Integrating Phonological Knowledge in ASR Systems for Spanish Language

Javier Mikel Olaso and María Inés Torres

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

Abstract. In this paper we undertake the use of phonological features applied to speech recognition in Spanish language. We investigate two different ways to integrate these phonological features into an HMM based speech recognition system. We also propose a method to integrate these features 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.

Keywords: 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], [12], phonation differences due to the diversity of speakers, coarticulation effects, spontaneous speech, problems with pronunciation dictionaries, mainly in the English language or 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 need to be represented in terms of these articulatory gestures, and on the other hand a system is needed to interpret such representation. Some studies have attempted to solve these problems. The seemingly most successful method has been the use of Recurrent Time Delay Neural Networks (RTDNN) [5] for articulatory gestures detection, and the re-scoring of lattices obtained using a system based on HMMs defined over Mel Frequency Cepstral Coefficients (MFCC) [1].

This paper is twofold. On one hand, we want to undertake the use of phonological features applied to the Castilian variety of Spanish, investigating two methods to integrate these features into an HMM based speech recognition system. The first method used vectors representing phonological information as

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 136–143, 2010.

observation vectors of HMM models and the second used acoustic vectors based on MFCC. On the other hand, we propose a method to integrate these features into a speech recognition system.

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 5, 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], 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 [13], among others.

We used neural networks in this study, and more specifically, RTDNN [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. Basing on 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, where we can see 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, see [15] for a more detailed description.

3 Speech Recognizer Architecture

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.

Table 1. Theoretical classification for phonemes in spanish language

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 2. Classification used for the phonological features

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

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*

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*

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 introduce phonological information in the system architecture. 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 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, that compliments:

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

Well, now we propose to use an architecture with independent feature streams.[1] Let S be the number of independent feature streams 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. 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, which may be different in each stream.

Likewise, in the case of using discrete symbol streams, the matrix, $b_j(o_t)$, can be rewritten 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 Evaluation

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 are given.

4.1 Database Description

The speech corpus used in this paper was Albayzin [14]. 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 consists of sentences of read text and 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.

[1] Most speech recognition systems use as observation vectors a concatenation of different types of feature vectors (e.g. MFCC, energy, and it's first and second derivatives). We propose to treat the different types of feature vectors independently and denote each independent feature as a *feature stream*.

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

	Speakers	Sentences	Words	Different Words	Phonemes
Training	164	4800	42144	712	187848
Test	40	2000	21052	1856	93696

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

For the case of use a phonological representation space we need a way to obtain such representation. Based on the study in [5], we used RTDNN for phonological feature detection. Five neural networks were used to detect each of the features presented in Table 2, that is, sonority, manner and place of articulation, vowel-nonvowel in front-central-back axis and vowel-nonvowel in open-close-midclose axis. 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 MFCC 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 observation vectors.

4.3 Comparing Phonological and Acoustic Representation Spaces

To compare the different representation spaces used we made three different experiments.

In the first experiment, we used an acoustic representation space in which the observation vectors were a concatenation of MFCC, energy and it's first and second derivatives. To include the phonological knowledge we proceeded as follows: we used a Gaussian function to model each of the classes presented in Table 2. To construct each of the Gaussian functions we obtained the mean and variance vectors of all the vectors belonging to each of the classes and used these as mean and variance for each of the Gaussian function. We used two different implementations. The first with a feature stream for each of the phonological features and the second integrating all the phonological features in an unique stream. Resulting in mixtures of 2, 5, 5, 9 and 8 Gaussians, respectively, for each of the streams in the case of independent feature streams and a mixture of 24 Gaussians in the case of an unique stream. Finally, and to maintain the phonological information in the training process we keep the values of the Gaussian functions fixed and only reestimate it's weights in the mixture.

For the second experiment, we used a phonological representation space based on using as observation vectors for the HMM states those obtained as outputs of the phonological feature detectors, see Section 4.2. Two different implementations were used. The first used independent feature streams for each of the phonological features with mixtures of 2, 5, 5 ,9 and 8 Gaussians respectively, and the second used a unique feature stream resulting from the concatenation of the vectors of each of the independent streams and which used a mixture of 128 Gaussians.

Finally, we made a last experiment combining both phonological and acoustic information. In this case, we used the same two types of observation vectors of the second experiment for the phonological space, and for the acoustic space we used four independent feature streams for each of the following features: 12 first MFCC, it's first and second derivative, and energy and it's first and second derivative. Mixtures of 32 Gaussians were used for each of the acoustic streams.

We also used discrete models when using phonological representation space only and combination of phonological and acoustic spaces. To obtain the codebooks for the phonological space, in the case of an unique stream, it was generated using the LBG algorithm to the concatenation of the independent feature vectors. For the various streams case, 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.

Finally, say that the topology of the HMM models used was the classical left-to-right of three states with transitions from one state to itself and to the adjacent one.

We then proceeded to train and test the models. Table 4 contains the phone recognition accuracies (PRA) obtained, together with the PRA for the acoustic based baseline system. The topology of this baseline system was identical to the topology of the system presented, with four independent feature streams corresponding to MFCC, it's first and second derivatives, and energy and it's first derivative, respectively. A codebook of 1024 classes in the case of discrete models and 32 mixture Gaussians in the case of continuous models were used for each of the streams.

When using the phonological representation space, we can see that better recognition accuracies was obtained in the case of discrete HMM models than in the case of continuous models. 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, when using the acoustic representation space the results obtained are not as good as in the previous case, and we can conclude that is better to use the phonological representation space.

Also can see that only when combining phonological and acoustic information we obtain recognition accuracies similar to the baseline system. 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.

Table 4. Phone recognition accuracies for baseline and presented systems. S is the number of independent feature streams. When Ph.+Ac. we have $S = S_{ph} + S_{ac}$ and $Sac = 4$.

	Ac. Space		Ph. Space				Ph. + Ac. Space			
	CHMM		DHMM		CHMM		DHMM		CHMM	
	$S=1$	$S=5$	$S=1$	$S=5$	$S=1$	$S=5$	$S_{ph}=1$	$S_{ph}=5$	$S_{ph}=1$	$S_{ph}=5$
Ac. baseline	75.15		69.40		75.15		69.40		75.15	
PRA	48.92	47.67	72.93	72.46	70.35	70.23	75.83	75.72	75.06	74.24

Table 5. Normalized computation times for baseline and discrete HMM models

	$S_{ph}=1$	$S_{ph}=5$
DHMM	0.13	0.03
BASELINE	1	

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 5 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 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.

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 used two different representation spaces, phonological and acoustic, to integrate phonological features in a speech recognition system and have found that is better to use the phonological space. Also 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 accuracy rate and in computational cost.

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.

References

1. Rose, R., Momayyez, P.: Integration of multiple feature sets for reducing ambiguity in ASR. In: ICASSP 2007, vol. 4, pp. 325–328 (2007)
2. Rose, R., et al.: An investigation of the potential role of speech production models in automatic speech recognition. In: Proceedings ICSLP 1994, pp. 575–578 (1994)
3. Koreman, J., 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 al.: 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 1(#5) (1997)
6. Blackburn, C.S., Young, S.J.: Pseudo-Articulatory speech synthesis for recognition using automatic feature extraction from X-Ray data. In: Proceedings ICSLP 1996, pp. 969–972 (1996)
7. Saenko, K., et al.: Articulatory features for robust visual speech recognition. In: ICMI 2004 (2004)
8. King, S., Taylor, P.: Detection of phonological features in continuous speech using neural networks. Computer Speech & Language, 333–353 (2000)
9. Frankel, J., et al.: Articulatory feature recognition using dynamic Bayesian networks. Computer Speech and Language Archive 21(4), 620–640 (2007)
10. Parya, M., et al.: Exploiting complementary aspects of phonological features in automatic speech recognition. In: IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 47–52 (2007)
11. Stouten, F., Martens, J.P.: On the use of phonological features for pronunciation scoring. In: Proceedings ICASSP, pp. 229–232 (2006)
12. BenZeghiba, M., et al.: Automatic speech recognition and intrinsic speech variation. In: 31st International Conference on Acoustics, Speech, and Signal Processing ICASSP 2006, May 14-19 (2006)
13. Abu-Amer, T., Carson-Berndsen, J.: HARTFEX: A multi-dimentional system of HMM based recognisers for articulatory features extraction. In: Proceedings NOLISP 2003, paper009 (2003)
14. Casacuberta, F., et al.: Desarrollo de corpus para investigación en tecnologías del habla (Albayzin). Procesamiento del Lenguage Natural 12, 35–42 (1992)
15. Olaso, J.M., Torres, M.I.: Speech production models for ASR in Spanish language. Paper Submmited to FALA 2010 (2010)