Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 165-168.

Acoustic-phonetic decoding of different types of spontaneous speech in Spanish

Doroteo T. Toledano[a], *Antonio Moreno Sandoval*[b], *José Colás Pasamontes*[a], *Javier Garrido Salas*[a]

[a]Human-Computer Technologies Laboratory (HCTLab), Escuela Politécnica Superior.
[b]Laboratorio de Lingüística Informática (LLI-UAM)

Universidad Autónoma de Madrid

Abstract

This paper presents preliminary acoustic-phonetic decoding results for Spanish on the spontaneous speech corpus C-ORAL-ROM. These results are compared with results on the read speech corpus ALBAYZIN. We also compare the decoding results obtained with the different types of spontaneous speech in C-ORAL-ROM. As the most important conclusions, the experiments show that the type of spontaneous speech has a deep impact on spontaneous speech recognition results. Best speech recognition results are those obtained on speech captured from the media.

1. Introduction

Currently, spontaneous speech processing is one of the most active research lines in speech technology, and in particular in speech recognition. In the last years, the National Institute of Standards and Technology (NIST) [1] has launched a new series of competitive evaluations under the name Rich Transcription in which spontaneous speech processing (and particularly disfluency detection) is a key topic. Unfortunately that program does not include Spanish as one of the languages of interest. There are a few Spanish research groups that are conducting research in the field of spontaneous speech processing [2]. However, this field is still largely unexplored in Spanish.

In this paper we present initial results of our research using the C-ORAL-ROM corpus [3]. This corpus is a spontaneous speech corpus covering several languages. One of the main features of the corpus is that spontaneous speech is classified under several categories. Section 2 describes this corpus as well as some adaptations that have been necessary to automatically process it. The experiments described are acoustic-phonetic decodings performed using Hidden Markov Model (HMM) as acoustic models. These models have been trained using a read speech corpus in Spanish, ALBAYZIN [4]. Section 3 describes the training of the HMMs used. Section 4 presents the acoustic-phonetic decoding results, both for read speech and spontaneous speech and compares them. This section also compares results on different types of spontaneous speech, as defined in the C-ORAL-ROM corpus. Finally, section 5 summarizes the most important conclusions as well as future research lines.

2. Description of the C-ORAL-ROM corpus

C-ORAL-ROM is a multilingual corpus that comprises four romance languages: Italian, French, Portuguese and Spanish. In our work we have used the Spanish sub-corpus, which contains around 300.000 spoken words. From a sociolinguistic point of view, speakers are characterized by their age, gender, place of birth, educational level and profession. From a textual point of view the corpus is divided into the parts shown on Table 1 [5].

Table 1: Distribution of words in C-ORAL-ROM.

Informal 150.000 words				Formal 150.000 words
Familiar 113.000		Public 37.000		Formal in natural context 65.000
Monologs 33.000	Dialogs/ Convers. 80.000	Monologs 6.000	Dialogs/ Convers. 31.000	Formal on the media 60.000
				Telephone conversations 25.000

Table 1 shows that the main division is balanced between formal speech and informal speech. For informal speech a division is considered between speech in a familiar/private context and speech in a public context. The first group is further classified into monologs, dialogs and conversations with three or more speakers. The second group is similarly classified into monologs, dialogs and conversations. Regarding formal speech, a division has been made between speech in natural context and speech on the media. The former includes political speeches, political debates, preaching, teaching, professional expositions, conferences, speech in business contexts and speech in legal contexts. Speech on the media (also referenced in this article as *broadcast news*, which is the usual name for this kind of speech in the context of automatic speech recognition) includes news, sports, interviews, meteorology, science, reports and talk shows. Telephone conversations, although initially considered under the formal speech category in C-ORAL-ROM, have very particular features and is more similar to informal speech than to formal speech. For these reasons we have considered telephone conversations under the category of informal speech on a subdivision of its own.

These divisions and subdivisions of C-ORAL-ROM will allow us to compare the acoustic-phonetic decoding results using different types of spontaneous speech.

C-ORAL-ROM contains 183 recordings totaling over 40 hours of speech. There are basically three type of recordings depending on their duration: 7-10 minutes, 15 minutes and 30 minutes. These recordings were too long for their automatic processing. For that reason, we extracted each spoken utterance (between pauses) on a separate file using the existing C-ORAL-ROM manual segmentation. This manual segmentation has been essential to perform the experiments described in this paper.

Table 2: Divisions of C-ORAL-ROM.

Informal	*Familiar/Private*	*Monolog*
		Dialog
	Public	*Conversation*

Formal		
Formal in natural context	*Media (Broadcast News)*	*Telephone*
Political speech	*News*	
Political debate	*Sports*	
Preaching	*Interviews*	
Teaching	*Meteorology*	
Professional exposition	*Scientific*	
Conferences	*Reports*	
Business	*Talk shows*	

2.1. Phonological transcription

In order to compare acoustic-phonetic decoding results, a reference phonological transcription is required in advance. C-ORAL-ROM did not include that phonological transcription, including only an orthographic one. For that reason, the phonological transcription was generated from the orthographic one, making use of a simple phonological transcriptor based on rules. This transcriptor uses a minimum set of phonemes for Spanish (23 phonemes). Obviously, such a simple transcriptor does not allow to obtain a correct transcription in all cases. However, we consider that the precision achieved is good enough to obtain significant acoustic-phonetic decoding results.

3. Training of the HMMs for acoustic-phonetic decoding

The Hidden Markov Models used to perform the acoustic-phonetic decoding were trained on the ALBAYZIN corpus using the Hidden Markov Model ToolKit (HTK) software [6]. The front-end used for feature extraction was the advanced distributed speech recognition front-end defined by the ETSI standard ETSI ES 202 050 [7]. This front-end includes mechanisms for robustness against channel (convolutive) distortion and additive noise. Basically the mechanism used for noise robustness is a double Wiener filter that estimates and substracts the noise spectrum. The one used against convolutive distortion is cepstral mean normalization (CMN).

The set of phonemes used in all experiments is the minimum set of 23 phonemes in Spanish. We also consider models for initial, final and intermediate silences. We trained both context-dependent and context-independent models. We started training seed context-independent models using 600 of the 1200 utterances of ALBAYZIN that were phonetically labeled and segmented by hand (the other 600 were reserved for adjustment and evaluation purposes). Next we used those seed models to train context-independent models with 3500 utterances from the training set of ALBAYZIN. We trained models with up to 150 Gaussians per state. However, we observed that results improved very slightly using over 65 Gaussians, so we decided to use that number of Gaussians per state. From the context-independent models, we trained context-dependent models and then performed state tying making use of an algorithm based on a decision tree. The models resulting from the state tying contained a total number of states of 2079. Given that the context-independent models contained a total of 26x3x65=5070 Gaussians, we chose to use context-dependent models with a complexity similar to that used in the context-independent ones. This way we can compare context-independent and context-dependent models. Following this reasoning we chose to use context-dependent models using 2 Gaussians per state, which implied a total of 2079x2=4158 Gaussians.

4. Acoustic-phonetic decoding results

The test we performed consisted of the evaluation of the accuracy of the acoustic phonetic decoding achieved with the models. In other words, we tried to determine the phonemic recognition accuracy using just the acoustic models, without any other kind of lexical or grammatical restriction. The only restrictions imposed were that each utterance should start and end with a silence. For the case of the context-dependent models, we also imposed that the contexts should be respected.

In order to evaluate the results, we aligned the phonemic string obtained from the decoder and the reference phonemic string obtained from the phonological transcriber (section 2.1). Using this alignment the percentage of phones correctly recognized (%C) and the phonemic decoding accuracy (%A) were computed. The phonemic decoding accuracy is the percentage of phones correctly detected minus the percentage of inserted phones.

4.1. Acoustic-phonetic decoding of read speech

It is important, before proceeding to further analysis of the results, to have an idea of the precision reached by the acoustic models under optimal conditions. These optimal conditions mean in our case read speech recorded under the same acoustical environment and conditions as the training speech. To assess that optimal performance we have made an acoustic-phonetic test on a subset of 300 utterances of the ALBAYZIN corpus. These utterances were phonetically segmented and labelled by hand and were not used in the training phase.

Using this test set, the acoustic-phonetic decoding with context-independent models reached %C = 81.07% correct phonemes and A = 76.56% phonemic accuracy. These results were evaluated using as reference phoneme strings the phoneme labels produced by the automatic transcriber based on the orthographic transcription. In order to check the validity of this phonemic string as reference string we also evaluated the same results comparing against the manually annotated reference phonemic labels. These results (%C = 81.36% and %A = 76.24%) are very similar to those using the automatically generated phonemic transcription. This justifies our evaluation of the acoustic-phonetic transcription of the C-ORAL-ROM corpus using an automatically generated reference phomemic labelling (there is not manually verified phonemic annotation for the C-ORAL-ROM corpus yet).

The former results were always using context-independent HMMs. If we use context-dependent models we obtain %C = 83.88% and %A = 74.55% when comparing against the automatic phonemic transcription. If we compare against the manual phonemic transcription results are very similar, %C = 83.79%, %A = 73.01%. Results obtained with context-dependent HMMs are also very similar to those obtained with context-independent HMMs.

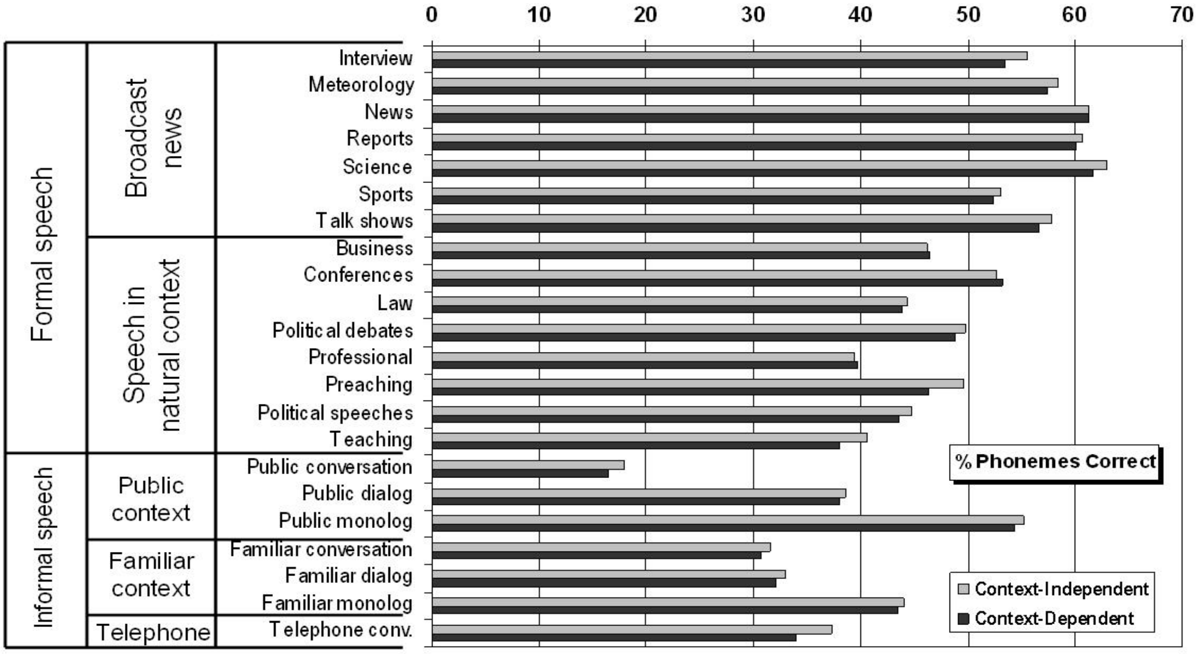

Figure 1: Phonemic decoding results (percentage of phones correct, %C) by subtype of spontaneous speech (see Section 2) of C-ORAL-ROM.

4.2. Acoustic-phonetic decoding of spontaneous speech

Once acoustic-phonetic decoding has been evaluated on read speech, we evaluate in this subsection the acoustic-phonetic decoding of spontaneous speech. If we perform the same test on the whole C-ORAL-ROM corpus using as reference phonemic labeling the automatically generated transcription, we get more modest results, as expected. In the case of using context-independent HMMs we obtain %C = 44.00% and %A = 25.71%. For the case of context-dependent HMMs results are again very similar, %C = 43.06% and %A = 25.07%.

Such a reduction in acoustic-phonetic decoding performance may be mainly due to the inherent difficulty that spontaneous speech presents for automatic processing. However, it would be misleading to consider that this is the only factor causing such a drastic decrease in phonetic decoding accuracy. Other factors that have an important impact on that reduction are the following:

- The channel mismatch between the speech used for training the HMMs and the speech on which the decoding was performed. ALBAYZIN is a head-mounted microphone, clean speech corpus, while C-ORAL-ROM is a corpus that includes speech recorded with different microphones on different acoustic environments (more or less noisy), speech taken from the media and even speech taken from telephone conversations. This channel mismatch is mitigated partially by the mechanisms of robustness against channel distortion and additive noise provided by the feature extraction front-end used [7]. However, its influence on the decoding results may still be important.

- The presence of noise with different characteristics and levels in C-ORAL-ROM. This effect is also mitigated, but not avoided, by the use of a front-end with mechanisms of robustness against noise [7].

- The mismatch between the characteristics of the speech used for training (ALBAYZIN) and testing (C-ORAL-ROM), both in type of speech and noise levels. It could be possible to perform a retraining or adaptation of the models using speech from C-ORAL-ROM. In this way, the acoustic models would be more adapted to the speed and level of the speech and the noise in C-ORAL-ROM, and presumably the phonetic decoding accuracy would increase.

All these factors limit the utility of the comparison between the phonetic decoding accuracy on read (Section 4.1) and spontaneous (Section 4.2) speech. However, even more interesting than this comparison is the comparison between the acoustic phonetic decoding accuracy on the different types of spontaneous speech in C-ORAL-ROM.

4.3. Comparison of phonetic decoding results on different types of spontaneous speech

Figure 1 shows the percentage of phones correctly recognized for each of the different types of spontaneous speech considered in C-ORAL-ROM, and briefly described in Section 2.

The comparison between context-dependent and context-independent acoustic models shows that results for both are very similar, although the context-independent ones are slightly better. This result could be due to the fact that our context-independent models are slightly more complex, since they include a larger overall number of Gaussians than the context-dependent models. Very likely an increase in the complexity of the context dependent models would produce an important improvement on the results shown here.

It is very interesting to see that there is a wide range of variation between the different types of spontaneous speech considered: from less than 20% phonemes correct for informal conversations in public context to over 60% phonemes correct for science programmes on the media.

In general, it can be observed that for conversations and dialogues results are among the worse obtained (around 30% phonemes correct for the whole group). Another subset related to them (in that it also contains dialogues and conversations) is the subset of telephone conversations for which results are similar. In all these cases it seems obvious that the interaction (with frequent overlappings) among the speakers is the cause of the reduced phonetic decoding performance. In the case of the telephone conversations there also exists a clear mismatch between the characteristics of the speech used to train the acoustic models and that used to perform the phonetic decodings.

Regarding the informal monologs, it can be observed that in familiar context results are slightly better than for dialogs and conversations (slightly over 40%), while in public context results are clearly superior (close to 55% phonemes correct).

The subsets mentioned in the former paragraphs correspond to informal speech. It can be observed that, with the only exception of the monologs in public context (epubmn), results are always worse than those obtained with formal speech, both in natural contexts and on the media (broadcast news). Comparing these two big groups it can be realized that speech from formal situations in natural context tend to produce results worse (around 40% or 50% phonemes correct) than those observed on formal speech on the media, for which phonetic decoding results tend to be between 50% and 60% phonemes correct.

Comparing the different subsets within the formal speech on the media, interesting differences may be observed. Worse results are obtained with sports programmes, probably due to a less careful use of the language and exaggerated articulations as well as more overlappings between different speakers. Slightly better are the results obtained on interviews, where overlappings might also be very frequent. Following, and with intermediate results, are the results on meteorology programmes and talk shows. Finally, best results are attained on news programmes, reports and scientific programmes. It might be argued that this kind of programmes have a reduced number of overlappings as well as a more careful use of language, presumably with less difluencies.

As a final experiment, we have compared the results of the automatic phonetic decoding with the problems found by human experts when transcribing the recordings in C-ORAL-ROM. These difficulties were analyzed in [8]. In doing this comparison we can observe very significative coincidences. In particular, human transcribers found serious difficulties with the typical features of interaction in a spontaneous communication: overlappings, number of words per turn, and speaking rate. In this case the following intuition applied

Scale 1: Degree of formality
informal media formal
+difficult ___________________________ - difficult

Scale 2: Number of speakers
conversation dialog monolog
+ difficult ___________________________ - difficult

On the first scale, the more formal the speech type, the easier to transcribe it because more rethoric and discursive conventions are followed. Speech is more predictable and pronunciation is more careful.

On the second scale, the more speakers talking on a recording, the more difficult the transcription because of the need of distinguish among the different turns and speakers and of the need to take care of overlappings. With monologs this difficulty is reduced to the minimum.

These findings coincide basically with those obtained in phonetic recognition experiments: the easier recordings to transcribe are those on the media, produced by professional speakers that combine a good diction with experience of fluid elaboration within the linguistic rules. The more we move towards informal speech with more speakers, the more complex are both manual and automatic transcription.

5. Conclusions and future work

In this paper we have presented phonetic decoding results on spontaneous speech and have compared them to the results obtained on read speech with the same characteristics as the speech used to train the acoustic models. This comparison shows an overall relative reduction of the percentage of phones correctly recognized of about 50% when we move from read speech to spontaneous speech. Although the influence of the characteristics of the type of speech (read vs. spontaneous) on the results of phonetic decoding is clear, it is also true that these experiments are also influenced by other factors like channel mismatch and noise level mismatch between training and testing. This makes the comparison between phonetic decoding accuracy for read and spontaneous speech of limited utility.

Much more interesting is the comparison of phonetic decoding results on different types of spontaneous speech. Among the different types of spontaneous speech analyzed, best results are those with speech taken from the media. For this kind of spontaneous speech results show a relative worsening of only 25% (approximately) from the results obtained on read speech with the same characteristics used to train the acoustic models. This means that this kind of spontaneous speech is the easiest to process automatically among those analyzed. Following in order of complexity are the formal speech in natural contexts, the informal monologs, and finally the informal dialogs and conversations, for which overlappings and interruptions make the complexity of the automatic processing of this kind of speech much higher than for the former types. These results largely coincide with the experience of human transcribers.

As future work, we would like to extend this study to take into more detailed consideration the influence of aspects like the frequency of overlappings, interruptions and disfluencies in acoustic-phonetic decoding results.

6. References

[1] http://nist.gov/speech/tests/rt/

[2] Luis Javier Rodríguez Fuentes. Estudio y modelización acústica del habla espontánea en diálogos hombre-máquina y entre personas. Tesis Doctoral. Facultad de Ciencia y Tecnología Universidad del País Vasco.

[3] Cresti & Moneglia (eds.), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, Amsterdam, John Benjamins, 2004.

[4] Climent Nadeu. ALBAYZIN, Universitat Politecnica de Catalunya, ETSET New Jersey, 1993.

[5] Moreno Sandoval, A. La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM. Actas de la II Jornadas en Tecnologías del Habla, diciembre 2002, Granada.

[6] Young, S. et al., The HTK Book (for HTK Version 3.1), Microsoft Corporation, July 2000.

[7] Aurora Front-End manual. ETSI ES 202 050 V1.1.3 (2003-11).

[8] González Ledesma, A.; De la Madrid, G.; Alcántara Plá, M.; De la Torre, R.; Moreno-Sandoval, A. Orality and Difficulties in the Transcription of Spoken Corpora. Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, 2004, Lisbon.