ACOUSTIC MODELLING FOR SPEECH PROCESSING IN COMPLEX ENVIRONMENTS

Nora Barroso, Karmele López de Ipiña and Aitzol Ezeiza
Polytechnic College, Department of Systems Engineering and Automation, University of the Basque Country,
Europa Plaza 1, Donostia 20018, Spain

Keywords: Multivariate Hidden Markov Models, Automatic Speech Recognition, Acoustic Modelling, Complex Environments, Speech Processing.

Abstract: Automatic Speech Recognition (ASR) is one of the classical multivariate statistical modelling applications that involves dealing with issues such as Acoustic Modelling (AM) or Language Modelling (LM). These tasks are generally very language-dependent and require very large resources. This work is focused on the selection of appropriate acoustic models for Speech Processing in a complex environment (a multilingual context in under-resourced and noisy conditions) oriented to general ASR tasks. The work has been carried out with a small trilingual speech database with very low audio quality. Thus, in order to decrease the negative impact that the lack of resources has in this task there have been selected two techniques: In the one hand, Hidden Markov Models have been enhanced using hybrid topologies and parameters as acoustic models of the sublexical units. In the other hand, an optimum configuration has been developed for the Acoustic Phonetic Decoding system, based on multivariate Gaussian numbers and the insertion penalty.

1 INTRODUCTION

Appropriate speech signal resources are required for multivariate statistical modelling applications, such as Hidden Markov Models for Automatic Speech Recognition (ASR). These applications require very large or optimum resources for training all the components of the system. Most of the ASR applications have to be developed for complex environments and in under-resourced conditions. Therefore, the development of a robust system for under-resourced languages, even if they are integrated in multilingual regions or coexist geographically with languages in best conditions needs high inversions (Le and Besacier, 2009; Seng et al., 2008; Barroso et al., 2007; Schultz. and Waibel, 1998)

These languages have the required resources in a very limited quantity and quality, and new strategies must be explored to tackle the challenge of creating robust systems in this area. Automatic Speech Recognition is a broad research area that absorbs many efforts from the research community. Indeed, many applications related to ASR have progressed quickly in recent years, but these applications are generally very language-dependent. Moreover, the creation of a robust system is a much tougher task for under-resourced languages, even if they count with powerful languages beside it. The development of these systems involves issues such as Acoustic Modelling, Language Modelling, and the development of Language Resources.

Figure 1 shows the classical scheme of an ASR system. During the Acoustic Phonetic Decoding (APD) stage, the speech signal is segmented into fundamental acoustic units that will be integrated into other system components. Classically, words have been considered the most natural unit for speech recognition, but the large number of potential words in a single language, including all inflected and derived forms, might become intractable in some cases for the definition of the basic components. In these cases, smaller phonologic recognition units like phonemes, triphonemes or syllables are used to overcome this problem. These recognition units, which are shorter than complete words are the so-called sublexical units (SLU) or sub-word units.

Multivariate Hidden Markov Models are undoubtedly the most employed technique for ASR construction, but when the data is affected by a high noise level, ASR systems are sometimes far from an

Barroso N., López de Ipiña K. and Ezeiza A..
ACOUSTIC MODELLING FOR SPEECH PROCESSING IN COMPLEX ENVIRONMENTS.
DOI: 10.5220/0003894105070516
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (MPBS-2012), pages 507-516
ISBN: 978-989-8425-89-8

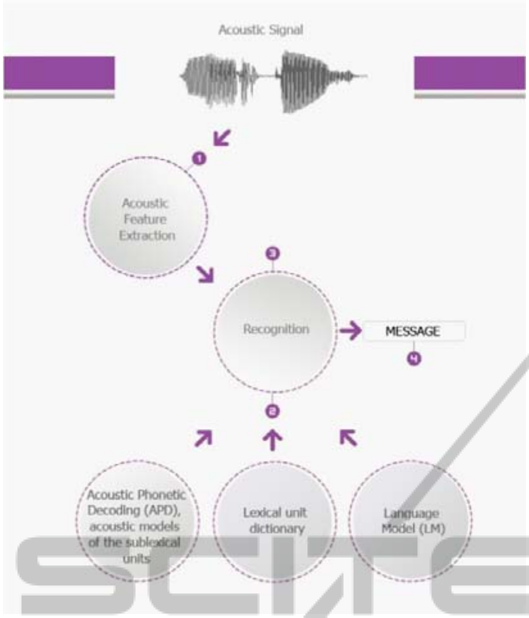
Acoustic Signal
1
Acoustic Feature Extraction
3
Recognition
MESSAGE
4
2
Acoustic Phonetic Decoding (APD), acoustic models of the sublexical units
Lexical unit dictionary
Language Model (LM)

Figure 1: Scheme of the process of Speech Recognition.

acceptable performance (Le and Besacier, 2009; Smith, 2000). Several hybrid proposals can be found in the bibliography (Smith, 2000; Cosi, 2000, Friedman, 1989; Martinez, 2001) dealing with this problem based on Machine Learning paradigms such as: Support Vector Machines or Neural Networks.

Indeed, the long term goal of this project is the development of robust ASR systems in real and complex environments, but this work is focused to the selection of appropriate acoustic models with under-resourced and noisy conditions. Specifically, the current work is oriented to Broadcast News (BN) in the Basque context. Thus, in order to decrease the negative impact that the lack of resources has in this issue several hybrid methodologies are applied for Acoustic Modelling.

The paper is organised as follows: next section describes the resources and the features of the languages. Section 4 presents the used methods, mainly oriented to the selection of appropriate acoustic models in complex environments. Section 5 gives the experimentation results, and finally, some conclusions are explained in section 6.

2 SPEECH RESOURCES

The basic audio resources used in this work have been mainly provided by Broadcast News sources. Specifically *Infozazpi* radio (Infozazpi, 2011), a trilingual (Basque (BS), Spanish (SP), and French (FR)) has provided 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 distinct speakers, noises, and sections of the speech files and transcriptions. The transcriptions for Basque also include morphological information such as each word's lemma and Part-Of-Speech tag. The Resources Inventory is summarised in table 1.

Table 1: Resources inventory.

Languages	BCN Audio hh:mm:ss
BS	2:55:00
FR	2:58:00
SP	3:02:00
Total	7:55:00

In the audio for French and Spanish there is a high background noise from the signature tunes of the programmes. For Basque, there is no signature tune mixing with the speakers' voices. In the other hand, they are radio study recordings, so in consequence there are very few instantaneous noises. Most of the noises are common ones. In the transcription process, there have been mainly labelled the inspirations produced by the speakers, the instantaneous noises from the microphone, and the noises of the movement of papers.

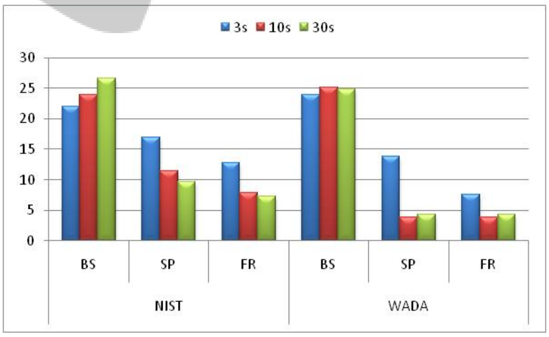

Figure 2: NIST and WADA noisy level with regard to the signal length.

The noise level and its effect over the signal is a crucial factor. In order to measure it, the standards NIST and WADA measures have been employed (Ellis, 2011). Figure 2 presents the results of these measures for Basque, Spanish, and French. It is straightforward to infer a higher level of noise for these last two corpora, being the French the noisiest corpus. For Spanish and French, the cleanest segments are those of 3 seconds. For Basque, the longest segments have the lowest level of noise. The two methodologies, NIST and WADA, use different calculation methods, so the results are different in level, but the general trends are kept similar in both.

Table 2: Sound Inventories for Basque, French and Spanish in the SAMPA notation.

Sound Type	Basque	French	Spanish
Plosives	p b t d k g c	p b t d k g	p b t d k g
Affricates	ts ts´ tS		ts
Fricatives	gj jj f B T D s s´ S x G Z v h	f v s z S Z	gj jj F B T D s x G
Nasals	m n J	m n J N	m n J
Liquids	l L r rr	l R	l L r rr
Vowel glides	w j	w H j	w j
Vowels	i e a o u @	i e E a A O o u y 2 9 @ e~ a~ o~ 9~,	i e a o u

3 FEATURES OF THE LANGUAGES

The analysis of the features of the languages chosen is a crucial issue because they have a clear influence on both the performance of the APD and on the vocabulary size of the system. In order to develop the APD, an inventory of the sounds of each language was necessary. Table 2 summarises the sound inventories for the three languages expressed in the SAMPA notation. Each sound would be taken into account in the phonetic transcription tools used in the training process.

In order to get an insight of the phonemes system of these three languages, we would like to remark some of the features mentioned above. In the one hand, Basque and Spanish have very similar vowels if not the same. The Basque language itself has many odd occurrences of other vocals, but many of them have fallen into disuse or they are used only in very local environments.

For example, only Basque speakers from the Northern side (bilingual Basque and French speakers) are used to pronouncing the Basque “@” (i.e. Sorrapürü). This vowel's pronunciation is between the Basque vocals “u” and “i”. In comparison to Basque or Spanish, French has a very much richer vocal system, but it is fair to say that some of their older forms have fallen into disuse too. Anyway, they keep on being different to those in Basque or Spanish, especially in the case of nasal vowels.

In the other hand, some of the consonants that are rare in French such as “L” (i.e. Feuille) are very common in Basque or Spanish. Therefore, a cross-lingual Acoustic Model could be very useful in these cases. Another special feature in this experiment is the richness of affricates and fricatives present in Basque.

These sounds will be very difficult to differ and the cross-lingual approach won't work for them, but it has to be said that even some native Basque speakers don't make differences between some affricates and fricatives due to dialectal issues. Consequently, the Acoustic decoder would have difficulties in these cases and further Language Modelling would be needed in order to get accurate results.

Finally, some sounds that are differentiated theoretically are very difficult to model, and many state-of-the-art approaches cluster these cases as the same sound. This is the case of the plosives in the three languages; there is little acoustic difference between “b”, “B”, “p”, and “P” depending on the context, and the Language Model should be able to manage the ambiguity derived of not separating those phonemes in this first stage.

4 METHODS

4.1 Multivariate Hidden Markov Models

In ASR classical Acoustic Modelling is carried out by Hidden Markov Models (HMMs) (Baum and Eagon, 1967; Baum et al., 1970; Baker, 1975; Jelinek, 1976).

The basic components of the HMMs are the states and the transitions between states. A Markov chain is, in short, a set of states and a set of transitions between them. Every state has a symbol and every transition has a probability associated to it. In each instant *t*, the system is in a certain state, and to regular intervals of time it goes on from one state to another as the transitions indicates. Finally, a symbol sequence is obtained from each of the states which have been crossed. HMMs are similar to Markov chains but, in this case, every state isn’t associated to a fixed symbol, but the event provided by the state forms a part of a probabilistic function. Thus, all the symbols are possible in every state and each one has its own probability. Consequently, an HMM consists of a not observable "hidden" process (Markov chain), and an observable process, which connects the input with the state of the hidden process. In order to do this, an HMM has to be a process doubly stochastic since it has, on the one hand, a set of transition probability coefficients that determine state sequence to continue and, on the other hand, there are defined probability functions associated with every state that determine the output that is observed in this state.

An HMM is defined as (Rabiner, 1989):

- q_t: state as the *t* time.
- *N*: the number of states in the model. The most used HMMs have 5 states. However both 1 state and 5 states do not generate any output.
- *S*: The individual states, $S=\{s_1,s_2,...,s_N\}$.
- *M*: the number of distinct observation symbols per state, i.e., the discrete alphabet size.
- *V*: The individual symbols, $V=\{v_1,v_2,...,v_M\}$.
- $A=\{a_{ij}\}$:The state transition probability distribution where, $a_{ij}=P(q_t=s_j \mid q_t-1=s_i)\quad 1\le i,j\le N$.
- $B=\{b_j(k)\}$: The observation symbol probability distribution in where, $b_j(O_k)=P(O_k \mid q_t=s_j)\quad 1\le j\le N, 1\le k\le M$, where O_k is a symbol of the observation set *V*.
- $\pi=\{\pi_i\}$: The initial state distribution, where: $\pi_i=P(q_0=s_i)\quad 1\le i\le N$.

Thus, an HMM is described by:

$$\lambda=\{A,B,\pi\}.$$

In the acoustic modelling, the likelihood of the observed feature vectors is computed given the linguistic units. Gaussian Mixture Model (GMM) classifiers can be used to compute for each HMM state *q*, corresponding to a SLU unit, the likelihood of a given feature vector given this phone $p(o|q)$ where Gaussians are multivariate. A way of thinking about the output of this stage is as a sequence of probability vectors, one for each time frame, and each vector at each time frame contains the likelihoods that each unit generated the acoustic feature vector observation at that time. Finally, in the APD phase, the acoustic model (AM) is employed. The AM consists of the sequence of acoustic likelihoods integrated in an HMM dictionary of Lexical Unit (LU) pronunciations, and then it is combined with the language model (LM). The output of the APD system is the most likely sequence of LUs. An HMM dictionary, is a list of LUs pronunciations, each pronunciation represented by a string of phones. Each LU can then be thought of as an HMM, where the SLUs are states in the HMM, and the Gaussian likelihood estimators supply the HMM output likelihood.

4.2 Selection of Sublexical Units and Acoustic Models

The lack of resources produces unwished effects in SLUs with very few samples. In consequence, it is necessary to define new HMM topologies that could optimize the own internal structure of the different sounds of the language. Two different HMM structure configurations were tested:

1. The HMMs had the same state number (SN) for all of the SLUs (EEK allSN)
2. Then it was analysed the effect of assigning different SNs to each SLUs with regard to its nature. For this purpose the classification was based on (Puertas, 2000).
    a. Vowels: 5 states
    b. Semivowels: 4 states
    c. Unvoiced plosives: 5 states
    d. Voiced plosives: 3 states
    e. /B/, /D/,/G/, /l/,/z_a/,/r\/: 4 states
    f. /s_a/, /R\/: 5 states

Table 3 shows the defined topologies for each language: Basque, Spanish, and French (1st column). In the second column the nomenclature is defined, in the third the SN and in the last one the allophones of each set.

Several HMM structure proposals were tested, not only by different model topologies, but also by the Gaussians Number (GN). In a first phase, only one Gaussian is used. Then, a range from 1 to 50 Gaussians was explored. Finally, the unit insertion penalty *p* was also adjusted implementing experiments for a range from -1 to -60.

Usually, an expert defines the broad sub-word classes needed during Acoustic Modelling. This method becomes very complex in the case of multilingual ASR tasks with incomplete or small databases. In search of alternatives, it is interesting to explore the data-driven methods that generate SLUs broad classes based on the confusion matrices of the SLUs. The similarity measure is defined using the number of confusions between the master sub-

Table 3: HMM topologies for the SLUs of the three languages with different topologies and State Number (SN).

		Description	
Language	**Top.**	**SN**	**SLU**
BS	EEK 1	3	/b/,/d/,/g/
		4	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		5	/a/,/e/,/i/,/o/,/u/,/p/,/t/,/k/,/x/,/f/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		6	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 2	4	/b/,/d/,/g/
		5	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		6	/a/,/e/,/i/,/o/,/u/,/p/,/t/,/k/,/x/,/f/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		7	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 3	5	/j/,/w/,/p/,/t/,/k/,/b/,/d/,/f/,/x/,/m/,/n/,/J/,/l/,/c/,/J\/,/r\/,/s_a/,/z_a/,/s_m/, /S/
		6	/a/,/e/,/i/,/o/,/u/,/B/,/D/,/g/,/G/,/F/,/N/,/n_d/,/L/,/j\/,/r/,/T/
		7	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 4	5	/j/,/w/,/p/,/t/,/k/,/b/,/d/,/f/,/x/,/m/,/n/,/J/,/l/,/c/,/J\/,/r\/,/s_a/,/z_a/
		6	/a/,/e/,/i/,/o/,/u/,/B/,/D/,/g/,/G/,/F/,/N/,/n_d/,/L/,/j\/,/r/
		7	/s_m/,/S/,/T/,/ts_a/,/ts_m/,/tS/,/INS/
	EEK 5	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	EEK 1	3	/b/,/d/,/g/
		4	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		5	/a/,/e/,/i/,/o/,/u/,/y/,/p/,/t/,/k/,/x/,/f/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		6	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 2	4	/b/,/d/,/g/
		5	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		6	/a/,/e/,/i/,/o/,/u/,/y,/p/,/t/,/k/,/x/,/f/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		7	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 3	4	/p/,/t/,/k/,/B/,/f/
		5	/a/,/e/,/i/,/o/,/u/,/y,/j/,/w/,/b/,/d/,/g/,/D/,/G/,/x/,/m/,/F/,/n/,/N/,/n_d/,/J/,/l/,/L/,/j\/,/c/,/J\/,/r\/,/r/,/z_a/,/s_a/,/S/,/s_m/, /T/
		6	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 4	5	/j/,/w/,/p/,/t/,/k/,/b/,/d/,/g/
		6	/a/,/e/,/i/,/o/,/u/,/y,/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/
	EEK 5	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	EEK 1	3	/b/,/d/,/g/
		4	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		5	/a/,/e/,/i/,/o/,/u/,/y/,/A/,/E/,/O/,/@/,/2/,/9/,/9~/,/a~/,/e~/,/o~/,/p/,/t/,/k/,/x/,/f/,/v/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		6	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 2	3	/b/,/d/,/g/
		4	/j/,/w/,/B/,/D/,/G/,/l/,/r\/,/z_a/
		5	/a/,/e/,/i/,/o/,/u/,/y/,/A/,/E/,/O/,/@/,/2/,/9/,/9~/,/a~/,/e~/,/o~/,/p/,/t/,/k/,/x/,/f/,/v/,/m/,/F/,/n/,/N/,/n_d/,/J/,/L/,/j\/,/c/,/J\/, /r/, /s_a/,/S/,/s_m/,/T/
		6	/ts_a/,/ts_m/,/tS/,/INS/
	EEK 3	6	/p/,/t/,/k/,/B/,/f/
		7	/a/,/e/,/i/,/o/,/u/,/y/,/A/,/E/,/O/,/@/,/2/,/9,/j/,/w/,/b/,/d/,/g/,/D/,/G/,/x/,/m/,/F/,/n/,/N/,/n_d/,/J/,/l/,/L/,/j\/,/c/,/J\/,/r\/,/r/,/z_a/,/s_a/,/S/,/s_m/,/T/
		8	/9~/,/a~/,/e~/,/o~/,/ts_a/,/ts_m/,/tS/,/INS/
	EEK 4	6	/j/,/w/,/p/,/t/,/k/,/b/,/d/,/g/
		7	/a/,/e/,/i/,/o/,/u/,/y/,/A/,/E/,/O/,/@/,/B/,/D/,/G/,/x/,/f/,,/m/,/F/,/n/,/N/,/n_d/,/J/,/l/,/L/,/j\/,/c/,/J\/,/r\/,/r/,/z_a/,/s_a/
		8	/2/,/9/,/9~/,/a~/,/e~/,/o~/,/T/,/S/,/s_m/,/ts_a/,/ts_m/,/tS/,/INS/
	EEK 5	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/

word unit and all other units included in the set, in this case the global *Phone Error Rate* (*PER*). In our approach the confusion matrices were calculated by several methods oriented to data optimization with small databases (Barroso, 2011a). The grouping for the three languages is summarized in the table 4.

Finally, the training and testing methodology was K-fold Cross Validation. This is one way to improve the results of classical train/test methods when the data set is small. The data set is divided into k subsets, and the training and test method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. Leave-One-Out (LOO), cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. In this work more than one value of K has been used, but the results show the case of K=10.

5 EXPERIMENTATION

The input signal is transformed and characterized with a set of 13 Mel Frequency Cepstral (MFCC), energy and their dynamic components, taken into account the high level of noise in the signal (42 features). The frame period was 10 milliseconds, the FFT uses a Hamming window and the signal had first order pre-emphasis applied using a coefficient of 0.97. The filter-bank had 26 channels. Then, automatic segmentation of the SLU units (phonemes) is generated by Semi Continuous HMM (SC-HMM). In a first stage, the Gaussian Number is 1 and no SLU data-driven or unit fusion is used. The set of units is the described in the Table 2.

When the SN of the HMM (with regard to the SLUs nature) are changed (configurations EEK1-EEK5), it can be observed an ascending progression in the value of Accuracy (Acc). EEK1 is the configuration based on the proposal of (Puertas, 2000). EEK2 is similar to EEK1 adding one more

Table 4: Description of the different groups of Sublexical Unit for the three languages.

Language	Group Type	Description			
	A	/i/=/i/+/j/ /u/=/u/+/w/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	
BS	**B**	/i/=/i/+/j/ /u/=/u/+/w/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a/=/s_a/+/z_a/ /j\/=/j\/+/J\/+/c/ /l/=/l/+/L/
	C	/i/=/i/+/j/ /u/=/u/+/w/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a/=/s_a/+/z_a/ /j\/=/j\/+/J\/+/c/+/L/
SP	**A**	/i/=/i/+/j/ /u/=/u/+/w/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	
	B	/i/=/i/+/j/ /u/=/u/+/w/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a/=/s_a/+/z_a/ /j\/=/j\/+/J\/+/c/+/L/
	A	/e/=/e/+/E/ /i/=/i/+/j/ /u/=/u/+/w/ /o/=/o/+/O/ /y/=/y/+/H/ /@/=/@/+/2/+/9/			
	B	/a/=/a/+/a~/ /i/=/i/+/j/ /e/=/e/+/E/+/e~/+/9~/ /o/=/o/+/O/+/o~/ /@/=/@/+/2/+/9/ /u/=/u/+/w/+/y/+/H/	/b/=/b/+/B/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a$_{nFR}$/=/s_a$_{nFR}$/+/z_a$_{nFR}$/ /R\/=/R\/+/r/+/r\/
FR	**C**	/a/=/a/+/a~/ e/=/e/+/E/+/e~/+/9~//i/=/i/+/j/ o=o+O+o~ u=u+w+y+H+@+2+9	/b/=/b/+/B/+/v/ /d/=/d/+/D/ /g/=/g/+/G/	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a/=/s_a/+/z_a/ /s_a$_{nFR}$/=/s_a$_{nFR}$/+/z_a$_{nFR}$/ /R\/=/R\/+/r/+/r\/
	D	/a/=/a/+/a~/ /e/=/e/+/E/+/e~/+/9~/ /i/=/i/+/j/ /o/=/o/+/O/+/o~/ /u/=/u/+/w/+/y/+/H/+/@/+/2/+/9/	/b/=/b/+/B/+/v/ /d/=/d/+/D/ /g/=/g/+/G	/m/=/m/+/F/ /n/=/n/+/N/+/n_d/	/s_a/=/s_a/+/z_a/ /s_a$_{nFR}$/=/s_a$_{nFR}$/+/z_a$_{nFR}$/ /R\/=/R\/+/r/+/r\/

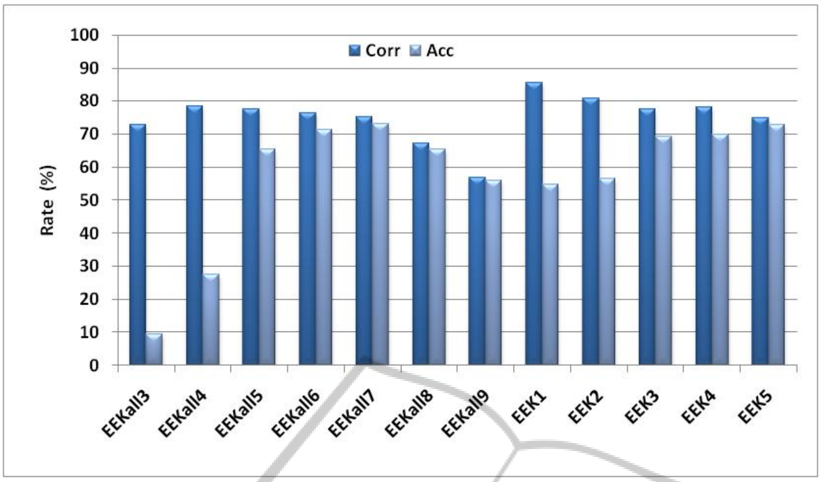

Figure 3: Results with different HMM topologies for Basque.

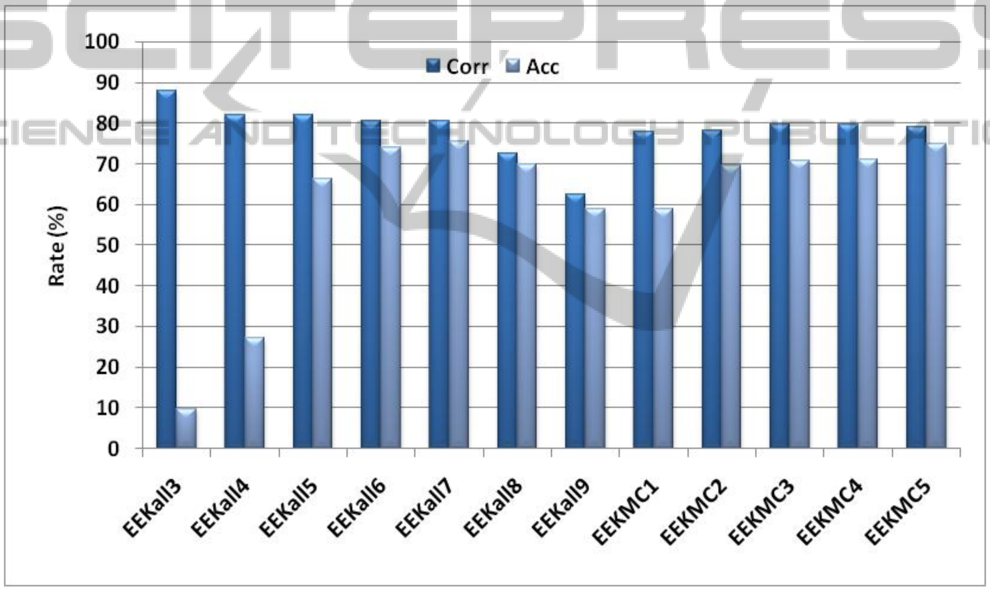

Figure 4: Results with different HMM topologies and the group type C for Basque.

stage to each SLU. Then, following stage distributions of the SLUs were adjusting taken into account the recognition rates and the insertion value for every unit obtained in the previous analysis (EEKall1 to EEKall9, EEK1 and EEK2). EEK5 configuration proposal provides the best results (figure 3). In this case, the weakest units have been modelled with many states. Moreover, the unit definition tries to support cohesion among the articulatory characteristics of the units. EEK5 does not overcome the recognition rate of EEKall7, but it approaches to it very much while using less global state numbers: Corr = 75 % and Acc = 72.85 %. The same effect appears for all allophone groups in the case of Basque. The best results are obtained with the type C. In figure 4 the results obtained for this unit grouping can be seen.

The best rates are obtained for the EEKall7 configuration: Corr =80.60 % and Acc = 75.62 %, the next best result is obtained by EEK5 with these rates: Corr = 79.16 % and Acc = 75.05 %. Though the values of Corr and Acc change, the evolution of the results with regard to the different configurations is kept for types A and B. In the figure 5 it can see the results for Spanish without using SLU groups. The analysis carried out for Spanish and French is the same. Nevertheless, because both languages are different with regard to phonetic characteristics and the audio signal, they have evolved towards different topologies (figure 5).

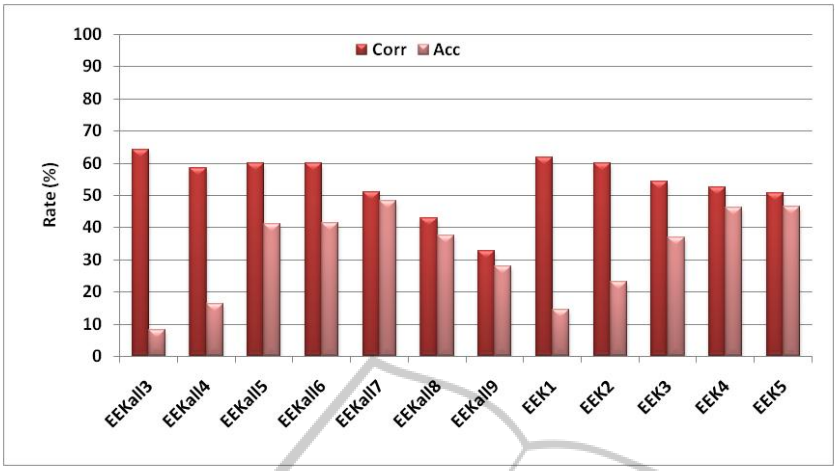

Figure 5: Results with different HMM topologies for Spanish.

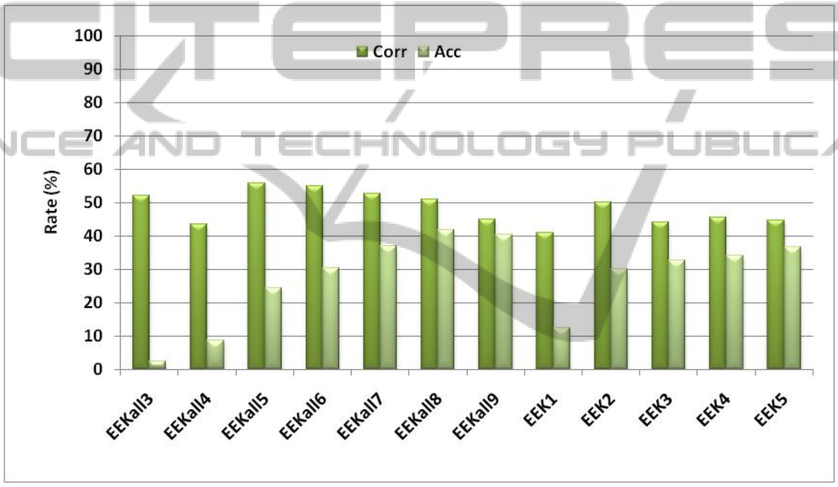

Figure 6: Results with different HMM topologies for French.

As with Basque, with regard to the evolution of topologies in which all the SLUs have the same SN the maximum value of Acc is obtained by 7 state: Corr = 51.04 % and Acc =48.39 %. With (Puertas, 2000) configuration the results are very poor but after several tests the following results are obtained for EEK5: Corr = 50.63 % and Acc = 46.63 %. In this case, it does not manage to overcome the rate of EEKall7, but the configuration EEK5 has less computational cost since fewer states are used. The same effect appears in for the SLU groups proposed for Spanish (table 3). Here the best result is provided for the Type B and EEKall7: Corr = 53.33 % and Acc = 50.58 %. For the EEK5 the following results are obtained: Corr = 51.08 % and Acc = 49.41 %. The French database presents the highest level of noise among the three languages. Figure 6 shows these results. For French, the best result with regard to Accuracy is obtained with the EEKall8 topology: Corr = 52.65 % and Acc =36.97 %.

Table 5: Summary of the best results obtained with regard to the topologies and the languages.

LG	Rate (%)	*SLU Type*	
		EEKall7	EEK5
BS	Corr (%)	80.60	79.16
	Acc (%)	75.62	75.05
		EEKall7	EEK5
SP	Corr (%)	53.33	51.08
	Acc (%)	50.58	49.41
		EEKall8	EEK5
FR	Corr (%)	58.54	50.41
	Acc (%)	49.46	46.08

Under noisy conditions, the increment of SN improves the results but anyway for French the results are very poor due to the under-resourced conditions. EEK5 configuration does not overcome the EEKall8 and obtains similar results: Corr = 44.75 % and Acc = 36.67 %. The SLU groups in general provide better results. The maximum value for French with regard to Acc is obtained for the type C and the configuration EEKall8: Corr = 58.54 % and Acc = 49.46 %. For EEK5: Corr = 50.41 % and Acc = 46.08 %.

Then we can conclude that:

1. The best results are obtained for Basque.
2. Noise effects are better absorbed for topologies with more states.
3. Each language has its own configuration guided for both the phonetic features and the noise level of the data.
4. The SLU groups provide better results for the three languages.
5. The configurations EEK5 of each language provides very interesting results, since the Acc values are close to the maximum values with fewer SN.

The most outstanding results are summarized in table 5 with regard to the topologies and the languages.

Finally, a new analysis has been carried out with regard to the insertion penalty, *p*, the Gaussian Number (GN), the topologies and the languages. Table 6 presents the obtained results. The worst results are obtained for French. In general it can be observed that the models with high SN need a lower GN and lower value for *p* because of the definition of appropriately configurations fitted to the needs of the system. In some cases Acc rates for APD experiments are very small, but this weakness can be absorbed in a later phase by a powerful LM as the ontologies (Barroso, 2011b).

6 CONCLUDING REMARKS

The present work is focused on the selection of appropriate Acoustic Models for Speech Processing in a complex environment (multilingual context and under-resourced and noisy conditions) oriented to general ASR tasks. The work has been carried out with a small trilingual speech database with very low audio quality. In order to decrease the negative impact that the lack of resources has in this task there were selected as acoustic models of the sublexical units several options such as hybrid HMM topologies and parameters, and optimum configuration for the APD system (Multivariate Gaussian Number or the insertion penalty). With the new Acoustic Modelling noise effects are better absorbed for topologies with more states. Moreover, each language has its own configuration guided for both the phonetic features and the noise level of the data. In some cases rates for APD experiments are very poor, but this weakness can be absorbed in a later phase by a powerful LM.

Table 6: Summary of the best results obtained with regard to the topologies, languages an APD configuration.

LG		SLU	GN	*p*	*Corr*	*PER*
BS	EEK5	Allophone	24	-20	81.50	21.10
	EEK5	Type A	24	-20	78.84	24.10
	EEKall7	Type B	8	-10	78.90	25.50
	EEKall7	Type B	4	-25	82.10	20.20
	EEKall7	Type B	4	-10	79.70	24.00
	EEK5	Type B	22	-20	80.50	23.95
SP	EEK5	Type A	24	-20	53.25	52.5
	EEKall7	Type B	8	-10	54.33	52.80
	EEKall7	Type B	4	-25	58.80	46.04
	EEKall7	Type B	4	-10	62.18	42.80
	EEK5	Type B	22	-20	59.90	45.94
FR	EEK5	Allophone	24	-20	46.33	58.28
	EEK5	Type A	24	-20	46.50	56.28
	EEKall7	Type B	8	-10	50.80	50.58
	EEKall7	Type B	4	-25	58.59	48.80
	EEKall7	Type B	4	-10	52.86	45.94
	EEK5	Type B	22	-20	81.50	21.10

In future lines of work, non-linear approaches, low-level ontologies and new methodologies for automatic features selection will be developed.

ACKNOWLEDGEMENTS

This work has been partially supported by the programmes *IKERTU* and *SAIOTEK* of the Basque Community Government. We are also very grateful for the collaboration of *Infozazpi* radio and *Irunweb* enterprise.

REFERENCES

Baker, J., 1975, Stochastic Modeling for Automatic Speech Recognition, Speech Recognition, Reddy, *Academic Press*.

Barroso, N. Ezeiza A., Gilisagasti, N., L Ipiña K., López A. and López J. M.,2007, Development of Multimodal Resources for Multilingual Information Retrieval in the Basque context., *INTERSPEECH* Antwerp, Belgium, 2007.

Barroso, N., Lopez De Ipiña K., Hernandez C. and Ezeiza A., 2011a. Matrix covariance estimation methods for robust security speech recognition with under-resourced conditions, *45th IEEE International Carnahan Conference on Security Technology*, Mataro Barcelona

Barroso, N., López de Ipiña, K., Ezeiza, A., Hernández, C., Ezeiza, N., Barroso, O., Susperregi, U. and Barroso, S., 2011. GorUp: an ontology-driven Audio Information Retrieval system that suits the requirements of under-resourced languages, *INTERSPEECH*, Florence Italy.

Baum, E., Petrie, T., Soules, G., & Weiss, N., 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistics*, vol. 41, no. 1, pp. 164–171.

Baum, L. E., and Eagon, J. A., 1967, An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology., *In Bulletin of the American Mathematical Society*, vol. 73, pp. 360-370.

Cosi P. “Hybrid HMM-NN architectures for connected digit recognition”. *Proc. of the IJC on Neural Networks*, vol. 5, 2000

Ellis, D., 2011, http://labrosa.ee.columbia.edu/

Friedman J. H., 1989, Regularized discriminant analysis. *Journal of the American Statistical Association*, vol. 84, pp. 165-175, 1989.

infozazpi radio www.infozazpi.com

Jelinek., 1976, Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532-556.

Le V. B. and Besacier L., 2009 Automatic speech recognition for under-resourced languages: application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 17, Issue 8, pp 1471-1482, 2

Martinez A. and Kak A., 2001, PCA versus LDA, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No.2, 228-233

Puertas, I., 2000, Robustez de Reconocimiento fonético de voz para aplicaciones telefónicas. Madrid: *Tesis doctoral*.

Rabiner, H. R., & Juang, B. H., 1993, Fundamentals of Speech Recognition, USA: *Prentice Hall*

Schultz, T. and Waibel, A., 1998, Multilingual and Crosslingual Speech Recognition, *Proceedings of the DARPA BC. Workshop.*

Seng S., Sam S., Le V. B., Bigi B. and Besacier L., 2008, Which Units For Acoustic and Language Modeling For Khmer Automatic Speech Recognition., *1st International Conference on Spoken Language Processing for Under-resourced languages* Hanoi, Vietnam

Smith N., Gales M. “Speech recognition using SVMs”, *Advances in Neural Information Processing Systems* 14. MIT Press, 2002.