Acoustical characterization of vocalic fillers in European Portuguese

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

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

Abstract

This study attempts to acoustically characterize the most common filled pause vocalizations (or vocalic fillers) in spontaneous speech in European Portuguese: the near-open central vowel [ɐ] and the mid-central vowel [ə]. For this purpose we analyzed the spectral information of the vocalic fillers by estimating their first two formant frequencies as well as their duration properties. The vocalic fillers are taken from a large corpus of European Portuguese broadcast news' speech. We also compared the vocalic fillers with lexical vowels possessing similar timbre. No formant variation trend was attained for the vocalic fillers and a great overlap of formant values is observed. These results provide a base of information for understanding the most common vocalic fillers in European Portuguese spontaneous speech.

Index Terms: filled pauses, vocalic fillers, formant estimation, spontaneous speech, hesitations.

1. Introduction

The interest in studying events that characterize the spontaneity of the speech has been increasing as the development of speech technologies grows. In this context, several studies on hesitations (so-called disfluencies) as well as vowel reductions have gained importance over the last years ([2,7,5,8], and [17,18] as examples, respectively). Among various hesitation phenomena, such as repetitions, truncated words or word extensions, filled pauses are, the ones more widely encountered in world's languages, mainly on spontaneous speech, [10], [19]. Relatively stable vocalic segments mostly fulfill these pauses. Occurring commonly without any lexical support, we refer to this type of fillers as vocalic fillers (VFs). Representing an insertion at any moment during spontaneous speech, VFs carry multiple functions, such as announcing upcoming discursive topics or planning and delaying speech, [2–4,9].

Although some works on VFs can already be found for Portuguese [6,10,11,13] and other languages [5,12,19], the most part of studies conducted on large spontaneous speech corpora comprises English or French languages. This paper presents a study which attempts to acoustically characterize the two most common vocalic fillers that occur in European Portuguese: the vocalic filler representing a similar timbre to the near-open central vowel [ɐ] and the one close to the timbre of the mid-central vowel [ə].

We chose the two first formant frequencies, F1 and F2, and filler duration as phonetic and prosodic parameters in search of reliable patterns of this type of fillers. Although all the studies mostly agree on characterizing these fillers as long and stable vocalic segments [10,19], their characterization for European Portuguese still needs further study. Additionally, we compare the vocalic fillers with the corresponding lexical vowels (LVs) possessing similar timbre, which are vowels produced in a context of complete words, in a similar way that was done for different languages, such as French, American English and European Spanish [16], [20,21]. Even though it has been noted that vocalic fillers manifest acoustical language-dependent characteristics, in fact they may not be necessarily acoustically equal to the lexical vowels with similar timbre, and, at least in some contexts, they appear to possess slightly different average positions in the triangle vowel area.

On the characterization of vocalic fillers, this study is an extension of our previous study [10] for European Portuguese where fundamental frequency and energy of VFs are presented. A better understanding of the structure of speech as well as insights on how to obtain filler acoustic models for use in automatic spontaneous speech recognition are ultimate goals of this work. We also believe that this work also promotes awareness of vocalic fillers in the phonetic studies of the language.

The rest of the paper is organized as follows. In the next section we briefly describe database selection. In section 3 we show how the estimation of formant frequencies of the sounds [ɐ] and [ə] belonging either to fillers or to lexicon was performed. In section 4 we present the main discussion of the achieved results. Finally, in section 5 the main conclusions are drawn and envisioned work is foreseen.

2. Database

For the present study we used two corpora: a corpus of hesitation events, the HESITA Database [22], and a European Portuguese corpus with no hesitations collected for control.

The corpus of hesitations was used to study filled pauses, which were manually annotated. This corpus comprises 30 daily news programs collected from a European Portuguese television channel podcast (about 27 hours of speech). It consists of 1152 vocalic fillers, exemplifying sounds similar to the near-open central vowel [ɐ] (808) and to the mid-central vowel [ə] (344). These two VFs were the most common in the database and the ones chosen for analysis. The next most common was the nasal [ɐ̃], with 155 occurrences.

The control corpus was used to estimate the acoustic characteristics of the vocalic sounds [ɐ] and [ə] occurring in a context of a complete word (the LVs): e.g. [ɐ] in <para> [pɐɾɐ], ('for' in English) or [ə] in <devolver> [dəvolver] ('to give back'). It consists of recordings from 7 European Portuguese native adult speakers of sentences and command words, taken in a small office room. In each recording session, the same common set-up was used, which consists of a laptop computer and three microphones. For each session, a segmentation and phone-level transcription were automatically performed through forced alignment using in-house tools. The total number of extracted vocalic segments was 7426, in which we count 4411 [ɐ] and 3015 [ə].

Table 1 summarizes the overall [ɐ] and [ə] distribution by database and gender.

Table 1: *Relative frequency (#) of VFs and LVs by gender and timbre.*

Type	Gender	#[ɐ]	#[ə]
VFs	Male	605	301
	Female	203	43
LVs	Male	2674	1771
	Female	1737	1244

3. Formant frequencies determination

The first (F1) and second (F2) formant frequencies of the [ɐ] and [ə] vocalic fillers were automatically extracted using the Praat tool [14]. The base recommended ceilings for estimating five formants are 5500 Hz for female speakers and 5000Hz for male speakers but, through observation, these values can't always successfully estimate F1 and F2. A similar method to [15] was then applied, given the foreknowledge that different vowels and speakers need different formant ceilings for the automatic calculation. An iterative calculation of formants was performed in 10 ms steps using ceilings in the 4000–5500Hz range (for males) or 4800–6500 Hz (for females) in 50 Hz steps, followed by a selection of the optimal ceiling. As we do not keep speaker information for most of the news broadcast VFs, they were considered as if each belonged to a different speaker. Therefore, the ceiling that was selected as optimal for a given VF was the one that provided the smallest variance of the F1 and F2 pairs of values of that VF, calculated as the sum of the variances of 20log(F1) and 20log(F2). We empirically observed that wrong ceilings would usually fail on F1 or F2 for a couple of points, leading to sudden jumps and a larger variance than intrinsic F1 and F2 variations. One problem of this method is that a very high ceiling would provide smooth F2 values where the third formant frequency (F3) should be, and pass as optimal; but this was already taken into consideration and countered with the selected ranges of ceilings.

Other restrictions were applied before and after formant calculation. Utterances with high clipping of the audio signal were discarded. For each utterance, only the formant values where the energy level was above 10% of the maximum energy were considered, to specifically remove unvoiced transcription limits. Finally, utterances with highly variant formant values, probably indicating a failure in detecting F1 and F2 were not considered.

The same analysis was conducted for the lexical vowels [ɐ] and [ə] with an additional restriction of only considering segments of duration larger than 50 ms.

4. Results and discussion

After applying the restrictions mentioned in section 3, the number of vocalic segments kept was 520 [ɐ] and 244 [ə] VFs and 1517 [ɐ] and 385 [ə] LVs. A very large number of lexical vowels were cut from analysis; mostly the small-duration or low-energy segments, barely recognized during alignment and more drastically occurring for [ə], as a consequence of the nature of continuous speech that almost eliminates the already reduced vowel [ə] in certain cases. The durations of VFs and LVs are shown in Figure 1 and, as expected, VFs are generally longer.

Figures 2 and 3 place the extracted F1 and F2 mean values on a logarithmical scale for male and female speakers respectively, as usually done in similar cases [15]. The 'triangle' of [i], [ɛ], [a], [ɔ] and [u] vowels come from the median values of F1 and F2 taken from the read speech corpus. These values were calculated in a similar fashion to the method described, although with much more restrictive ceilings. They were included to show the centrality of [ɐ] and [ə]. Averagely, F1 is higher and F2 is lower for VFs than for LVs, but their distributions overlap. LVs show the biggest variances, which could be explained by the high dependence of phonetic context (and related coarticulation phenomenon) in both word and sentence production.

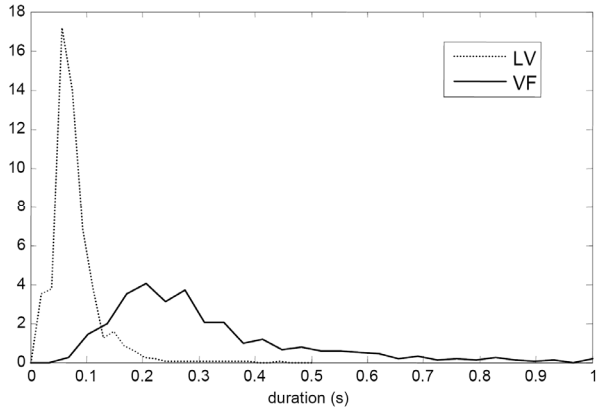

Figure 1: *Normalized histogram of the duration of VFs and LVs.*

The variation trend of F1 and F2 during each vocalic segment was also analyzed. Although their change can be non-linear, a linear fit was applied to each sequence of values and the variation rate was extracted from this fit. Fig. 4 shows these rates for F1 and F2, and it is observable that behaviors are highly variant, either positive or negative. Although the average is for a small negative change, no trend can be discerned. Furthermore, no correlation exists between F1 and F2 simultaneous variation. LVs also prove to be less stable than the noted VFs, as they achieve much higher variation rates.

The presented results also point out that [ɐ] and [ə] are often hard to distinguish. Each speaker could have its own personal preference on how to fill a pause vocalically. Choosing mainly sounds of the central vowels system, speakers appear to adapt the production with their own specific production, possibly even in a middle point of [ɐ] and [ə]. It would be interesting to perform an in-depth perceptual study to confirm that some vocalic fillers can be understood differently with and without context or for different listeners, which could coincide with most of the cases that overlap so far.

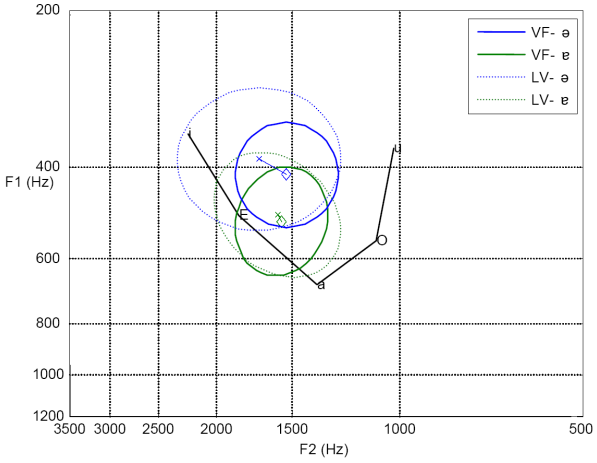

Figure 2: *Male speakers: F1 and F2 means and concentration ellipsoids for [ə] (blue) and [ɐ] (green) for VF and LV, including the male vowel 'triangle' from the LVs database.*

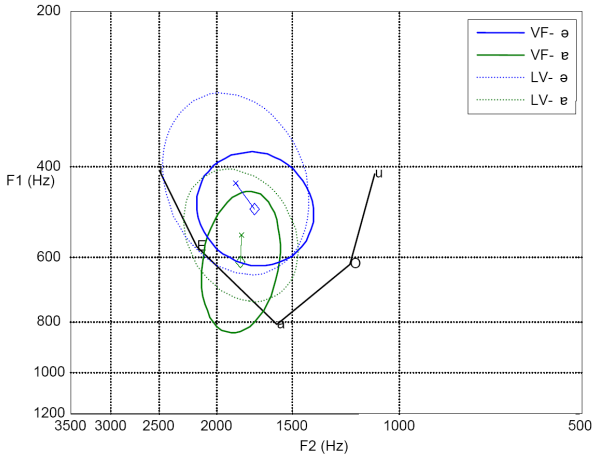

Figure 3: *Female speakers: F1 and F2 means and concentration ellipsoids for [ɐ] (blue) and [ə] (green) for VF and LV, including the female vowel 'triangle' from the LVs database.*

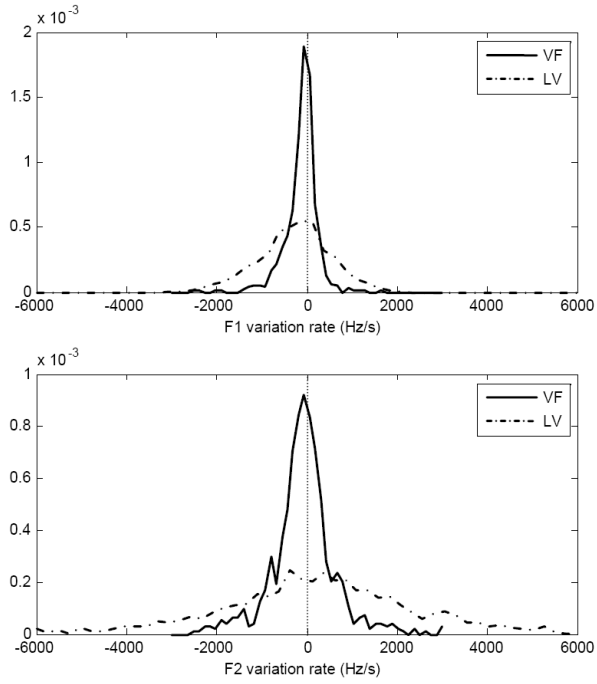

Figure 4: *Normalized histogram of the variation rates of F1 (top) and F2 (bottom) from a linear fit to each utterance, for VFs and LVs.*

5. Conclusions

The most common filled pause vocalizations (or vocalic fillers) in European Portuguese, [ɐ] and [ə], were characterized concurrently with their corresponding intra-word vowel productions (lexical vowels). Vocalic fillers were taken from a large corpus of European Portuguese broadcast news' speech (about 27 hours).

As expected, vocalic fillers are of longer duration and the lexical vowels are very short. Although the average of formant variation is for a small negative change, no specific trend was observed. Still, the variations of F1 and F2 indicate a higher stability of the fillers in comparison to the vowels. As to formant values, there is a small tendency for higher F1 and lower F2 in VFs. Fillers [ɐ] and [ə] are often hard to distinguish and each speaker could have its own specific production and may be strongly dependent on linguistic context. A perceptual study to confirm that some vocalic fillers can be understood differently with and without context or for different listeners would be interesting as future work.

We plan on extending this study to vocalic extensions, which are another class of filled pauses, as they may have similar behaviors in duration and formant frequencies as vocalic fillers. Based on the knowledge attained from this study, we also intend to develop an automatic detector of fillers and extensions from continuous speech.

6. 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.

7. References

[1] W. J. M. Levelt, *Speaking. From Intention to articulation*, Cambridge, Massachusetts, The MIT Press 1993, 1989.

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

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

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

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

[6] 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.

[7] 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.

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

[9] E. Shriberg, "Spontaneous speech: how people really talk, and why engineers should care", in *Proc. Eurospeech*, Lisboa, 2005.

[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, 2012, 7243, pp. 421–426.

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

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

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", [Computer program, Version 5.3.42]. Available: http://www.praat.org/, retrieved on 2 March 2013.

[15] P. Escudero, et al., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese", *J. Acoustical Society of America*, 126(3), pp. 1379–1393, 2009.

[16] 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.

[17] S. Candeias, and F. Perdigão, “A realização do schwa no português europeu”, in II workshop on Portuguese description-JDP, 8th Symposium in Information and Human Language Technology (STIL), 2011.

[18] D. Braga, et al., “Back close non-syllabic vowel [u] behavior in European Portuguese: reduction or suppression,” in *Proc. ICSLP*, Seoul, 2001.

[19] R. Eklund, and E. Shriberg, E., “Crosslinguistic disfluency modeling: a comparative analysis of Swedish and American English human–human and human–machine dialogues”, in *Proc. Int. Conf. on Spoken Language Processing*, vol. 6, pp. 2631–2634, Sydney: Australian Speech Science and Technology Association, 1998.

[20] I. Vasilescu, and M. Adda-Decker, “On the Acoustic and Prosodic Characteristics of Vocalic Hesitations across Languages”. Available: http://perso.limsi.fr/Individu/madda/publications/PDF/IOS.pdf, retrieved on 2 March 2013.

[21] M. Candea, et al., “Inter- and intra-language acoustic analysis of autonomous fillers”, in *Proc. DiSS 2005*, Aix-en-Provence, France, pp. 47–51, 2005.

[22] Candeias, S., Celorico, D., Proença, J., Veiga, A. and Perdigão, F., “HESITA(tions) in Portuguese: a database”, *Proc. DiSS 2013*, ISCA endorsed Interspeech 2013 satellite workshop, August 21–23, 2013, KTH Royal Institute of Technology, Stockholm, Sweden.