HESITA(tions) in Portuguese: a database

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

[1] Instituto de Telecomunicações, Coimbra, Portugal
[2] Electrical and Computer Engineering Departement, University of Coimbra, Portugal

Abstract

With this paper we present a European Portuguese database of hesitations in speech. Under the name of HESITA, this database contains annotations of hesitation events, such as filled pauses, vocalic extensions, truncated words, repetitions and substitutions. The hesitations were found over 30 daily news programs collected from podcasts of a Portuguese television channel. The database also includes speaking style classification as well as acoustical information and other speech events. Statistic analysis of the hesitation events in terms of their occurrence is presented. Insights into the process of human speech communication can be extracted from this database, which encloses relevant information about how Portuguese speakers hesitate. The HESITA database is freely available online to the research community.

Index Terms: hesitations, disfluency, prepared speech, spontaneous speech, annotation, hesitation corpus

1. Introduction

It is commonly agreed that hesitations (synonym here for disfluencies) characterize spontaneous speech and play a fundamental role in its structure, reflecting aspects of the language production and the management of inter-communication [1], [2] and [3]. Across several corpora, studies as [2], [5], [8] have shown that hesitation-like events occur frequently at high rates per word during the speech production. In the last decade, a growing number of works on language processing have focused on hesitation events underlining the importance of gathering knowledge on these type of events for successful speech technology development (see [2], [6], [11], [17] and [31], as examples). Regular features of those events have been accepted as an important parameter to take into account both in automatic speech recognition (for more robust language and acoustic models [10], [11], [12]) and in speech synthesis (to improve the naturalness of the speech [13]).

Although some theories and models have arisen in an attempt to explain the phenomenon and its benefits for communication purposes, hesitation phenomenon remains as a linguistic challenge. Hence, they appear to be regulated by language specific constraints and they perform a linguistic universal role in the speech structure, systematically and meaningfully [14–17].

Since hesitation events are crucial to facilitate natural language processing tasks, several studies have attempted to verify which properties may provide clues to their recognition. Phonetic and prosodic properties and contextual distributions are shown to give significant cues in [11], [15], [16], [23] and [18], respectively. Studies on different languages, such as English [19], [20], Swedish [5], Mandarin [21] and French [8], have attempted to distinguish linguistic properties between filled pauses and extension events, mainly in order to pursuit the linguistic reasons of why extensions cannot be eliminated at a pre-processing module. Others, e.g. [22], point out lexical and syntactic principles, which may link up repetitions with word cut-offs. To detect repetitions, acoustic features including duration [23] and some syntactic cues [24] have been frequently used.

For European Portuguese there are also various linguistic studies on hesitations that have attempted to provide significant knowledge on the topic and claiming the regular trend of it. Regarding filled pauses, works such as [25–27] can be mentioned as first works on the subject. In [7] and in [10], 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. Other works on the topic for European Portuguese can be found in [9] and in [6], [11]. Although the classification of filled pauses is not the main topic of these last two works, it shows that such hesitation events are responsible for the distinction between unplanned versus planned speech.

With this paper we intend to present a European Portuguese database of hesitations in speech. Under the name of HESITA, this database contains annotations of hesitation events, such as filled pauses, vocalic extensions, truncated words, repetitions and substitutions. Additionally, other acoustical characteristics such as environment condition, speaking styles and speaker were annotated as well. We believe that these multiple annotation layers provide a wide range of opportunities for studying the structure of the human speech communication process, under the domain of either speech technology development or linguistic descriptive works.

In section 2 we concisely describe the components of the HESITA database. Section 3 provides statistics about the distribution of the hesitation events, illustrating their phonetic forms and relation with speaking styles. Section 4 presents a brief discussion, mainly focusing on the HESITA application

2. The HESITA database

The HESITA database comprises manually annotated hesitation events in 30 daily news programs collected from podcasts of a European Portuguese television channel (about 27 hours of speech). The audio was downsampled from 44.1 kHz to 16 kHz sampling rate and the video information was discarded. It contains studio and out of studio recordings as well as some telephone sessions. Prepared (read) speaking style is dominant, since most of the speech encompasses utterances of anchors and professional speakers (14 hours). However we can frequently find spontaneous

speech segments in commentators, reporters, interviewers and interviewees (10 hours). Lombard speech appears as well, but with low representativeness (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 [2]:

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

Only the speech segments were annotated in terms of hesitation events. Filled pause vocalizations were transcribed using the SAMPA phonetic alphabet for European Portuguese [28]. HESITA database (hereafter HESITA DB) also encompasses information regarding audio characteristics (background environments, such as studio, street, speech overlapping, noise and music) and acoustic events (non-speech events, such as music, jingles, laughter, coughing or clapping). Respiratory and other events, such as noise from cars or wind, were also taken in consideration in the annotation procedure. Speaking style and speaker information are included in the annotation labels as well.

All the annotations were performed by using the Transcriber software tool [29]. See an example in Fig.1, in which (SP_STU_E1_JM) exemplifies an annotation of speech with noise-free environment (SP), in a spontaneous speaking style (STU) with low level of spontaneity (E1) and from a male journalist (JM). (SP_STU_E3_M) exemplifies 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). (SP_OVR_E3_M) represents the annotation of an audio segment speech with noise-free environment (SP) but with overlapping (OVR), in a spontaneous speaking style with high level of spontaneity (E3) and from a male speaker (M). We can also verify the annotation of some hesitation events, including repetitions (r), extensions within a word (w+) and filled pauses (f). Phonetic symbols attest extended vowel sounds or vocalic fillers. The presence of a respiratory event is annotated as (res).

In the database, each news program is associated with a WAV audio file and a TRS text file (containing the manual transcriptions in the Transcriber format).

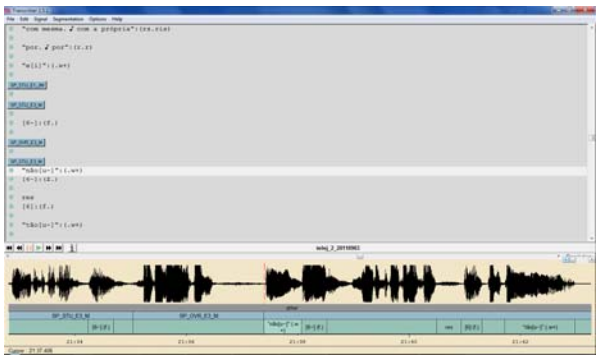

Figure 1: *Examples of audio segments annotation, using the Transcriber software tool.*

3. Hesitation patterns

Considering all the segments that were annotated accordingly to the presence of hesitations, we can see in Table 1 how the hesitation patterns are distributed. A total of 4608 events were observed in which filled pauses (f.) and vocalic extensions within a word (.w+) are the most common, achieving 36.5% and 23.4% of the hesitation events, respectively. The most common hesitation events are somewhat similar to what has been observed for English or Swedish, as reported in [2] and [5], respectively. A similar distribution was previously stated for European Portuguese in [4] and [10].

In Tables 1 and 2, figures are given for the relative occurrence of the most frequent hesitation events observed in the HESITA DB. The left column in Table 1 and Table 2 gives the information about each hesitation pattern. Pattern models display the way that the hesitation occurs, indicating the order of the words before and after the so-called "repair-point". This point marks the place from which the hesitation is corrected and the fluency is restored. For instance, the pattern (r.r) indicates that a word r was repeated as repair or reinforcement; in the pattern (s-.s), the word s was cut and then substituted; in (r2.r) the same word r was repeated twice and finally restored; in (rs-.rs) the word r was repeated and word s was cut and, then substituted with correction.

More complex hesitation patterns are present in the HESITA DB, although not very frequently. For instance, the hesitation with the transcription "*que vo-.que.que.que voltam.que.que possam*" has the following pattern which shows embedded hesitations: ((rs-.(r2.r)s).(r.r)s).

3.1. Hesitations across speaking styles

According to overall figures that have been given for other languages, in general hesitation events occur mainly in spontaneous speech [2], [5], [8], [9]. The same trend is observed in the present database, in which the occurrences of such events in spontaneous speech count 4406 against 188 in read (prepared) speech and 12 in Lombard speech (2 additional events were in noisy segments and thus not further classified).

The total of 188 hesitations observed in 14 hours for read (prepared) speaking style results in a rate of 0.22 hesitations per minute. The 4406 hesitation events in 10 hours of spontaneous speech result in a rate of 7.34 hesitations per minute, which reveals a tendency in fluency also verified for other languages (see [30], for instance).

Considering the gender in our database, this variable appears to not affect fluency rates in spontaneous speech: female and male speakers produce similar rates of hesitations, with 7.72 and 7.26 hesitations per minute, respectively.

It has also been noted by [14] for other languages, that the density of hesitations in speech varies with the speaking style, and in prepared speech corpora vocalic fillers are relatively infrequent, not exceeding 0.7% of the speech data. We corroborate this finding; in our case, in read (prepared) speech, fillers accounts for less than 0.2% of the speech duration.

Table 1.*Top 10 most frequent hesitation patterns.*

Patterns	# Events	% Events
(f.)	1681	36.5
(.w+)	1078	23.4
(r.r)	376	8.16
(p.)	213	4.62
(r+.r)	165	3.58
(s-.s)	95	2.06
(s.s)	84	1.82
(rr.rr)	72	1.56
(r2.r)	45	0.98
(rs-.rs)	43	0.93
others	756	16.4

Table 2 shows the distribution of the 5 most common hesitation patterns in the read (prepared) speech, with the high frequency of vocalic expressions (.w+) (39.36%) just followed by filled pauses (f.) (32.45%). Although the difference between the occurrences of vocalic extensions and filled pauses is not so expressive, it is possible that the choice for the extensions reflects the fact that vocalic fillers tend to be more stigmatized in a prepared speech context. Repetitions in read or prepared speech become residual. The relative frequency rates for substitutions are higher in the prepared speech than in spontaneous speech (9.57% vs. 3.61%), proving that they are more adequate for communicative strategy mainly in what the fluency of speaking is concerned.

Table 2. Top 5 most frequent *hesitation patterns for read (prepared) speech.*

Patterns	# Events	% Events
(.w+)	74	39,36
(f.)	61	32,45
(s-.s)	10	5,32
(s.s)	8	4,26
(rs-.rs)	4	2,13

3.2. Phonetic form of filled pauses

During the annotation procedure, we found the two most common phonetic forms for filled pauses: the near-open central vowel [ɐ] ([6] in SAMPA) and the mid-central vowel [əə] ([@] in SAMPA), representing 48% and 20% of the all filled pauses, respectively. Table 3 shows the ranking of the ten most used in HESITA DB.

Table 3. *Phone distribution of filled pauses (top10 most frequent).*

Phone	N. events	% (Perct.)
[ɐ]	808	48,07
[ə]	344	20,46
[ɐ̃]	155	9,22
[ɐm]	73	4,34
[ũ]	46	2,74
[ɐũ]	31	1,84
[ɐə]	28	1,67
[ɐɐ]	21	1,25
[əɐ]	13	0,77
[ɐ̃m]	9	0,54

The distributon shown in Table 3 also supports the view that the vocalizations preferred by Portuguese speakers are around central vowels, corresponding to the reduced vowels in an unstressed position (/a/ vs. /i/, /e/, /ɛ/, respectively). There is also a slight inclination for the high back rounded nasal vowel [ũ] as well (around 3%). Also a nasal preference is evident in the DB: see [ɐ̃], [ɐm] and [ɐ̃m] or [ũ] in Table 3. Our point here is not to associate a meaning to the filler sounds. However, there is strong empirical evidence that speakers use all of them for playing a structuring role in the speech. The choice for a vocalic sound rather than other appears to be, at least in some contexts, motivated by the behavior of neighbor phonetic segments, neutralizing in some way the phonetic difference of the vocalic fillers.

3.3. Segmentation of hesitations

The annotation of the hesitation events closely follows [2]. It encompasses the initial and final temporal marks and the corresponding label contains the pattern and the orthographic transcription. The repair-point was also marked temporally, showing the instant were the hesitation is corrected and when the fluency on speech is recovered. It has been verified that the period of time that corresponds to the beginning of the hesitation to its repair-point is much larger (0.61 seconds in average) than the period of time between the repair point and the end of the hesitation correction (0.34 seconds in average). This matches what was found in previous studies, such as in [31]. These trends concerning the distribution and duration of hesitation events may be analyzed as manifestations of planning effort as well.

4. Discussion and conclusion

Browsing the literature in the area (e.g. [2], [5], [8]), it has been strongly evidenced that speakers use hesitations as part of their speech structure and in order to achieve a better synchronization with the interlocutors. Various scientific domains more directly interested in gathering knowledge for better identifying salient information in human speech communication, such as the linguistic or clinical/therapeutic areas covering speech fluency, can beneficiate of the analysis of the hesitation distribution along the speech, matching the complementary distribution of such events with the speaking styles, speakers or with the acoustical environment as example. Also, some relevant observation about the temporal characteristics (duration of segments) can be pointed out from the HESITA DB. Different vocalic choices for filled pauses can also be associated, at least in some contexts, with different functions and meanings of expressiveness, such as doubt, denial or agreement.

Studies on hesitations have recently gained in importance to increase the usability of speech systems, by overpassing the challenges proposed by the presence of such phenomena in continuous speech. We also believe that automatic language processing could benefit from a richer representation of the audio signal that incorporates speaking styles information, in order to breakdown errors in the automatic speech recognition or to improve automatic conversational speech summarization, for instance. Detection of hesitation events also provides the segmentation of multimedia data into consistent parts, as

claimed in [11]. It leads also to important applications such as the identification of speech segments to train acoustic models for speech recognition in a more cost-effective way.

Our goal in this paper was to present a database for European Portuguese, which contains a large and rich variety of speech data events and is mainly focused on the hesitations. We thus expect that this database can be a relevant base of work for further studies regarding a variety of speech phenomena. The HESITA database is available through the Meta-Net [32] as well as in the project page [33].

5. Acknowledgements

This work is funded by FCT and QREN projects (PTDC/CLE- LIN/11 2411/2009; TICE.Healty13842) and partially supported by FCT (Instituto de Telecomunicações multiannual funding PEst-OE/EEI/LA0008/2011).

Sara Candeias is supported by the FCT grant SFRH/BPD/36584/2007.

6. References

[1] W.J.M. Levelt, *Speaking. From Intention to articulation*, Cambridge, Massachusetts: The MIT Press, 1989.

[2] E. Shriberg, "Preliminaries to a theory of speech disfluencies", Ph.D. dissertation, University of California, 1994.

[3] H. Clark, *Using Language*, Cambridge, MA: Cambridge University Press, 1996.

[4] H. Moniz, et al., "On Filled Pauses and Prolongations in European Portuguese", in *Interspeech '07*, ISCA, Antwerp, Belgium, pp. 2645–2648, 2007.

[5] R. Eklund, "Disfluency in Swedish human–human and human–machine travel booking dialogues", PhD dissertation, Institute of Technology, Linköping University, 2004.

[6] A. Veiga, et al., "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.), 2012. vol. 328, pp. 89–98, Springer.

[7] A. I. Mata, "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, 1999.

[8] M. Candea, "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, Universite Paris III – Sorbonne Nouvelle, 2000.

[9] H. Moniz, "Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 2006.

[10] A. Veiga, et al., "Characterization of hesitations using acoustic models", in *Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII*, Hong Kong, pp. 2054–2057, 2011.

[11] A. Veiga, et al., "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, 2012.

[12] Y. Liu, et al., "Enriched speech recognition with automatic detection of sentence boundaries and disfluencies", in *IEEE Transaction on Audio, Speech, and Language Processing* 14, pp. 1526–1540, 2006.

[13] J. Adell, et al., "On the Generation of Synthetic Disfluent Speech: Local Prosodic Modifications used by the Insertion of Editing Terms", in *Interspeech '08*, Brisbane, Australia, pp. 2278–2281, 2008.

[14] I. Vasilescu, et al., "Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech". in *Proc. Interspeech '06*, Pittsburgh, PA, USA, pp. 1850–1853, 2006.

[15] I. Vasilescu, et al., "Perceptual Salience Of Language-Specific Acoustic Differences In Autonomous Fillers Across Eight Languages," in *Interspeech'05*, Lisboa, pp. 1773–1776, 2005.

[16] M. Candea, et al., "Inter- and Intra-Language Acoustic Analysis Of Autonomous Fillers." in Proc. DISS'05, Aix-en-Provence, France, pp. 47–51, 2005.

[17] R. Eklund and E. Shriberg, "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, 1998.

[18] H.H. Clark and J.E. Fox Tree, "Using uh and um in spontaneous speaking", *Cognition* 84, pp. 73–111, 2002.

[19] J. E. Fox Tree and H. H. Clark, "Pronouncing "the" as "thee" to signal problems in speaking", *Cognition* 62, pp. 151–167, 1997.

[20] A. Bell, et al., "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, 2003.

[21] T.-L. Lee, et al., "Prolongation in spontaneous Mandarin", in *Interspeech'04*, Jeju Island, Korea, pp. 2181–2184, 2004.

[22] S. Henry and B. Pallaud, "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, 2003.

[23] E. Shriberg, "Acoustic properties of disfluent repetitions", in *Proc. ICPhS*, Stockholm, Sweden 4, pp. 384–387, 1995.

[24] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech", *Cognitive Psychology* 37, pp. 201–242, 1998.

[25] M. J. R. Freitas, "Estratégias de Organização Temporal do Discurso", M.S. thesis, Faculdade de Letras, Universidade de Lisboa, 1990.

[26] M. R. Delgado-Martins and M. J. Freitas, "Temporal structures of speech: reading news on TV", in *ETRW '91*, Barcelona, 19-1–19-5, 1991.

[27] M. C. Viana, "Para a Síntese da Entoação do Português", Graduate research thesis, Universidade de Lisboa, 1987.

[28] J.C. Wells, "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, 1997. (http://www.phon.ucl.ac.uk/home/sampa/)

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

[30] H. Bortfeld Leon, et al., "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender", Language and Speech, 2001, 44, pp. 123–147.

[31] H. Moniz, et al., "Analysis of disfluencies in a corpus of university lectures", in *Proc. of ExLing, Athens*, Greece, 2012

[32] http://metanet4u.l2f.inesc-id.pt/repository/search/

[33] http://lsi.co.it.pt/spl/hesitation/downloads.html