A cross-language study of acoustic and prosodic characteristics of vocalic hesitations

Ioana VASILESCU [a] and Martine ADDA-DECKER [a]

[a] *Spoken Language Processing Group, LIMSI-CNRS, BP 133*
91403 Orsay Cedex, FRANCE

Abstract. This contribution provides a cross-language study on the acoustic and prosodic characteristics of vocalic hesitations.One aim of the presented work is to use large corpora to investigate whether some language universals can be found. A complementary point of view is to determine if vocalic hesitations can be considered as bearing language-specific information. An additional point of interest concerns the link between vocalic hesitations and the vowels in the phonemic inventory of each language. Finally, the gained insights are of interest to research in acoustic modeling in for automatic speech, speaker and language recognition.

Hesitations have been automatically extracted from large corpora of journalistic broadcast speech and parliamentary debates in three languages (French, American English and European Spanish). Duration, fundamental frequency and formant values were measured and compared. Results confirm that vocalic hesitations share (potentially universal) properties across languages, characterized by longer durations and lower fundamental frequency than are observed for intra-lexical vowels in the three languages investigated here. The results on vocalic timbre show that while the measures on hesitations are close to existing vowels of the language, they do not necessarily coincide with them. The measured average timbre of vocalic hesitations in French is slightly more open than its closest neighbor (/œ/). For American English, the average F1 and F2 formant values position the vocalic hesitation as a mid-open vowel somewhere between /ʌ/ and /æ/. The Spanish vocalic hesitation almost completely overlaps with the mid-closed front vowel /e/.

Keywords. Vocalic hesitation, Filler words, Automatic language identification, Formant extraction, Speech alignment.

Introduction

Substantial progress in speech processing research over the past decade has led to a variety of successful demonstrators involving spontaneous, or at least unprepared speech. To increase the usability of speech systems the challenges of multilinguality and spontaneous speech must be addressed efficiently. In this context there has been growing intrest in studies on hesitation phenomena. Hesitation phenomena have been initially viewed as "speech disfluencies", globally noisy and irregular events, in opposition with "fluent" or "well formed" speech [13]. However, recent corpus-based studies have shown that hesitation phenomena exhibit regularities, carry multiple functions in speech and contribute to the elaboration of the verbal message [6,14].

Hesitations can result in different vocalic realizations, varying from simple events such as silent and filled pauses, and word lengthening, to more complex phenomena involving repetitions and speech repairs. In this study we focus on the very common phenomenon of filled pauses, which correspond to autonomous vocalic hesitations. Such vocalic hesitations occur without lexical support and thus are to be distinguished from vocal lengthening of segments belonging to lexical items (generally function words). According to Clark and Fox Tree [6] the main role of vocalic hesitations is *to announce the initiation of a delay in speaking*. As for other types of hesitations such as silent pauses, repairs, restarts or repetitions, the vocalic hesitation may mark a speaker's effort to build a verbal message and can carry different additional functions in speech interactions: keeping or passing the floor, indicating the affective state of the speaker, etc. Vocalic hesitations are widely encountered in the world's languages and consist of the insertion of a relatively long, stable vocalic segment, considered as a type of 'filler'. The orthographic transcription of vocalic hesitations varies across languages suggesting differences in their perception by native speakers, for example *uh/um* in American English, *er* in British English, *euh* in French, *eh* in Spanish. Fillers can have other possible realizations, as for instance lengthened nasal consonants (*mm* in Mandarin Chinese [18]) or demonstratives (*ano*, *eto* in Japanese [17]). The density in speech of vocalic hesitations varies with the speaking style: in semi-prepared speech corpora (broadcast news-type) they are relatively infrequent, rarely exceeding 1% of the speech data whereas in corpora collected in more variable conditions (i.e. speakers under stress) they can easily reach more than 5% of the data [16].

If the role of vocalic hesitations in speech is now commonly accepted, comparative studies on their language-specific characteristics are still lacking. While most of the recent studies conducted on large speech corpora have focused on English or French [2,7,8,14,15], descriptions can now be found for other languages [17,18]. The acoustic and prosodic parameters generally studied are vocalic quality, pitch and duration. Pitch and duration appear to exhibit reliable patterns across the analyzed languages, i.e. flat and stable pitch and long duration compared to intra-lexical vowels [4,7,14]. Observations on the vocalic quality are less consistent: according to [14] the timbre of vocalic hesitations varies predictably with the vowel inventory of the language. For instance, in American English the hesitation vowel has a neutral (close to schwa) timbre and appears to be almost homophonic with the determiner *a*. The neutral quality of the hesitation vowel has been correlated with a possible "rest" position and articulatory settings of a language. The universal or language-specific properties of the articulatory settings are still unclear, even though recent experimental manipulations with X-ray data seem to support the second option [11]. If the language-specific hypothesis of the articulatory settings is confirmed, the vocalic hesitations might have articulatory targets of their own which differ across languages.

In some of our earlier studies comparing hesitations in 8 languages (American English, Middle Oriental Arabic, Mandarin Chinese, French, German, Italian, South-American Spanish and European Portuguese [5]) we have shown that vocalic hesitations can be observed in all these languages and that they can be characterized by duration, pitch and timbre. The present study aims to contribute both to speech processing research as well as to linguistics. From a linguistics perspective, this work addresses the question of whether some language universals can be found for hesitation vowels across languages. Seen from a complementary perspective, this study also investigates whether hesitation vowels bear language-specific information. An additional point of interest is then the link between hesitation vowels and intra-lexical vowels for each language. Does the timbre of the vocalic hesitations correspond to a vowel of the phonemic inventory of the languages or does the vocalic hesitation tend to a given, potentially universal schwa-like "rest" position? The gained insights may also be of interest to research in speech technology fields such as acoustic modeling for automatic speech, speaker, language and accent recognition.

1. Corpus and methodology

Some of the most readily available sources of transcribed speech corpora in different languages are journalistic shows and parliamentary debates. These data are available at LIMSI in the framework of different projects. For French, 20 hours of speech from several national radio and TV broadcast channels (*France Inter, France Info, France2*...) have been selected. The American English data is comprised of about 10h of recordings from broadcast channels (CNN, VOA, ABC, etc.) distributed by LDC. For Spanish, 10h of European Parliamentary debates are used. For all three languages, several dozens of different speakers contribute the majority of the audio data. Male speakers are roughly twice as frequent as female speakers in these types of corpora [11]. Although the Spanish corpus is of a different type than the French and English ones, the global acoustic properties investigated throughout this paper should not be significantly affected by this change.

In journalistic shows and parliamentary debates, the articulation is usually quite careful, so that speech can be understood by a broad audience. It is comprised mostly of prepared speech, so the speaking style cannot be qualified as fully spontaneous. There are substantially fewer hesitations here than as are observed in more informal, spontaneous speaking style corpora. However, the observed hesitations can be considered as relatively standard and prototypical, which is an interesting property for first cross-language comparisons.

The LIMSI speech transcription system [9] has been used for speech alignment. Using orthographic transcriptions and pronunciation dictionaries, the alignment system locates phones, hesitations, silences, breath and other noisy segments. For the class of extra-lexical events, such as silence, breath and filler words, the alignment system permitted an insertion at any word boundary. Since all aligned filler words are potential vocalic hesitations, before considering them as autonomous hesitation vowels, they were manually verified in order to avoid selection errors.

Parameter analysis (pitch, duration, timbre with first, second and third formants F1, F2, F3) was conducted using Praat [3]. For pitch and formant values, measures have been carried out on a frame by frame basis every 5ms. For each segment, a voicing ratio was computed as the ratio between the number of voiced frames and the total number of frames. Only segments with a voicing ratio > 0.4 are used in the analysis. For each eligible segment, mean values have been computed for all parameters using different strategies: (1) over all voiced frames of the segment; (2) over the three central voiced frames; (3) over three voiced frames in the first half (respectively the last half) of the segment. If not specified otherwise, the described measures correspond to the mean of the segment center (strategy 2). Differences between segment initial and final measures allow

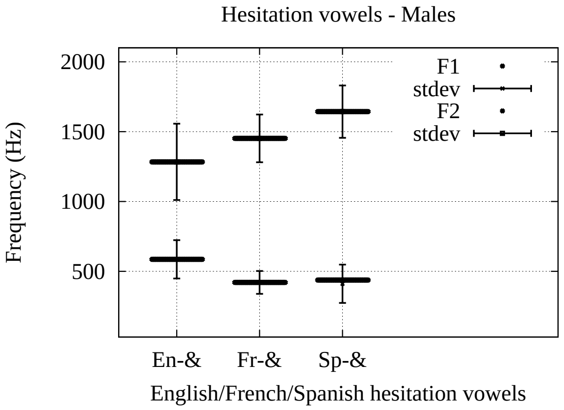

Figure 1. Average F1/F2 measures (and standard deviations) of hesitation vowels ("&") in English, French and Spanish (male speakers).

the slope of pitch and formants of hesitation and intra-lexical vowels to be estimated. In particular they have been used to measure the slope of F0 in hesitations.

2. Timbre

In this section we first investigate the timbre of hesitation vowels across languages in terms of F1 and F2 measures. Next, for each language, the hesitation vowel is examined with respect to its intra-lexical vowel system. Vocalic hesitation selection using automatic alignment resulted in approximatively 1300 items in French (respectively 2000 in English and 1700 in Spanish). Results here are presented for male speakers since they are better represented in the data, however the observations remain valid for female speakers.

2.1. Timbre of hesitation vowels

As already mentioned in the introduction, the transcription of vocalic hesitations varies across languages, for example *uh/um* in American English, *euh* in French or *eh* in Spanish. In order to measure acoustic differences between vocalic hesitations, formant values have been extracted using PRAAT. These differences can be seen in Figure 1, which presents the mean F1/F2 values obtained for American English, French and Spanish, suggesting distinct timbres for each language. The standard deviations of vocalic hesitations are very similar to those measured on intra-lexical vowels. Vowel timbre moves from mid-open in English to mid-closed in Spanish with a /œ/-like realization in French. The measured F1/F2 values are consistent with the expected formant values given the orthographic transcription of hesitation vowels in each language.

2.2. Link between hesitation vowels and vocalic systems

The formant measures given here show important differences between hesitations across languages. In this section we aim to clarify how the hesitation vowels are positioned in

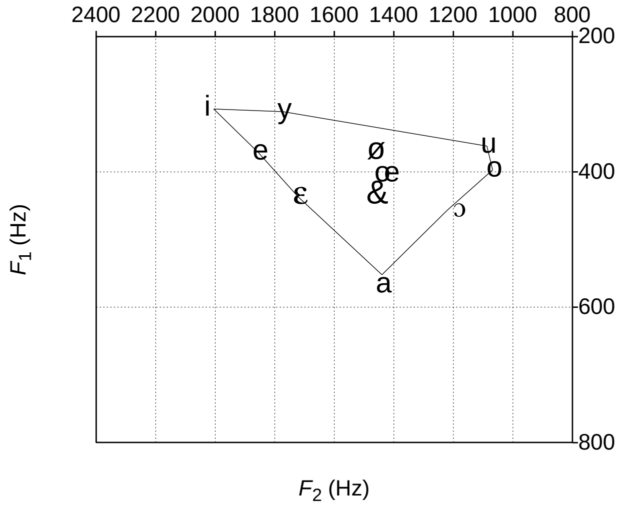

Figure 2. French F1/F2 mean values for intra-lexical vowels and hesitation vowel (“&”). The hesitation corresponds to a central mid-open vowel close to the French /œ/.

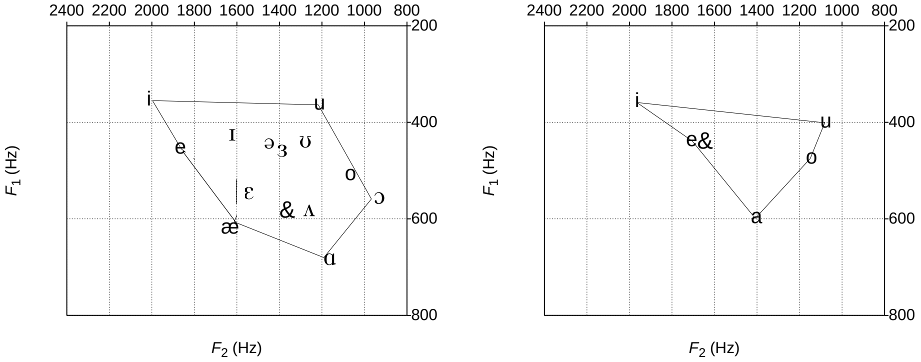

Figure 3. English and Spanish (from left to right) F1/F2 mean values for intra-lexical vowels and the hesitation vowels (“&”). For American English the mean position of the hesitation vowel corresponds to an open vowel, close to /ʌ/, whereas the Spanish hesitation is closest to the mid-closed front vowel /e/.

their vocalic system. To do so, formant values were extracted for all intra-lexical vowels in the three languages, and average formant values computed for each vowel.

Figure 2 shows the vocalic system of French using the average intra-lexical vowel formant values in a F1/F2 space. The vocalic hesitation is represented with “&” in the F1/F2 space along with the other French oral vowels. The resulting representation allows the distance between hesitation vowels and the phonemic vowels to be visualized. Similarly Figure 3 shows hesitation vowels in the vocalic systems for English and Spanish.

The comparison between vocalic hesitations and the vocalic systems reveals interesting differences across the 3 languages [12]. For European Spanish and to a lesser extent for French, vocalic hesitations exhibit timbres coinciding with another vowel in the system. In French the timbre is very close to a central /œ/ vowel, but slightly more

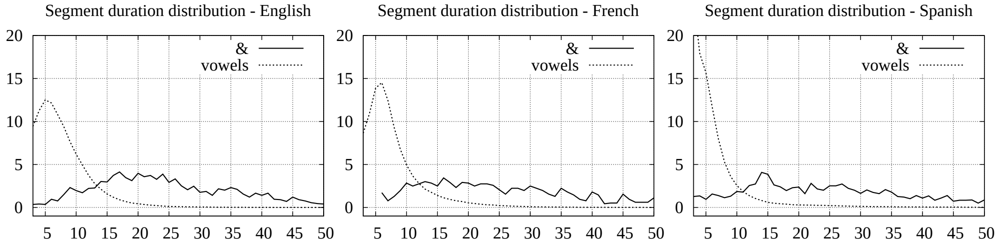

Figure 4. English, French and Spanish (from left to right) segment duration distributions for intra-lexical (vowels) and hesitation (&) vowels. X-axis : segment duration in centiseconds; Y-axis : percentage of segments in population.

open. Differences in the degree of opening might be correlated with durational aspects: as shown in [10], acoustic vowel spaces are more centralized, as the duration of the segments decreases. In contrast, for American English, the average timbre of the vocalic hesitation does not directly correspond to a vowel of the language inventory, as measured in our corpora (Figure 3). The average timbre of the hesitation vowel is closest to /ʌ/ and thus to a central position, thus it is more open than the vocalic hesitation in French.

As can be observed in Figures 2 and 3 the relation between the vocalic hesitation and language inventory varies across languages. Among the languages analyzed here, French and European Spanish vocalic hesitations exhibit timbres which are very close to existing vowels in the vocalic system. The average timbre of vocalic hesitations in American English is rather close to a central realization, however it does not coincide with an existing vowel. The average timbres of vocalic hesitations, close to central, mid-open realizations in French and American English and closer to a front mid-closed vowel in European Spanish, suggest that vocalic hesitations do probably not correspond to a universal speech "rest" position but that they are language-dependent. These results bring new acoustic evidence to the question of language-dependent articulatory settings supporting the hypothesis made by Gick et al. [11]. This also suggests that vocalic hesitations might be considered as a speech item carrying language-dependent information for automatic language identification. Consequently it may be interesting to use language-specific acoustic models which take into account timbre variability across languages.

3. Duration

Vocalic hesitations are widely described as being perceptually and objectively longer than intra-lexical vocalic segments [5,7,15]. Phonemic alignment of large speech corpora in different languages allows the comparison of segment durations of intra-lexical vowels and vocalic hesitations.

Figure 4 shows duration distributions of vocalic hesitations and intra-lexical vowels. Whereas the vowels have a modal value around 60 ms (even shorter for Spanish) vocalic hesitations exhibit rather flat distributions with a large proportion in the range of 150 to 250 ms. Intra-lexical segments are rarely lasting for more than 150ms.

In a recent study we have shown that duration can furthermore be correlated with speaking style: under stress vocalic hesitations tend to be even longer than in more con-

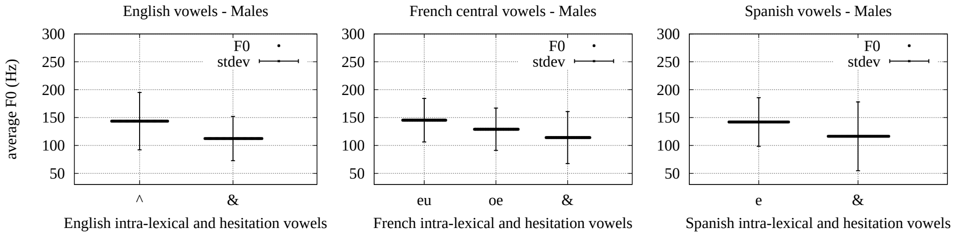

Figure 5. English, French and Spanish (from left to right) average F0 measures and standard deviations for intra-lexical vowels and hesitation vowel (&). For each language the selected intra-lexical vowels are closest neighbors to the hesitation vowel: for French /ø/ and /œ/; for Spanish /e/; for English /æ/ and /ʌ/.

trolled conditions such as semi-prepared journalistic speech [16]. These factors have been neglected here.

4. Fundamental frequency (pitch)

As for duration, fundamental frequency exhibits common patterns across languages, i.e. relatively stable, flat, slightly descending contours with low average values. Average F0 measures are shown in Figure 5 for the hesitation vowel and the closest intra-lexical vowels in each language. In order to give an idea, average pitch values for French male speakers were computed using 27k segments of the central mid-open vowel /œ/ [1], 3k segments for the central mid-closed vowel /ø/ and 1k for the hesitation vowel. Whereas significantly less data were used for female speakers the observed tendencies are globally the same. The average fundamental frequency of hesitation vowels is consistently lower than for intra-lexical vowels.

As for duration pitch can also vary with speaking style, for example speech produced under stress and language proficiency, i.e. the expression in native vs. foreign language [16].

5. Conclusion & Perspectives

This paper has studied the acoustic and prosodic characteristics of vocalic hesitations in a cross-language perspective. We aimed at establishing whether some language universals as for instance the concept of speech "rest" position can be related to the hesitation phenomena. From another point of view, the relation between the vocalic hesitations and the languages' vowel inventories was explored. The gained insights may also interest the acoustic modelling research community for application to automatic speech and language recognition.

Vocalic hesitations in French, American English and European Spanish were automatically extracted from large journalistic broadcast and parliamentary debate corpora. Duration, fundamental frequency and formant values were measured and compared both from an intra- and inter-language perspective. The results on timbre quality show that

[1] The average F0 measure for vowel /œ/ is computed using segments from both realized schwas (by far the most frequent ones) and [œ] segments.

vocalic hesitations are realized differently across languages. This suggests that, firstly, hesitations do not necessarily result from neutral realizations close to a rest position of speech, and secondly, this position, if salient, is language-dependent. Furthermore, the vocalic hesitations exhibit similarities with some intra-lexical vowels but the degree of similarity varies across languages. In European Spanish the vocalic hesitation almost coincides with the mid-closed front-vowel /e/ whereas in American English and French, it is central slightly more open than /œ/ (in French) and mid-open central between /ʌ/ and /æ/ (in American English). Finally, in the framework of language identification approaches, the results suggest that vocalic hesitations require language-specific acoustic models.

Fundamental frequency and duration exhibit common patterns across languages. The average fundamental frequency of the intra-lexical segments is consistently higher than that of vocalic hesitations, whereas the duration of vocalic hesitations is usually significatively longer than intra-lexical vowels.

Many more detailed studies can be carried out in the future, including analyses depending on different hesitation functions. Extensions to other languages, regional accents and to more spontaneous speech collected in various interaction contexts, are certainly promising future research directions which can contribute to an extensive description and more in-depth understanding of hesitation phenomena in speech.

6. Acknowledgements

This work has been partially financed by the French CNRS MIDL and European EC TC-STAR and CHIL projects. The authors would like to thank Maria Candea (Paris 3 - Sorbonne Nouvelle), for her collaboration to parts of this work; and Rena Nemoto for her work comparing the hesitation vowels to the vocalic systems in French, American English and European Spanish.

References

[1] M. Adda-Decker and L. Lamel, "Pronunciation variants across system configuration, language and speaking style", Speech Communication, vol. 29, pp. 83-98, 1999.

[2] M. Adda-Decker and al., "A Disfluency study for cleaning spontaneous automatic transcripts and improving speech language models", In DiSS-2003, Papers in Theoretical Linguistics, vol. 90, pp. 67-70, 2003.

[3] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer", Institute of Phonetic Sciences of the University of Amsterdam", pp. 132-182, 1999.

[4] M. Candea, Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané. Phd dissertation, University of Paris 3,2000.

[5] M. Candea, I. Vasilescu and M. Adda-Decker, "Inter- and intra-language acoustic analysis of autonomous fillers", In DiSS-2005, pp. 47-51, 2005.

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

[7] Duez, D., "Caractéristiques acoustiques et phonétiques des pauses remplies dans la conversation en français", Travaux Interdisciplinaires du Laboratoire Parole et Langage, vol. 20, pp. 31-48, 2001.

[8] D. Duez, "Modelling Aspects of Reduction and Assimilation in Spontaneous French Speech", In Proc. of IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition, Tokyo, 2003.

[9] J.L. Gauvain, L.F. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System", Speech Communication, vol.37(1-2), pp. 89-108, 2002.

[10] C. Gendrot, M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German", Proc. of Eurospeech-Interspeech, Lisboa, Portugal, 2005.

[11] B. Gick and al., "Language specific articulatory settings: evidence from inter-utterance rest position", Phonetica, vol. 61(4), pp. 220-233, 2004.

[12] R. Nemoto, "Hésitation vocalique autonome vs. système phonétique de la langue : étude acoustique en plusieurs langues", Mémoire de DESS (Master dissertation), LIMSI-CNRS, 2006.

[13] H. Maclay and C.E. Osgood,"Hesitation phenomena in spontaneous English speech", Word, vol. 15, pp. 19-44, 1959.

[14] E. E. Shriberg, Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley,1994.

[15] E.E. Shriberg, "The 'errrr' is human: ecology and acoustics of speech disfluencies", Journal of the International Phonetic Association, vol. 31(1), 2001.

[16] I. Vasilescu and M. Adda-Decker, "Language, gender, speaking style and language proficiency as factors characterizing autonomous vocalic fillers in spontaneous speech", Proc. of ICSLP 2006, Pittsburgh, USA, 2006.

[17] M. Watanabe,Y. Den, K. Hirose and N. Minematsu, "The effects of filled pauses on native and non-native listeners speech processing", In DiSS-2005, pp.169-172, 2005.

[18] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses", In DiSS-2005, pp.179-182, 2005.