken by the automatic system [72]. Some recent works also argue that explanations must be related to the perceptual process from cognitive psychology [73]. In brief, XAI is still a domain to be explored by AI researchers in relation to the domain addressed. Due to the intuition of vision and the availability of data, much of the XAI research has focused on image prediction tasks [10]. On the contrary, few techniques have been developed for audio or speech prediction [6,74,75]. In this work, an architecture based on CNNs and inspired by computer vision was designed. Moreover, employing the spectrogram as the input allows us to represent the audio as an image and apply XAI image processing techniques.

In summary, we propose an emotion detection system capable of providing two emotion representation levels: a categorical one and another one based on three-dimensional VAD space. The proposed model is simple enough to allow a detailed analysis of the network behaviour layer by layer while providing accurate classification results. In order to increase the level of explainability of the decisions made by the network, the Visual Model Classification XAI technique was selected. To this end, the spectrogram was selected as the input of the system along with a CNN-based deep neural architecture.

3. Emotion Detection

In this section, we describe the selected neural model capable of detecting emotions in TV debates. This model is inspired by previous works [2,39], but it has been adapted for the joint classification in terms of both categories and emotional dimensions.

3.1. Task and Corpus

This task consists of human–human spontaneous conversations gathered from the La Sexta Noche Spanish TV program. This TV show addresses the hot news of the week in social and political debate panels led by two moderators. A very wide range of talk-show guests (politicians, journalists, etc.) analyse social topics from their perspectives. Given that the topics under discussion are usually controversial, it is expected to have emotionally rich interactions. However, the participants are used to speaking in public so they mostly do not lose control of the situation. Nevertheless, even if participants might overreact sometimes, it is a real scenario in which emotions are subtle.

In order to build a corpus, the programs of La Sexta Noche broadcasted during the electoral campaign of the Spanish general elections in December 2015 were selected. Then, speech signals were extracted from the videos of the TV shows. Then, they were split into shorter segments or chunks. The segments have to be short enough to avoid changes in emotional content but long enough to allow for their identification. Thus, the speech signal was divided into clauses. A clause was defined as "a sequence of words grouped together on semantic or functional basis" [76], and it can be hypothesised that the emotional state does not change inside a clause. An algorithm that considered silences and pauses, as well as text transcriptions, was designed to identify the utterances compatible with the clauses [39]. This produced a set of 5500 audio chunks in Spanish, ranging from two to five seconds long that was used as our data set. Regarding the speaker features, the resulting gender distribution in the processed data was 30% female and 70% male, with a total number of 238 different speakers and an age ranging from 35 to 65. These data just reflect the nature of the described TV shows.

The corpus was emotionally annotated in the framework of the AMIC "Affective multimedia analytics with inclusive and natural communication" project [77], as described in [2,39]. The annotation was carried out through perception experiments in which crowd annotators were asked to identify both emotional categories and Valence–Arousal–Dominance dimensions [78,79]. A crowdsourcing platform [38] was used to gather five annotations for each audio chunk. All the annotators filled out a questionnaire related to the perceived emotions in each audio chunk, and an agreement higher than 60% was required as a quality guarantee for the categorical annotation [35,39]. The questionnaire related to the dimensional model considered discrete labels to facilitate the crowd annotation process. However, instead of asking a 60% of agreement again, a consensus of the different annotations was achieved by converting the labels to real values, as shown below:

- Valence: Positive = 1, Neutral = 0.5, Negative = 0;
- Arousal: Excited = 1, Slightly excited = 0.5, Neutral = 0;
- Dominance: Rather Dominant = 1, Neutral = 0.5, Rather intimidated = 0.

Then, the average values attached to each audio chunk by different annotators were computed. In this way, each label of the VAD model corresponds to a real value. This scenario suggests a classification problem for the categorical model and a regression problem for the VAD. However, previous works [39] showed that the regression problem might be too ambitious for this task, and if it is addressed as a classification task, a better performance might be achieved. Specifically, the distribution of the annotated data for each of the VAD dimensions was analysed. According to these distributions, a discretisation of each VAD dimension was carried out in order to learn a categorical classifier to predict each of the discretised classes. This procedure led to a set of 4118 annotated chunks distributed, as shown in Table 1. Let us note that Table 1 shows a high imbalance between classes. The tendency to neutrality that is observed is related to the spontaneity conditions in which the corpus was acquired.

Table 1. Class distribution of the annotated data for categorical and VAD model.

Categorical Model (%)	Dimensional Model		
	Arousal (%)	**Valence (%)**	**Dominace (%)**
Angry: 30.2 Happy: 15.3 Calm: 54.5	Excited: 25.5 Neutral: 74.5	Positive: 29.0 Neutral: 54.4 Negative: 16.6	Dominant: 26.2 Neutral: 73.8

3.2. Convolutional Neural Model

Convolutional neural architectures have become a standard in image processing over the last few years. However, other types of tasks have also taken advantage of these architectures by adapting the problem and addressing it with computer vision techniques. For example, audio analysis can be performed with computer vision techniques if the audio is represented by a spectrogram. In addition, other areas such as speech and natural language processing have also taken advantage of the potential of convolutions for the analysis of temporal sequences [80].

In this work, we propose a simple and light convolutional network architecture (a network with 43K parameters) and compare it with the VGG16 convolutional network [9], a model widely used in the literature but that consists of 134M parameters, which makes it difficult to understand its behaviour.

Both network architectures are designed to obtain, from the speech spectrogram, the joint classification of emotional state in terms of both representations, the categorical model and VAD dimensions, as shown in Figure 1. On the one hand, after a number of convolutional and pooling layers are applied, three scalar values corresponding to each dimension of the VAD model are computed through three linear layers with a sigmoid activation function at point A. These values are then converted to discrete VAD predictions linearly, at point B. On the other hand, the categorical model is inferred in two ways. First, at point C, the categories are predicted based on the output of the CNNs. Second, at point D, the scalar predictions of the VAD model are used instead. Intuitively, this second prediction might perform better as it could explicitly take advantage of the multi-tasking capabilities of the networks.

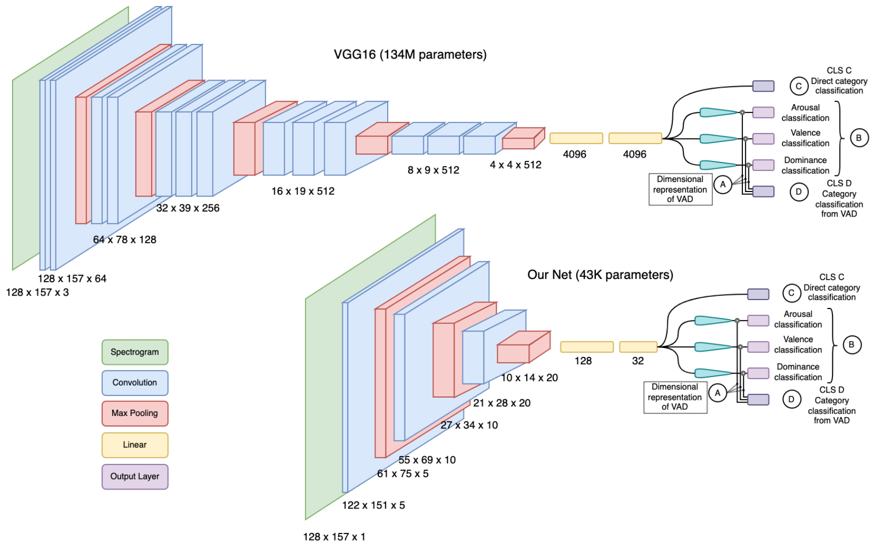


Figure 1. Representation of the structure of the VGG16 [9] network and our proposal. In both cases, the convolutions (blue boxes) and max-poolings (red boxes) help to extract a set of appropriate features, while the linear ones are in charge of performing the class discrimination. Point A represents a dimensional prediction of the VAD model, while point B provides the discretisation of each of the dimensions. On the other hand, the categorical model is inferred directly in point C, and from the scalar VAD model predictions, in point D.

As for the training procedure, a resampling strategy is used when training both networks to deal with the imbalance of the data, as observed in Table 1. The employed method consisted of selecting the samples inside a batch using a random function that can take into account the weight given to each sample. The weight W_x for each sample x was computed, as Equation (1) shows

$$W_x = min\left\{ \frac{|X|}{|X_c|}, \beta \cdot \min_{\forall c \in C}\left\{ \frac{|X|}{|X_c|} \right\} \right\} \tag{1}$$

where $|X|$ is the number of samples in the corpus, $|X_c|$ is the number of samples in the class which sample x belongs to, and β is the oversampling coefficient (in this work it was chosen a value $\beta = 2$). In this way, the samples of the minority classes appear proportionally more times, but never more than twice as much as a sample from the majority class.

Regarding the optimisation hyperparameters, the Adam optimiser was chosen with a learning rate of 10^{-4} and with a batch size of 16. The models were trained throughout 7K iterations. Note that our network is trained from scratch, whereas VGG16 is fine-tuned from the publicly available pretrained checkpoint. The cross-entropy loss function is used for each classification task. The five losses (two for the categorical model, and three for the VAD model) are then averaged (with different weights) to obtain the final loss. The weights for the categorical losses were half the weights for the VAD model, since this led to the best results, empirically.

3.3. Classification Results

The performance of the proposed classification system for both emotion categories and VAD dimensions are compared to the VGG16 model in Tables 2 and 3. All results were obtained after a 10-fold cross-validation procedure. For the categorical model, two different results are given, one corresponding to the direct categorical classification associated with output B in Figure 1 (CLS C) and another one making use of the predicted VAD floating point values associated with output D (CLS D). The average and standard deviation of five metrics commonly employed to evaluate emotion classification systems [81,82] are reported: F1 score, Unweighted Accuracy (UA, also known as balanced accuracy or unweighted average recall), average precision, Matthews Correlation Coefficient and Area Under the ROC Curve (AUC). Additionally, paired *t*-tests were computed to assess the statistical significance of the performance differences between our model and the VGG16 network.

Table 2. Classification performance of our proposal and VGG16 for the categorical model prediction task. Two ways of predicting the emotion categories were tested. Comparisons where a network significantly outperforms the other (i.e., p-value < 0.05) are marked with the symbol *. The best result for each comparison is highlighted in bold.

Our Net/VGG16	CLS C	CLS D
F1	**0.58** ± **0.04**/0.57 ± 0.06	**0.59** ± **0.05**/0.57 ± 0.05
UA	**0.57** ± **0.04**/0.54 ± 0.06	**0.58** ± **0.04**/0.54 ± 0.05
Average precision	0.60 ± 0.05/**0.63** ± **0.07**	0.60 ± 0.06/**0.63** ± **0.07**
Matthews corr. coef.	**0.39** ± **0.06**/0.38 ± 0.08	**0.41** ± **0.05**/0.38 ± 0.06
AUC	**0.80 *** ± **0.03**/0.75 ± 0.03	**0.81 *** ± **0.02**/0.74 ± 0.04

Table 3. Classification performance of our proposal and VGG16 for the VAD model prediction task. Comparisons where a network significantly outperforms the other (i.e., p-value < 0.05) are marked with the symbol *. The best result for each comparison is highlighted in bold.

Our Net/VGG16	Arousal	Valence	Dominance
F1	**0.67** ± **0.11**/0.67 ± 0.02	**0.42** ± **0.03**/0.41 ± 0.05	**0.57** ± **0.05**/0.56 ± 0.03
UA	**0.67** ± **0.09**/0.66 ± 0.03	**0.45 *** ± **0.04**/0.41 ± 0.03	**0.57** ± **0.03**/0.56 ± 0.03
Average precision	0.69 ± 0.13/**0.69** ± **0.02**	0.44 ± 0.05/**0.45** ± **0.09**	0.58 ± 0.07/**0.58** ± **0.04**
Matthews corr. coef.	0.35 ± 0.17/**0.35** ± **0.03**	**0.14** ± **0.04**/0.12 ± 0.04	**0.15** ± **0.06**/0.14 ± 0.06
AUC	**0.74** ± **0.11**/0.72 ± 0.02	**0.65 *** ± **0.03**/0.60 ± 0.02	**0.63** ± **0.07**/0.60 ± 0.03

First, and most importantly, our network performs slightly better than the fine-tuned VGG16 network for most classification tasks and metrics. This is already remarkable because our architecture uses around 3000 times fewer parameters than the VGG16 CNN. Furthermore, the differences in performance are statistically significant in four comparisons: when measuring the AUC for CLS C, CLS D and arousal, and also for the Unweighted Accuracy in the arousal prediction task. Importantly, these results support the use of our light network to apply XAI techniques.

If we further analyse the results, Table 2 shows a slight tendency towards a better performance of CLS D, i.e., the classifier that uses the predicted VAD values before classifying the categories, particularly in the case of our proposed network. Looking at the results for the VAD dimensional classification of Table 3, it can be concluded that the best results are achieved for Arousal. However, it should be noticed that, in this case, there are only two different categories (Excited and Neutral), whereas, in Valence, there are three different ones (Positive, Neutral and Negative).

For a better understanding of the VAD results, Figure 2 shows the VAD predicted values vs. the annotated, i.e., perceived, values. Straight lines in the figure show the borderlines learnt to discretise the problem, i.e., to transform the regression problem into a categorisation one. This figure shows the good performance of our proposal. In fact, opposite diagonals are very low-density regions. When it comes to Arousal, for instance, it seems to be easier to predict accurately higher values than lower ones that are more scattered in the lower part of Figure 2. In Valence, it can be concluded again that positive and negative categories are sometimes mixed with Neutral, but rarely among each other (see secondary diagonal in the figure). Finally, Dominance shows a lower correlation between the predicted and annotated values.

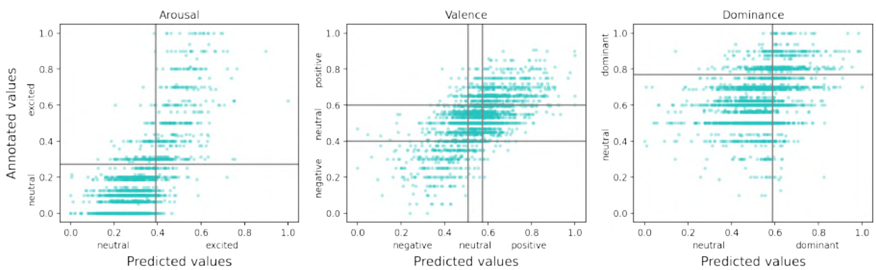


Figure 2. Comparison between the samples annotated by humans and predicted by the network in the dimensional model. Each of the plots depicts a VAD dimension. Each sample is placed at a point based on the actual value of the annotations (y-axis) and the actual value of the predictions of the network (x-axis). The lines show how the samples have been separated for the classification problem on each dimension (y-axis) and how the network has separated them (x-axis).

4. Interpreting the Model Behaviour

This section aims to explain the proposed model's decisions, providing a better understanding of the work of each layer, as well as an analysis of the input features learnt by the model.

4.1. Analysis of the Classification Results

Figure 3 shows the three projections of the VAD values in a 2D space. The colours of the points show the category they belong to. In the first row, both the VAD points and the categories are the ones perceived by annotators. In the second row, the VAD points are the ones predicted by the network (Point A) and coloured according to the perceived categories. Finally, in the third row, the colour of the points corresponds to the predicted categories. Specifically, the output of the network at Point D (CLS D) has been used, since it outperforms CLS C in our experiments.

The first row of Figure 3 shows mixed VAD points. In fact, the subjectivity of human perception witnessed during the annotation procedure makes it difficult to obtain clear boundaries between classes. However, some patterns can still be observed. For example, “Angry” samples present higher arousal than “Calm” and “Happy”, which can be seen in the first and second plots. In terms of valence, the distinction between the three categories is a bit more clear: “Happy” is the most positive emotion, followed by “Calm” which is neutral, and “Angry”, which indicates negative arousal. This result is clearly aligned with the literature on emotion theory. Finally, we would like to mention that the dominance axis does not show clear boundaries between the categories.

Figure 3. Correlation between the VAD and categorical models, according to both the annotated and predicted data.

A transformation of the space is observed in rows two and three; points are no longer located in the same place as in the first row of Figure 3. Instead, they correspond to the model's predictions. The difference between the second and third rows is that in the third one we can clearly see the boundaries learnt by the network to decide to which category a sample belongs to. In this new space, the categories can be better separated, even the annotated ones in the second row. It can be hypothesised that this fact is due to the joint training of the categories and VAD dimensions. The simultaneous VAD and categorical classifications result in the collaboration of both models in decision-making. Thus, the regions associated with categories are well-defined in the VAD projection space. Finally, similar relations of the categories and the VAD axes can be seen in the last two rows. Moreover, in this case, the dominance axis shows that, as expected, "Calm" is less dominant than "Happy" and "Angry".

4.2. Evolution of the Model

In this section, we show a representation of the work that each layer of the deep network is carrying out. To this end, the progress in the training stage can be explained as a fine-tuning process of the data representation, which can lead to a good classification. For this purpose, we present the output provided by each layer for each training sample in a bidimensional space, by applying a dimensionality reduction method such as PCA. Assuming that $X = \{x_1, x_2, \ldots, x_n\}$ is the set of training samples, where x_i is the spectrogram associated with each audio chunk presented in Section 3.1, the output of the first convolutional layer for each sample can be defined as $y_i = conv1(x_i)$. This y_i output is transformed into a flattened vector that can be visualised in a 2D space by applying a decomposition in Principal Components as $y_i' = (z_1, z_2)$, where z_1 and z_2 are the two first principal components in the PCA analysis of y_i. This representation can be replicated for the output for each convolutional network (conv1, conv2 and conv3) and also for the two dense layers: linear1 and linear2.

The visualisation of the aforementioned representations is displayed in Figure 4. Points in the picture stand for the decomposition in Principal Components of the outputs of each convolutional layer. The colour of each point represents the category it belongs to. In the first row, the colours of the categories correspond to the annotated labels (ground truth), whereas in the second row the colours represent the predicted categories.

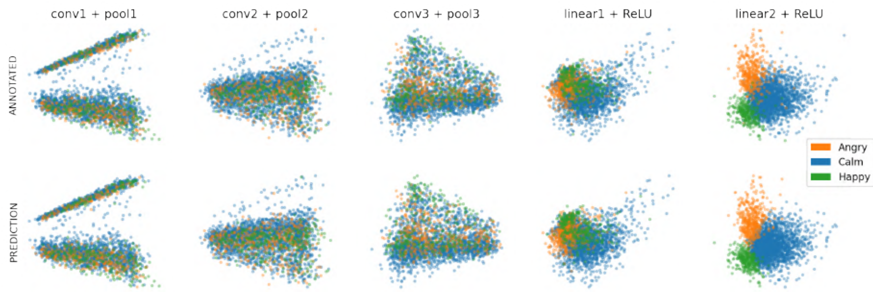


Figure 4. Two-dimensional representation obtained for each sample over different layers of the network using the PCA technique. The colour of each point represents the category to which it belongs, showing in the first row the annotated categories (ground truth) and in the second row the predicted categories.

Samples are mixed in the first stages, but as we go deep into the network, the categories are better-defined, so that the figure shows how the network learns how to differentiate them. Let us note that the dimensionality reduction in conv1 + pool1 (left image) is from 3416 to 2, whereas in linear2 + ReLU (right image) this is from 32 to 2, which could also lead to regions being better delimited, as the last picture of the right-hand side of Figure 4 shows.

In the second row, the colours are associated with the predicted categories. Thus, the regions of each category are well-defined at the final stages by clear boundaries, which supports the network's decision-making. It is interesting to note that, comparing the right-hand side pictures of annotated and predicted categories, a tendency toward "Angry" and "Happy" is shown in the predicted values. This correlates well with the values of the confusion matrices of the previous section, where the network estimates some "Calm" samples as "Angry" or "Happy". This seems to be due to the oversampling method that makes the minority classes more relevant. A more accurate sweep of the β coefficient might be useful in future work.

4.3. Class Model Visualisation

Class Model Visualisation is a global method within the Explainable Artificial Intelligence (XAI) framework, the goal of which is to generate image visualisations of each of the classes or categories the system is trying to predict [10]. We selected this method because it can provide insights into the features that the system takes into account when making those predictions in a visual way. At this point, we take advantage of the fact that we use the spectrogram to parametrise the voice signal by processing it as an image. This can be very useful for understanding the performance of the system and acting. For instance, it might be used for analysing the diversity of the samples in a category, which can be influenced by different factors such as the subjectivity in the annotation process. If a sample has a feature that the model relates to a different category that is not the predicted one, it might be due to a low agreement among the annotators, and it might be interesting to see what happens after a second annotation process.

Given a convolutional network f and a class of interest c, the goal is to generate an image visualisation I', which is representative of c. This is based on the scoring methods used to train f, which maximises the class probability score $S_c(I)$ for c, combined with a weighted (λ) L2 regularisation so that the image I' keeps regular values, such that:

$$I' = \arg\max_{I} S_c(I) - \lambda \|I\|_2^2 \quad (2)$$

Thus, the generated images (usually called Deep Dream) provide information related to what the black box model had learnt for a particular class or category in the dataset [83]. In this work, the Deep Dream images associated with the different classes, for both categorical and VAD models, are shown in Figure 5. A random spectrogram sample was selected for initialisation and an L2-regularisation method was employed to obtain the final images.

These deep dream images show different patterns for different categories. However, their interpretation is difficult since speech information is harder to interpret visually than deep dreams of images. First of all, it is worth noting that usually human speech is located in the 300–4000 Hz range, so the analysis will be focused on that interval. Focusing on the categorical model, it can be appreciated that, in "Calm", there is an intensity attenuation of around 1000 Hz, whereas, in "Happy", this is an intense interval, and the attenuation can be appreciated at lower frequencies, below 500 Hz. For the "Angry" category, the attenuation is observed below 1024 Hz, and above that frequency, there is an intense band (narrower than in the "Happy" class).

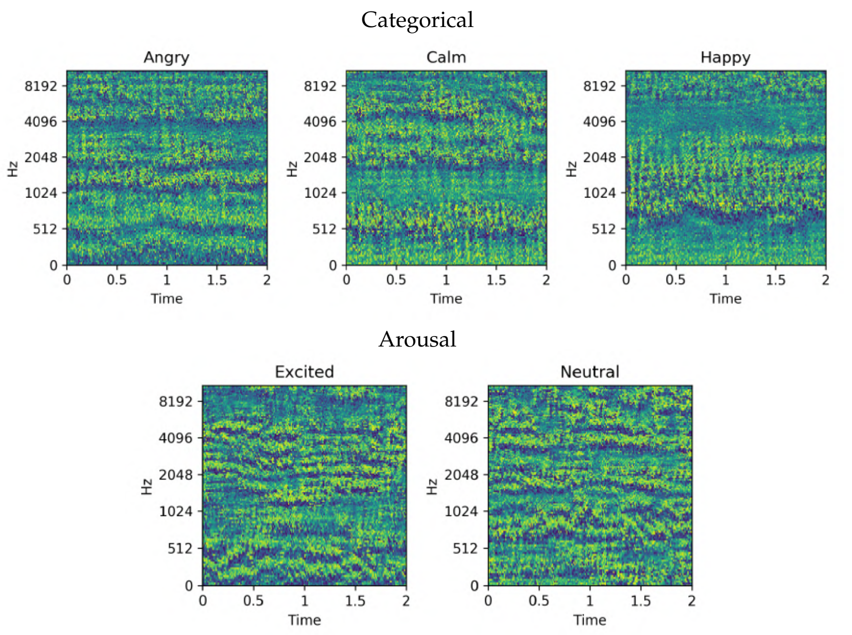


Figure 5. *Cont.*

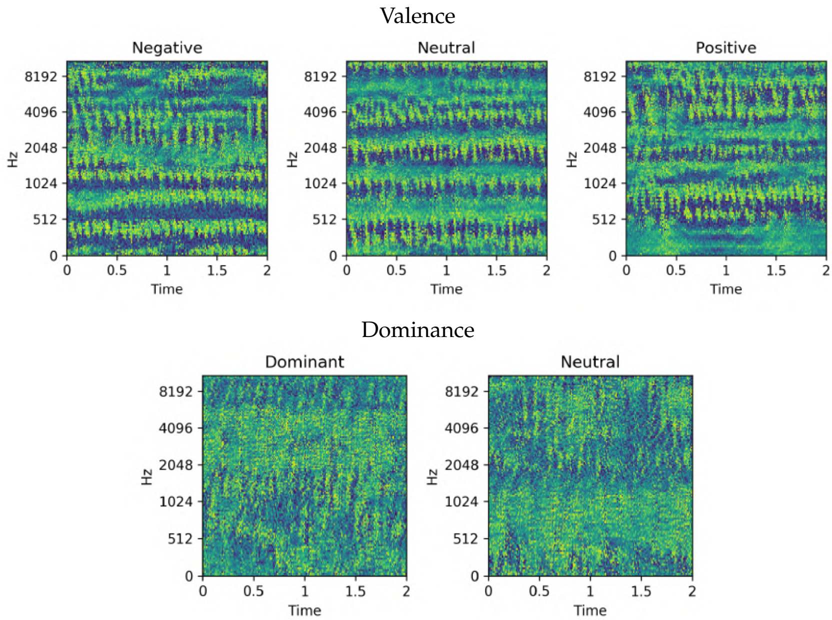


Figure 5. Extraction of the suitable spectrogram that maximises the classification output for each class using the DeepDream technique.

Switching to the VAD model, if we focus on Arousal, it can be concluded that, in "Neutral", there is a mix of high and low frequencies that are activated. However, the "Excited" category seems to be more activated at high frequencies (above 1024 Hz). Regarding Valence, the patterns are better-defined, and there seems to be much less noise here. Clearly differentiated bands emerge in this dimension, and are located in different places for the three different values. If we compare "Negative" with "Neutral", it can be seen that "Negative" has more defined blue bands between green or yellow bands, mainly at low frequencies. In "Neutral", the separation among bands becomes vaguer, and a pattern is replicated all over the frequencies, which might be considered as complementary to the one appreciated in "Negative" (see frequency bands above 512 Hz, blue in "Negative" and green/yellow in "Neutral"). Finally, the vagueness among bands increases in "Positive". Finally, the obtained images for the Dominance dimension are much noisier. However, the "Dominant" category seems to be activated at higher frequencies (2000–4000 Hz), while "Neutral" is activated at lower ones (below 1000 Hz).

In order to evaluate the effectiveness of these images to represent the categories, we tried to artificially transform all the samples in our test set into a specific category (using Deep Dream). To this end, the transformation consists of removing the profile of the deep dream image associated with the category it belongs to from each spectrogram, according to the system, and then adding the profile (deep dream) of the target category. We define the deep dream profile as a function that provides the average and standard deviation, over time, for each frequency, as Equation (3) shows. Assuming that x is the intensity associated with a time t and a given frequency f, the average and standard deviation for each frequency are computed as follows:

$$DD_{avg}(f) = \sum_{t \in \Delta t} \frac{x(f,t)}{\Delta t} \tag{3}$$

$$DD_{std}(f) = \sqrt{\left(\frac{\sum_{t\in\Delta t}(x(f,t) - DD_{avg}(f))^2}{\Delta t}\right)} \quad (4)$$

The transformation made to remove the profile of a specific category is described in Equation (5).

$$x'(f,t)_{\forall t\in\Delta t} = \frac{x(f,t) - DD_{avg}(f)}{DD_{std}(f)} \quad (5)$$

and the transformation made to add the profile of a new category is described in Equation (6):

$$x'(f,t)_{\forall t\in\Delta t} = (x(f,t) \cdot DD_{std}(f)) + DD_{avg}(f) \quad (6)$$

First, the samples that were correctly classified by the Neural Network were considered. These samples were transformed to a new category by applying the transformation in Equation (5) to each spectrogram, thus removing the profile of the category the sample belongs to. Then, the transformation on Equation (6) is applied to add the profile of the new category. These samples were firstly transformed to "Angry", then to "Calm" and finally to "Happy". Finally, the neural network classifies the transformed samples. The resulting confusion matrix is shown in Table 4.

Table 4. Confusion matrix with the percentage of correctly classified samples after profile transformation for each category (only correctly classified test samples). Each sample has been transformed to the profile of each class and therefore each row sums up to 100%.

	Angry	Calm	Happy
Angry	100	0	0
Calm	0	100	0
Happy	7.33	0	92.67

Table 4 shows that, when the transformations are applied, the system classifies the samples correctly in almost all the cases, i.e., only 7.33% of the samples transformed to "Happy" were wrongly classified as "Angry". Let us note that the transformation may introduce some noise that might lead to peak values that could be misinterpreted by the system, leading to errors. However, the good results suggest that the deep dream images are good representations of what the neural network learns for each category.

Then, all the samples of the test set, i.e., the ones not correctly classified, were considered and the process was replicated again. In this case, the predicted category was considered to remove the profile in the first step. The new results are shown in Table 5.

Table 5. Confusion matrix with the percentage of correctly classified samples after profile transformation for each category (all test samples). Each sample has been transformed to the profile of each class and therefore each row sums up to 100%.

	Angry	Calm	Happy
Angry	97.67	2.33	0
Calm	0	89.33	10.66
Happy	8.00	0	92.00

Table 5 shows some more misclassified samples for this experiment (8% of "Happy" that were classified as "Angry", 10% of "Calm" that were classified as "Happy" and 2% of "Angry" that were classified as "Calm"). In this case, there are more noisy samples because, when converting them to a new category, once the information about its prediction was removed, a higher error is achieved. For these samples, a reannotation process might be considered in order to see whether the noise comes from the subjectivity associated with the annotation procedure.

Moreover, the achieved results let us estimate which frequency bands are more relevant in each category. Let us focus, for instance, on the band above 512 Hz in the categorical model. We took a sample, labelled as “Happy”, which shows low-intensity values in the selected band. Then, the intensity values in that band were gradually increased until the spectrogram shown in Figure 6 was achieved. In this process, the values of the last layer of the network, from which the predictions were carried out, are represented in Figure 7. The figure shows how the prediction changes from “Happy” to “Calm”, which corresponds to a higher intensity band above 512 Hz in the Deep Dream image. This provides a hint to analyse samples that should be “Calm” and are predicted as “Happy” for instance. If the band above 512 Hz has low-intensity values, it might be a sample wrongly annotated as “Calm”.

Finally, it is worth mentioning that a qualitative comparison of the Deep Dream images shed some light on the achieved results. In fact, it is understandable why sometimes the system’s predictions are not accurate and some classes are mixed. If we focus on the categorical model, it is noticeable that Calm is something that is in between “Angry” and “Happy”, having similarities with both of them. However, “Happy” and “Angry” are much more different from each other. In the same way, regarding Valence, high similarity can be appreciated among “Negative”-“Neutral” and “Neutral”-“Positive”, while the differences between “Positive” and “Negative” are much more significant.

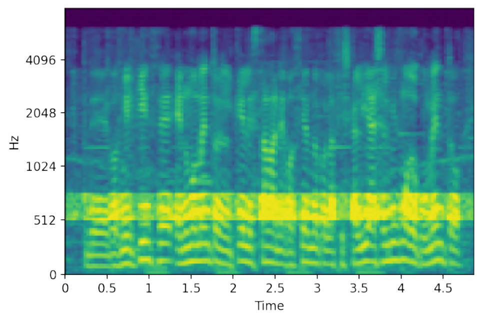


Figure 6. Spectrogram modified to alter the network prediction from “Happy” to “Calm”, intensifying frequencies above 512 Hz.

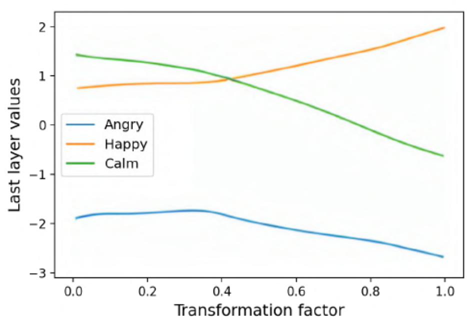


Figure 7. Representation of how the values of the network prediction change when applying changes by a factor in the spectrogram in Figure 6.

5. Conclusions

Here, we have presented a method to improve our understanding of the computational methods and decision-making involved in the identification of emotions in spontaneous speech. The selected task consists of Spanish TV debates, with a high level of complexity as well as additional subjectivity in the human perception-based annotation procedure.

Both categories of emotions and Valence–Arousal–Dominance dimensions were considered to represent emotional information. Then, a simple and light convolutional neural model was proposed to allow the joint identification of emotions using the VAD and the categorical model. The architecture of the model allows us to follow the decision-making process in order to understand where the outputs come from. The overall performance of our proposed model has also shown to be slightly higher than VGG16, a complex well-established CNN for image processing.

In this work, we focused on the understanding of the decision-making process—that is, where the errors come from and how the decisions are made. The evolution of the extracted patterns in the network layers that support their internal decisions was visualised and analysed. In addition, an XAI technique called Deep Dream was used to visualise the features related to the emotional categories. The experiments carried out show that the Deep Dream images might be an interesting tool when considering such complex neural network architectures for carrying out speech emotion detection over realistic tasks.

Author Contributions: Conceptualization, M.d.V., R.J., A.L.Z. and M.I.T.; methodology, M.d.V. and R.J.; software, M.d.V.; validation, M.d.V., A.L.Z. and R.J.; formal analysis, M.d.V., R.J., A.L.Z. and M.I.T.; investigation, M.d.V., A.L.Z. and R.J.; resources, M.d.V., R.J. and M.I.T.; data curation, M.d.V., R.J. and M.I.T.; writing—original draft preparation, M.d.V. and R.J.; writing—review and editing, M.d.V., R.J., A.L.Z. and M.I.T.; visualization, M.d.V. and R.J.; supervision, R.J. and M.I.T.; project administration, R.J. and M.I.T.; funding acquisition, R.J. and M.I.T. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this paper was conducted as part of the AMIC, AMIC-PdC, BEWORD and MENHIR projects, which received funding from the Spanish Minister of Science under grants TIN2017-85854-C4-3-R, PDC2021-120846-C43 and PID2021-126061OB-C42 and from the European Union's H2020-MSCA-RISE-2018 Research and Innovation Staff Exchange, under Grant No. 823907. The first author also received a PhD scholarship from the University of the Basque Country UPV/EHU, PIF17/310.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This corpus was developed by a consortium of Spanish Universities under the umbrella of AMIC, "Affective multimedia analytics with inclusive and natural communication" project. ATRESMEDIA, producer and owner of the copyright of LaSextaNoche program's contents, provided the consortium with the rights to use the audio files only for research purposes. The availability of the data are being considered by the consortium.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

VAD	Valence, Arousal, and Dominance
LSTM	Long Short-Term Memory
XAI	eXplainable Artificial Intelligence
NLP	Natural Language Processing
CNN	Convolutional Neural Network
PCA	Principal Component Analysis

References

1. Moors, A. Comparison of affect program theories, appraisal theories, and psychological construction theories. In *Categorical versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell*; John Benjamins: Amsterdam, The Netherlands, 2012; pp. 257–278.
2. de Velasco, M.; Justo, R.; Inés Torres, M. Automatic Identification of Emotional Information in Spanish TV Debates and Human-Machine Interactions. *Appl. Sci.* **2022**, *12*, 1902. [CrossRef]
3. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons: Hoboken, NJ, USA, 1999; Volume 98, p. 16.
4. Russell, J.A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **2003**, *110*, 145. [CrossRef]
5. Raheel, A.; Majid, M.; Alnowami, M.; Anwar, S.M. Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. *Sensors* **2020**, *20*, 4037. [CrossRef]
6. Egger, M.; Ley, M.; Hanke, S. Emotion recognition from physiological signal analysis: A review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55. [CrossRef]
7. Ekman, P.; Friesen, W.V.; Ellsworth, P. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 11.
8. Low, D.M.; Bentley, K.H.; Ghosh, S.S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig. Otolaryngol.* **2020**, *5*, 96–116. [CrossRef]
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
10. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
11. Brave, S.; Nass, C. Emotion in human-computer interaction. *Hum. Comput. Interact. Fundam.* **2009**, *20094635*, 53–68.
12. Richardson, S. Affective computing in the modern workplace. *Bus. Inf. Rev.* **2020**, *37*, 78–85. [CrossRef]
13. 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. [CrossRef]
14. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
15. Alharbi, M.; Huang, S. A Survey of Incorporating Affective Computing for Human-System Co-Adaptation. In *Proceedings of the 2020 The 2nd World Symposium on Software Engineering*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 72–79. [CrossRef]
16. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
17. Piana, S.; Stagliano, A.; Odone, F.; Verri, A.; Camurri, A. Real-time automatic emotion recognition from body gestures. *arXiv* **2014**, arXiv:1402.5047.
18. Liu, B. Sentiment analysis and subjectivity. *Handb. Nat. Lang. Process.* **2010**, *2*, 627–666.
19. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl. Based Syst.* **2022**, *235*, 107643. [CrossRef]
20. Deng, J.; Ren, F. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
21. Li, W.; Shao, W.; Ji, S.; Cambria, E. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* **2022**, *467*, 73–82. [CrossRef]
22. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
23. Zhang, K.; Li, Y.; Wang, J.; Cambria, E.; Li, X. Real-Time Video Emotion Recognition Based on Reinforcement Learning and Domain Knowledge. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1034–1047. [CrossRef]
24. Prinz, J. Which emotions are basic. *Emot. Evol. Ration.* **2004**, *69*, 88.
25. 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. [CrossRef]
26. Gunes, H.; Pantic, M. Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot. IJSE* **2010**, *1*, 68–99. [CrossRef]
27. Wöllmer, M.; Eyben, F.; Reiter, S.; Schuller, B.; Cox, C.; Douglas-Cowie, E.; Cowie, R. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In Proceedings of the 9th Interspeech 2008 Incorp 12th Australasian International Conference on Speech Science and Technology SST 2008, Brisbane, Australia, 22–26 September 2008; pp. 597–600.
28. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
29. Nicolaou, M.A.; Gunes, H.; Pantic, M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* **2011**, *2*, 92–105. [CrossRef]
30. Fontaine, J.R.; Scherer, K.R.; Roesch, E.B.; Ellsworth, P.C. The world of emotions is not two-dimensional. *Psychol. Sci.* **2007**, *18*, 1050–1057. [CrossRef] [PubMed]
31. Scherer, K.R. What are emotions? In addition, how can they be measured? *Soc. Sci. Inf.* **2005**, *44*, 695–729. [CrossRef]

32. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September, 2005.
33. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
34. Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; Pantic, M. AVEC 2012: The continuous audio/visual emotion challenge. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 2012; pp. 449–456.
35. Vázquez, M.D.; Justo, R.; Zorrilla, A.L.; Torres, M.I. Can Spontaneous Emotions be Detected from Speech on TV Political Debates? In Proceedings of the 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Naples, Italy, 23–25 October 2019; pp. 289–294.
36. Sen, T.; Naven, G.; Gerstner, L.M.; Bagley, D.K.; Baten, R.A.; Rahman, W.; Hasan, K.; Haut, K.; Mamun, A.A.; Samrose, S.; et al. DBATES: Dataset of DeBate Audio features, Text, and visual Expressions from competitive debate Speeches. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
37. Blanco, R.J.; Alcaide, J.M.; Torres, M.I.; Walker, M.A. Detection of Sarcasm and Nastiness: New Resources for Spanish Language. *Cogn. Comput.* **2018**, *10*, 1135–1151. [CrossRef]
38. Justo, R.; Torres, M.I.; Alcaide, J.M. Measuring the Quality of Annotations for a Subjective Crowdsourcing Task. In Proceedings of the Pattern Recognition and Image Analysis—8th Iberian Conference, IbPRIA 2017, Faro, Portugal, 20–23 June 2017; Lecture Notes in Computer Science; Alexandre, L.A., Sánchez, J.S., Rodrigues, J.M.F., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10255, pp. 58–68. [CrossRef]
39. deVelasco, M.; Justo, R.; López-Zorrilla, A.; Torres, M.I. Automatic Analysis of Emotions from the Voices/Speech in Spanish TV Debates. *Acta Polytech. Hung.* **2022**, *19*, 149–171. [CrossRef]
40. Panda, R.; Malheiro, R.M.; Paiva, R.P. Audio Features for Music Emotion Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
41. Latif, S.; Cuayáhuitl, H.; Pervez, F.; Shamshad, F.; Ali, H.S.; Cambria, E. A survey on deep reinforcement learning for audio-based applications. *arXiv* **2021**, arXiv:2101.00240.
42. Huang, K.; Wu, C.; Hong, Q.; Su, M.; Chen, Y. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5866–5870. [CrossRef]
43. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *arXiv* **2017**, arXiv:1706.00612.
44. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
45. Marazakis, M.; Papadakis, D.; Nikolaou, C.; Constanta, P. System-level infrastructure issues for controlled interactions among autonomous participants in electronic commerce processes. In Proceedings of the Tenth International Workshop on Database and Expert Systems Applications, DEXA 99, Florence, Italy, 3 September 1999; pp. 613–617. [CrossRef]
46. Parthasarathy, S.; Tashev, I. Convolutional Neural Network Techniques for Speech Emotion Recognition. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 121–125. [CrossRef]
47. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
48. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013.
49. Tian, L.; Moore, J.D.; Lai, C. Emotion recognition in spontaneous and acted dialogues. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 698–704.
50. Ocquaye, E.N.N.; Mao, Q.; Xue, Y.; Song, H. Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *Int. J. Intell. Syst.* **2021**, *36*, 53–71. [CrossRef]
51. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-Based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484. [CrossRef]
52. Zheng, L.; Li, Q.; Ban, H.; Liu, S. Speech emotion recognition based on convolution neural network combined with random forest. In Proceedings of the 2018 Chinese Control In addition, Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 4143–4147. [CrossRef]

53. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.
54. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
55. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093. [CrossRef]
56. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, pp. 2449–12460.
57. Peyser, C.; Mavandadi, S.; Sainath, T.N.; Apfel, J.; Pang, R.; Kumar, S. Improving tail performance of a deliberation e2e asr model using a large text corpus. *arXiv* **2020**, arXiv:2008.10491.
58. López Zorrilla, A.; Torres, M.I. A multilingual neural coaching model with enhanced long-term dialogue structure. *ACM Trans. Interact. Intell. Syst.* **2022**, *12*, 1–47. [CrossRef]
59. Boloor, A.; He, X.; Gill, C.; Vorobeychik, Y.; Zhang, X. Simple Physical Adversarial Examples against End-to-End Autonomous Driving Models. In Proceedings of the 2019 IEEE International Conference on Embedded Software and Systems (ICESS), Las Vegas, NV, USA, 2–3 June 2019; pp. 1–7. [CrossRef]
60. LeCun, Y. Generalization and network design strategies. *Connect. Perspect.* **1989**, *19*, 143–155.
61. Weng, J.; Ahuja, N.; Huang, T.S. Cresceptron: A self-organizing neural network which grows adaptively. In Proceedings of the 1992 IJCNN International Joint Conference on Neural Networks, Baltimore, MD, USA, 7–11 June 1992; Volume 1, pp. 576–581.
62. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
64. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. [CrossRef]
65. Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. In *Proceedings of the 29th ACM International Conference on Information*; Knowledge Management; Association for Computing Machinery: New York, NY, USA, 2020; pp. 105–114. [CrossRef]
66. Zubiaga, I.; Menchaca, I.; de Velasco, M.; Justo, R. Mental Health Monitoring from Speech and Language. In Proceedings of the Workshop on Speech, Music and Mind, Online, 15 September 2022; pp. 11–15. [CrossRef]
67. Patel, N.; Patel, S.; Mankad, S.H. Impact of autoencoder based compact representation on emotion detection from audio. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 867–885. [CrossRef] [PubMed]
68. Senthilkumar, N.; Karpakam, S.; Gayathri Devi, M.; Balakumaresan, R.; Dhilipkumar, P. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. *Mater. Today Proc.* **2022**, *57*, 2180–2184. [CrossRef]
69. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]
70. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
71. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 210–215.
72. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
73. Zhang, W.; Lim, B.Y. Towards Relatable Explainable AI with the Perceptual Process. *arXiv* **2022**, arXiv:2112.14005v3.
74. Das, A.; Mock, J.; Chacon, H.; Irani, F.; Golob, E.; Najafirad, P. Stuttering speech disfluency prediction using explainable attribution vectors of facial muscle movements. *arXiv* **2020**, arXiv:2010.01231.
75. Anand, A.; Negi, S.; Narendra, N. Filters Know How You Feel: Explaining Intermediate Speech Emotion Classification Representations. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 756–761.
76. Esposito, A.; Marinaro, M.; Palombo, G. Children Speech Pauses as Markers of Different Discourse Structures and Utterance Information Content. In *Proceedings of the International Conference: From Sound to Sense*; MIT: Cambridge, MA, USA, 2004.
77. Ortega Giménez, A.; Lleida Solano, E.; San Segundo Hernández, R.; Ferreiros López, J.; Hurtado Oliver, L.F.; Sanchis Arnal, E.; Torres Barañano, M.I.; Justo Blanco, R. AMIC: Affective multimedia analytics with inclusive and natural communication. *Proces. Leng. Nat.* **2018**, *61*, 147–150.
78. Calvo, R.; Kim, S. Emotions in text: Dimensional and categorical models. *Comput. Intell.* **2012**, *Early view*. [CrossRef]
79. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]

80. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
81. Letaifa, L.B.; Torres, M.I. Perceptual Borderline for Balancing Multi-Class Spontaneous Emotional Data. *IEEE Access* **2021**, *9*, 55939–55954. [CrossRef]
82. Pastor, M.; Ribas, D.; Ortega, A.; Miguel, A.; Solano, E.L. Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation. In Proceedings of the IberSPEECH 2022, Granada, Spain, 14–16 November 2022; pp. 76–80.
83. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371.