

Inter- and intra-language acoustic analysis of autonomous fillers

Maria Candea¹, Ioana Vasilescu², Martine Adda-Decker³

¹ Paris 3 – EA1483, 13 rue de Santeuil, bur.431, 75005 Paris, ²LTCl-ENST, 46, rue Barrault, 75634 Paris cedex 13, ³LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex
maria.candea@univ-paris3.fr, vasilesc@tsi.enst.fr, madda@limsi.fr

Abstract

The present work deals with autonomous fillers in a multilingual context. The question addressed here is whether fillers are carrying universal or language-specific characteristics. Fillers occur frequently in spontaneous speech and represent an interesting topic for improving language-specific models in automatic language processing. Most of the current studies focus on few languages such as English and French. We focus here on multilingual fillers resulting from eight languages (Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish). We propose thus an acoustic typology based on the vocalic peculiarities of the autonomous fillers. Three parameters are considered here: duration, pitch (F0) and timbre (F1/F2). We also compare the vocalic segments of the fillers with intra-lexical vowels possessing similar timbre. In this purpose, a preliminary study on French language is described.

1. Introduction

Among various hesitation or “edition” phenomena, the one we analyze here is widely encountered in world’s languages, i.e. the insertion at any moment within spontaneous speech of a long and stable vocalic segment, defined as a type of filler. The role of this item is “to announce the initiation of what is expected to be a [...] delay in speaking” [1]. Such elements have no lexical support and are hence distinguished from the lengthening of a vocalic segment belonging to a particular lexical item (most often a function word). Most of the studies conducted on large spontaneous speech corpora have focused on English or French [2], [3], [4], [5], [6], [7], even if recent description can be found in other languages (see for example [9] or the proceedings of the DiSS03 workshop [8]).

We address here the question whether the autonomous fillers are carrying universal acoustic characteristics or language-specific information. They occur frequently in spontaneous speech, i.e. about five percent in spontaneous corpora, and this proportion can increase according to the spontaneous speech communication situation. We are also interested in the modeling problem of these phenomena in a language identification context. The question is then whether autonomous fillers (such as *uh/um/er* in English and *eah* in French) deserve language-specific models or whether a language-independent filler model is more appropriate.

The vocalic segment of autonomous fillers is generally lengthened. This segment can occur alone or surrounded by additional segment as nasal coda in English (*um*) and represent in our terminology the *vocalic support* of the filler.

More precisely, we study in this paper the vocalic peculiarities of autonomous fillers in several languages, i.e. the realization of a central vs. non-central timbre of their vocalic support. In previous studies [10], [11] we observed acoustic differences among the vocalic supports of the multilingual fillers. We also conducted perceptual experiments in order to test listeners’ capacity at differentiating languages from isolated autonomous fillers without any context [12].

In the following section we describe the corpus and the methodology adopted. Section 3 is dedicated to the inter-language analysis of the acoustic characteristics of fillers. It will be followed by intra-language study carried on French (section 4). Finally, section 5 will summarize the current findings.

2. Corpus and methodology

A multilingual broadcast corpus has been gathered for the following eight languages: standard Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish. French and Arabic are French DGA resources, partially available via the ELDA linguistic resources agency. English, Spanish and Mandarin are excerpts from LDC Hub4 corpora. German, Portuguese and Italian BN data are resources acquired within various European FP5 LE projects (OLIVE, ALERT) or purchased from ELDA. The audio data correspond either to news data which is mainly prepared speech, or news-related shows containing more spontaneous speech specific items. From this multilingual corpus, a subcorpus of autonomous fillers has been extracted semi-automatically for the eight languages under consideration: fillers, which have been located automatically in aligned speech, are listened to and selected if the selection criteria are met.

Filler extraction is based on duration and autonomy criteria. 200ms has been considered as the minimum duration threshold. Items considered in this study as autonomous fillers are isolated from the speech context by silences in order to avoid lengthened words. Finally, 57 to 1889 occurrences per language have been selected for both genders (see table 1 below). The size of the present corpus allows exploring the questions mentioned in introduction. However some languages are less represented and as a general observation data from female speakers are less abundant. Current work conducted by the authors focuses on the size of the database on hesitations, which is progressively increased.

Table 1: Number of occurrences of hesitations per language (male and female speakers).

Language	Nb. of occurrences (M+F)
Arabic	246
Mandarin Chinese	89
French	1889
German	458
Italian	57
European Portuguese	64
American English	532
Lat. Amer. Spanish	93

A supplementary corpus has been extracted in French in order to conduct the intra-language analysis from about 6 hours various types of broadcast speech. In addition to the 1889 fillers (1509 from male speakers, 380 from female speakers) it contains also other intra-lexical vocalic segments with similar vocalic qualities, i.e. 1718 [ɛ-ɛ] (954 male, 754 female) and 1114 [ø] (923 male, 291 female). The intra-lexical vocalic segments have been extracted via the LIMSI speech alignment system. A duration criterion has been employed, i.e. we selected intra-lexical vocalic segments superior to 40 ms. As a general observation we can notice a higher representation of male compared with female speakers. The vowel proportion reflects the gender representation observed in broadcast-type corpora (about 70% male vs. 30% female speakers). Besides, the number of occurrences per vocalic timbre illustrates also the frequency of the analyzed segments in French. Indeed, the vowel [ø] is 10 times less frequent than [ɛ-ɛ].

The PRAAT software¹ has been used to extract the acoustic parameters comprising fundamental frequency (F0) and the first two formants (F1, F2).

3. Inter-language analysis: acoustic features of fillers in eight languages

Three parameters have been considered: the duration, the F1/F2 characteristics of the fillers' vocalic segments and the pitch (F0). Whereas pitch and duration are mainly useful to localize the fillers in the speech flow, the F1/F2 parameters potentially contain more language specific characteristics. The F0 and the duration of the fillers do not show significant differences among the eight languages confirming previous findings [1], [2], [3], [6], [7], i.e. fillers are significantly longer than intra-lexical vocalic segments (see below French example) and have a flat F0 contour. The behavior of the two parameters seems to be identical across the eight languages. This results answer affirmatively to the first question mentioned above, i.e. duration and pitch tend to be universal criteria.

In return, the acoustic analysis of F1/F2 peculiarities of the fillers' vocalic segments reveals language-dependent characteristics.

The figures 1 and 2 below provide the mean values for F1 and F2 per language.

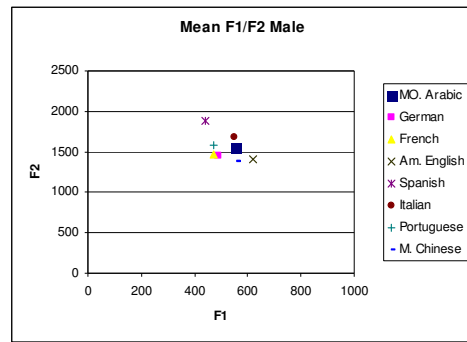


Figure 1: F1/F2 distribution of vocalic segments of autonomous fillers: all languages (mean values for male speakers)

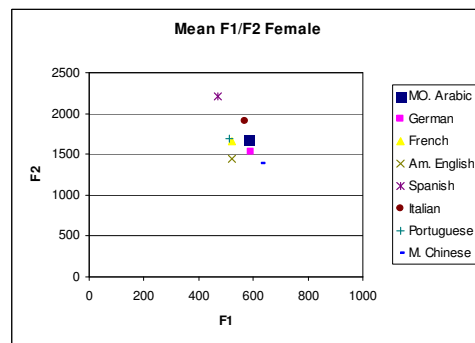


Figure 2: F1/F2 distribution of vocalic segments of autonomous fillers: all languages (mean values for female speakers).

Preliminary results strengthen the hypothesis of timbre differences across languages of the vocalic support of the autonomous fillers. Indeed, the central position does not seem to be a universal realization. These results tend to show that different languages analyzed here admit various vocalic realizations. The realizations can be central [ɛ] and/or correspond to other vocalic quality. We can hypothesize that the vocalic supports are vowels of the system. Spanish employs thus a mid closed vowel [e] and English makes use of low central vowels. Italian so far is the only language of the corpus with both central [ɛ] (which is not part of the Italian vocalic system) and non-central vocalic supports, i.e. the front mid open vowel [ɛ]. More data from other languages are needed to consolidate these hypotheses. Finally, the observed differences are not uniquely in terms of vocalic timbre. Language-specific features can be observed in the segmental structure of the fillers. French, for example, prefers a vocalic segment as filler realization, whereas English prefers vowels followed occasionally by a nasal coda consonant [m], which confirms observations made by [1] and [7]. In Portuguese as well, more complex diphthongized segments can be found.

To conclude, for some languages the vocalic support of the fillers might be a segment exterior to the vocalic system of the language (i.e. Italian in our corpus). However, all the eight languages seem to accept as fillers' vocalic support at least one of the vowels of their vocalic system. The vowel generally exhibits a timbre close to a quite central position.

¹ www.praat.org

However the central position does not seem to be universal “rest position”, but rather a language dependent realization.

In order to evaluate the relationship between the vocalic support of the filler and the vocalic system of the language we conducted a preliminary analysis on the French language. We compared thus the vocalic timbre of “*eah*” with the closest vowels of the system [ɔ], [ø] and [œ].

4. Intra-language analysis: vocalic support of French filler “*eah*” vs. vocalic system

As for the vocalic support of the fillers, the same parameters have been considered for the intra-lexical vowels of the system: duration, vocalic timbre (F1/F2) and pitch (F0).

Duration analysis confirms previous observation made by [2], [4], [5]. The distribution of the duration for fillers exceeds significantly the duration of intra-lexical segments as shown in Figures 3, 4 and 5 below. The duration criterion adopted here avoids fillers below 200ms. Fillers shorter than 200ms definitely exist in the spontaneous speech. However, as the extraction has been conducted automatically, we selected a threshold high enough to eliminate the potential confusion with intra-lexical segments.

In order to evaluate the amount of fillers potentially eliminated by the selected threshold, we proceeded to a listening of a 45 minutes speech sample and we manually extracted fillers shorter than 200ms. It appears that 14.7% of the fillers show a duration between 150 and 200ms and 11.6% show durations inferior to 150ms. This observations support the hypothesis that fillers are mainly longer than 200ms. The listening experiment confirmed though that by selecting a 200ms threshold we eliminated about 25% of real fillers from our analysis.

We can notice thus that duration of a very large part of the fillers varies from 200 till 650 ms whereas intra-lexical segments rarely extend beyond 200ms. Fig. 4 and Fig. 5 show differences in the duration distribution for intra-lexical vowels. The duration distribution for the vowels [ɔ-œ] (among them schwa segments are suppressible in speech) focuses close to the minimum segment duration whereas duration for [ø] has a broader distribution with a peak around 80ms. Schwa segments are suppressible in speech and belong to non accented syllables. The duration is thus the shortest among the analyzed segments, the exception being the lengthened schwa (i.e. the realization as fillers). The [ø] segments can occur in accented syllable as well, and the duration shows a more important variability in terms of realization.

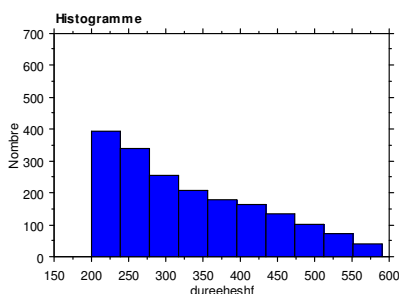


Figure 3: Duration above 200ms vs. nr of occurrences distribution for fillers (male/female)

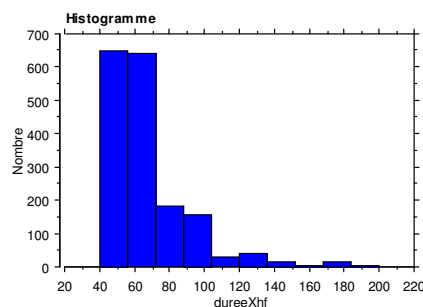


Figure 4: Duration vs. nr of occurrences distribution for intra-lexical [ɔ-œ] (male/female).

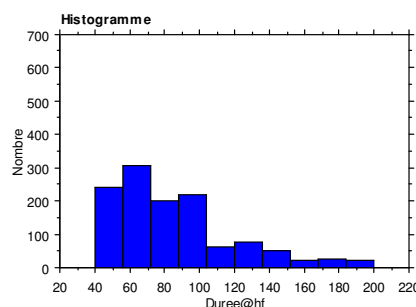


Figure 5: Duration vs. nr of occurrences distribution for intra-lexical [ø] (male/female).

Concerning **timbre** analysis of fillers and intra-lexical vowels, mean F1/F2 measures are shown in Table 1. The measured values are compared with reference values provided by [13] and by [14] for the French intra-lexical vowels.

Table 2: F1/F2 mean values and for male and female speakers of vocalic support of filler “*eah*” and of intra-lexical segments [ɔ-œ] and [ø] (Hz).

	F1 male/female	F2 male/female
Fillers	470/523	1464/1659
[ɔ-œ]	404/413	1421/1675
<i>Fant</i> [13]	500/550	1450/1650
<i>Gendrot&al.</i> [14]	400/437	1444/1659
[ø]	382/430	1465/1666
<i>Fant</i> [13]	400/450	1450/1650
<i>Gendrot&al.</i> [14]	375/417	1465/1677

Values presented in Table 2 do not show any notable difference in terms of back/front distribution, i.e. values on F2 for both filler and intra-lexical segments are analogous. On the F1 axis (open/close) a difference is observed in particular for male speakers, in vocalic support of the fillers vs. intra-vocalic segments: the first ones are more open. This difference could be related to the duration, as the intra-lexical vocalic segments are significantly shorter than the fillers. Among the intra-lexical vocalic segments, [ɔ-œ] are more open than [ø]. However, these differences are not statistically significant (independent t-test).

Mean values calculated for F1 for the intra-lexical vowels are similar to the values observed by [14], for male and female speakers. Mean values for F2 are similar to values observed by both [13] and [14].

We considered as well the F3 in order to evaluate if a distinction could be made in terms of rounded/non-rounded opposition. We compared thus the measurements automatically completed by [14]. They reach mean values of 2500 Hz for both [□-□] and [ø], consequently we do not have evidence of a timbre difference of the analyzed vowels.

Finally, measurements on the **pitch (F0)** have been calculated for both vocalic segments of the fillers and intra-lexical vowels. Table 3 shows mean values and standard deviations in F0 for the three types of segments. F0 differences among vowels are not significant (independent t-test) and the current data do not allow at concluding on the fillers mean F0 peculiarities compared with the intra-lexical vocalic segments.

Table 3: F0 mean values and standard deviation for F0 (male and female speakers) for vocalic support of the filler “*euh*” and for intra-lexical segments [□-□] and [ø] (Hz).

	F0mean	St.Dev.
	Male/Female	Male/Female
Fillers	155/179	97,3/68,5
[□-□]	114/192	48,6/38,3
[ø]	144/219	45,4/56

This preliminary result suggests that F0 distribution for the fillers correspond to a larger “ambitus” than for the intra-lexical segments. This variability is suggested by the standard deviation values globally higher for the fillers. These remarks concern more male than female speakers. However, if we listen to the fillers providing high F0 values (>300Hz), they seem to correspond to detection errors of PRAAT more than to extreme articulations. In addition, we could notice that number of fillers show both a low perceived F0 and an irregular voice quality (i.e. vocal fry). Most of the F0 detection errors stem from these type of segments and in a further work we intend to compute male F0 mean without these potential “erratic” segments.

In order to get an overall impression of the irregular voice quality in the production of the fillers, we compared the number of F0 detection errors for “*euh*”, [□-□] and [ø]. They correspond to the “undefined” values provided by Praat for the speech samples for which the software could not compute the F0. These detection errors concern more often the fillers than the intra-lexical vocalic segments. Detection errors for [□-□] and [ø] represent 3% of the production of the male speakers and 0,5% of the female speakers. In return, for the fillers, F0 detection errors represent 11,5% of the productions of the male speakers and 8,2% of the female speakers. These findings corroborate the hypothesis of an unstable voice quality in the production of the fillers, which could be either vocal fry, creaky or breathy. Besides, these observations confirm previous remarks made by [6] and [7] about the American English: the intra-lexical vowels are produced more often with a modal voice quality than fillers. Finally, differences in detection errors for male vs. female speakers suggest that women might tend to pay more attention to the degree of control of their “disfluent” productions than men. Our further studies will consider more deeply the voice quality aspect of the fillers compared with “fluent” speech. We hypothesise that fillers might be produced with a limited articulator and/or airflow.

5. Conclusions

In this paper we presented inter- and intra-language acoustic analysis of autonomous fillers. The question which has been

addressed is whether the fillers possess universal acoustic characteristics or if they are language-specific phenomena. The current work answers partially to the question.

Fillers are frequent phenomena in spontaneous speech and show regular patterns in terms of duration, fundamental frequency and vocalic timbre which tends to be central for most languages considered in this study. Whereas F0 and duration are similar among languages, language-specific acoustic characteristics can be observed in terms of vocalic quality and segmental structure of the fillers. Among the eight languages analyzed in this paper, non-central vocalic timbre characterize at least two of them, Spanish and to a lesser extend, English.

The comparison of the vocalic support of French fillers with the closest intra-lexical vocalic segments [□], [ø] and [□] allowed at observing differences in terms of duration, pitch and voice quality. Differences on the F1 axis were also observed, but they are not statistically significant.

Interesting questions still remain open. They concern for example the relationship between the vocalic timbre of the fillers and the vocalic systems. The comparison of the fillers’ vocalic supports and the intra-lexical vowels will be developed as well on other languages of the multilingual corpus in order to determine the relationship between the fillers and languages’ vocalic system.

Another aspect related to the universal characteristics of the fillers concerns the relationship between the filler and the so-called “articulatory rest position”. The question addressed is whether this position exists and whether the fillers are close to it and thus they might represent universal realizations across languages. The preliminary results presented in this paper seem to exclude the universal timbre of the fillers. In addition, a recent study on English and French articulatory settings [16] suggests that a clear correlation between “hesitation vowel” and the “articulatory rest position” could not be proved. The “articulatory rest position” itself seems to be significantly different for each language. Such studies are still in progress and may provide interesting information about an eventual influence, in a given language, of the global speech posture and articulatory rest position on vocalic fillers.

Finally, further studies should consider other aspects allowing at describing fillers in the context of the so called “disfluencies phenomena” which characterize the spontaneous speech. It would be thus interesting to observe those aspects which differentiate fillers from vocalic lengthening. The relationship between vocalic fillers such as *huh*, *hum*, *euh* and language word hesitation (as, for example in Japanese [9]) would be as well examined.

6. Acknowledgements

This research has been carried out in the context of the MIDL (Modélisations pour l’IDentification des Langues) project, supported by a CNRS interdisciplinary program and involving several French laboratories (LIMSI-CNRS, LTCI-ENST, CTA-DGA, LPP Paris3 and EA 1483 – Paris3). Its aim was to bring together linguistic and computer engineering knowledge in order to increase our knowledge in human language identification and to contribute to the domain of automatic language identification.

The authors want to thank Cédric Gendrot (ILPGA, Paris III) for his help in the acoustic analysis of the corpus.

7. References

- [1] Clark H.H., Fox Tree J.E. 2002. Using uh and um in spontaneous speaking, *Cognition* 84, 73-111.
- [2] Adda-Decker et al. 2003. A Disfluency study for cleaning spontaneous automatic transcripts and improving speech language models, *DISS'03, Göteborg, Sweden (Papers in Theoretical Linguistics 90)*: 67-70).
- [3] Shriberg, E., Bear, J., Dowding, J. 1992. Detection and correction of repairs in human-computer dialog, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Delaware.
- [4] Candea, M. 2000. *Contribution à l'étude des pauses silencieuses et phénomènes dits "d'hésitation" en français oral spontané*. PhD thesis, University of Paris 3-Sorbonne nouvelle.
- [5] Guaitella, I. 1991. Hésitations vocales en parole spontanée: réalisations acoustiques et fonctions rythmiques, *Travaux de l'Institut de Phonétique d'Aix*, vol.14: 113-130.
- [6] Shriberg, E. 1999. Phonetic consequences of speech disfluency, *ICPhS'99*, San Francisco.
- [7] Shriberg, E., 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, 31/1.
- [8] Eklund R. editor, 2003. Disfluencies in Spontaneous Speech, DiSS workshop, *Proceedings of DISS'03, (Papers in Theoretical Linguistics 90)*, Göteborg.
- [9] Watanabe, M. 2003. The constituent complexity and types of fillers in Japanese, *15th ICPhS*, Barcelone.
- [10] Clerc-Renaud J. Vasilescu I., Candea M., Adda-Decker M. 2004. Etude acoustique et perceptive des hésitations autonomes multilingues, *XXV^{es} JEP*, Fès Morocco.
- [11] Vasilescu I., Candea M., Adda-Decker M. 2004. Hésitations autonomes dans 8 langues : une étude acoustique et perceptive, *Workshop MIDL04*, Paris F.
- [12] Vasilescu I, Candea M., Adda-Decker M., 2005. Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages, *Interspeech 2005* Lisboa, Portugal.
- [13] Fant, G., 1973. *Speech sound and features*, MIT Press, Cambridge, USA.
- [14] Gendrot C., Adda-Decker M., 2004. Analyses formantiques automatiques de voyelles orales: évidence de la réduction vocalique en langues française et allemande, *Workshop MIDL04*, Paris France.
- [15] Calliope, 1989. *La parole et son traitement automatique*, Paris, Masson ed.
- [16] Gick B., Wilson I., Koch K., Cook C., 2004. Language specific articulatory settings: evidence from inter-utterance rest position, *Phonetica*, 61 (4), 220-233.