Speech production models for ASR in Spanish language

Javier Mikel Olaso, María Inés Torres

Universidad del País Vasco
javiermikel.olaso@ehu.es, manes.torres@ehu.es

Abstract

In this paper we undertake the extraction of phonological features applied to Spanish language. Also propose a method to integrate these features into an HMM based speech recognition system using an architecture that uses independent feature streams. In the experimental results we find that higher recognition accuracies and less computational cost can be obtained.

Index Terms: speech recognition, acoustic modeling, phonological features

1. Introduction

The majority of speech recognition systems are currently based on the use of the acoustic properties of speech to establish its characteristics. This method has to tackle various difficulties, such as, [2, 3, 13]:

- Phonation differences due to the diversity of speakers.
- Coarticulation effects.
- Spontaneous speech.
- Problems with pronunciation dictionaries, mainly in the English language.
- Ambient noise and interferences.

Other approaches have alternatively been proposed. One such approach seeks to incorporate information relating to the way speech is produced in terms of articulatory gestures. This approach is considered to be highly beneficial for automatic speech recognition systems, mainly due to the invariance of critical articulators, those mostly involved in sound production, and the lower susceptibility of the articulatory space to the effects of coarticulation, [1, 2]. This approach has to deal with two main problems. On the one hand, the speaker's utterances needs to be represented in terms of these articulatory gestures, and on the other hand a system is need that is able to interpret the articulatory gestures based representation. Some studies have attempted to solve these problems. The seemingly most successful method has been the use of Time Delay Neural Networks (TDNN) [5] for articulatory gestures detection, and the re-scoring of lattices obtained using a system based on HMMs defined over mel frequency cepstrum [1].

This paper is twofold. On one hand, we want to undertake the extraction of phonological features applied to the Castilian variety of Spanish. On the other hand, we propose a method to integrate these features into a speech recognition system. TDNN was used for the extraction of the features and an alternative method based on treating the vectors representing the phonological features as observation vectors of HMM models for the integration. Two types of experiments were carried out. The first only used articulatory information and the second combined both articulatory and acoustic information. We should point out that these experiments focused on the Castilian variety of Spanish and it was a challenge given that it was the first ever attempt to carry out this task.

The structure of the article is as follows. Section 2 provides a short description of the different methods studied to obtain articulatory information and describes how we decided to implement this phase. Section 3 describes the architecture of the speech recognizer used in our experiments. Section 4 contains the results of our experiments. And the paper ends with the concluding remarks and acknowledgements in Sections 5 and 6, respectively.

2. Phonological feature extraction

Several methods have been proposed for the extraction of the phonological features. These methods fall into one of two approaches. On the one hand, there are the methods based on extraction of information directly from the measurement of the positions or the articulatory organs responsible for speech generation, such as those presented in [6] where measures of the articulator's positions taken with X ray are used. On the other hand, there are the methods based on indirect measurements. Examples of the indirect methods can be found in [7], where visual information of the mouth is used, or in [8, 10, 11, 12], where the phonological information is taken from the surface waveform.

The most common of these two approaches seems to be the indirect one, and more specifically when information is taken from the surface waveform. This is mainly due to the fact that direct measurements require expensive and invasive devices, such as an electropalatograph. On the other hand, different methods are used to extract phonological information from the surface waveform, such as, the use of artificial neural networks [8, 10], dynamic Bayesian networks [4, 9] or Hidden Markov Models [14], among others.

We used neural networks in this study, and more specifically, Recurrent Time Delay Neural Networks [5], a type of neural networks that combines time-delay windows and recurrent connections to capture the dynamic information of the speech signal.

We therefore needed to define the set of sounds (phonemes) used in our experiments and how they were described in terms of articulatory features. Using the theoretical classification shown in Table 1, and after a set of tests to maximize the classification accuracy, we defined the articulatory feature sets shown in Table 2. It can be seen that it corresponds to the theoretical classification, plus a class *silence* in all features except sonority, a class *vowel* in manner and place of articulation, and a *non-vowel* class for vowel/non-vowel features.

Place of articulation	Manner of articulation							
	Plosive		Fricative	Affricate	Lateral	Trill	M. Trill	Nasal
	unvoiced	voiced	unvoiced		voiced			
Bilabial	p	b						m
Labiodental			f					
Linguodental			z					
Alveolar	t	d	s	ch	l	r	rr	n
Palatal					ll			ñ
Velar	k	g	j					

	Front	Central	Back
Close	i		u
Close-Mid	e		o
Open		a	

Table 1: Theoretical classification for phonemes in spanish language.

3. Speech recognizer based on phonological features

Different systems have been developed that make use of the phonological features. For example, a system is presented in [1], [10], that uses phonological features to re-score the lattices generated by a MFCC based HMM phone recognizer.

In this paper, we propose a system based on a classical acoustic speech recognition system, based on HMMs, with two main differences. On one hand, we replaced, or combined, the acoustic feature vectors with vectors representing the phonological information, that were obtained via the feature extractors mentioned in section 2. On the other hand, we followed an approach of integrating the feature vectors using independent feature streams.

Let,

$$O = o_1, o_2, ..., o_T \tag{1}$$

be a sequence of speech vectors or observations where o_t is the speech vector observed at time t. When o_t are elements of a continuous observation alphabet, and in case of using Gaussian mixtures as probability distribution function, the observation symbol probability matrix, $b_j(o_t)$, for an HMM can be written as:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}) \tag{2}$$

where $\mathcal{N}(o_t, \mu_{jm}, \Sigma_{jm})$ denotes m'th Gaussian, with μ_{jm} mean vector and Σ_{jm} variance matrix, for state j. M is the number of Gaussians in the mixture and c_{jm} is the weight of the m'th component in the mixture, satisfying:

$$\sum_{m=1}^{M} c_{jm} = 1 \tag{3}$$

Well, to integrate the features defined in Table 2 we propose to use an architecture with independent feature streams. Let S be the number of independent feature streams, e.g. those defined in Table 2, and O_{st} a vector defined as:

$$O_{st} = o_{st}^1, o_{st}^2, \ldots, o_{st}^n \tag{4}$$

that represents an observation in stream s and time t, and with n its dimension, which may vary for each feature stream.

Sonority	
Voiced	a,e,i,o,u,b,d,g,l,ll,r,rr,m,n,ñ
Unvoiced	p,t,k,f,z,s,j,ch

Manner	
Plosive	p,t,k,b,d,g
Fricative	f,z,s,j
Affricate	ch
Lateral	l,ll
Trill	r
M. Trill	rr
Nasal	m,n,ñ
Vowel	a,e,i,o,u
Silence	*SIL*

Place	
Bilabial	p,b,m
Labiodental	f
Linguodental	z
Albeolar	t,d,s,ch,l,r,rr,n
Palatal	ll,ñ
Velar	k,g,j
Vowel	a,e,i,o,u
Silence	*SIL*

Vowel - Non Vowel			
Front	i,e	Open	a
Central	a	Mid-Close	e,o
Back	o,u	Close	i,u
Non Vowel	*rest*	Non Vowel	*rest*
Silence	*SIL*	Silence	*SIL*

Table 2: Classification used for the phonological features.

With this approach the observation symbol probability matrix, $b_j(o_t)$, can be rewritten as:

$$b_j(o_t) = \prod_{s=1}^{S}(\sum_{m=1}^{M_s} c_{jms}\mathcal{N}(O_{st}; \mu_{jms}, \Sigma_{jms})) \tag{5}$$

where M_s is the number of Gaussians in the mixture of stream s.

Likewise, in the case of using discrete symbol streams, the matrix, $b_j(o_t)$, can be written as:

$$b_j(o_t) = \prod_{s=1}^{S} b_{js}(O_{st}) \tag{6}$$

where $b_{js}(O_{st})$ is the observation symbol probability matrix of stream s.

4. Experimental results

This section is dedicated to a more detailed description of the implementation of the system presented. First, we provide a short description of the corpus used. The process for the phonological feature extraction is then described, and finally the different configurations, and the recognition results of the speech recognition system used are given.

4.1. Database description

The speech corpus used in this paper was Albayzin [15]. This is a corpus in the Castilian variety of Spanish recorded at 16KHz divided in three sub-corpus: a phonetic corpus without syntactic-semantic restrictions, which was used in this study, a second corpus including those restrictions and a third corpus designed for noisy environments. The phonetic corpus is divided in a training set of 200 sentences pronounced by 4 speakers and 25 sentences more pronounced by 160 speakers, making a total of 4800 sentences, 42144 words (712 different) and 187848 phonemes, along with a test set with 50 sentences pronounced by 40 speakers, making a total of 2000 sentences, 21052 words (1856 different) and 93696 phonemes. Table 3 contains a short description of the phonetic corpus.

	Training	Test
Speakers	160	40
Sentences	4800	2000
Words	42144	21052
Different Words	712	1856
Phonemes	187848	93696

Table 3: Summary of the phonetic subcorpus of Albayzin speech corpus.

On the other hand, the representation of the corpus in terms of the phonological features needed to be obtained prior to training the HMM models. This representation was obtained by making previously trained networks, see section 4.2, act on the acoustic representation of the corpus.

Finally, the corpus was transcribed using a set of 24 phonetic units, 23 phonemes and 1 silence, and therefore 24 HMM models were trained.

4.2. Phonological feature extraction results

Based on the study in [5], we used Recurrent Time Delay Neural Networks for phonological feature detection. Five neural networks were used to detect each of the following features:

- Sonority
- Vowel-NonVowel (2)
- Articulation manner
- Place of articulation

These neural networks had multiple outputs and the classes to be detected for each feature were those described in Section 2. The inputs of all the neural networks were 12 first Mel Frequency Cepstral Coefficients plus energy, which were extracted in 25 ms Hamming windowed frames with an overlapping of 10 ms. The outputs of the neural networks were real values ranging from 0 to 1. Although these values could be treated as the posterior probabilities of the features, we applied a more basic implementation and used them as simple vectors, as if they were MFCCs.

Table 4 contains the detection accuracies of the different phonological feature sets. It can be seen very high detection accuracy was obtained in the case of both sonority and vowels detectors, given the very good vowel-nonvowel detector obtained. In the case of manner and place of articulation, good overall detection accuracies were obtained, but this poor detection in some classes. For example in the case of the articulation manner, poorer detection accuracies were obtained for the classes *trill*(r) and *multiple trill*(rr), which were mostly detected as vowels. This is due to the fact that these phonemes do not have their own spectrum in Spanish and they inherit the spectrum of the preceeding vowel and give it an intermittent pattern.

Sonority	
Class	% correct
Voiced	93.3
Unvoiced	96.0

Manner	
Class	% correct
Total	83.8

Place	
Class	% correct
Total	83.4

Vowel - Non Vowel			
Class	% correct	Class	% correct
Front	81.1	Open	78.3
Central	84.1	Mid-Close	79.0
Back	73.8	Close	75.1
Non Vowel	90.8	Non Vowel	90.3
Silence	96.0	Silence	93.6

Table 4: Classification accuracies for the different phonological features.

4.3. Recognition results

The HMM topology used was the classical left-to-right of three states with transitions from one state to itself and to the adjacent one. Two types of experiments were likewise carried out. On the one hand, the system was tested using phonological information only, and on the other hand, phonological and acoustic information was combined.

When using phonological information only, two ways of integrating the information were used. The first used independent feature streams. The second used a unique feature stream resulting from the concatenation of the vectors of each of the independent streams. When acoustic information was also integrated, it took place as four independent feature streams. These streams corresponded to the first twelve cepstral coefficients, their first and second derivatives, and an additional stream with the energy and first derivative of the energy, per frame.

We also used discrete models. In this case, codebooks needed to be generated both for the case of a unique feature stream and of various feature streams. In the case of a unique stream, it was generated using the LBG algorithm to the concatenation of the independent feature vectors. For the various streams case, the codebooks were generated as follows: for each independent feature, the representative vector for each class was obtained as the mean vector of all the vectors belonging to that class. And were these representative vectors what we used as the codebook's vectors.

We then proceeded to train and test the models. It should be noted that tests varying the number of Gaussians in the mixtures for continuous models and the number of vectors of the codebooks for discrete models were carried out. Table 5 contains the results obtained, together with the recognition results for the acoustic based system used as baseline. The topology of this baseline system was identical to the topology of the system presented, with the same four independent acoustic feature streams used in the combinations with the phonological features. On the other hand, a codebook of 1024 classes in the case of discrete models and 32 mixture Gaussians in the case of continuous models were used in the results given for this baseline system .

It can be seen that a recognition improvement was obtained in the case of discrete HMM models, however in the case of continuous HMM models only when combining phonological and acoustic information we obtain recognition accuracies similar to the baseline system. We believe that this could be due to the fact that the phonological space is highly discretized which favours the use of discrete models. On the other hand, and comparing the systems with just phonological information and with both phonological and acoustic information, it can be seen that the systems combining both types of information have better recognition accuracies.

	DHMM		CHMM	
BASELINE	69.40		75.15	
	$S = 1$	$S = 5$	$S = 1$	$S = 5$
PH.	72.93	72.46	70.35	70.23
	$S_{ph} = 1$	$S_{ph} = 5$	$S_{ph} = 1$	$S_{ph} = 5$
PH.+AC.	75.83	75.72	75.06	74.24

Table 5: Recognition results for DHMM and CHMM. When combining phonological (PH) and acoustic (AC) spaces, we have $S = S_{ph} + S_{ac}$ and $S_{ac} = 4$, being S_{ph} and S_{ac} the number of independent feature streams for phonological and acoustic spaces respectively.

We find that the results obtained for the discrete models are pretty good because they have proved to be computationally faster than continuous ones. In Table 6 we show computation times for the recognition process of the continuous HMM models based baseline system and the different implementations used with discrete HMM models, normalized with the value of the baseline system. It can be seen that using both phonological and acoustic features has higher computational cost than using only phonological features, although this cost is less than the cost of the baseline system and therefore is a reasonable cost because the gain in recognition accuracy is higher. Alternately, also can be seen that when speaking of computational cost is better to use phonological features in independent streams rather than concatenate them in one stream.

	PH.		PH. + AC.	
	$S = 1$	$S = 5$	$S_{ph} = 1$	$S_{ph} = 5$
DHMM	0.13	0.03	0.27	0.17
BASELINE	1			

Table 6: Normalized computation times for baseline and discrete HMM models. When combining phonological (PH) and acoustic (AC) spaces, we have $S = S_{ph} + S_{ac}$ and $S_{ac} = 4$, being S_{ph} and S_{ac} the number of independent feature streams for phonological and acoustic spaces respectively.

5. Concluding remarks

In this work we have undertaken the problem of using phonological features for speech recognition in Castilian variety of Spanish. Also we have proposed a method for integrate these features in a speech recognition system based on HMM models.

We have found, that the use of phonological features could be highly beneficial above all in the case of using discrete HMM models where we have obtained better results than the baseline system used, both in accuracie rate and in computational cost.

6. Acknowledgements

This work has been partially supported by the University of the Basque Country under grant GIU07/57, Spanish CICYT under grant TIN2008-06856-C05-01 and by the Spanish program Consolider-Ingenio 2019 under grant CSD2007-00018.

7. References

[1] Rose, R. and Momayyez, P., “Integration of multiple feature sets for reducing ambiguity in ASR.“, ICASSP 2007, Volume 4, 325-328, 2007.

[2] Rose, R. et ál, “An investigation of the potential role of speech production models in automatic speech recognition”, Proceedings ICSLP-94, pp. 575-578, 1994.

[3] Koreman, J. and Andreeva, B., “Can we use the linguistic information in the signal¿‘, Phonus (Institute of Phonetics, University of the Saarland) 5: 47-58, 2000.

[4] Livescu, K. et ál “Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report“, Technical Report, Center for Language and Speech Processing, Johns Hopkins University, 2007.

[5] Strom, N., “Phoneme probability estimation with dynamic sparsely connected artificial neural networks.“, The free speech journal, Vol 1, Issue #5, 1997.

[6] Blackburn, C.S. and Young, S.J., “Pseudo-Articulatory speech synthesis for recognition using automatic feature extraction from X-Ray data.“, In proceedings ICSLP 96, 969-972, 1996.

[7] Saenko, K. et ál, “Articulatory features for robust visual speech recognition.“, ICMI’04, 2004.

[8] King, S. and Taylor, P., “Detection of phonological features in continuous speech using neural networks.“, Computer Speech & Language, 333-353, 2000.

[9] Frankel, J. et ál, “Articulatory feature recognition using dynamic Bayesian networks“, Computer Speech and Language archive, Volume 21 , Issue 4, 620-640, 2007.

[10] Parya, M. at ál, “Exploiting complementary aspects of phonological features in automatic speech recognition.“, IEEE Workshop on Automatic Speech Recognition & Understanding”, 47-52, 2007.

[11] Stouten, F. and Martens, J.P., “On the use of phonological features for pronunciation scoring.“, In proceedings ICASSP, 229-232, 2006.

[12] Stouten, F. and Martens, J.P., “Speech Recognition with Phonological Features: Some issues to attend.“, Interspeech-06, paper 1081-Mon2BuP-4, 2006.

[13] BenZeghiba, M. et ál, “Automatic speech recognition and intrinsic speech variation“, ICASSP 2006, 31st International Conference on Acoustics, Speech, and Signal Processing, May 14-19, 2006.

[14] Abu-Amer, T. and Carson-Berndsen, J., “HARTFEX: A multidimentional system of HMM based recognisers for articulatory features extraction.“, In proceedings NOLISP-2003, paper009, 2003.

[15] Casacuberta, F. et ál, “Desarrollo de corpus para investigación en tecnologías del habla (Albayzin).”, Procesamiento del lenguage natural, 12:35-42, 1992.