Neural Computing and Applications (2020) 32:17869–17886
https://doi.org/10.1007/s00521-019-04618-7

DEVELOPING NATURE-INSPIRED INTELLIGENCE BY NEURAL SYSTEMS

Multilingual audio information management system based on semantic knowledge in complex environments

Karmele Lopez-de-Ipina[1,2,3] · Nora Barroso[1,3] · Pilar M. Calvo[3,4] · Carmen Hernandez[5] · Aitzol Ezeiza[1,3] · Unai Susperregi[3,6] · Elsa Fernández[1,3]

Received: 15 April 2019 / Accepted: 22 November 2019 / Published online: 3 February 2020

Abstract
This paper proposes a multilingual audio information management system based on semantic knowledge in complex environments. The complex environment is defined by the limited resources (financial, material, human, and audio resources); the poor quality of the audio signal taken from an internet radio channel; the multilingual context (Spanish, French, and Basque that is in under-resourced situation in some areas); and the regular appearance of cross-lingual elements between the three languages. In addition to this, the system is also constrained by the requirements of the local multilingual industrial sector. We present the first evolutionary system based on a scalable architecture that is able to fulfill these specifications with automatic adaptation based on automatic semantic speech recognition, folksonomies, automatic configuration selection, machine learning, neural computing methodologies, and collaborative networks. As a result, it can be said that the initial goals have been accomplished and the usability of the final application has been tested successfully, even with non-experienced users.

Keywords Evolutionary computing · Artificial neural networks · Internet information management · Management of complex systems

Abbreviations

AdiUP	Audio information management system
ANN	Artificial neural networks
APD	Acoustic phonetic decoding
ASR	Automatic speech recognition
Cl	Corrrect rates for classes
Co	Correct rates for concepts
FFT	Fast Fourier transform
FIP	Filler insertion penalty

✉ Karmele Lopez-de-Ipina
karmele.ipina@ehu.eus

Nora Barroso
nora.barroso@ehu.eus

Pilar M. Calvo
pilarmaria.calvo@ehu.eus

Carmen Hernandez
mamen.hernandez@ehu.eus

Aitzol Ezeiza
aitzol.ezeiza@ehu.eus

Unai Susperregi
unai@irunweb.com

Elsa Fernández
elsa.fernandez@ehu.eus

[1] Department of Systems Engineering and Automation, University of the Basque Country, Donostia-San Sebastián, Spain

[2] Department of Psychiatry, University of Cambridge, Cambridge, UK

[3] EleKin, Engineering and Society Research Group, University of the Basque Country, Europa Plaza 1, 20018 Donostia-San Sebastián, Spain

[4] Department of Computer Architecture and Technology, University of the Basque Country, Donostia-San Sebastián, Spain

[5] Department of Computer Science and Artificial Intelligence, University of the Basque Country, Donostia-San Sebastián, Spain

[6] Irunweb S.L. Company, Zentolen Gunea, 12, 1 Left, 20100 Renteriaz, Gipuzkoa, Spain

HMM	Hidden Markov model
LID	Language identification
LOOCV	Leave-one-out cross-validation
LU	Lexical units
MFCC	Mel frequency cepstral coefficients
NG	Number of Gaussians
NS	Number of states
OOV	Out of vocabulary
PCA	Principal components analysis
PER	Phone error rate
SC-HMM	Semicontinuous HMM
SLU	Sublexical units
SNR	Signal-to-noise ratio
SSG	Audio and speech segments
SVM	Support vector machines
VAD	Voice activity detection
WADA	Waveform amplitude distribution analysis
WER	Word error rate
WIP	Word insertion penalty
X-SAMPA	eXtended speech assessment methods phonetic alphabet

1 Introduction

In order to contribute to the development of a community, we should take into account its socio-cultural and economic context. In this paper, we propose a joint solution to the socio-cultural context of the Basque Country and the needs and characteristics of small and medium companies with limited staff and economic resources, which are one of the foundations of the economy of the area. The model is easily exportable to other similar environments that may appear in developing areas [1]. Thus, innovative technologies have to be adequately designed so that they suit the needs of the industrial fabric, and they can be integrated into real applications. For instance, those technologies should be designed including auto-changing and auto-learning abilities. These capabilities will allow systems to automatically learn from new data and to adapt its components to new conditions, thus increasing system robustness [1].

In the socio-cultural context of the Basque Country, the interest in multilingual systems [1–6] arises because there are three languages in coexistence (Basque, Spanish, and French). However, Basque is an under-resourced language and its situation is different depending on the geographical location: In the Spanish State, Basque and Spanish are both official languages, and the recovery of Basque is active with positive results; in the French State, Basque is not an official language and it shows a clear negative growth with a low number of speakers. In addition to this, there is a high cross-lingual interaction between them, even though the origins of Basque are completely different from the origins of the other two languages (Spanish and French). In this scenario, speakers tend to mix words and sentences that belong to the three languages in their discourse, and the acoustic interactions between the languages and the Basque dialects are a very interesting area of research [6]. Although some people speak the three languages like native speakers, people are more commonly fluent in only two of them, Basque–Spanish or Basque–French. This means that in their conversations there are mixed features of different languages, and for this reason cross-lingual ASR is necessary. Thus, the three languages have to be taken into account in order to develop an efficient ASR system, and it is essential to try to find new strategies for the development.

In this sense, audio information retrieval systems are increasingly used in different applications ranging from the extraction of musical information [7, 8], health-related applications [9, 10], to applications related to the acquisition of data in the mobile environment [11] or on the Internet of Things (IoT). New trends in the field of information and communications technologies are committed to humanize not only information, but also ways of access by means of new techniques such as concept-based semantic web, or access to information by more user-friendly interfaces [7]. In this research area, within the field of speech processing, multilingual automatic speech recognition (ASR) systems provide users with improved interaction by means of various languages. This increases comfort and naturalness because users can express themselves in their own language. Machine learning and neural computing-based approaches provide flexible solutions for complex systems [1].

In this context, this paper proposes an audio information management system (the *adiUP* system, available online: [12]) that can be directly applied to radio broadcasting and that has been tested using the trilingual (Basque, Spanish, and French) internet radio channel *Info7* [13]. The system provides automatic tools to manage all the information included in the audio records by semantic knowledge. One of the objectives of the development is to allow to introduce new information resources in a more open and collaborative way, where the user is an active part of the system. In this trilingual environment with an under-resourced language, there are other constrains regarding limited resources (financial, material, human, and audio resources) which, in addition to the poor quality of the signal, increase the complexity of the development. The developed *adiUP* system is an ASR system for complex environments that fulfills these requirements with under-resourced languages. The proposed system has been built using universal design, so that it can provide barrier-free access for example for people with hearing loss, cognitive impairments, or deafness.

This document is organized as follows: Sect. 2 presents the environment and the requirements of the system. Section 3 describes the resources. Section 4 sets forth the design of the proposed system, its architecture and components. In Sect. 5, a new methodology for automatic selection of the system configuration is proposed. Section 6 comprises the experimentation and discussion. Finally, conclusions are drawn in Sect. 7.

2 Description of the environment and requirements of the system

In the field of speech processing, extraction and indexing of audio information from internet mass media is a research area of interest to the international scientific community [14, 15]. The complexity of this task is closely related to factors such as the type of oral communication (isolated words, continuous speech, spontaneous speech, dialogue), the quality of the signal, the environment in which recognition takes place, the amount of terminology, the ability to generate confusion, language, and speakers [16]. There is a wide variety of solutions that addresses these problems in different ways, ranging from the detection of large vocabularies [17], through the detection of spoken numbers for telephone applications [18], to the detection of segments, spoken or not spoken [19].

When working in complex environments with limited amount of data, multilingual contexts, nonlinearities, or uncontrollable noise, some possibilities are based on: enriching poor resources of a language with resources from another powerful language beside it, approaches oriented to the lack of resources, cross-lingual approaches [20], training of acoustic models for a new language using results from other languages [21], data optimization methods, collaborative systems, or open configuration systems. However, the development of a robust ASR system is very tough when there are under-resourced languages involved, even if there are powerful languages beside them, and the classic techniques perform poorly with regard to correct rates [22, 23]. In fact, the statistical nature of most of the approaches used in ASR systems requires large amounts of language resources in order to perform properly. In the case of under-resourced languages, the availability of resources is very limited, most of which are of very poor quality. Therefore, nowadays there is an increasing interest in multilingual systems and under-resourced languages. There are interesting applications focused on these fields such as Babel [24] and meetings and publications whose main goal is to search for new methodologies oriented to this kind of complex environments [25–27]. Current research lines include: proposals for generating language resources; acoustic and language modeling; multilingual approaches; or management of audio information [28, 29].

Within this complex environment, our proposal aims to: automatically generate an optimal corpus from an initial corpus; a corpus is said to be optimal when the recognition results measured by means of cross-validation methods are optimal for the task; reduce the dimensionality of data through the convergence of redundant information by means of principal components analysis (PCA); estimate the accuracy of the system by means of the leave-one-out cross-validation technique (LOOCV); reduce unwished effects in sublexical units (SLU) due to the lack of resources and the low number of samples by choosing properly sublexical units and possible groupings of them; and extract information by means to folksonomies.

In this context, the requirements and novelty of the system lie in the complex environment in which it has to work, that is defined by:

- The trilingual context that comprises the three languages of the Basque Country (Basque, Spanish, and French), which requires the design of the first multilingual ASR system for these three languages.
- The regular appearance of cross-lingual elements between the three languages.
- The fact that within the geographical area of this research project, the variety of the Basque language in the French region is in a critical under-resourced situation because it is not an official language in the French State.
- The limited resources (financial, material, human, and audio resources), specially the audio corpora for training is only about 0.5–1% of the usual corpora for such systems, for example, in [14] and [30] they use 100–200 h for training, while in our system we use only 1 h.
- The extremely poor quality of the audio signal, taken from an internet radio channel, which increases the complexity of the development.

3 Materials

In this section, the resources for the development of the system are presented. Furthermore, the main linguistic features are analyzed because they have a clear impact both on the performance of the acoustic phonetic decoding (APD) system, and on the size of the vocabulary of the system.

3.1 Phonetic features of the languages

The three languages involved in our study are: Basque, Spanish, and French. Basque is a Pre-Indo-European language of unknown origin and has circa 1,000,000 speakers

in the Basque Country, which spreads over the international border between France and Spain. The Basque language has a wide range of dialects, there are six main dialects and several variations, and this dialectal variety entails phonetic, phonologic, and morphologic differences. In order to develop the APD system, a sound inventory of each language is necessary. Table 1 summarizes the sound inventories for the three languages, expressed in the eXtended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) notation, and the usage of phonemes in the three cases.

3.2 Lexical structure

A further challenge for developing an ASR system is that Basque is an agglutinative language with a special intra-word morphosyntactic structure [11] that may, depending on the complexity of the task, lead to intractable vocabularies. Inside Basque words, there is not only semantic information, but also grammatical elements as it can be seen in Table 2.

A plausible approach to the problem would be to use lemmas and morphemes instead of words when defining the system vocabulary [11]. However, this could lead to a recognition problem for the shortest morphemes, especially those that are not lemmas, for instance: "ak," "ko," "go," "k", etc. In this project, a robust proposal based on lemmas and pseudo-morphemes was used [31].

Table 1 Sound inventories in the X-SAMPA

Sound type		BS	FR	SP
Plosive	p b t d k g	X	X	X
	J\ c p_ht_hk_h	X	–	–
Affricates	tS J\j\	X	–	X
	ts_mts_adZ	X	–	–
Fricatives	B f s_az_a	X	X	X
	S Z	X	X	–
	D x G	X	–	X
	S_mz_m j\ h	X	–	–
	v	–	X	–
Nasals	m n J	X	X	X
	F n_d N	X	–	X
Liquids	l	X	X	X
	R\	X	X	–
	L r\ r	X	–	X
Vowel glides	w j	X	X	X
	H	–	X	–
Vowels	i e a o u	X	X	X
	y @	X	X	–
	A E O 2 9 a~ e~ o~ 9~	–	X	–

BS Basque, *FR* French, *SP* Spanish, and usage of phonemes across languages

3.3 Cross-lingual effects

Most speakers in the Basque Country are bilingual, and they commonly mix two of the three languages in their speech, particularly in spontaneous speech. The two languages that are mixed depend on the region of the country where the speaker lives: most Basque speakers living in the Spanish side also use Spanish, while Basque speakers in the French side also use French. Moreover, mixing all the three languages also occurs. Indeed, the acoustic interactions between these three languages with the addition of Basque dialects are very strong, because speakers naturally and spontaneously mix sounds and vocabulary, and sometimes they also add other influences, such as English. Some speakers are able to use the three languages consecutively in the same sentence with native pronunciation. All the resources contain numerous instances of cross-lingual material at words, sentences, and pronunciation levels (Tables 1 and 3). The strongest effects appear in Spanish and French recordings, which unfortunately also have the highest background noise level. The inventory of allophones for the languages of the system and the matches between each of them are shown in Table 1 [32].

3.4 Resources and materials

The basic audio resources used in this project have been mainly provided by a small local news radio channel, *Info7* [13], that is trilingual (Basque-BS-, Spanish-SP-, and French-FR-). It has provided the audio and text data from their news bulletins for each language (semi-parallel corpus). The texts have been processed to create XML files which include information of different speakers, noises, and parts of the speech files and transcriptions. The transcriptions for the Basque language also include morphological information such as the lemma of each word and Part-Of-Speech tag.

In order to correctly implement an ASR system, it is crucial to design and obtain appropriate linguistic resources. A speech corpus is a collection of audio recordings tagged at different levels, which contains phrases, words, and common expressions of a certain language. This type of corpus is a database that stores implicitly various properties of the language, and this information lays the foundation for building voice recognition systems.

When only limited resources are available, the appropriate choice of the training corpus is a fundamental part of the design of the application. This is because on the one hand, depending on the circumstances a larger corpus does not imply better performance results [33], and on the other

Table 2 Examples of the agglutinative structure of the Basque language

Example in Basque	Lemma+morphemes	Translation
Etxekoarenak	Etxe+ko+aren+ak	The people from home
Parisekoak	Paris+eko+ak	People from Paris
Miguelek txakur hori ikusiko du	Miguel+ek txakur hori ikusi+ko du	Miguel will see that dog

Table 3 Examples of cross-lingual appearance in the language resources

Primary language	Text
Basque	*Luis* [Spanish] *Scola* [Italian] (*Argentina* [Spanish], 30 urte) Baskoniako jokalari ohia 2007tik dago *Houston* [English] *Rockets* [English] NBAko taldean, eta hauxe izan du denboraldirik onena: liga erregularrean 16.2 puntu eta 8.6 errebote lortu ditu partiduko
French	*Euskaltzaleen* [Basque] *Biltzarra* [Basque] organise l'Assemblée Générale annuelle de ses membres le dimanche 25 avril, à partir de 9 heures, à *Donostia* [Basque]
Spanish	*Koldo* [Basque] *Landaluze* [Basque] habla de Más Allá del Tiempo dirigido por *Robert* [English] *Schwentke* [English], Que Se Mueran los Feos de Nacho G. Velilla y de El Fantástico *Mr.*[English] *Fox* [English] película de animación dirigida por *Wes* [English] *Anderson* [English]

hand, the errors within the training corpus cause deterioration in the performance of the recognition system, or in other words, recordings with errors at various levels can cause the results not to be appropriate. The methodology for choosing optimal corpus is called *HobeCor*. Thanks to the *HobeCor* system, files that contain extreme anomalies can be automatically removed, thus decreasing the number of errors in the recognition process, for example: if an extreme noise interrupts communication, which causes an extreme distortion of the signal that prevents readability; if background music covers the speech; if the audio file is empty although there is a phonetic transcription, etc. The *HobeCor* system optimizes the recognition results, even after removing part of the corpus. Therefore, it makes possible to reduce even more the training corpus and to balance the corpuses of the different languages with regard to time duration. The *HobeCor* system has only been applied to speech segments.

The resources inventory is summarized in Table 4, which shows the duration of the overall corpus provided by *Info7* radio channel (second column, audio), and the duration of the speech segments (third column, SSG). There is a difference between them because in the broadcasted signal there are also segments that only contain music. The third column of Table 4 (SSG used for training) shows the time features of the training corpus used in this project after using the *HobeCor* system.

Table 4 Summary of the resources inventory, audio, and speech segments (SSG)

Languages	Audio (hh:mm:ss)	SSG (hh:mm:ss)	SSG used for training (hh:mm:ss)
BS	2:47:27	2:10:37	0:55:09
FR	3:17:23	1:22:54	0:55:38
SP	2:10:15	1:01:13	0:55:37
Total	8:15:05	4:34:44	2:46:24

In the audio for French and Spanish, there is a high background noise due to signature tunes, which can be noted in Fig. 1 that shows NIST signal-to-noise ratio (SNR) and the waveform amplitude distribution analysis (WADA) SNR of speech signals with regard to the signal length [34, 35].

4 Development of the proposed system

The *adiUP* multilingual tool [30] developed for the *Info7* internet radio channel provides users with information such as identification of language, subject of speech, terms of interest, or duration of audio files in news broadcasted in Basque, Spanish, and French. Thus, by reading the information added to the audio, the user can select topics of interest, or search among the audio files within the radio repository by subject, keywords, or language. In Sect. 4.1, the architecture of the system is analyzed, and in the following subsections the components of the system are described: folksonomies, sublexical units and acoustic phonetic models, and lexical units (LU).

4.1 Architecture of the system

The architecture of the *adiUP* system is based on three layers: the interface layer or user interaction layer, the domain layer, and the database layer, as it is outlined in Fig. 2.

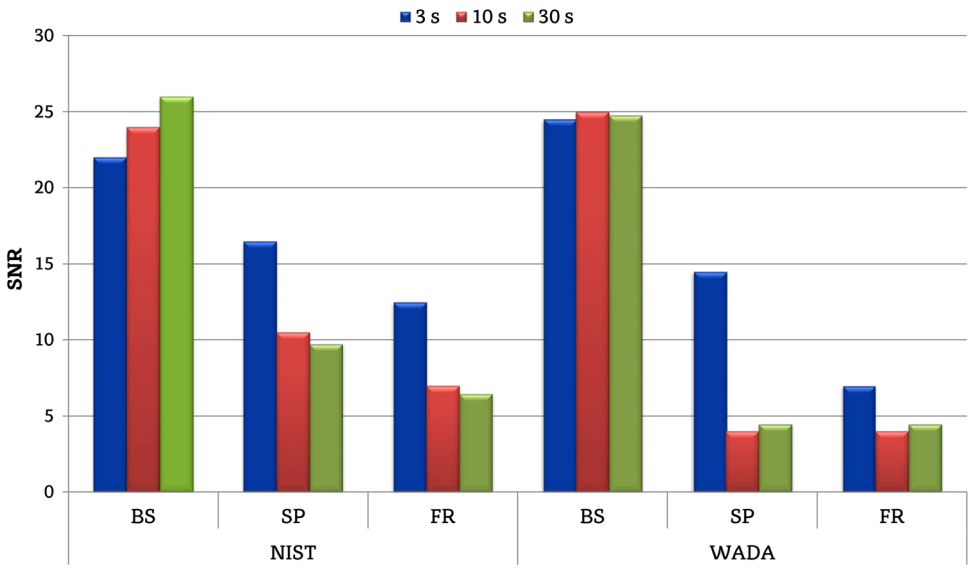

Fig. 1 NIST SNR and WADA SNR of speech signals with regard to the signal length

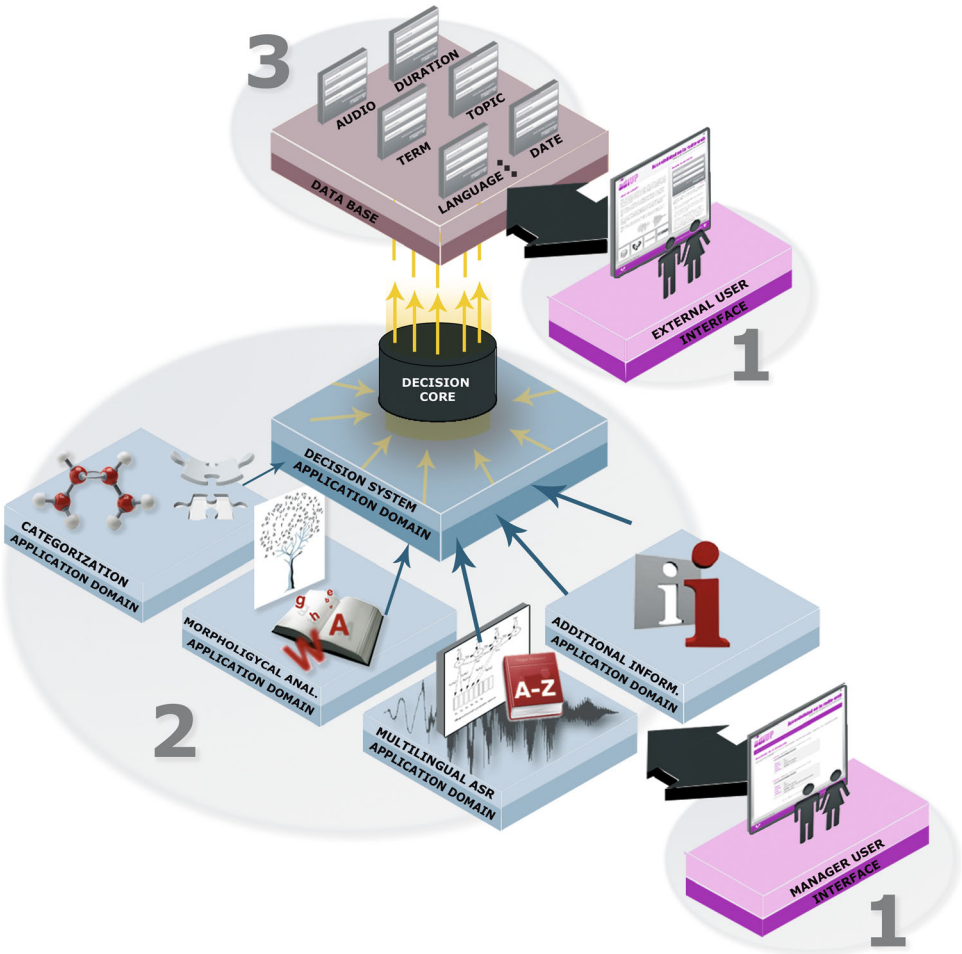

Fig. 2 Architecture of the *adiUP* system

4.1.1 Interface layer

This layer, labeled as 1 in Fig. 2, comprises all aspects of software and design related to interfaces for different kinds of users. In addition to the super-user, there are two types of users that can access the application: external users and system-management collaborators:

- *Super-user* This user is responsible for the supervision and control of the whole system, manages external users and system-management collaborators, and grants permissions.
- *External users* The so-called external users only use the application for information retrieval, but they cannot insert or modify any feature of the resources of the *adiUP* system.
- *System-management collaborators* These special external users are not necessarily experts. They are part of a collaboration network of the radio channel intended to improve and evolve the initial prototype, by adding new information and knowledge. These members are actively involved in the *adiUP* system and have the potential to enrich the resources of the system by inserting new audio files with related information (such as audio data, the name of the speaker, or the language), and by correcting errors at different stages. These active users can influence and improve both the results of the ASR module and the categorization of concepts of the audio files.

4.1.2 Domain layer

The domain layer consists of different components which provide all the information to fill the database of the system. This layer includes the Language Identification (LID) tool [32] that uses a hybrid system based on triphonemes and cross-lingual approaches which achieves an optimal and stable LID recognition rate despite the complexity of the problem. The domain layer is divided into some modules related to: the semantic management of audio information, the morpheme extraction task, the characteristics of folksonomies (categorization and extra information provided by the super-user), and the decision process. The modules of this layer are labeled as 2 in Fig. 2:

Multilingual ASR Audio files inserted through the interface are processed by a multilingual ASR system. The engine is a keyword-spotting system based on concepts (one word, several words, or slogans) and fillers of different types. The aim of the fillers is to absorb the words out of vocabulary (OOV) according to folksonomies.

Morphological analysis The output provided by the recognition system is analyzed morphologically in order to extract lemmas, which are used for concept specification in the folksonomy. The morphological analyzer has been developed by a partner company, *Insima Teknologia* [36], and it is multilingual: English, French, Spanish, and Basque.

Categorization Three folksonomies have been inserted in the *adiUP* system, one for each language, and concepts and terms have been grouped into superclasses, classes, and subclasses. The results of the morphological analyzer are the input to the categorization module, which calculates a value that defines their membership to a class.

Decision system In this module, the final assignment of concepts to the audio file is done by using a decision equation.

Additional information Extra information supplied by the super-user is inserted through this module.

4.1.3 Database layer

The database contains all the information of the application, which is related to users, audio, concepts, and conceptual information. This information is divided into tables and developed in the management system MySQL. This layer is labeled as 3 in Fig. 2. In order to index the results of the domain layer in the database, there is also a step called *Indexing* that transmits the information with the results between the domain layer and the database layer. Note that it is important to supply the system with a high flexibility level, so that the application can provide a powerful way to generate a network of collaborators, who will assist in improving the initial system. The parts that may be changed are the application dictionary, the expansion and creation of folksonomies, the categorization method, and the thresholds of thematic decision.

4.2 Folksonomies

For each language, a folksonomy has been defined. Figure 3 shows an example of the structure of these folksonomies.

4.2.1 Components of the folksonomy

The folksonomy of each language consist of:

- *Concepts* The concepts defined in the folksonomy of each language are politics, international, culture, general, sports, and headlines.
- *Superclasses, classes, and subclasses* One concept, for example politics, is subdivided into superclasses which

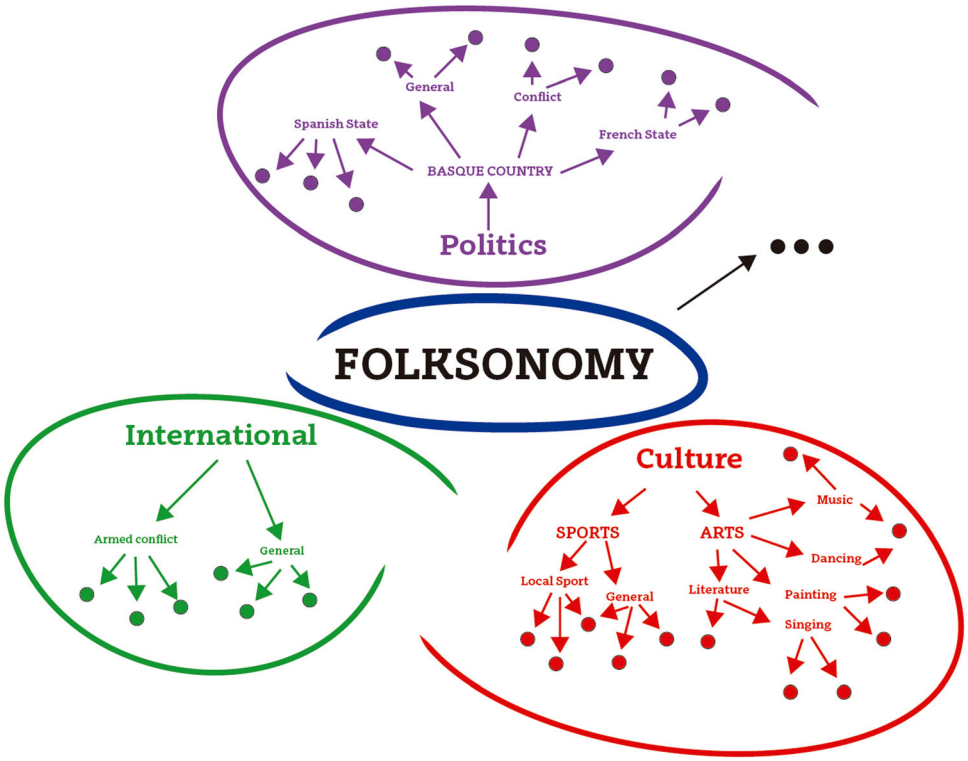

Fig. 3 Example of the structure of the folksonomy. Showing only some of the concepts (culture, politics, international, …), superclasses (sports, arts, …), classes (local sports, general…), etc

are more specific, for example Basque Country and State. Superclasses are divided into classes as well, and subsequently, classes are divided into subclasses. Each level provides more accurate information.

- *Instances* Instances are the last elements of the folksonomy, and they are words, lemmas, or terms made of groups of words.
- *Attributes* The attribute is a numeric value that indicates the relevance or weight of a subclass within a certain class. In general terms, there is a single attribute for each subclass. However, some subclasses can be shared by two or more classes. The relevance or weight (attribute) given to the subclass within each class makes the difference.
- *Axioms* An axiom is a parameter that takes into account the relevance and the number of instances of each class that appears in a certain text. Thus, the class where a certain text belongs to can be automatically chosen by comparing the different axioms of the text computed for each class.

4.2.2 Construction of the folksonomy

In order to determine the class where a certain text belongs to, the following procedure is followed, where several variables have been defined:

WT_c It is the total weight for a class c. For each class, this value is calculated by adding the weights of the n instances in the text that belong to that class:

$$WT_C = \sum_{J=1}^{n} w_{jsch_v} \tag{1}$$

where w_{jsch_v} is the weight of the instance j within a subclass s, that belongs to that particular class c, that can also belong to a superclass h, which can belong to a particular concept ν as well. The value of the weight ranges from 1 (the lowest relevance) to 5 (the highest relevance).

WR_c The so-called ratified weight for a certain class c is a numerical value which takes into account the relevance of the subclasses that belong to that class. Within each class, there are defined five weight values W_{cr}, $r = 1,\ldots,5$, where W_{c5} represents the weight of the most relevant subclasses, and W_{c1} represents the weight of the less relevant subclasses. These parameters (W_{cr},) depend on the subjectivity and experience of the creator of the folksonomy. For each class, the value of the ratified weight WR_c is calculated as the sum of these five weights, this is:

$$WR_C = \sum_{r=1}^{5} W_{cr} \tag{2}$$

W_c It is the final weight value or axiom of the text that is computed for each class c by multiplying the total weight by the ratified weight:

$$W_C = WT_C \cdot WR_C \quad (3)$$

Finally, the class where the text belongs to is the one that fulfills the following optimization criterion:

$$C_t = \max_c \{W_C\} \quad (4)$$

4.3 Modeling of sublexical units

Although usually an expert defines the set of SLUs, this method becomes very complex when the application is multilingual, resources are limited, or in shortage or under-resourced conditions Fig. 4. The lack of resources produces unwished effects in SLUs with very few samples. In consequence, it is necessary to define new hidden Markov model (HMM) topologies that can optimize the internal structure of the different sounds of the language. Two different HMM structure configurations have been tested:

1. In the first configuration, the HMMs have the same number of states (NS) for all the SLUs (NS-X, where X is the number of states used).
2. In the second configuration, the effect of assigning different number of states to each SLU is analyzed, where the selection of the number of states depends on the nature of the allophones (a similar approach can be found in [37]) as it is shown in Table 5. There are also groups of SLUs defined according to the sound class for the three languages. In some cases, certain allophones are grouped (joining acoustic models) by considering that different allophones are the same sound unit. Table 6 shows for each language: Basque-BS, Spanish-SP, and French-FR, the different groups taken into account for each of the languages under study (language is stated in the first column). The second column shows the name of the group type, and the third column describes the groups of allophones and their nomenclature. For example, /i/ = /i/+/j/ represents that a new /i/ SLU is created by joining the /i/ vowel model and the /j/ semi-vowel model. Thus, a new acoustic model /i/ will be trained with speech samples of the previous /i/ and /j/.

Finally, sets of triphonemes are created for all the selected allophonic options, and discrete and semicontinuous HMM are generated [38]. These triphonemes will be used in the acoustic phonetic decoding system as SLUs.

4.4 Modeling of lexical units

The *adiUP* system uses LUs both for creating the vocabulary for the ASR engine and for other elements of the system, such as the components of the folksonomy. These LUs are selected taking into account the nature of each language. In the case of the Basque language, the fundamental LU will be the lemma or a combination of lemmas, based on the pseudo-morphemes proposed in [31]. In the case of Spanish and French, the fundamental LU will be the word (simple or compound).

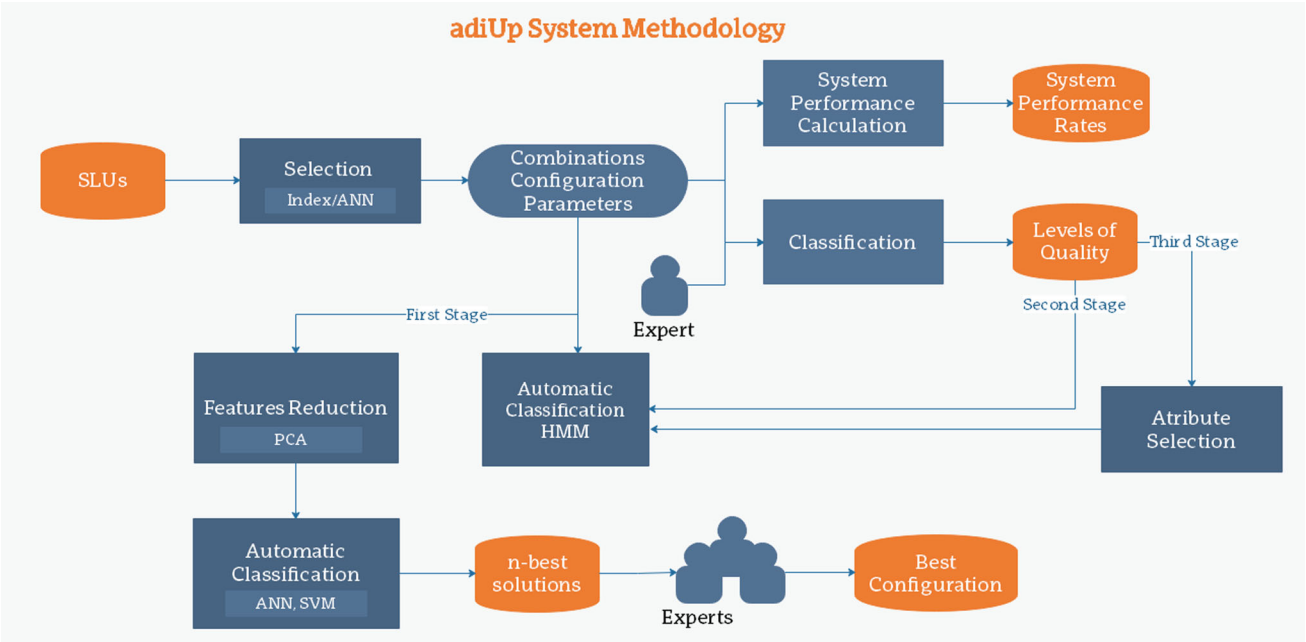

Fig. 4 General diagram of *adiUP* system

Table 5 Examples of HMM topologies for the SLUs of the three languages with different topologies and number of states (NS)

Description			
Languages	Topology	NS	SLU
BS	Bs-T5	5	/j/,/w/,/p/,/t/,/b/,/f/,/m/,/n/,/J/,/l/,/c/,/J\/,/r\/,/s_a/
		6	/a/,/e/,/i/,/o/,/u/,/d/,/F/,/N/,/L/,/r/
		7	/k/,/B/,/D/,/g /,/G/,/x/,/n_d/,/j\/,/z_a/,/s_m/,/S/,/T/,/ts_a/,/ts_m/,/tS/, /INS/
SP	Sp-T5	5	/p/,/t/,/k/
		6	/a/,/e/,/i/,/o/,/u/,/j/,/w/,/b/,/d/,/g/,/B/,/D/,/G/,/x/,/f/,,/m/,/F/,/n/,/N/,/n_d/,/J/,/l/,/L/,/j\/,/c/,/J\/,/r\/,/r/,/z_a/,/s_a/
		7	/T/,/S/,/s_m/,/ts_a/,/ts_m/,/tS/,/INS/
FR	Fr-T5	6	/p/,/t/,/k/,/m/,/n/
		7	/a/,/e/,/i/,/o/,/u/,/y/,/j/,/w/,/b/,/d/,/g/,/x/,/f/,/F/,/N/,/n_d/,/J/,/l/,/L/,/j\/,/c/,/J\/,/r\/,/r/,/z_a/,/s_a/
		8	/A/,/E/,/O/,/@/,/2/,/9/,/9～/,/a～/,/e～/,/o～/,/B/,/D/,/G/,/T/,/S/,/s_m/,/ts_a/,/ts_m/,/tS/,/INS/

Table 6 Description of examples of groups of SLUs for the three languages

Language	Group type	Description of different SLU groups			
BS	G-C	/i/=/i/+/j/	/b/=/b/+/B/	/m/=/m/+/F/	/s_a/=/s_a/+/z_a/
		/u/=/u/+/w/	/d/=/d/+/D/	/n/=/n/+/N/+/n_d/	/j\/=/j\/+/J\/+/c/+/L/
			/g/=/g/+/G/		
SP	G-B	/i/=/i/+/j/	/b/=/b/+/B/	/m/=/m/+/F/	/s_a/=/s_a/+/z_a/
		/u/=/u/+/w/	/d/=/d/+/D/	/n/=/n/+/N/+/n_d/	/j\/=/j\/+/J\/+/c/+/L/
			/g/=/g/+/G/		
FR	G-C	/a/=/a/+/a～/	/b/=/b/+/B/+/v/	/m/=/m/+/F/	/s_a/=/s_a/+/z_a/
		/e/=/e/+/E/+/e～/+/9～/	/d/=/d/+/D/	/n/=/n/+/N/+/n_d/	/s_anFR/=/s_anFR/+/z_anFR/
		/i/=/i/+/j/	/g/=/g/+/G/		/R\/=/R\/+/r/+/r\/
		/o/=/o/+/O/+/o～/			
		/u/=/u/+/w/+/y/+/H/+/@/+/2/+/9/			

4.5 Modeling of fillers

In a keyword-spotting system, the presence of words out of vocabulary (OOV) in the speech signal during the recognition process can produce unwanted performance results. One of the classic methods for solving this problem consists in using filler models to absorb this part of unwanted signal. Furthermore, these fillers can provide valuable information for the construction of hypothetical new vocabulary words or in vocabulary. This approach is really interesting in the case of under-resourced languages because there is a lack of words in vocabulary. On the one hand, the *adiUP* system tests fillers of type: *Allophonic*, allophone, triphoneme, etc.; *Syllabic*, syllables or combinations; *SLUs*, most frequent sublexicals or subwords based on words or lemmas, and morphemes of the language. On the other hand, in the recognition process a parameter is used in order to adjust word insertion penalty (WIP) and filler insertion penalty (FIP). The folksonomy completes the recognition process by extracting relevant semantic information in the form of concepts or conceptual phrases built from lexical unit. Moreover, the strong cross-lingual effect requires the creation of folksonomy elements which integrate concepts of the three languages.

5 Evolutionary and adaptive methodology for automatic selection of the system configuration

5.1 Analysis

The characteristics of a system determine its performance and its ability to evolve and improve, and to become more general in the long term. Nevertheless, the definition of these characteristics becomes tedious when multiple possibilities must be tested in order to reach optimal configurations Fig. 4. This task is crucial in the case of complex

environments where, due to minimal resources, experimental tests cannot objectively measure the performance of the system in the future. New data will be available during the auto-learning process in the collaborative network, and the system components could change, for instance, the SLUs, folksonomies, and LUs, which could lead to an increase in the robustness and performance of the system. Optimization methods for automatic selection of the system configuration provide a powerful tool that can forecast long-term changes. The *adiUP* system allows multiple system configuration options that could be generated by changing the components of the systems described in Table 7.

In our case, for the three languages there are about 6000 possible system configurations. Therefore, our goal is to automatically find optimal combinations of parameters (optimal configurations), while ensuring the quality of the system. In the literature, phone error rate (PER) and word error rate (WER) and word correct rate (WCR) have usually been used in order to assess the quality of ASR systems.

$$\text{PER/WER} = \frac{S + I + D}{N} \cdot 100 \tag{5}$$

$$\text{WCR} = \frac{C}{N} \cdot 100 \tag{6}$$

where S is the number of incorrect words substituted, I is the number of extra words inserted, D is the number of words deleted, C the number of correct words, and N is the total number of words in the correct transcription.

However, these measures strongly depend on the vocabulary and they only analyze some features of recognition, without measuring any features related to the quality of unit partitions, the mutual interaction between units, combined semantic quality features, or a combination of these factors. Therefore, there is not a comprehensive quality measurement of the system that would be very useful in the future, when there could be significant changes in the original components of the system. In this paper, we propose a new method for the selection and ranking of the features of the system in order to adjust the performance of the system, which is already in use in other fields [1, 39, 40], where we take into account not only the components of the system, but also their integration.

5.2 Methodology

The proposed methodology based on optimization algorithms, fuzzy indices, PCA, ANN and SVM is described below:

1. In a first stage, the quality of the sets of SLUs is ensured by an automatic selection based on several

Table 7 Inventory of components involved in the system configuration

Component	Type	Description
SLU type	Allophone	Allophones of the language (Table 1)
	Triphoneme	Triphonemes based on allophones
	G-X	Group of allophones or triphonemes based on Table D
HMM structure	DC	Discrete
	SC	Semicontinuous
HMM topology	XLG-TX	XLG = Language, TX, number of states based on Table C proposals
	NS-X	All SLUs the same number of states, X
	NG	Number of Gaussians
LUs type	Words	Word type unit
	Pseudo-morphemes	Unit based on morphemes, Lemmas and morphemes
Fillers	Allophonic	Allophone, triphoneme
	Syllabic	Syllables
	Subwords	Most frequent filler words in the languages, automatically calculated. Length
	Subword-2	subwords of length 2
	SYL-subword-2	Syllabic type and subwords of length 2
Recogn-adjustment	WIP	Word insertion penalty
	FIP	Filler insertion penalty
Folksonomies	WTc	Total weight for a class c
	WRc	Ratified weight for a class c

Table 8 System performance rates

Rate	Component involved	Description
APD	*Triph-PER*	PER using triphonemes as SLUs
	Triph-PCR	PCR using triphonemes as SLUs
adiUP	adiUP-Co	Concept correct rate in the *adiUP* system
	adiUP-Cl	Class correct rate in the *adiUP* system
LEXICON	w-WCR	WCR of the words present in the folksonomy
	w-WER	WER of the word present in the folksonomy
	WCR-mean	Mean of w-WCR
	WER-mean	Mean of w-WER
Folksonomy	kc-WCR	WCR of the key concepts defined in the folksonomy
	kc-WER	WER of the key concepts defined in the folksonomy
	kc-WCR-mean	Mean of kc-WCR
	kc-WER-mean	Mean of kc-WER

indices which analyze the quality of unit partitions, the mutual interaction between units, false positive and true negative rates, entropy, and similarities. More than 80 external and fuzzy indices have been analyzed, and the best eight are chosen for the selection process described in [41]. In this reference, the optimum ones are selected: IR, ISTL, IMAC, IIMN, IMN, IKAPPA, IHP, and IPE.

2. Then, for each of the combinations of configuration parameters, twelve system performance rates are calculated (described in Table 8). These rates are relative to the APD, the *adiUP* system, the lexicons, and the folksonomies.
3. During the second stage, an expert classifies all configurations in five levels of quality with regard to an Objective Function which consists in a linear combination of the aforementioned rates (Table 8). Several rates of the classes of the system are used in order to automatically classify the combinations of configuration parameters, because they play a key role according to the experts and the staff of the *Info7* radio channel. Finally, an automatic classification by using the so-called decision tree (DT) machine learning algorithm [42] is carried out. This classification is used as a reference.

In ANN case, multi-layer perceptron (MLP) with neuron number in hidden layer (NNHL) = (attribute number + classes/2) and training step (TS) NNHL*10. In SVM case, PolyKernel option has been used. For the training and validation steps, we used k-fold cross-validation with $k = 10$. Cross-validation is a robust validation for variable selection [1].

4. The third stage consists in the automatic selection of rates in order to get the optimal set of parameters. The most relevant features are selected by means of attribute selection algorithms, which are available in the WEKA Program [43]. In general, these algorithms can be classified by several criteria: filters or wrappers, and analysis of each attribute or group of attributes [44]. Then, in order to determine the most influential factors in the process (ASR or performance of the *adiUP* application), the rates are sorted according to the number of times that they are chosen by these selection methods, and a ranking of the system performance rates is obtained taking into account their repetitiveness in the appearance in the WEKA selection methods.
5. Once the ranking is obtained, the system performance rates outside the ranking are discarded, and the classification is repeated by using the DT machine learning algorithm in order to analyze significant changes in the results, and overall quality of the system.
6. Then the number of features (rates) is reduced by applying the principal component analysis (PCA) algorithm to the ten selected rates, and the *n-best* solutions for the configuration are selected by employing an Objective Function which is a linear combination of the PCA components.
7. In this step, artificial neural networks (ANN) and support vector machines (SVM) are used for the final automatic model selection.
8. Finally, a group of experts (developers, linguists, and staff of the *Info7* radio channel) supervises and selects the best configuration for each language from the *n-best* solutions provided by the previous stage.

6 Experimentation and discussion

The main aim of the experiments is to improve the performance of the designed system using the current resources, but keeping in mind the possibility of

incorporating new resources in the future. The experimentation is divided into two tasks: the automatic selection of the system configuration by using the methodology described in Sect. 5; and the analysis of usability with the *Info7* internet radio channel in order to assess the performance of the system under the design requirements.

6.1 Automatic configuration and performance of the system

Prior to evaluating the system by users, it is necessary to select the optimal configuration of the system, that is, to choose the best set of parameters in order to optimize the system performance. These parameters include the optimal unit set, the folksonomy, the topology, the LUs, the SLUs, the number of Gaussians, the word insertion penalty, the fillers, and the acoustic models, among others. The selection process is carried out by using the methodology described in Sect. 5. The parameters of the configuration of the system have been automatically selected and tuned by using 1000 sentences from recordings of the *Info7* radio channel [13], and the validation method was a 10-fold cross-validation. In some cases, for the selection of SLUs the leave-one-out cross-validation technique has been used (LOOCV).

6.1.1 Automatic selection of sublexical units

In the first stage, the optimal sets of SLUs are selected from the proposals described in Sect. 4. All parameters are also selected pursuing optimal performance of the system under noise conditions. Semicontinuous hidden Markov models with different topologies were used. The input signal is transformed and characterized with a set of 13 Mel Frequency Cepstral Coefficients (MFCC), energy and their dynamic components, by taking into account the high noise level of the signal (42 features). The frame period is 10 ms, the FFT uses a Hamming window, and the signal has first-order pre-emphasis applied using a coefficient of 0.97. The filter-bank has 26 channels. By using the methodology described in Sect. 5.2 and also voice activity detection (VAD) methodologies [45], the most robust sets of SLUs are selected for the three languages among 10,000 options. These best sets are based on SC-HMM and triphonemes.

6.1.2 Automatic selection of the system configuration

Then, the combinations of system configuration parameters are generated. There are about 6000 different combinations for the three languages of the *adiUP* system. The optimal set of parameters is obtained by using the methodology described in Sect. 5. That is, for each one of the combinations of configuration parameters, the twelve system performance rates described in Sect. 5 are calculated (Table 8). Afterward, an expert classifies all the configurations in five levels of quality with regard to the radio objectives, and the classification reference is calculated. The ranking of rates is carried out by using the DT machine learning algorithm, and then, the attribute selection algorithms yield Table 9 that shows the ten rates that have the greatest effect on the classification.

The other system performance rates are discarded, and the classification is repeated by using the DT machine learning algorithm without significant changes in the results, so we decided to keep using only the best rates.

In the next stage, new combinations of rates are generated by PCA, and the number of criteria is reduced. The result of the reduction by PCA is the creation of five groups that integrate different performance features. These groups are shown in Table 10. The group PC1 comprises information related to the WER of LUs; PC2 describes the general performance of the *adiUP* system; PC3 is related to the performance of SLUs; PC4 defines the performance of words; PC5 measures the performance of key concepts.

For each language, an Objective Function is created by a linear combination of PCA components, and the six best solutions are selected. Afterward, an automatic selection by ANNs and SVM is carried out and the best selected

Table 9 Rates that have the greatest effect on the classification, according to the attribute-selection algorithms

System performance rates	Repetitiveness
adiUP-Cl	8
adiUP-Co	7
kc-WER	7
w-WER	6
w-WCR	6
kc-WCR	6
kc-WER-mean	6
kc-WCR-mean	6
Triph-WER	6
Triph-WCR	6

Table 10 Results of the PCA algorithm

Principal component	Definition
PC1	c_{11} kc-WER-mean +…
PC2	c_{21} adiUP-Cl + c_{22} adiUP-Co +…
PC3	c_{31} Triph-WER + c_{32} Triph-WCR +…
PC4	c_{41} w-WER + c_{42} w-WCR +…
PC5	c_{51} kc-WER + c_{52} kc-WCR+…

Table 11 Three of the best configuration options selected for each language and their system performance % rates

Features						System rates							
Language	Topology	SLU	NG	WIP	Filler type	APD		A*diUP*		Key concepts		Words	
						Triph-PCR	Triph-PER	Cl	Co	Kc-WCR	Kc-WER	WCR	WER
BS	Bs-T5	Triphonemes	26	− 25	SYL-subword-2	**81,50**	21,10	80,50	84,67	**81,22**	35,47	82,80	**19,81**
	NS-6	G-C	10	− 10	subword-2	78,90	25,50	77,89	80,65	64,12	37,60	**86,29**	22,42
	Bs-T5	G-C	6	− 25	subword-2	80,50	23,95	**81,17**	**84,83**	**81,22**	34,50	85,98	21,74
SP	Sp-T5	Triphonemes	24	− 20	subword-2	53,25	52,5	**77,50**	79,88	63,36	54,48	82,97	24,78
	NS-7	G-B	8	− 10	subword-2	58,80	46,04	72,89	75,65	59,12	52,60	**83,97**	**21,48**
	NS-7	G-B	4	− 25	SYL-subword-2	**62,18**	**42,80**	76,14	**79,92**	**67,76**	**54,93**	81,90	23,12
FR	NS-8	Triphonemes	10	− 25	subword-2	46,33	58,28	65,30	66,88	53,26	**54,48**	61,07	40,78
	NS-7	G-C	8	− 10	SYL-subword-2	52,86	45,94	62,24	**69,20**	56,65	52,81	60,75	39,27
	Fr-T5	G-C	26	− 25	SYL-subword-2	**58,07**	**45,20**	**66,89**	67,85	**57,95**	52,11	**62,79**	**37,54**

solutions are supervised by a group of experts according to their performance on: traditional classification paradigms; experiments with semicontinuous HMM (SC-HMM) ASR systems; experiments with the *adiUP* system; and industrial interests. Finally, the optimal solution is implemented in the *adiUP* system. Table 11 shows the results of the success rate in percentage for three of the best options for each language (Basque-BS, Spanish-SP, and French-FR) using the chosen system configuration parameters selected from the 6000 possible options available at the beginning. Several proposals for the HMM structure were tested, not only by using different model topologies, but also with different numbers of Gaussians, and with different word insertion penalties. The features shown in Table 11 are the topology (as described in Table 5), or NS-X if the same number of states—X—has been used for all the SLUs); the type of SLU (as described in Tables 6 and 7); the number of Gaussians (NG); the value of the WIP parameter, and the type of filler in use (as described in Table 7); the WCR and the WER of the APD system, of the key concepts, and of the words (as described in Table 8); and the correct rates for classes (Cl) and concepts (Co) of the *adiUP* system (as described in Table 8). The best rates are marked in bold in Table 11.

It can be seen that the best results were obtained for Basque. Although Spanish has weaker acoustic models, the success rates in the *adiUP* system for words and concepts are good. The worst results are obtained for French, as it was expected due to the cross-lingual influence from the other two languages (French is not the native language of the radio-speaker). Note that the models that have a lot of states need fewer Gaussians, and in general they need fewer unit insertions because their configuration is appropriately adjusted. In some cases, the WER value of the APD system is very high, but this weakness is compensated by the folksonomy, achieving reasonable rates for the *adiUP* system and acceptable WER values for key concepts.

6.1.3 Evaluation oriented to users

An expert supervises the proposed solutions, and selects the final combination of configuration parameters for each of the three languages. Table 12 shows for each language the final configuration selected for the *adiUP* system.

The assessment of the final performance of the system was carried out both automatically by using the tags created during the analysis stage and manually by experts, mainly journalists. Figure 5 shows the results obtained from tests in the laboratory. These results of the performance of the *adiUP* system in terms of concepts (Co) and classes (Cl) fulfill the requirements defined by the clients.

Table 12 Configurations implemented in the *adiUP system*

Language	Features				
	Topology	SLU	NG	WIP	Filler
BS	Bs-T5	G-C	6	− 25	subword-2
SP	NS-7	G-B	4	− 25	SYL-subword-2
FR	Fr-T5	G-C	26	− 25	SYL-subword-2

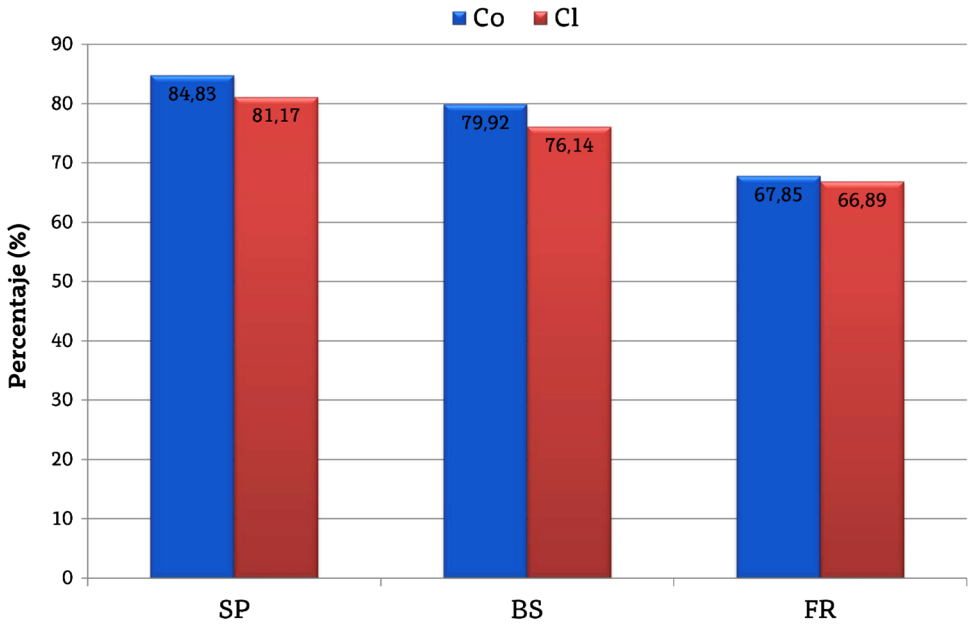

Fig. 5 Performance of the final system in terms of concept correct rate (*adiUP*-Co) and class correct rate (*adiUP*-Cl)

6.2 Usability

Finally, in order to evaluate the usability of the system, it was also tested by people not familiarized with this kind of systems. Tests have been carried out with four types of users: One person that uses the system as a super-user; three expert workers of the internet radio channel; three collaborators of the project; and three external real users of the system. These people have used the system and have subjectively graded it with regard to: the interface of the application; the answers provided by the search engine; the updating of the system; the flexibility of the system; and the general information provided by the system. Real users have only been inquired about this last general issue. The subjective parameters represent the degree of satisfaction with regard to the outcome of the system and the interface. This satisfaction was measured in a range of 0–10, where 0 is the lowest satisfaction level, and 10 is the highest satisfaction level with regard to the application performance: interface, perplexity, level of confusion, information, search usefulness, updating-adaptation, and information level. The tests and usability heuristics were designed by following the methodology described in [46–48]. Figure 6 summarizes the results of the tests that fulfill the design requirements.

7 Concluding remarks

Neural computing-based approaches provide flexible solutions for complex systems. This paper presents the design and development of the *adiUP* system, a multilingual audio information management system in complex environments based on a scalable architecture of automatic methodologies for semantic knowledge analysis. The neural computing-based approaches integrate: artificial neural networks (ANN), support vector machines (SVM), and hidden Markov models (HMM). The complex environment of internet mass media is given by: the trilingual context (Basque, Spanish, and French); the regular appearance of cross-lingual elements between the three languages; the under-resourced situation of Basque because the Basque language in the French State is not an official language and has a low number of speakers; the limited resources (financial, material, human, and audio resources), specially the audio corpora for training is only about 0.5–1% of the usual corpora for such systems; and the extremely poor quality of the audio signal, taken from an internet radio channel, that increases the complexity of the development.

The *adiUP* system has been tested with a local radio channel, the trilingual *Info7* radio channel (Basque, Spanish, and French) [13], and the design specifications have been successfully fulfilled. As a result, a multilingual ASR system has been created which is able to: handle under-resourced languages with a very small audio corpus; select automatically attributes by reducing the redundant data; reduce unwished effects in sublexical units (SLU) due to the low number of samples; and extract information by means to folksonomies. Additionally, this is an evolutionary system where new information resources can be introduced in a more open and collaborative way by users, that has been developed using universal design criteria. In

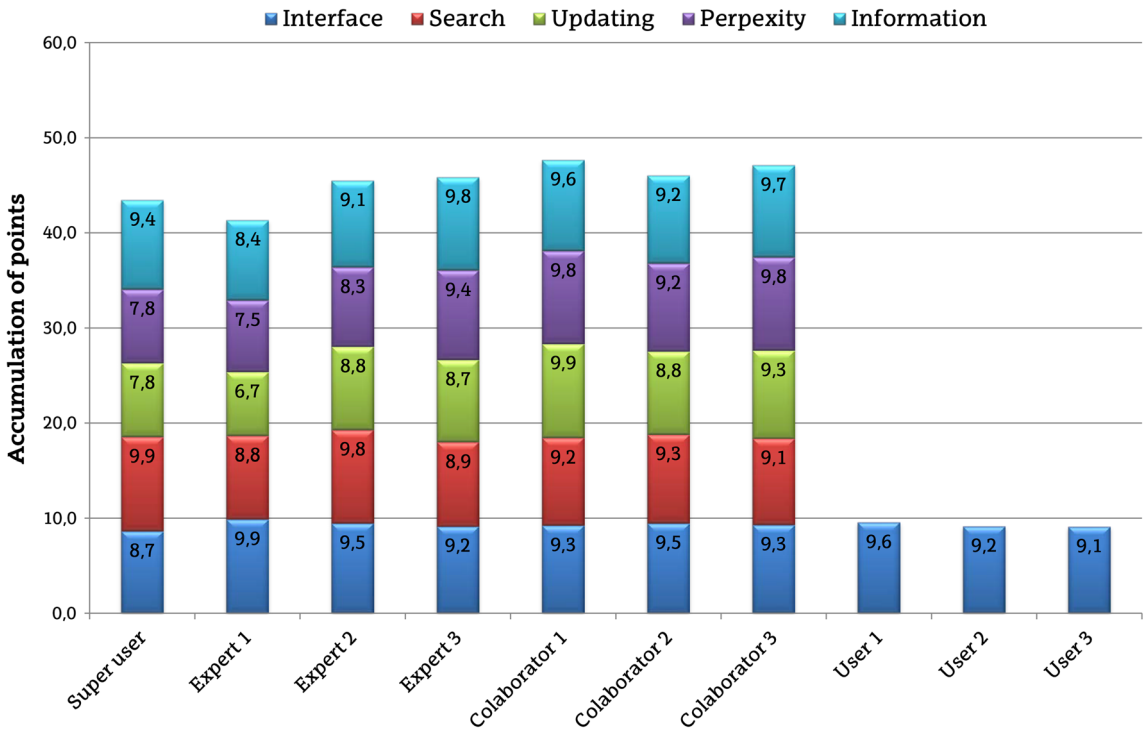

Fig. 6 Summary of the results of the usability tests for each user profile

future research lines, other under-resourced languages and media can be integrated.

Acknowledgements This work is being funded by Grants: TEC2016-77791-C4 from Plan Nacional de I + D + i, Ministry of Economic Affairs and Competitiveness of Spain and from the DomusVi Foundation Kms para recorder, the Basque Government (ELKARTEK KK-2018/00114, GEJ IT1189-19, the Government of Gipuzkoa (DG18/14 DG17/16), UPV/EHU (GIU19/090), COST ACTION (CA18106, CA15225).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Barroso N (2011) Ph.D. Thesis in Basque: contributions to the management of semantic information in complex audio environments. Department of Systems Engineering and Automation, University of the Basque Country (UPV/EHU), Donostia, Basque Country
2. Lopez de Ipiña K, Torres I, Oñederra L, Varona A, Ezeiza N (2000) First selection of lexical units for continuous speech recognition of Basque. In: Proceedings of ICSLP, vol 2, pp 531–535. Beijing
3. Ezeiza A, Lopez-de-Ipiña K, Hernández C, Barroso N (2013) Enhancing the feature extraction process for automatic speech recognition with fractal dimensions. Cogn Comput 5(4):545–550
4. Lopez-de-Ipiña K, Alonso JB, Solé-Casals J, Barroso N, Henriquez P, Faundez-Zanuy M, Travieso CM, Ecay-Torres M, Martinez-Lage P, Eguiraun H (2015) On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. Cogn Comput 7(1):44–55
5. Faundez-Zanuy M, Hussain A, Mekyska J, Sesa-Nogueras E, Monte-Moreno E, Esposito A, Chetouani M, Garre-Olmo J, Abel A, Smekal Z, Lopez-de-Ipiña K (2013) Biometric applications related to human beings: there is life beyond security. Cogn Comput 5(1):136–151
6. Lopez-de-Ipiña K (2013) Ph.D Thesis in Basque: automatic continuous speech recognition for Basque by means of stochastic

models. Department of Computational Science and Artificial Intelligence, University of the Basque Country (UPV/EHU). Donostia, Basque Country
7. Kim J, Urbano J, Liem CCS, Hanjalic A (2019) One deep music representation to rule them all? A comparative analysis of different representation learning strategies. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04076-1
8. Tran SN, Ngo S, d'Avila A (2019) Probabilistic approaches for music similarity using restricted Boltzmann machines. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04106-y
9. Guruler H (2016) A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. Neural Comput Appl. https://doi.org/10.1007/s00521-015-2142-2
10. López-de-Ipiña K, Martinez-de-Lizarduy U, Calvo PM, Beitia B, García-Melero J, Fernández E, Ecay-Torres M, Faundez-Zanuy M, Sanz P (2018) On the analysis of speech and dysfluencies for automatic detection of mild cognitive impairment. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3494-1
11. Mustafa MK, Allen T, Appiah K (2017) A comparative review of dynamic neural networks and hidden Markov model methods for mobile on- device speech recognition. Neural Comput Appl 1:2–3. https://doi.org/10.1007/s00521-017-3028-2
12. The adiUP system. The application and generated resources. Uni. of the Basque Country. http://www.adiUP.info. Accessed 9 Nov 2019
13. Info7. Internet radio channel of Basque country. Available: http://www.info7.com/. Accessed 9 Nov 2019
14. Gauvain JL, Lamel L (2002) Adda G (2002) The LIMSI broadcast news transcription system. Speech Commun 37(1–2):89–108
15. Barroso N, Lopez-de-Ipiña K, Ezeiza A, Hernandez C, Ezeiza N, Barroso O, Susperregi U, Barroso S (2011) GorUp: an ontology-driven audio information retrieval system that suits the requirements of under-resourced languages. In: Proceedings of Interspeech2011. Florencia, Italia
16. Anusuya M, Katti S (2011) Front end analysis of speech recognition: a review. Int J Speech Technol 14(2):99–145
17. Beyerlein P, Aubert XL, Haeb-Umbach R, Harris M, Klakow D, Wendemuth A, Molau S, Pitz M, Sixtus A (2002) Largevocabulary continuous speech recognition of broadcast news—thephilips/RWTH approach. Speech Commun 37(1–2):109–131
18. Lin H, Ou Z (2006) Partial-tied-mixture auxiliary chain models for speech recognition based on dynamic bayesian networks. In: IEEE international conference on systems, man and cybernetics 2006, p 4415–4419. Taipei, Taiwan
19. Huijbregts M, de Jong F (2011) Robust speech/non-speech classification in heterogeneous multimedia content. Speech Commun 53(2):143–153
20. Schepens J, Dijkstra T, Grootjen F, van Heuven WJB (2013) Cross-language distributions of high frequency and phonetically similar cognates. PLoS ONE 8(5):e63006. https://doi.org/10.1371/journal.pone.0063006
21. Kanthak S, Ney H (2001) Context dependent acoustic modelling using graphemes for large vocabulary speech recognition. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing 2001, p. 845–848. Orlando, Florida, US
22. Le VB (2009) Besacier L (2009) Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans Audio Speech Lang Process 17(8):1471–1482
23. Seng S, Sam S, Le VB, Bigi B, Besacier L (2008) Which units for acoustic and language modelling for Khmer automatic speech recognition. In: Proceedings of 1st international workshop on spoken languages technologies for under-resourced languages 2008. Hanoi, Vietnam
24. Gales MJF, Knil KM, Ragni A, Rath SP (2014) Speech recognition and keyword spotting for low resource languages: Babel project research at CUED. In: Proceedings of 4th international workshop on spoken languages technologies for under-resourced languages 2014, pp 16–23. St. Petersburg, Russia
25. Schlippe T, Quaschningk W, Schultz T (2014) Combining grapheme-to-phoneme convertor outputs for enhanced pronunciation generation in low-resource scenarios. In: Proceedings of 4th international workshop on spoken languages technologies for under-resourced languages 2014, pp 139–145. St. Petersburg, Russia
26. Barnard E, Davel M, Van Heerden C, de Wet F,Badenhorst J (2014) The NCHLT speech corpus of the South African languages. In: Proceedings of 4th international workshop on spoken languages technologies for under-resourced languages 2014, pp. 194–200. St. Petersburg, Russia
27. Vakil A, Palmer A (2014) Cross-language mapping for small-vocabulary ASR in under-resourced languages: investigating the impact of source language choice. In: Proceedings of 4th international workshop on spoken languages technologies for under-resourced languages 2014, pp 169–175. St. Petersburg, Russia
28. STLU 2014. In: The 4th international workshop on spoken languages technologies for under-resourced languages, 2014. St Petersburg, Russia. http://www.mica.edu.vn/sltu2014. Accessed 9 Nov 2019
29. Besacier L, Barnard E, Karpov A (2014) Schultz T (2014) Introduction to the special issue on processing under-resourced languages. Speech Commun 56:83–84
30. Rousseau A, Deléglise P, Estève Y (2012) TED-LIUM: an automatic speech recognition dedicated corpus. In: Proceedings of 8th international conference on language resources and evaluation, pp 125–129, 2012. Istanbul, Turkey
31. Lopez-de-Ipiña K, Torres I, Oñederra L, Varona A, Ezeiza N, Peñagarikano M, Hernández M, Rodriguez LJ (2000) First selection of lexical units for continuous speech recognition of Basque. In: Proceedings of 7th international conference on spoken language processing, InterSpeech 2000, vol 2, pp 531–535. Beijing, China
32. Barroso N, Lopez-de-Ipiña K, Ezeiza A, Hernandez C (2013) Language identification for internet security in the Basque context: a cross-lingual approach. Aerosp Electron Syst Mag 28(8):24–31
33. Silipo R, Berthold MR (2000) Input features' impact on fuzzy decision processes. IEEE Int Conf Syst Man Cybern B Cybern 30(6):821–834
34. Fillinger A (2008) NIST speech signal to noise ratio measurements. Technical Report Information Technology Laboratory, National Institute of Standards and Technology, US Department of Commerce; 2008, http://www.nist.gov/smartspace/nist_speech_snr_measurement.html. Accessed 9 Nov 2019
35. Kim C, Stern RM (2008) Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Proceedings of 9th annual conference of the international speech communication association, Interspeech 2008. Springer, Berlin, pp 2598–2601
36. Insima Teknologia S.L.L. Company of Donostia-San Sebastian, Basque country. http://www.yildun-backup-remoto.com. Accessed 9 Nov 2019
37. Puertas JI (2000) Ph.D. Thesis in Spanish: robustness of phonetic speech recognition for telephone applications. Department of Signals, Systems and Radiocommunication. Madrid Polytechnic University (UPM). Madrid, Spain
38. HTK - Hidden Markov Model Toolkit - Speech Recognition toolkit. Cambridge University Engineering Department (CUED). http://htk.eng.cam.ac.uk. Accessed 9 Nov 2019

39. Tadjudin S, Landgrebe D (1999) Covariance estimation with limited training samples. In: IEEE transactions on geoscience and remote sensing symposium, Seattle, WA, vol 37, no 4, pp 2113–2118
40. Martinez A, Kak A (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
41. Barroso N, Lopez-de-Ipiña K, Hernandez C, Ezeiza A (2011) Design of multi-feature class models for speech recognition security systems with under-resourced languages. In: Proceedings of 45th IEEE Carnahan conference on security technology 2011. Mataro, Spain
42. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
43. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software [Internet]. University of Waikato. http://www.cs.waikato.ac.nz/ml/weka. Accessed 9 Nov 2019
44. Quinlan JOR (1993) C4.5: programs for machine learning. Morgan Kaufman Publishers, Boston
45. Solé J, Zaiats V (2010) A non-linear VAD for noisy environment. Cogn Comput 2(3):191–198
46. Witten I, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers, Boston
47. Hix D, Hartson H (1993) Developing user interfaces: ensuring usability through product and process. Wiley, New York
48. Nielsen J (1993) Usability engineering. AP professional. Academic Press, Boston