CHARACTERIZATION OF HESITATIONS USING ACOUSTIC MODELS

Arlindo Veiga[a,c]; *Sara Candeias*[a]; *Carla Lopes*[a,b,c] *& Fernando Perdigão*[a,c]

[a]Polo de Coimbra, Instituto de Telecomunicações, Coimbra, Portugal;
[b]ESTG, Campus 2, Instituto Politécnico de Leiria, Leiria, Portugal;
[c]DEEC, Polo II, Universidade de Coimbra, Coimbra, Portugal
aveiga@co.it.pt; saracandeias@co.it.pt; calopes@co.it.pt; fp@co.it.pt

ABSTRACT

Spontaneous speech is full of hesitations, such as fillers, word cut-offs, repetitions and segmental extensions. Automatic identification of such hesitations has several applications; however, it is a challenging research problem. In this paper acoustic-phonetic properties of hesitation phenomena are explored in order to identify and annotate some of these events in a spontaneous speech corpus of Portuguese broadcast television news. Based on pitch, energy, spectral and durational characteristics of the filled pauses and segmental extensions during their production, we intend to characterize the acoustic-phonetic regularity of the phenomena. A speech recognition sys-tem was used to help locating the filled pauses and extensions. The events detected were then manually validated. Our preliminary results suggest that there are regular trends in the production of these hesitation events, which could distinguish them from other events within the structure of Portuguese. Our purpose with this work is to improve acoustic modeling for spontaneous speech recognition systems. Some insights into the process of human speech communication for Portuguese are gained as well.

Keywords: filled pauses, extensions, acoustic-phonetic features, Portuguese spontaneous speech

1. INTRODUCTION

Spontaneous and read speech have diverse structures both acoustically and syntactically. The presence of hesitations such as filled pauses, extensions, repetitions and word cut-offs is very common in spontaneous speech and plays an important role in the structuring of speech [14]14, 24]. Hesitation events can be used to identify the idiosyncrasy of the speakers and also to improve the performance of automatic speech recognition systems. In this study we concentrated on both the filled pauses (FPs) and extensions (EXs), present in spontaneous Portuguese speech. FPs comprise all sounds that phonetically belong to the Portuguese language but do not occur in the context of a complete word (e.g., *uum*, *aaa*, *eee*). With EXs we mean the phonetic prolongation into both functional and lexical words (e.g. [ɐ] in <*para*> or the [u] in <*do*>). FPs and EXs for English and other languages is an issue that has been widely addressed by the scientific community (e.g. in [4, 5, 6, 8 12, 24]). Nevertheless for Portuguese, only a few language studies focus on this problem. Although the main topic of the work by Freitas [9] and Delgado-Martins [7] is not FPs phenomena, they show that duration features and syntactic information are responsible for the distinction between spontaneous speech, oral and reading presentations. In Mata [15], following Moniz and her colleagues [18], FPs characteristics are presented to demonstrate the contribution of the fundamental frequency trend for on-line planning efforts both in spontaneous speech and in oral reading. An important report about the distinction between fluency and disfluencies in the communication process within a teaching context is presented by Moniz in [19]. Despite interesting conclusions, the study was based on a sample that is scarcely representative of the phenomenon. Updates made in [18] and [21], exploring prosodic cues in an attempt to classify (dis)fluency, seem to confirm the limited representativeness of the phenomenon.

The acoustic-phonetic characterization of FPs and EXs will certainly lead to improved speech recognition systems. In fact, the presence of hesitations in speech signals negatively affects the performance of the automatic speech recognition (ASR) systems. Dealing with this problem becomes a challenge to the recognizers and various techniques have been proposed to find hesitations in the speech signal, including FPs. Some studies deal with identifying the strict location in time of the hesitation event (e.g. [2, 16, 28]), while others

analyze specific properties of the hesitations [1, 11, 27, 30]. The study reported here intends to study the acoustic-phonetic cues that could be considered to detect FPs and EXs, such as pitch, energy, spectral and durational characteristics, as well as their relation with phones.

The remainder of the paper is organized as follows. In section 2 the acoustic-phonetics characteristics described in this study are presented. Section 3 presents data and shows how hesitation events were automatically detected. The achievements found related to the acoustic parameters studied are also described. Finally, conclusions are drawn in section 4.

2. ACOUSTIC-PHONETIC CHARACTERISTICS

This section describes the acoustic-phonetic cues for each hesitation type. Vocal tract resonances (i.e. the formants), pitch, spectral and durational structure of FPs and EXs were investigated in a first stage and analyzed at a second stage. These acoustic-phonetic characteristics are used in most studies for detecting hesitation events. Shriberg et al, [25, 26], have obtained information about the variation in the F0 value and other duration and voicing based features. The work carried out by Masataka and al. [13] and Garg and al. [10] provided some specific features of FPs as general indicators of their presence in speech: they showed that FPs have a flat pitch and a constant energy which falls towards the end of the utterance. Audhkhasi, [2], argued that formant based features are better identifiers than pitch based features for FP detection. Kaushik and his colleagues [11] proposed the computation of previous features including duration, pitch and spectral and formant based features to carry out an algorithm for identifying (and removing) FPs in spontaneous speech. In the study reported here the phonetic properties are derived from the European Portuguese phonology background [18, 22]. The analysis of the hesitation events into the three-region surface structure, such as the reparandum, the editing phase and the repair (terms adapted from Levelt [13] and used by Shriberg in her studies) have not been considered yet.

Several hesitation detectors have been proposed (e.g. [2, 10, 11, 31]) using acoustical cues. The performance of the detectors depends on a suitable characterization of the hesitation units in acoustical terms and can only be achieved if a suitable database of the events is available. We start by using a Portuguese phone multi speaker recognizer to provide some cues about probable localization of FPs and EXs in the speech signal.

3. FILLED PAUSES AND EXTENSION CORPUS

In order to explore the acoustical-phonetic features of FPs and EXs, a large number of examples of both occurrences is necessary. Since there is no public European Portuguese database with this kind of annotated events, we collected podcasted television news, resulting in around 22 hours of non-annotated speech. Because hesitations occur mainly in spontaneous speech, almost only the parts with interviews have hesitations. However, annotating FPs and EXs present in the audio signal by an expert is a time-consuming task and so a semi-automatic procedure was employed. For that we used a phone recognizer with several restrictions in terms of phone sequences and durations in order to give probable hypothesis of FP or EX events. These events were later manually accepted or rejected.

3.1. Data collection and annotation

Multimedia signals from podcasts of television news were collected and the audio converted to a 16 kHz sampling rate. In order to recognize events, the speech was analyzed according to the recognizer front-end: 12 Mel-frequency cepstral coefficients, plus log energy, and their first and second order regression coefficients, at a frame-rate of 100Hz. The phone acoustic models are defined in terms of Hidden Markov Models (HMM), which were previously built using HTK 3.4 [29] and the TECNOVOZ database, [15]. A HMM decoder is applied to speech segments between detected pauses, providing an optimal phone sequence. The phone sequence is then analyzed to locate FPs or EXs. A hesitation candidate was hypothesized when:

- a vocalic phone was longer than a pre defined threshold;
- sequences of similar phones occur, such as:
  - [õ], [w̃], [ũ], [ẽ], [ĩ], [j̃];
  - [o], [ɔ], [õ], [ũ], [w̃];
  - [e], [ɛ], [ẽ];
  - [ɐm];
- unvoiced phones may also occur among voiced ones if their duration is short.

A 350ms threshold for vocalic phone duration was chosen considering the average duration of Portuguese vowels. This duration threshold agrees to the value suggested in [3]. The event detector implements also a confidence measure based on the phone durations in each candidate segment. The detector is built such that an insertion (false alarm) will be preferred to a miss. The target is not a detector with very good performance, but one that does not miss FPs or EXs present in the signal. The output of this system is a sequence of possible hesitation events. A human expert then validates or rejects the hesitation proposals and assigns a phonetic label to each validated event. Our estimation is that this semi-automatic procedure reduced the duration of the annotation task by at least 4 times, compared to the completely manual annotation process.

Table 1: EX (extensions) and FP (filled pauses) occurrences.

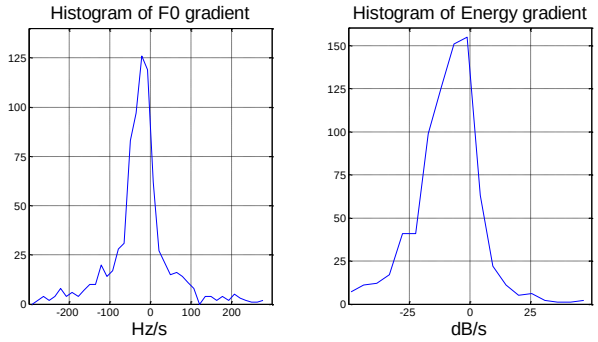
Hesitation	Type		Label	#Occurrences
FPs	vowels	oral	ə	53
			ɐ	198
		nasal	ɐm	21
EXs	vowel	oral	ə	70
			ɐ	61
			a	25
			ɛ	37
			e	13
			i	58
			ɔ	20
			u	61
		nasal	o	18
			õ	14
			ũ	35
	diphthong	oral	jɐ	6
		nasal	ɐ̃j̃	8
			ɐ̃w̃	18

The FP and EX corpus of transcriptions thus obtained includes about 800 event annotations. Extensions are more frequent than FPs, covering 62% of the annotated events. 15 different labels for FPs, were found but they rely mainly on two FPs: [ə] and [ɐ], representing respectively 17.8% and 66.4% of the total number of occurrences. The remaining FPs include [ɐe], [ɐj], [a], [ɛ], [e], [ɛm], [i], [ɔ], [u], although there are only a few examples of each. 33 different labels were set to EX; nevertheless, we discarded labels with few occurrences, so for this study only 14 were considered. Table 1 presents the EXs and FPs based on the phonetic label assigned to the event.

The most frequent EX is [ə], followed by [ɐ] and [u]. The extension of the [i] is also common in spontaneous Portuguese, representing the fourth more frequent EX. The open and open-mid vowels, such as [a], [ɛ] and [ɔ] were not recognized as so frequent. An interesting fact is the lengthening of the diphthongs (both oral and nasal), in which the most frequent is the diphthong [ɐ̃w̃]. In the diphthong category, we considered both the rising and falling diphthongs.

We have verified that EXs occur mainly in prepositions and on the last syllable. Sometimes the difference from FP or EX is not obvious and is distinguished only in the phonetic context.

Figure 1: Histograms of the gradient of F0 and energy during hesitations.



3.2. Data analysis

From the hesitation events we computed several acoustic parameters to characterize hesitation segments, namely F0 (pitch), energy and spectrum. We are interested mainly to confirm characteristics known for other languages, namely the constancy of these parameters during the events. For each event an average value of pitch and energy, as well its deviation is computed. The gradient of these parameters was also considered as the linear regression coefficient of its variation within the segment. Equivalent values for spectrum were computed using 32 frequency bands (on a mel scale) and its deviation from the average spectrum in the segment.

The gradients of F0 and energy in hesitation segments present, most of the times, negative values, which means that they decay smoothly during hesitations (Figure 1). However, these values have small variation: the standard deviation of F0 is on average around 15 Hz and standard deviation of energy is on average around 2.7 dB. The parameter based on standard deviation of spectral band energies show a similar behavior.

Also observed is that these characteristics do not separate well between FP and EX hesitations, which agrees with the fact that perceptually their distinction is also ambiguous without a context.

4. CONCLUSION

This paper addresses the problem of collecting a database of filled pauses and extensions in spontaneous speech from broadcast news using a phone recognizer. Although it is not the optimal method, it proved to be useful for semi-automatic annotation. The detected events were characterized phonetically and acoustically. In the near future we intend to explore hesitations within utterances using the three-region surface structure.

5. ACKNOWLEDGMENTS

The three first authors acknowledge *Instituto de Telecomunicações* (Arlindo Veiga), *Science and Technology Foundation-FCT* (Sara Candeias, SFRH/ BPD/36584/2007) and (Carla Lopes, SFRH/BD/27966/2006) for their scholarships.

6. REFERENCES

[1] Audhkhasi K. 2009. Automatic evaluation of fluency in spoken language. *IETE Tech Rev.* 26, 108-114.

[2] Audhkhasi, K., Kandhway, K., Deshmukh, O.D., Verma, A. 2009. Formant-based technique for automatic filled-pause detection in spontaneous spoken English, *ICASSP-IEEE* 4857-4860.

[3] Bartkova, K. 2005. Prosodic cues of spontaneous speech in French. *Proc. of DiSS'05, Disfluency in Spontaneous Speech Workshop* Aix-en-Provence, France, 21-25.

[4] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *J. Acoust. Soc. Am.* 113(2), 1001-1024.

[5] Candea, M. 2000. *Contribution à l'Etude des Pauses Silencieuses et des Phenomenes Dits «d'Hesitation» en Français Oral Spontané – Etude sur un Corpus de Récit en Classe de Français.* Ph.D. diss., Universite Paris III – Sorbonne Nouvelle.

[6] Clark, H.H., Fox Tree, J.E. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 73-111.

[7] Delgado,M.R., Freitas, M.J. 1991. Temporal structures of speech: reading news on TV. *ETRW' 91* Barcelona.

[8] Eklund, R. 2004. *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues.* Ph.D. diss., Institute of Technology, Linköping University.

[9] Freitas, M.J.R. 1990. *Estratégias de Organização Temporal do Discurso.* MSc thesis. Faculdade de Letras da Universidade de Lisboa.

[10] Garg, G., Ward, N., 2006. *Detecting Filled Pause in Tutorial Dialogs.* The University of Texas at El Paso.

[11] Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A. 2010. Automatic detection and removal of disfluencies from spontaneous speech. *Proc. 13th Australasian Int. Conf. on Speech Science and Technology* Melbourne, 98-101.

[12] Lee, T-L, He, Y-F, Huang, Y-J, Tseng, S-C, Eklund R. 2004. Prolongation in spontaneous Mandarin. *Inter-speech' 04* Jeju Island, Korea, 2181-2184.

[13] Levelt, W.J.M. 1983. Monitoring and self-repair in speech. *Cognition* 14, 41-104.

[14] Levelt, W.J.M. 1989. *Speaking: From Intention to Articulation.* Cambridge, Massachusetts: MIT Press.

[15] Lopes, J., Neves, C., Veiga, A., Maciel, A., Lopes, C., Perdigão, F., Sá, L. 2008. Development of a speech recognizer with the Tecnovoz database. *Propor 2008, Int. Conf. on Computational Processing of Portuguese* Aveiro, Portugal, 260-263.

[16] Masataka, G., Katsunobu, I., Satoru, H. 2000. A real-time system detecting filled pauses for spontaneous speech. *IEICE Trans. on Information and Systems*, Pt.2, Vol.J83-D-2, No.11, 2330-2340.

[17] Mata, A.I. 1999. *Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didácticas*. Ph.D. diss., Universidade de Lisboa, Faculdade de Letras.

[18] Mateus, M.H., d'Andrade, E. 2000. *The Phonology of Portuguese.* Oxford: Oxford University Press.

[19] Moniz, H. 2006. *Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu*. MSc thesis, Univ. Lisboa, Faculdade de Letras.

[20] Moniz, H., Mata A.I., Viana, M.C. 2007. On Filled Pauses and Prolongations in European Portuguese. *Interspeech' 07, ISCA* Antwerp, Belgium, 2645-2648.

[21] Moniz, H., Trancoso, I., Mata, A.I. 2009. Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. *Interspeech' 09, ISCA* Brighton, UK, 1719-1722.

[22] Morais-Barbosa, J. 1965 *Etudes de Phonologie Portugaise.* Lisboa: JIU.

[23] O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. *ICASSP'92*, I-521-524.

[24] Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. Diss., University of California.

[25] Shriberg, E. 1999. Phonetic consequences of speech disfluency. *Proc. of the International Congress of Phonetic Sciences* San Francisco, 1, 619-622.

[26] Shriberg, E., Bates, R., Stolcke, A. 1999. A prosody-only decision-tree model for disfluency detection. *Proc. of Eurospeech* Rhodes, Greece, 2383-2386.

[27] Shriberg, E., Stolcke, A. 2002. Prosody modeling for automatic speech recognition and understanding. *Proc. Workshop on Mathematical Foundations of Natural Language Modeling*.

[28] Snover, M., Dorr, B., Schwartz, R. 2004. A lexically-driven algorithm for disfluency detection. *Proc. North American Chapter of the Association of Computational Linguistics.* Boston, 157-160.

[29] Young, S., *et al*, 2006. *The HTK book. Revised for HTK version 3.4*. Cambridge University Engineering Department, Cambridge.

[30] Zechner, K., Bejar, I. 2006. Towards automatic scoring of non-native spontaneous speech. *Proc. Hu-man Language Technologies Conference* New York, 216-223.

[31] Žgank,A., Rotovnik, T., Maučec, M., 2008. Modeling filled pauses for spontaneous speech recognition applications, *Proc. 7th WSEAS International Conference on Application of Electrical Engineering* Norway, 42-47.