Evaluation of Sublexical and Lexical Models of Acoustic Disfluencies for Spontaneous Speech Recognition in Spanish

L.J. Rodríguez, I. Torres, A. Varona

Departamento de Electricidad y Electrónica. Facultad de Ciencias. UPV/EHU.
Apartado 644. 48080 Bilbao. SPAIN.
{luisja,manes,amparo}@we.lc.ehu.es

Abstract

Spontaneous speech is full of acoustic disfluencies that rarely appear in read or laboratory speech. A very simple and straightforward approach is presented, in which acoustic disfluences are modelled by augmenting the inventory of sublexical units, which originally consisted of 23 context independent phones plus a special unit for silent pauses. This set was augmented with 12 additional units accounting for lengthenings of sounds, filled pauses and noises. Two speech databases, both in Spanish, were used in the experiments. A phonetically balanced database was used for initializing the acoustic models. A spontaneous speech database consisting of 227 dialogues was used both for training and testing purposes. Recognition rates, in terms of acoustic-phonetic accuracy and word accuracy, with and without filtering acoustic disfluencies prior to alignments, were obtained to evaluate the contribution of these models to speech recognition. Also, some specific but significant examples were explored and discussed. Experimental results showed that using explicit models of acoustic disfluencies clearly improved the performance of a spontaneous speech recognition system.

1. Introduction

In the last years many interactive applications are emerging which use a dialogue strategy and natural language to get their purposes [1, 2, 3]. Applications like remote access to databases or interactive problem solving usually involve a great number of interactions (turns) between the user and the system. In that context users tend to speak in a more natural and spontaneous way than that of read-speech databases, not exactly as they would with a human, but introducing many acoustic, lexical and syntactic disfluencies: silent pauses, filled pauses, lengthenings, unfinished or mispronounced words, false starts, repetitions, etc.

An immediate consequence of this change to spontaneous input is that current speech recognition techniques must be improved to be robust against it. Improvements should include more robust acoustic models, more flexible lexical models and maybe a new specific language model for syntactic disfluencies. Additionally, in the framework of a dialogue system, a strategy must be defined for the interface between the speech recognizer and the speech understanding module, because disfluencies, once detected, could be just filtered or passed to upper levels.

A secondary consequence when using stochastic models is that, to train and test them, databases are needed with annotations of disfluencies at all levels. The creation of such databases is a tedious and time-consuming, but a necessary task –which we recently accomplished for a dialogue task in Spanish [4]. In fact, it has been demonstrated that annotating and characterizing disfluencies improves the performance of speech recognition systems remarkably [5].

Acoustic disfluencies cause a great variability in segment durations, intonation, articulation quality, etc. [6]. Many strategies have been applied in the literature to specifically deal with them [7, 8]. Additionally, works related to pronunciation modelling, durational modelling and automatic inference of topologies point at the same objective.

Here we present a first approach to the problem of modelling acoustic disfluencies in Spanish, developed in the framework of the dialogue task mentioned above. The approach is based on defining specific acoustic and lexical units for acoustic disfluencies, and then experimentally measuring the contribution of such models to recognition. The paper is organized as follows. Section 2 briefly describes the main features of the spontaneous speech database, including the inventory of acoustic disfluencies. Section 3 presents the methodology. Section 4 gives the conditions, the results and a brief discussion of phonetic and word-level experiments. Finally, Section 5 gives some conclusions.

2. Description of the spontaneous speech database

A spontaneous speech database containing dialogues in Spanish was recorded at 8 kHz across telephone lines applying the well known *Wizard of Oz* (*WoZ*) mechanism [9]: a human operator simulated the desired behaviour of the dialogue system, including some recognition and/or understanding errors, so that callers could think they were interacting with a real system. Callers (*users*, henceforth), who were in fact recruited volunteers, were given each one three scenarios with dates and times for departures and/or arrivals, preferred routes and other conditions for long-distance travels by train between two spanish cities. They were told to get as much information as they wanted (timetables, prices, etc.) from the dialogue system, doing it in a natural manner, just as they would in a real call. The *wizard* looked for the requested information into a true database and typed the answer according to a predefined set of templates [10]. Finally, the answer was converted to speech by a synthesis program, and then sent back to the user.

The database includes 227 dialogues (75 users $\times$ 3 scenarios + 2 repeated scenarios), which amounts to 1657 non-empty

This work was partially supported by the Spanish CICYT, under project TIC98-0423-C06-03.

user turns or about 150 minutes of speech. So, an average dialogue had seven non-empty user turns, each one lasting around five seconds. In fact, many user turns consisted of simple answers like "*yes*" or "*no, thanks*", but in contrast many others consisted of several sentences, full of spontaneous speech events, lasting up to 55 seconds. Ortographic transcriptions were obtained by listening the speech signals corresponding to user turns and taking the answers typed by the *wizard*. Noises and filled pauses –at least the most obvious– were coarsely annotated. These transcriptions were subsequently augmented with more specific annotations, corresponding to disfluencies, semantic frames and speech acts.

2.1. The annotation of disfluencies

After careful listenings, the original ortographic transcriptions of user turns were corrected and augmented with acoustic, lexical and syntactic disfluencies. A XML-like annotation format and the corresponding procedure were defined, and a simple text editor used to insert the marks [4].

Acoustic disfluencies included filled and silent pauses, lengthenings of sounds –especially vowels–, and noises. These latter should not strictly be considered as disfluencies; in fact, channel and environmental noises do not depend on the linguistic component, but on the recording conditions, and are randomly distributed. It's their pervasiveness in spontaneous speech what suggests to annotate these acoustic events, because more accurate annotations should produce more accurate acoustic models. Lexical disfluencies –unfinished and mispronounced words–, and syntactical disfluencies –mostly repetitions and reformulations– were annotated too. Finally, a special category was defined for discourse markers, which –like noises– were annotated due to their pervasiveness in spontaneous speech. The inventory of acoustic disfluencies defined for this work is showed in Table 1.

Table 1: Inventory of acoustic disfluencies

Category	Sub-category
Noises	environmental isolated environmental affecting words speaker aspiration speaker lips speaker cough
Silent pauses	
Filled pauses	/a/ /e/ /m/ trash
Lengthenings	/a/ /e/ /i/ /l/ /m/ /n/ /o/ /s/ /u/

3. Methodology proposed to model acoustic disfluencies

The aim of this work was to test the performance of a speech recognizer, with and without modelling acoustic disfluencies. Before attempting more sophisticated approaches, a straightforward methodology was set, by defining one sublexical unit for each acoustic event identified as disfluency.

We accepted the hipothesis that filled pauses and lengthenings are basically the same acoustic event, in the first case as independent non-lexical events, in the latter as part of the acoustic realization of a word. So, for instance, the samples corresponding to lengthenings of the sound */m/* can be put together with those of the filled pauses identified as */m/*, thus allowing the joint training of just one model. Once gathered the samples of the filled pauses and the corresponding lengthenings, those events with not enough training samples, in particular the lengthenings of vowel */u/* and the speaker coughs, were discarded. With regard to environmental noises affecting words, for the moment we decided to handle noisy phrases just as they were free of noise.

Finally, the set of sublexical units was formed by 23 context independent phones –widely used for continuous speech recognition in Spanish [11]–, that we will call *Phone-Like Units* (PLU), and 13 additional units corresponding to acoustic events happening in spontaneous speech: three for noises, one for silent pauses, three for filled pauses gathered with equivalent lengthenings, one for not clearly identified filled pauses (*trash* in Table 1), and five more for lengthenings. We will refer to the resulting set of 36 units as *Acoustic Units* (AU).

The dictionary used for the word-level recognition experiments was augmented with 13 new entries, corresponding to the acoustic events mentioned above. Lexical models for these units were built just by copying the corresponding acoustic models. Since these models did not represent true words, we called them *pseudo-words*. On the other hand, those lengthenings appearing at the beginning or at the end of a word were replaced by a normal phone in the word transcription plus an additional word –representing the lengthening– before or after, respectively, the original word transcription. Consider the following example:

el tren qu⟨lengthening:e⟩ llega a las ⟨lengthening:o⟩cho [1]

After replacing the lengthenings that appear at the borders of words, the sentence would result in:

el tren que ⟨lengthening:e⟩ llega a las ⟨lengthening:o⟩ ocho

This allows to model such lengthenings as independent pseudo-words. A new language model was trained taking into account the new lexical models. The inclusion of pseudo-words in the language model makes sense since some of them –like lengthenings, filled and silent pauses– show regular patterns and can be used as clues for detecting disfluencies at higher levels [12, 13].

4. Experimental evaluation

The significance of modelling acoustic disfluencies was experimentally evaluated over the database of Spanish dialogues described in Section 2. Two sets of sublexical units were evaluated: the first one consisted of PLU's plus one auxiliary model for silent pauses; the second one consisted of the whole set of AU's, as defined above.

4.1. Acoustic-phonetic decoding experiments

Multiple codebook discrete hidden Markov models were used to represent sublexical units. Each model had a simple left to right topology with three states, each of them including a loop but without transition between non-consecutive states.

[1] translated to English as: *the train that arrives at eight*.

Four codebooks of 256 codewords were used in these experiments: Cepstrum, δCepstrum, ΔCepstrum and Energy + δEnergy. To initialize the set of PLU's and the acoustic model for silent pauses, Baum-Welch and Viterbi training algorithms were sequentially applied over the training set of a phonetically-balanced medium-size read-speech database in Spanish. This database consisted of 1529 sentences, uttered by 47 speakers, and involved around 60000 phones. The resulting models were used to initialize the acoustic models for lengthenings and filled pauses. The acoustic models corresponding to noises and filled pauses not clearly identified (*trash* in Table 1) were initialized randomly. The whole set of AU's were further refined by iteratively applying the Viterbi procedure over the database of spontaneous speech dialogues. The training set consisted of 191 dialogues carried out by 63 different speakers, giving a total amount of 1349 user turns, containing around 64500 AU's. The test set used in these experiments consisted of 36 dialogues (308 user turns) corrresponding to 12 speakers not included in the training set, and involved around 14200 AU's.

For each experiment, after optimal alignments between the recognized and the correct transcriptions, the following values were obtained: the number of units that were correctly recognized (*c*), the number of units that were inserted (*i*), deleted (*d*) and substituted (*s*). From these values, the *Unit Error Rate* (UER) was computed as follows:

$$\%UER = \frac{i + d + s}{i + d + s + c} \times 100$$

Table 2 shows the %UER obtained through a series of acoustic-phonetic decoding experiments. When using AU's, the %UER was also obtained in terms of PLU's. In this latter case, the acoustic disfluencies were filtered before alignments.

Table 2: Experimental results obtained through acoustic-phonetic decoding experiments for PLU's and AU's.

units	%UER	
	AU	PLU
PLU	–	**47.69**
AU	50.30	**45.62**

Table 2 shows that an explicit modelling of acoustic disfluencies led to an improvement of system performance, when measured in terms of PLU error rates. A detailed analysis of the confusion matrix showed that most times acoustic disfluencies were correctly recognized; in fact, the AU's corresponding to acoustic disfluencies obtained lower %UER than PLU's themselves. Moreover, the confusion among units in the same category, noises for example, was quite low.

Consider the following example, which shows the correct phonetic transcription and the output of the acoustic-phonetic decoder, both in terms of the AU's, with PLU's in SAMPA format [14]:

k u a l e s e l o r a r i o d e l o s t r e n e s ⟨pause⟩ ⟨filled:m⟩ d e b a r T e l o n a ⟨pause⟩ a b a L ⟨lengthening:s⟩ ⟨pause⟩ p a r a e l d i a k i n T e d e x u n i o [2]

⟨pause⟩ p x u a m e s r e b a i e b o s t e n e s ⟨pause⟩ ⟨filled:m⟩ t e b a x l o m a ⟨pause⟩ a b a i n ⟨lengthening:s⟩ ⟨pause⟩ p a rr a n t L e k i n T d e x u m i ⟨pause⟩ ⟨filled:trash⟩

[2]The translation to English, without disfluencies, of the original sentence would be more or less: *what is the timetable for trains from Barcelona to Valls on June the fifteenth*.

Note that sublexical units corresponding to acoustic disfluencies were recognized better than PLU's, and were correctly placed in the recognized string. Note also the insertion of silent pauses, for instance at the beginning and at the end of the sentence. In fact, they should not be considered as errors, because only relatively long silent pauses were annotated in the original transcriptions, which usually did not include neither initial nor final silences –as in the latter example.

4.2. Word-level experiments

For lexical decoding experiments, a language model was generated by a k-testable in the strict sense (k-TSS) grammar. k-TSS language models can be considered as a syntactic approach to the well known n-grams, where k plays the same role as n does in n-grams [15]. The database of spontaneous speech dialogues is a medium-size vocabulary task of 2000 words. The ortographic transcriptions corresponding to the same training set defined for acoustic-phonetic decoding experiments, containing around 16500 words, was used to train several k-TSS language models, with $k = 2$, 3 and 4. Then, the language model was integrated in the lexical decoder, where each word was represented by a linear chain of hidden Markov models corresponding to its standard phonetic transcription.

Table 3: Lexical decoding results when using the PLU's and the AU's to build the lexicon and the extended lexicon, respectively.

units	%WER					
	extended lexicon			lexicon		
	$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
PLU	–	–	–	**43.70**	**41.74**	**41.91**
AU	42.75	41.15	41.20	**40.27**	**38.01**	**37.84**

For a baseline series of experiments, PLU's were used to build word models. The test set included 3500 words. Table 3 shows the %WER (*Word Error Rate*, defined the same way as UER) for $k = 2$, 3 and 4 language models. In a second series of experiments, the lexicon was extended with 13 pseudo-words representing acoustic disfluencies. The inclusion of pseudo-words augmented the training set up to 20000 samples, and the test set up to 4600 samples. Three k-TSS language models, with $k = 2$, 3 and 4, were trained again, this time over the database with the extended lexicon. Table 3 shows the %WER for these experiments. The %WER was also obtained in terms of the original –not extended– lexicon. In this latter case, the pseudo-words were filtered before alignments.

Table 3 shows that the use of pseudo-words to represent acoustic disfluencies improved the performance of the lexical decoder, even when measured in terms of %WER over the extended lexicon. Once again, acoustic disfluencies were quite well recognized and correctly placed in the resulting string of words. When the %WER was measured over the not extended lexicon, i.e. when filtering pseudo-words prior to alignments, the increase of the system performance was still more significant. Consider the reference transcription, in terms of words and pseudo-words, for the same example showed in Section 4.1:

cuál es el horario de los trenes ⟨pause⟩ ⟨filled:m⟩ de Barcelona ⟨pause⟩ a Valls ⟨lengthening:s⟩ ⟨pause⟩ para el día quince de junio

When using PLU's to build the word models in the not-extended lexicon, which also included an auxiliary model for silent pauses, the output of the lexical decoder for the previous sentence was:

> cuál es de de los trenes de Barcelona para estar el día quince de junio ⟨pause⟩

Various errors can be observed: "*los horarios*" is replaced by "*de*", "*a Valls*" is deleted and "*estar*" inserted. This reveals that acoustic disfluencies force the search algorithm to deviate from the correct path to accommodate them. We argue that these errors could be avoided by including in the lexicon specific models for acoustic disfluencies. In fact, when using pseudo-words to build the extended lexicon, the output of the lexical decoder was very close to the reference transcription, showing a high recognition accuracy for the acoustic disfluencies:

> cuál es el horario de los trenes ⟨pause⟩ ⟨filled:m⟩ de Barcelona ⟨pause⟩ a París ⟨pause⟩ para el día quince de junio

As shown by these examples, the use of pseudo-words representing acoustic disfluencies helped the system to improve its output. We argue that acoustic disfluencies, due to their relatively large duration –or to the fact that the availability of many samples produces more robust models– are easily detected, acting as anchors for the recognition procedure, as proved by comparing the output string of the lexical decoder with that of the acoustic-phonetic decoder.

It has been found in a previous work [4] that acoustic, lexical and syntactic disfluencies tend to appear all together. At least, whenever a lexical or syntactic disfluency takes place, acoustic disfluencies appear in its surroundings. So, the accuracy observed in recognizing acoustic disfluencies is an interesting feature to be considered by a system aiming at recognizing and processing lexical and/or syntactic disfluencies. Alternatively, acoustic disfluencies could be filtered, as we did in some of the experiments, to supply an acoustically clean – but disfluent at lexical and syntactic levels– string of words to a hypothetical semantic decoder in the framework of a speech understanding system.

5. Conclusions

This paper presented a straightforward approach to model acoustic disfluencies in a spontaneous speech recognition system. It was based on augmenting the inventory of sublexical units with new ones corresponding to a series of disfluent acoustic events, including noises, filled and silent pauses and lengthenings of sounds. Phonetic and word-level recognition experiments over a spontaneous speech database in Spanish proved the goodness of such a simple approach. Unit and word error rates were reduced in around 4% in the first case, and around 9% in the second. On the other hand, acoustic disfluencies were quite well recognized and correctly placed in the output string. As acoustic, lexical and syntactic disfluencies tend to appear all together, we conclude that this reliable recognition of acoustic disfluencies should help to detect the presence of lexical and/or syntactic disfluencies.

6. References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 100–112, January 2000.

[2] L. Lamel, "Spoken language dialog system development and evaluation at LIMSI," in *Proceedings of the International Symposium on Spoken Dialogue*, Sydney, Australia, November 1998.

[3] A.L. Gorin, G. Riccardi, and J.H. Wright, "How May I Help You," *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, October 1997.

[4] L.J. Rodríguez, I. Torres, and A. Varona, "Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish," in *Proceedings of the Workshop on Disfluency in Spontaneous Speech*, University of Edinburgh, Scotland, August 29-31 2001.

[5] R.C. Rose and G. Riccardi, "Modeling disfluency and background events in ASR for a natural language understanding task," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1999, pp. 1709–1712.

[6] E. Shriberg, "Phonetic consequences of speech disfluency," in *Proceedings of the International Congress of Phonetic Sciences*, 1999, vol. 1, pp. 619–622.

[7] D. O'Shaughnessy, "Better detection of hesitations in spontaneous speech," in *Proceedings of the Workshop on Disfluency in Spontaneous Speech*, U.C. Berkeley, July 30 1999, pp. 39–42, Satellite Meeting of ICPhS99.

[8] D. Liu, L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, and R. Schwartz, "Improvements in spontaneous speech recognition," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, February 8-11 1998, URL: http://www.nist.gov/speech/publications/darpa98/.

[9] J.B. Mariño and J. Hernando, "Especificación de las grabaciones mediante Mago de Oz," Research Note BS16AV10, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, November 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[10] I. Esquerra, A. Sesma, and J.B. Mariño, "Generación de respuesta para el Mago de Oz," Research Note BS61AV23, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad Politécnica de Cataluña, December 1999, URL: http://gps-tsc.upc.es/veu/basurde/Home.htm.

[11] I. Torres and F. Casacuberta, "Spanish phone recognition using semicontinuous hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1993, vol. II, pp. 515–518.

[12] E. Shriberg, J. Bear, and J. Dowding, "Automatic detection and correction of repairs in human-computer dialog," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992, pp. 419–424.

[13] E. Shriberg and A. Stolcke, "Word predictability after hesitations: A corpus-based study," in *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, 1996, pp. 1868–1871.

[14] SAM Phonetic Alphabet (SAM: Speech Assessment Methods, ESPRIT Project 1541), URL: http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[15] I. Torres and A. Varona, "k-TSS Language Models in Speech Recognition Systems," *Computer, Speech and Language*, vol. 15, no. 2, April 2001.