HESITA(te) in Portuguese

Sara Candeias[1], Dirce Celorico[1], Jorge Proença[1], Arlindo Veiga[1,2], Carla Lopes[1,3], Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Coimbra, Portugal

[2] Electrical and Computer Eng. Department, University of Coimbra, Portugal

[3] Instituto Politécnico de Leiria, Leiria, Portugal

IT, Department of Electrical and Computer Engineering, University of Coimbra - Pole II, 3030-290, Coimbra, Portugal
E-mail: {saracandeias,dircelorico,jproenca,aveiga,calopes,fp}@co.it.pt

Abstract

Hesitations, so-called disfluencies, are a characteristic of spontaneous speech, playing a primary role in its structure, reflecting aspects of the language production and the management of inter-communication. In this paper we intend to present a database of hesitations in European Portuguese speech - HESITA - as a relevant base of work to study a variety of speech phenomena. Patterns of hesitations, hesitation distribution according to speaking style, and phonetic properties of the fillers are some of the characteristics we extrapolated from the HESITA database. This database also represents an important resource for improvement in synthetic speech naturalness as well as in robust acoustic modelling for automatic speech recognition. The HESITA database is the output of a project in the speech-processing field for European Portuguese held by an interdisciplinary group in intimate articulation between engineering tools and experience and the linguistic approach.
Keywords: Disfluencies, Hesitations, European Portuguese

1. Introduction

Filled pauses with non-lexical segments, such as *uum*, *mm*, *amm* or *aa*, fillers like *pois*, *bem* ('well' in English), vocalic extensions within words, such as *deeeee* ('of' in English), cut words like *pa- para a* ('fo- for' in English) and repetitions (*de de*, 'of of') are prevalent linguistic events in spontaneous spoken language which fall under the category of hesitations, employed here as a synonym for disfluencies (Levelt, 1989; Shriberg, 1994; Clark, 1996). Several works in the last decade have underlined the importance of acquiring knowledge on hesitation events for the successful development of speech technology and to facilitate natural language processing tasks (Shriberg, 1994; Eklund & Shriberg, 1998; Veiga et al., 2012a; Veiga et al., 2012b; Moniz et al. 2012). Automatic speech recognition benefits from the consideration of hesitations for more robust language and acoustic models (Veiga et al., 2012b; Liu et al., 2006) as well as speech synthesis by improving the naturalness of speech (Adell et al., 2008). Detection of hesitation events also enables the segmentation of multimedia data into consistent parts, as claimed in Veiga et al. (2012b). It leads to important applications such as the identification of speech segments to train acoustic models for speech recognition in a more cost-effective way.

Several studies have attempted to pinpoint which properties provide clues for robust automatic recognition of hesitations. Phonetic and prosodic properties and contextual distributions are shown to be significant in (Veiga et al., 2012a; Vasilescu et al., 2005; Candea et al., 2005; Shriberg, 1995; Clark & Fox Tree, 2002). Studies on several languages, such as English (Fox Tree & Clark, 1997; Bell et al., 2003), Swedish (Eklund, 2004), Mandarin (Lee et al., 2004) and French (Candea, 2000), have attempted to identify linguistic properties from filled pauses and extension events. Others point out lexical and syntactic principles, which may link repetitions with word cut-offs (Henry & Pallaud, 2003). For the detection of repetitions, features such as duration (Shriberg, 1995) and syntactic cues (Clark & Wasow, 1998) have been frequently used.

There are also various linguistic studies on hesitations for European Portuguese (EP). Works such as Viana (1987), Freitas (1990) and Delgado-Martins & Freitas (1991) are some of the first to classify filled pauses. In Mata (1999), fundamental frequency and duration of filled pauses are presented as characteristics that contribute for on-line planning efforts either in spontaneous speech or in oral reading. Our previous studies on hesitations and speaking styles have already used the same speech source database that culminated in HESITA (Veiga et al., 2011; Veiga et al., 2012a; Veiga et al., 2012b; Proença et al., 2013).

2. HESITA Database

The HESITA database consists of manually annotated hesitation events in 30 daily news programs collected from podcasts of a European Portuguese television channel, amounting to approximately 27 hours of speech. The video information was not included and the audio was downsampled from 44.1 kHz to 16 kHz. It contains studio, indoor and outdoor recordings including a few telephone sessions. The dominant speaking style is prepared (read) speech, as most utterances are of anchors and professional speakers (14 hours). However, commentators, reporters, interviewers and interviewees

provide frequent samples of spontaneous speech (10 hours). Lombard speech also appears, but with a low frequency (18 minutes, with only 12 events of hesitation). Under the term of hesitation, the following categories were identified and annotated, closely following the notation presented in Shriberg (1994), with identifying symbol in parentheses:

- filled pauses (f),
- vocalic extensions (+),
- repetitions (r),
- substitutions (s),
- filler words (p),
- deletions (d) and
- insertions (i).

Table 1 describes all the symbols used in the annotation of hesitations, showing the classes of the events (syntactic, extra-syntactic diacritics) as well as hesitations' pattern examples. RP and IP indicate Repair Point and Interruption Point respectively, along the lines of (Shriberg, 1994).

Syntactic word symbol	**Meaning**	**Example of hesitation pattern**
r	Repeated word	"que.que" → (r.r)
s	Substituted word	"esta.este"→(s.s)
i	Inserted word	"khad-.de khadafi"→ (s-.is)
d	Deleted word	"dado que.|podemos dizer" → (dd.)
Extra-syntactic word symbol	**Meaning**	**Example**
f	Filled pause	[6] → (f.)
p	Filled word	"que portanto. que" → (rp.r)
Diacritics	**Meaning**	**Example**
-	Cut word	"pod-.possam" → (s-.s)
^	Reduced word	"que 'tá.que era" → (rs^.rs)
~	Misarticulated word	"me(s)mo~ que.sempre que"→ (s~r.sr) "uma novia~.uma nova"→ (rs~.rs)
+/w+	Vocalic extension	"este[@]"→ (.w+) "este[@] é. este era" → (r+s.rs)
¨	Respiration inside the hesitation	"já não(res).já não":(rr¨.rr)
; .	IP and RP	

Table 1: Symbols used for hesitation annotation in HESITA, with accompanying examples.

The SAMPA phonetic alphabet (Wells, 1997) expanded for European Portuguese was employed to transcribe filled pause vocalizations. The HESITA database is also tagged for certain audio characteristics (background environments for speech, such as studio, street, speech overlapping, noise and music) and acoustic events (non-speech events, such as music, jingles, laughter, coughing or clapping). Respiration and events such as noise from cars or wind were also accounted for in the annotation procedure. Speaking style and speaker information are included in the annotation labels as well.

All annotations were carried out with the Transcriber software tool (Barras et al., 1998), of which Fig.1 shows an example of its use.

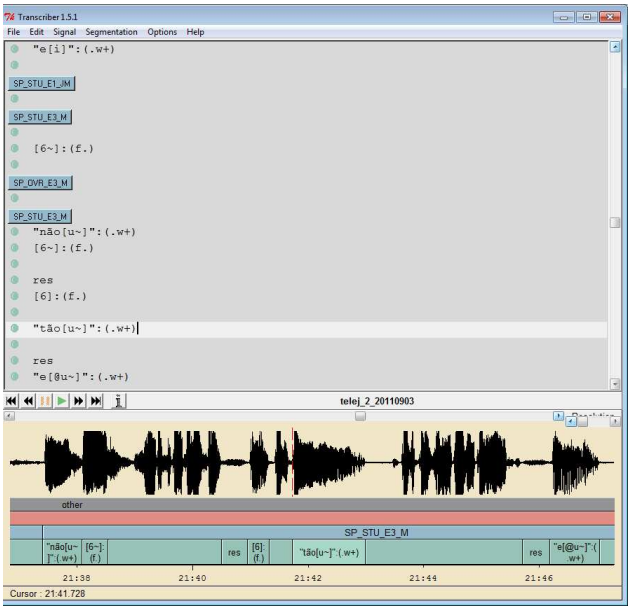

Figure 1: The Transcriber software tool with examples of annotated audio segments.

An example of annotation from Fig.1, 'SP_STU_E1_JM' relates to speech (SP) with noise-free environment (STU), in a spontaneous speaking style with low level of spontaneity (E1) and from a male journalist (JM). Another example, 'SP_STU_E3_M' shows an annotation of speech with noise-free environment (STU), in a spontaneous speaking style with high level of spontaneity (E3) and from a male speaker (M). Repetitions (r), extensions within a word (w+) and filled pauses (f) are some of the hesitation events which are annotated. Extended vowel sounds or vocalic fillers are accompanied by a transcription with phonetic symbols. Respiratory events are marked as 'res'.

The database contains WAV audio files and corresponding TRS text files (containing the manual transcriptions in the Transcriber format) for each individual news program (often separated in two parts). HESITA is available through Meta-Net[1] as well as in the project page[2].

[1] http://metanet4u.l2f.inesc-id.pt/repository/search/
[2] http://lsi.co.it.pt/spl/hesitation/downloads.html

2.1 Hesitations through speaking styles

The trend of hesitations occurring most frequently in spontaneous speech, according to overall figures for other languages (Shriberg, 1994; Candea, 2000; Eklund, 2004), is also observed in HESITA, in which the occurrences of such events in spontaneous speech amount to 4406 against 188 in read (prepared) speech and 12 in Lombard speech (2 additional events were marked in noisy segments and not further classified).

Spontaneous speech has a rate of 7.34 hesitations per minute while read (prepared) speech has a rate of 0.22 hesitations per minute. These levels of fluency were also verified for other languages (Bortfeld et al., 2002).
Gender does not appear to importantly influence the rate of hesitating in spontaneous speech, as female and male speakers generate similar rates, with 7.72 and 7.26 hesitations per minute, respectively.

In read (prepared) speech, the hesitations with the highest frequency are vocalic extensions (.w+) (39.36%) followed closely by filled pauses (f.) (32.45%). This shows a higher tendency for vocalic extensions contrarily to global figures, reflecting their contextual preference during prepared speech. The same conclusion applies to substitutions, which appear at a higher rate in prepared speech (9.57% vs. 3.61% in spontaneous speech). Repetitions are residual in prepared speech.

2.2 Duration statistics

Some relevant observation about the temporal characteristics (duration of segments) can be pointed out from the HESITA database and may be analyzed as manifestations of planning effort as well. The annotation of the hesitation events includes the initial and final temporal marks and the corresponding label contains the pattern and the orthographic transcription, closely following Shriberg (1994). The timing of the repair-point was also included, marking the instant when the hesitation is corrected and the fluency of speech is recovered. It is verified that the initial interval corresponding to the beginning of the hesitation until its repair-point is much larger (average of 0.61 seconds) than the period of time between the repair-point and the end of the correction (average of 0.34 seconds). This matches earlier studies, such as Moniz et al. (2012).

For the most common hesitations that are not corrected, filled pauses are shorter than vocalic extensions. The duration statistics are as follows (mean ± standard deviation): filled pauses (f.) last 0.412 ± 0.260 seconds and vocalic extensions (.w+) 0.698 ± 0.263 seconds. Of the most frequent filled pauses we have [6] with 0.315 ± 0.164 s, [@] 0.337 ± 0.204 s, [6~] 0.546 ± 0.226 s, [6m] 0.686 ± 0.276 s, [u~] 0.443 ± 0.280 s. Of the most frequent vocalic extensions, "que[@]" 0.553 ± 0.190 s, "e[i]" 0.538 ± 0.191 s, "de[@]" 0.520 ± 0.191 s, "com[o~]" 0.529 ± 0.130 s, "o[u]" 0.467 ± 0.165 s, "um[u~]" 0.530 ± 0.220 s.

3. Final Remarks

From browsing the literature (e.g. [2], [16], [18]), there is strong evidence that hesitations are used as a part of the speaker's speech structure, in order to achieve an improved synchronization with interlocutors. Scientific domains that try to identify significant information in human speech, such as the linguistic or clinical/therapeutic areas dealing with speech fluency, can benefit from an analysis of the distribution of hesitations along the speech, matching the complementary distribution of such events with speaking styles, speakers or acoustical environments.

4. Acknowledgements

This work was funded by FCT project (PTDC/CLE-.-LIN/112411/2009) and partially supported by FCT, Instituto de Telecomunicações multiannual funding (PEst-OE/EEI/LA0008/2011). Sara Candeias and Jorge Proença are supported by the SFRH/BPD/36584/2007 and SFRH/BD/97204/2013 FCT grants, respectively.

5. References

Adell, J., et al. (2008). On the Generation of Synthetic Disfluent Speech: Local Prosodic Modifications used by the Insertion of Editing Terms. In *Interspeech'08*, Brismane, Australia.

Barras, C., et al. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proc. 1st International Conf. on Language Resources and Evaluation (LREC)*, pp. 1373-1376. (http://trans.sourceforge.net/)

Bell, A., et al. (2003). Effects Of Disfluencies, Predictability, And Utterance Position On Word Form Variation In English Conversation. In *Journal of the Acoustical Society of America*, 113 (2), pp. 1001-1024.

Bortfeld, H., et al. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44, pp. 123-147.

Candea, M. (2000). Contribution à l'Étude des Pauses Silencieuses et des Phenomenes Dits «d'Hesitation» en Français Oral Spontané – Étude sur un Corpus de Récit en Classe de Français. Ph.D. dissertation, Université Paris III –Sorbonne Nouvelle.

Candea, M., et al. (2005). Inter- And Intra-Language Acoustic Analysis Of Autonomous Fillers. In *DISS'05*, Aixen-Provence, France.

Clark, H. (1996). *Using Language*. Cambridge University, Press, Cambridge.

Clark, H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, 37, 201-242.

Clark, H. and Fox Tree, J.E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84, pp. 73-111.

Delgado-Martins, M. R. and Freitas, M. J. (1991). Temporal structures of speech: reading news on TV. In *ETRW'91*, Barcelona.

Eklund, R. and Shriberg, E. (1998). Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human-human and human-machine dialogs. In *International Conf. on Spoken Language Processing*, Sydney, Australia, 6, pp. 2631-2634.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. *PhD dissertation*, Institute of Technology, Linköping University.

Fox Tree, J. E. and Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62, pp.151-167.

Freitas, M. J. R. (1990). Estratégias de Organização Temporal do Discurso. M.S. thesis, Faculdade de Letras, Universidade de Lisboa.

Henry, S. and Pallaud, B. (2003). Word fragment and repeats in spontaneous spoken French. In *Disfluency in spontaneous speech workshop, DiSS'03*, R. Eklund (Ed.), Göteborg University, 3-8 Sept., 2003, pp. 77-80.

Lee, T.-L., et al. (2004). Prolongation in spontaneous Mandarin. In *Interspeech'04*, Jeju Island, Korea, pp. 2181-2184.

Levelt, W.J.M. (1989). *Speaking. From Intention to articulation*. Cambridge, Massachusetts, The MIT Press.

Liu, Y., et al. (2006). Enriched speech recognition with automatic detection of sentence boundaries and disfluencies. In *IEEE Transaction on Audio, Speech, and Language Processing*, 14, pp. 1526-1540.

Mata, A. I. (1999). Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didáticas. Ph.D. dissertation, Faculdade de Letras, Universidade de Lisboa.

Moniz, H., et al. (2012). Analysis of disfluencies in a corpus of university lectures. In *Proc. of ExLing*, Athens, Greece.

Proença, J., et al. (2013). Acoustical Characterization of Vocalic Fillers in European Portuguese. In *DiSS 2013, ISCA endorsed Interspeech 2013 satellite workshop*, KTH Royal Institute of Technology, Stockholm, Sweden, pp. 63-65.

Shriberg, E. (1994). Preliminaries to a theory of speech disfluencies. Ph.D. dissertation, University of California.

Shriberg, E. (1995). Acoustic properties of disfluent repetitions. In *ICPhS*, Stockholm, Sweden, 4, pp.384-387.

Vasilescu, I., et al. (2005). Perceptual Salience Of Language-Specific Acoustic Differences In Autonomous Fillers Across Eight Languages. In *Interspeech'05*, Lisboa, pp. 1773-1776.

Veiga, A., et al. (2011). Characterization of hesitations using acoustic models. In *Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII*, Hong Kong, pp. 2054-2057.

Veiga, A. et al. (2012a). Prosodic and Phonetic Features for Speaking Styles Classification and Detection. In *Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science*, Toledano, D.T., Ortega, A., Teixeira, A., Gonzalez-Rodriguez, J., Hernandez-Gomez, L., San-Segundo, R., Ramos, D. (eds.), vol. 328, pp. 89-98, Springer.

Veiga, A. et al. (2012b). Towards Automatic Classification of Speech Styles. *In Lecture Notes in Artificial Intelligence* (LNAI), H. Caseli et al. (Eds.), Springer-Verlag Berlin Heidelberg, 7243, pp. 421-426.

Viana, M. C. (1987). Para a Síntese da Entoação do Português. Graduate research thesis, Universidade de Lisboa.

Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems*, Gibbon, D., Moore, R. and Winski, R. (eds.), Berlin and New York: Mouton de Gruyter, Part IV, section B. (http://www.phon.ucl.ac.uk/home/sampa/)