Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages

Ioana Vasilescu, Maria Candea, Martine Adda-Decker

Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages

Ioana Vasilescu[1], Maria Candea[2], Martine Adda-Decker[3]

[1]LTCI-ENST, 46, rue Barrault, 75013 Paris - France, [2] Paris 3 – EA1483, 13 rue de Santeuil, bur.431, 75005 Paris - France, [3]LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex
vasilesc@tsi.enst.fr, maria.candea@univ-paris3.fr, madda@limsi.fr

Abstract

Are acoustic differences in autonomous fillers salient for the human perception ? Acoustic measurements have been carried out on autonomous fillers from eight languages (Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English and Latin American Spanish). They exhibit timbre differences of the support vowel of autonomous fillers across languages. In order to evaluate their salience for human perception, two discrimination experiments have been conducted, French/L2 and Portuguese/L2. These experiments test the capacity to discriminate languages by listening to isolated autonomous fillers without any lexical support and without any context. As listeners have been native French speakers, the Portuguese/L2 experiments aim at evaluating a potential mother tongue bias. Results show that the perceptual discrimination performance depends on the language pair identity and they are consistent with the acoustic measures. The present study aims at providing some insight for acoustic filler models used in automatic speech processing in a multilingual context.

1. Introduction

Among various hesitation or editing markers, 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], even if recent descriptions can be found in other languages (see for example the proceedings of the DiSS03 workshop [7]).

We address here the general question whether the autonomous fillers possess universal acoustic characteristics or whether they are carrying language-specific and perceptually salient information. They occur frequently in spontaneous speech, i.e. about five percent in spontaneous corpora, and this proportion can increase depending on the spontaneous speech communication situation. The vocalic segment of the autonomous filler generally represents a lengthened central segment (i.e. schwa or other proximal vocalic unit). 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.

In previous studies [9], [10] we observed acoustic differences among the vocalic supports of the multilingual fillers (i.e. the realization of a central vs. non-central segments). Given these observed inter-language differences in terms of vocalic timbre of the fillers, we focus in this paper on the perceptual salience of their language-specific particularities. If autonomous fillers (such as *uh/um/er* in English and *euh* in French) deserve language-specific models, these information could be useful in a multilingual speech processing context.

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 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, 30 to 200 occurrences per language and per gender have been selected. An important inter-language variability in terms of number of occurrences per language has been observed. In particular our corpus is relatively poor in hesitations for Mandarin Chinese. For German only a small amount of hesitations could be gathered for women, whereas male fillers are rather frequent. However, the size of the present corpus allows exploring the hypothesis mentioned above.

The PRAAT software[1] has been used to extract the acoustic parameters comprising fundamental frequency (F0) and the first two formants (F1, F2). Perceptual tests have been conducted via an interface created with the EPRIME software[2].

[1] www.praat.org
[2] www.pstnet.com/e-prime.

3. Acoustic features of fillers in eight languages

Three parameters have been considered: the F1/F2 characteristics of the vocalic segment of each hesitation, the pitch of the hesitation (F0) and the duration. Whereas pitch and duration are mainly useful to localize hesitations in the speech flow, the F1/F2 parameters potentially contain more language specific characteristics. The F0 and the duration do not show significant differences among the eight languages confirming previous findings [2], [3], [4], [6], i.e. significantly longer than the intra-lexical vocalic segments and stable F0 contour. The acoustic analysis of F1/F2 peculiarities of the vocalic segment of fillers is more interesting in terms of language-dependent characteristics.

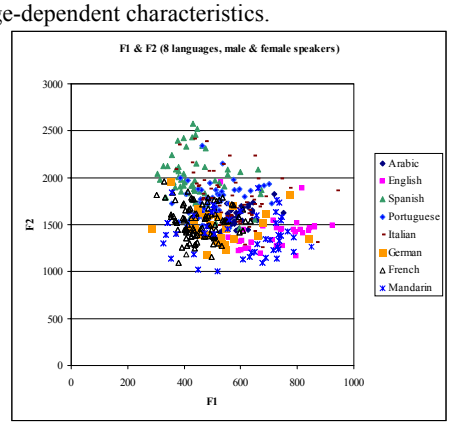

Figure 1: F1/F2 distribution of vocalic segments of autonomous fillers: all languages, male/female

The preliminary results strengthen the hypothesis of timbre differences of the support vowel of the autonomous fillers across languages. Indeed, the central position does not seem to be a universal realization. These results show that the different languages analyzed here admit various vocalic realizations which are either vowels of their vocalic system *or* a central realization [ə]. English mainly makes use of low central vowels existing in the vocalic system whereas in Italian both central [ə] (which is not part of the Italian vocalic system) and the front mid open vowel [ɛ] can be observed. Spanish employs a non-central segment [e]. Therefore, a central vocalic position or a close position seems to be preferred by all the languages we analyzed here, excluding thus high front and back vowels. However, more data are needed to consolidate these hypotheses. Finally, the observed differences are not uniquely in terms of vocalic timbre. They concern also the segmental structure of the autonomous fillers. French, for example, prefers a vocalic segment as filler realization, whereas English prefers vowels followed frequently by a nasal coda consonant [m], which confirms observations made by [6]. In order to find perceptual evidence of the language-specific peculiarities of the fillers we conducted two perceptual experiments, described below.

4. Perceptual evidence for language-specific features of fillers

The perceptual experiments aim at testing listeners' capacity at differentiating languages from autonomous fillers without any lexical support and without any context.
Two discrimination tasks have been chosen: French (**L1**)// other language (**L2**) and Portuguese (**L1'**) // other language (**L2**), as French and Portuguese fillers present similar vocalic qualities. The two target languages, French and Portuguese, have been selected in order to verify the hypothesis of the language specific differences among fillers. As the listeners were for both tests native French speakers, the choice of two target languages according to a [+/- native language] criterion should allow at eliminating a potential effect of the mother tongue bias. Autonomous fillers have been selected in order to illustrate prototypical realizations for each of the eight languages.

42 native French listeners (20 for **L1/L2** and 22 for **L1'/L2**) have been participating in the experiments. Each experiment opposed the target language (French L1 or Portuguese L1') to the remaining 7 languages of the broadcast corpus presented in section 2. Accordingly, the experiments consisted in 7 L1/L2 and 7 L1'/L2 subtests presented *via* an interface created with EPRIME.

A familiarization phase preceded each subtest and consisted in listening to several speech samples containing fillers in a lexical context, extracted from the two languages of a given subtest. During each subtest, subjects listen to 24 (12 x 2) autonomous fillers without any context (12 per language). They decided if the language of extraction of each stimulus was French or L2 (where L2 is one of Italian, Spanish, Portuguese, American English, German, Arabic, and Mandarin Chinese) for the subtests L1/L2. The same question have been addressed for the Portuguese/L2 subtests (where L2 is one of Italian, Spanish, French, American English, German, Arabic, and Mandarin Chinese).

Stimuli were different for each test, pronounced both by male and female speakers. Particular individually-marked voices (i.e. high creakiness or breathiness) and particular recordings qualities (i.e. noisy) have been eliminated. The duration of the stimuli was varying from 200ms to 600ms. The stimuli selected for the perceptual tests present the F1/F2 characteristics illustrated by figures 2 and 3 below. They support thus the peculiarities described in the previous section.

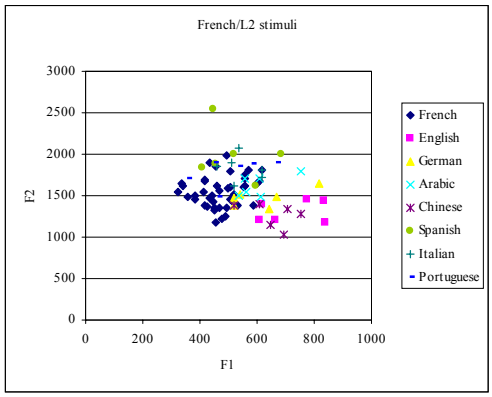

Figure 2: F1/F2 distribution of vocalic segments of autonomous fillers employed as stimuli in the French/L2 perceptual test.

The distribution of the vocalic segment of the multilingual stimuli in the F1/F2 plane show similar vocalic

qualities for French and Portuguese. Those qualities are shared with some stimuli in other L2 languages (such as German, Arabic, Italian), and are definitely different from those of the stimuli in English and Spanish. The question addressed by the perceptual tests concerns the perceptual salience of those differences. However, we have to keep in mind that other peculiarities could play a role in the correct discrimination (general language/gender-specific voice quality, duration, other segmental peculiarities as the nasal coda in English, etc.). Results are presented in subsections 4.1. and 4.2. below.

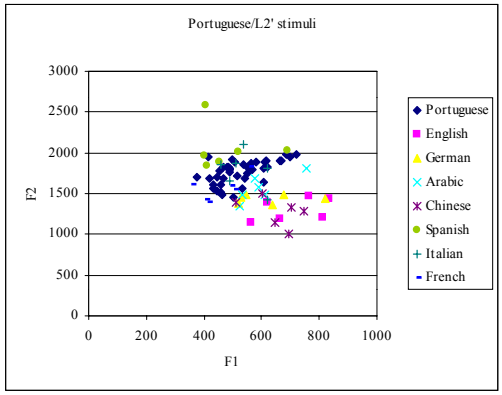

Figure 3: F1/F2 distribution of vocalic segments of autonomous fillers employed as stimuli in the Portuguese/L2 perceptual test.

4.1. French vs. L2 (L1/L2)

We present here the measured discrimination scores for each subtest French/L2 (Figure 4). The average discrimination score is 75%. A general observation is that French autonomous fillers have been discriminated from the L2 ones with scores which are significantly above chance level. Two groups of L2 emerge according to the correct identification: Arabic, Portuguese, and German, close to chance level (<65%) vs. Italian, Spanish, English and Mandarin Chinese (>80%). Differences in correct discrimination scores between the 2 groups are statistically significant (t-test, $p<0.0001$). Moreover, given the factors "mother tongue" and "L2", a significant effect of the factor "L2" has been observed (ANOVA, $p<0.001$). "L2" has been globally and statistically better recognized than French (ANOVA, $p=0.0198$), more particularly for the subtests French / Mandarin Chinese, French / Arabic and French / German. We hypothesize that French listeners evaluated the stimuli according to the *acoustic differences* compared to the "prototypical" stimuli from their native language.

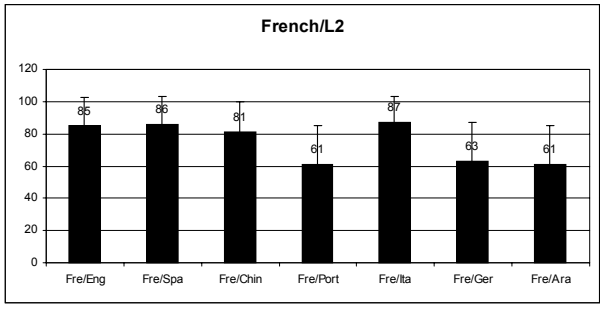

Figure 4: Percentages of correct discrimination for French/L2.

These results seem to confirm the hypothesis that acoustic particularities of fillers are responsible for the discrimination in a higher proportion than the presence of the mother tongue among the languages of the test. In order to validate this hypothesis we conducted the second test Portuguese/L2.

4.2. Portuguese vs. L2 (L1'/L2)

We present here the measured discrimination scores for each subtest Portuguese/L2 (Figure 5).

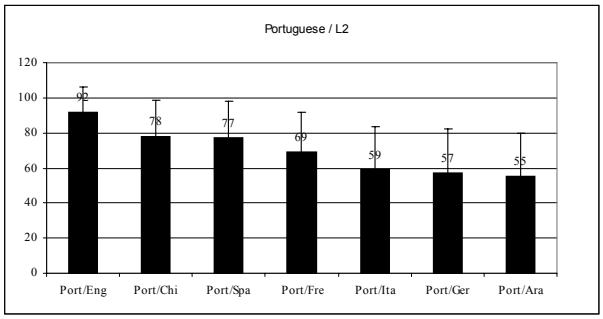

Figure 5: Percentages of correct discrimination for Portuguese/L2.

English, Spanish, Mandarin Chinese and French fillers have been discriminated from the Portuguese ones with scores which are significantly above chance level ($p<0.0001$); answers for the other three languages are not different from the chance level. The average discrimination score is 70% (79% for the subtests Portuguese/L2 providing scores above the chance level, i.e. Portuguese/English, Portuguese/Spanish, Portuguese/Chinese and Portuguese/ French). Differences in correct discrimination scores between the 2 groups (French, Spanish and Mandarin Chinese vs. Arabic, German and Italian) are statistically significant (t-test, $p<0.0001$). As for the test French/L2, we statistically evaluated the role of the different factors involved in the correct discrimination. We analyzed thus the role of the factors "target language" (Portuguese) and L2. As for the previous French/L2 test, L2 has been slightly better identified than L1' (ANOVA, $p=0.0011$). This finding confirms the discrimination strategy employed for the previous test, i.e. listeners are searching for acoustic differences of the L2 compared to the target language.

As for the test French/L2, two groups of L2 emerge in terms of correct identification: Arabic, German and Italian (<60%) vs. French, Spanish, English and Mandarin Chinese (>%69). The groups are similar as for the previous French/L2 test, except for Italian language. French has been

discriminated from Italian with scores superior to the chance level, whereas Portuguese/Italian subtest received chance answers. Italian is so far the only language providing two vocalic qualities for the fillers, [ə] and [ɛ]. More data are needed in order to define the role of the respective realizations among the Italian disfluencies; however, we decided to keep the two timbres in the perceptual tests, both French/L2 and Portuguese/L2. We hypothesize that the better discrimination of French from Italian could be analyzed as the illustration of a "mother tongue bias". The Table 1 below allows at better evaluating the role of different factors involved in the correct discrimination.

Table 1: comparison of the percentages of correct discrimination for French/L2 (L1/L2) and Portuguese/L2 (L1'/L2)

LggePair/ %	French-L1/L2	Portuguese-L1'/L2
English	85%	**92%**
Chin. Mand.	**81%**	78%
Spanish	**86%**	77%
French	X	69%
Portuguese	61%	X
Italian	**87%**	59%
German	**63%**	57%
Arabic	**61%**	55%

The same tendencies among the results are observed for both French/L2 and Portuguese/L2 tests. Thus higher scores are obtained for language pairs with different stimuli in terms of vocalic qualities. Results comfort thus the hypothesis of a perceptual salience of the acoustic differences among the fillers of the eight considered languages.

However, other factors played a role in discrimination too. As a general observation, we can notice the higher results obtained for the test French/L2. Consequently, the scores obtained for the subtest French/Italian, German and Arabic are higher than those for Portuguese/Italian, German and Arabic, particularly for Italian, and above the chance level. We hypothesize that the global difference could be interpreted as the effect of the "mother tongue bias".
English has been less accurately discriminated from L1 than from L1', and Spanish has been better discriminated from L1 than Chinese. For both results, we hypothesize that a potential "first subtest" effect has been played a role. Indeed, listeners started the test French/L2 with the subtest French/English and the test Portuguese/L2 with the subtest Portuguese/Spanish. The loss of some points in the detection scores could be related to the familiarization with the type of experiment.

5. Conclusions

We described here two perceptual experiments conducted in order to evaluate the perceptual salience of the acoustic differences among multilingual fillers in eight languages. Two discrimination tests, French/L2 and Portuguese/L2, have been thus conducted with native French subjects. The experiments aimed at testing their capacity at differentiating languages from autonomous fillers without any lexical support and without any context. The present results sustain the hypothesis concerning the inter-language acoustic and perceptual differences of the vocalic supports of autonomous fillers. More precisely, the vocalic support does not correspond systematically to a "central" segment (i.e. schwa-type) and the acoustic variability across the eight languages helps the listeners to discriminate isolated fillers even if the languages are unknown. Further work will consider a larger corpus and more languages in order to validate this hypothesis.

6. Acknowledgements

This research has been carried out in the context of the MIDL project (Modélisations pour l'IDentification des Langues), supported by a CNRS interdisciplinary program and involving several French laboratories (LIMSI-CNRS, LTCI-ENST, CTA-DGA, ILPGA Paris3 and EA1483 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 speech processing methods.

The authors want to thank Cédric Gendrot (ILPGA, Paris 3) for his help in the acoustic analysis of the corpus.

7. References

[1] Clark H.H., Fox Tree J.E., Using uh and um in spontaneous speaking, *Cognition* 84, 73-111, 2002.

[2] Adda-Decker et al., A Disfluency study for cleaning spontaneous automatic transcripts and improving speech language models, DISS'03, Göteborg, Sweden (*Papers in Theoretical Linguistics* 90: 67-70), 2003.

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

[4] Guaïtella, I., Hésitations vocales en parole spontanée: réalisations acoustiques et fonctions rythmiques, *Travaux de l'Institut de Phonétique d'Aix*, vol.14: 113-130, 1991.

[5] Shriberg, E., Phonetic consequences of speech disfluency, *ICPhS'99*, San Francisco, 1999.

[6] Shriberg, E., The 'errrr' is human: ecology and acoustics of speech disfluencies, *Journal of the International Phonetic Association*, 31/1, 2001.

[7] Eklund R. editor, Disfluencies in Spontaneous Speech, DISS'03, Göteborg, Sweden (*Papers in Theoretical Linguistics 90*), 2003.

[8] Clerc-Renaud, J, Vasilescu, I. Candea, M., Adda-Decker, M., Etude acoustique et perceptive des hésitations autonomes multilingues, *XXV^{es} JEP*, Fès Morocco 2004.

[9] Vasilescu, I. Candea, M., Adda-Decker, M., Hésitations autonomes dans 8 langues : une étude acoustique et perceptive, *Workshop MIDL04*, Paris France, 2004.

[10] Candea, M., Vasilescu, I., Adda-Decker, M., Inter- and intra-language acoustic analysis of autonomous fillers, *Diss05*, Aix-en-Provence, France, 2005.