Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech

Ioana Vasilescu & Martine Adda-Decker

LIMSI-CNRS, Bât. 508, BP 133, 91403 Orsay cedex, France
ioana, madda@limsi.fr

10.21437/Interspeech.2006-510

Abstract

This paper deals with the factors characterizing the production of autonomous vocalic filled pauses in large spontaneous speech corpora, namely language, gender, speaking style and language proficiency. Two types of corpora are analyzed: a corpus of broadcast news in French and American English and a corpus of short talks in a conference in English spoken by native and non-native speakers. Several acoustic and prosodic parameters are evaluated and correlated with each factor, namely timbre, pitch, duration and density. Results presented here show that the timbre is correlated with language and language proficiency, whereas the duration is linked both to gender and speaking style, the latter conditioning also the hesitation density in speech.
Index terms: speech disfluencies, autonomous filled pauses, L1/L2, emotional state.

1. Introduction

This paper focuses on autonomous vocalic filled pauses in spontaneous speech corpora. Among the phenomena described as "disfluencies", filled pauses represent one of the most frequently encountered across languages. Autonomous vocalic hesitations as a type of filled pause are widely represented and consist in the insertion "at any moment" in the speech flow of a lengthened vocalic segment, alone or in combination with other segments (such as a nasal coda in English). Its aim is "to announce the initiation of what is expected to be a […] delay in speaking" [1]. Autonomous vocalic hesitations occur without lexical support and are thus to be distinguished from vocal lengthening of segments belonging to lexical items (generally function words). Filled pauses have however other possible realizations, as for instance lengthened nasal consonants ("mm" in Mandarin Chinese) or demonstratives ("ano", "eto" in Japanese) [2,3].
For the present study we consider vocalic hesitations in French ("euh") and English ("uh", "um" in American English; "er" in British English).

Previously autonomous vocalic hesitations have been studied in intra- and inter-language perspectives with no particular consideration of the role of the context on their acoustic and prosodic characteristics. In our former studies, we have compared autonomous vocalic hesitations in 8 languages: American English, Middle Oriental Arabic, Mandarin Chinese, French, Italian, South-American Spanish, and European Portuguese. We have focused on the support vowel of the hesitations in each considered language. The support vowel has been defined as the main vocalic segment of a hesitation, i.e. the longest and most stable realization of each item. This vowel occurs in isolation (as unique realization of the hesitation), in a diphthong or followed by a nasal consonant as in English. Among the parameters characterizing the support vowel, duration, pitch and timbre have received a particular attention. Analysis revealed that the timbre is the most language-dependent parameter characterizing vocalic hesitations. Pitch and duration help both at differentiating the hesitation vowel from vowels with similar timbre within a given language. Pitch and duration seem to show universal patterns, i.e. the main vowel of a hesitation is significantly longer than other similar intra-lexical vowels and exhibits a flat and stable F0 contour [4]. Consequently, the hypothesis has been made that timbre is a language-dependent parameter, whereas pitch and duration could be considered as language-independent features.

In this study, we consider 4 factors which may play a role in the production of vocalic hesitation in spontaneous speech corpora: language; gender; spoken style and language proficiency (mother tongue vs. second language).

2. Corpus and methodology

For the present study we make use of 2 corpora: a corpus of broadcast news in American English and in French and a corpus of talks recorded in a conference, in English.
The broadcast news corpora in American English (hereafter, BNE) and French (BNF) consist in recordings of native speakers of the 2 languages, both male and female. BNE consists in the recordings of 6 American English broadcast channels (such as CNN, VOA, ABC, etc.) and 150 different speakers (about 100 male speakers and 50 female speakers). French resources in BNF are represented by 4 broadcast channels (France Inter, France Info, France2, and France3) and about 130 speakers (100 male, 30 female). The total duration of BNF is averaging 200 hours vs. 100 hours for the BNE corpus.

The corpus of talks in a conference is "Translanguage English Database" (hereafter, TED) consisting in recordings of 10 minutes talks of native and non-native speakers of English in the Eurospeech conference in 1993. The total duration of the corpus used here is averaging 1.5 hours. For this preliminary study, we selected 8 French speakers (hereafter, TEDF) and 3 English speakers (TEDE). All speakers are giving their talks in English. So far we do not consider female speakers.

Hesitations have been extracted automatically and manually verified in order to avoid potential selection errors, i.e.

hesitations surrounded by non-verbal events such as laughs, mouth noises, music, telephone ring etc. In former studies, we made use of a duration threshold for the automatic selection of hesitations, i.e. higher than 200ms [4,5]. The threshold allowed to exclude word-final schwa vowels in French or other short noise phenomena which could have been automatically aligned with the hesitation model. In this study, all hesitation segments have been considered without applying a minimum duration threshold. This choice has been done in order to obtain a realistic estimation of the duration density of hesitations.

Table 1. *Number of hesitations for French and English speakers in native language (L1) or second language (L2).*

Corp./Loc.	**French**	**English**
BN	L1: 1640 (m)/270 (f)	L1:4455 (m)/491(f)
TED	L2:762 (m)	L1:439 (m)

We focus here on parameters usually considered as characterizing autonomous vocalic hesitations: timbre (F1/F2), pitch (F0), duration and density (measured as ratio of total duration of hesitation events from the total duration of the corpus).

3. *Language* and *gender* factors

Previous studies have shown that *timbre* represents a language-dependent feature, whereas duration and pitch reveal universally observed patterns. Hesitations extracted and analyzed here from broadcast news data in English and French confirm those findings. Timbre is estimated via the distribution of the two vowels in a F1/F2 space. The comparison of the timbres of the support vowels in French and in English confirms inter-language differences. The hesitation vowel in French is significantly less open and less fronted than its English counterpart (independent t-test, $p<0.0001$). Differences concern both male and female productions.

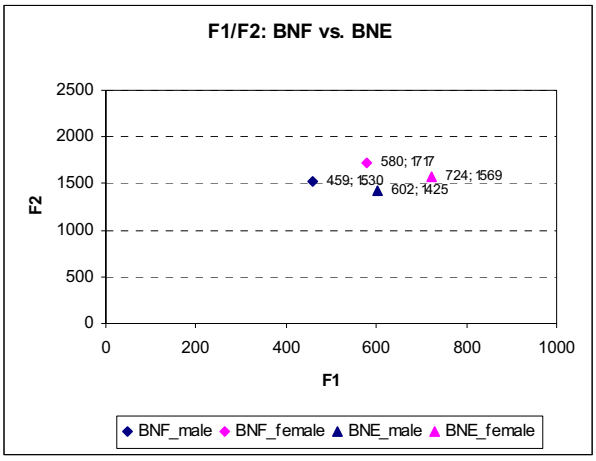

Figure 1 *Mean value distribution for F1 and F2 of hesitation vowels in French and American English (male and female speakers).*

The analysis of *pitch* patterns (F0) does not show differences according to language. F0 values are ranging similarly for French and American English speakers and more particularly for male speakers (Figure 2). Concerning the female speakers, Figure 2 highlights that the range is more important for French female speakers and with more extreme values, which could be a language specific feature. However, this observation needs to be considered cautiously, as the difference could be due to a potential corpus effect.

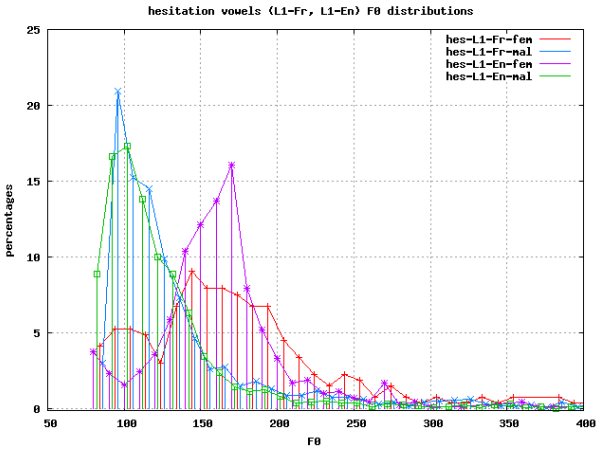

Figure 2 *F0 values distribution for hesitation vowels in BN corpora (French and American English, male and female speakers).*

We also measured the temporal evolution of the vocalic hesitations, i.e. the pitch trajectory in all the corpora (BN and TED). It appears that globally, the pitch contour is stable and describes a falling pattern (about 75% of hesitations are showing this pattern). The falling pattern is language and gender-independent and seems to be a universal one (it is encountered across data from all the corpora analyzed here, i.e. broadcast news but also conference talks).

The third considered parameter is *duration*. In this study we selected hesitations without considering a duration threshold. In a former study on French language, we found that about 25% of fillers with duration lower than 200ms have been eliminated by making use of this threshold as a selection criterion [5].
Although data analyzed here confirm former trends, i.e. the hesitation vowel is significantly longer than an intra-lexical vowel (i.e. about 80ms mean duration), they highlight other interesting characteristics. Concerning the *language/gender* factor, data show that there are differences in duration. Hesitations are significantly longer in French spoken by male than in American English (ANOVA, $F=237.102$, $p<0.0001$). However, as this observation concerns only male speakers, it might be related to corpus specificities, thus it needs to better be more extensively investigated in further studies.

Table 2. *Mean and median duration values for hesitations in BN corpora (French and American English, male and female speakers).*

Corpus/gender (ms)	**Male (mean/median)**	**Female (mean/median)**
BNF	343 / 300	262 / 220
BNE	267 / 230	266 / 230

The last parameter considered in terms of correlation with factors *language* and *gender* is the *segmental structure* of the hesitations. French language exhibits a unique widely employed hesitation, "euh", thus a lengthened central vowel, whereas in American English two types of hesitations are frequently encountered (though in different discourse contexts [1]), "uh" and "um". The American English hesitation vowel is more open and more fronted than the French one, and may be followed or not by a nasal coda. Important differences are noticed when considering the ratio of hesitations according to the presence of the nasal coda. The BNE corpus shows that realizations with a nasal coda are less frequent than the vocalic ones, as only 23% of hesitations could be transcribed as "um". Besides, those realizations are more frequent in female (45%) than in male speech (19%) and the difference is statistically significant. This difference suggests that the segmental structure of hesitations in American English follow some gender specific trends. Finally, the acoustic segmental analysis of the support vowels does not reveal important intra-language timbre variability. The acoustic analysis of F1 and F2 evolution has shown that support vowels are globally stable. If diphthongs occur among hesitations, they are quantitatively non significant neither in French, nor in American English.

As a preliminary conclusion, language and gender factors estimation shows that F0 values and trajectories, duration of hesitations vs. of intra-lexical vowels and F1/F2 evolution follow universal trends, whereas timbre and segmental structure are language-dependent. Finally, segmental structure shows different patterns according to gender in American English.

4. *Speaking style* factor

In order to evaluate the influence of the *speaking style* on vocalic hesitations, we compare here two types of corpora obtained in two different conditions, i.e. broadcast news (BN) vs. conference talks (TED). We contrast thus hesitations produced in two speech conditions which potentially influence their acoustic and prosodic characteristics as well as their frequency in the speech corpora.

BN corpora illustrate the news style, i.e. semi-prepared speech, given in a limited amount of time by professional speakers accustomed to read or relate events in radio/television shows. TED corpus is made of short oral presentations of researchers in speech and language processing given in a conference. They talk to a public of colleagues who potentially evaluate and judge their work. We can hypothesize that conference presentations condition might be a more stressful one, and in addition, the stress might be higher when the talks are given in a foreign language (L2). This is the case for 8 of the 11 male speakers analyzed here (8 French natives, 3 British English natives). This change in the emotional state of the speaker (i.e. stress) might influence the disfluency production in speech.

In order to compare broadcast news vs. conference presentation hesitations, we consider here only items produced by male speakers from BN data.

The *density* of hesitations appears to be influenced by the factor *speaking style*. The total duration of hesitations represents 0.7% from BNE and 0.1% from BNF, whereas in TED, native English speakers employ 5.8% of their 10 minutes talk for hesitations vs. 5.7% for their native French colleagues speaking in L2. These findings seem to comfort the hypothesis that conference represents a more stressful speech condition which influences the amount of disfluencies present in speech. However, inside TED corpus we did not notice differences related to the expression in L1 vs. L2. For this preliminary data, it appears that the expression in L2 does not increase the ratio of disfluencies phenomena and thus stress seems to be, if not the unique, at least the major explanation of the inter corpora dissimilarities.

Concerning *pitch* (F0), we have compared mean values for F0 in different corpora. An ANOVA analysis shows an important "corpus" effect (ANOVA, $f=27.978$, $p<0.001$). This effect could be correlated with the *speaking style,* however as data are less important for the conference presentation condition, this hypothesis has to be considered carefully (Table 3).

Table 3. *Mean and standard deviation for F0 in BN and TED corpora (male speakers)*

Corpus/F0_Mean (Hz)	**F0_Mean**	**STDEV**
BNF	142	86
BNE	129	71
TEDF	129	42
TEDE	141	67

We also measured the "speaker" effect in TED corpus, as each speaker presentation is well delimited and labeled. An ANOVA statistical analysis shows a "speaker" effect in both TEDE and TEDF corpora (TEDF: ANOVA, $F=9.7111$, $p<0.0001$; TEDE: ANOVA, $F=23.151$, $p<0.0001$). However as only 11 speakers have been analyzed here, more data are necessary to validate those findings and to better define the role of stress in the production of speech disfluencies.

Finally, the hypothesis of the role of stress in the hesitation production may possibly be linked to the patterns observed in *duration*. D*uration* is a parameter correlated with the *speaking style* factor. More precisely, hesitations in conference talks (TED) are significantly longer than hesitations in broadcast news (BN). These findings concern only male speakers in both corpora and are language-independent (ANOVA, $F=268.897$, $p<0.0001$) (Table 4).

Table 4. *Mean and median duration for vocalic hesitations in BN and TED corpora (male speakers).*

Corpus/Duration (ms)	**Mean**	**Median**
BNF	342	300
BNE	266	230
TEDE	429	420
TEDF	415	380

Consequently, the analysis of acoustic and prosodic parameters according to speaking style highlights the stress effect on hesitations production. Stress seems to particularly influence duration and density, and potentially pitch.

5. Factor *language proficiency*

We analyze here the *timbre* of hesitations produced by French speakers within TED in English as a measure of their *proficiency* in L2. The timbre of L2 hesitations is compared with the timbre of hesitations in L1 French and English. Figure 4 below illustrates mean values of F1 and F2 for support vowels of hesitations in BN corpora and for each of the 11 speakers analyzed in TED.

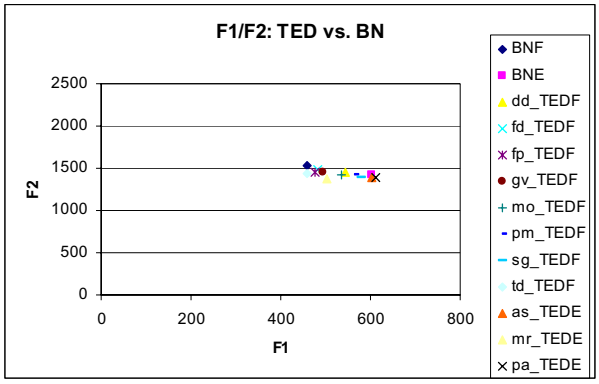

Figure 3 *Mean values distribution for F1 and F2 of hesitation vowels in BN and TED corpora (BNE, BNF: mean value per corpus; TEDF mean value per speaker: dd, fd, fp, gv, mo, pm, sg, td; TEDE mean value per speaker: as, mr, pa).*

Figure 4 shows that L2 hesitation vowels correspond mainly to intermediary values between the L1 French (native language of the speakers) and the L2 English vowels spoken by natives (both TEDE and BNE). This distribution concerns the F1 axis, we did not notice significant differences on the front/back axis. In terms of degree of opening (F1), some of the French speakers produce the "euh" hesitations of their mother tongue when speaking in English, others produce "uh"/"um" hesitations as in English, but most of them hesitate with intermediate realizations. The present findings are particularly interesting as they allow at estimating language proficiency via a particular speech phenomenon, i.e. vocalic hesitations. However, in order to validate the hypothesis of a gradual proficiency illustrated by intermediary realizations of hesitation vowels, more data are needed for a further analysis.

6. Discussion

Speech disfluencies and in particular hesitations have received a particular focus in recent studies. Attention has been paid particularly to their acoustic and prosodic characteristics as well as their role and distribution in the discourse.
In this study, we considered the acoustic and prosodic characteristics of autonomous vocalic fillers in relationship with factors characterizing speech corpora. Factors *language*, *gender*, *speaking style* and *language proficiency* have been evaluated via acoustic and prosodic parameters such as timbre, duration, pitch and density of autonomous vocalic hesitations.

The *language* factor confirmed previous remarks, i.e. the most language-dependent parameter is the timbre. *Duration* can be related mainly to the *speaking style*, but also, to a lesser extent, to *language* and *gender* factors. *Speaking style* is also characterized by the density of the analyzed phenomenon, as a stressed emotional state results in an increase of disfluency productions. Finally, language proficiency has been evaluated via the timbre pattern in L1 French and English vs. L2 English. For most of the considered speakers we found that hesitations in L2 represent intermediary realizations between French and English as L1. These latter observations leave open interesting questions concerning the role and the modalities of acquisition of disfluencies in L2.

7. References

[1] Clark H.H., Fox Tree J.E., "Using uh and um in spontaneous speaking", Cognition 84, 73-111, 2002.
[2] Zhao, Y., Jurafsky, D., A preliminary study of Mandarin filled pauses, In *DiSS-2005*, 179-182, 2005.
[3] Watanabe, M., Den, Y., Hirose, K., Minematsu, N., "The effects of filled pauses on native and non-native listeners speech processing", In *DiSS-2005*, 169-172, 2005.
[4] Shriberg, E., "The 'errrr' is human: ecology and acoustics of speech disfluencies", Journal of the International Phonetic Association, 31/1, 2001.
[5] Candea, M., Vasilescu, I., Adda-Decker, M., Inter- and intra-language acoustic analysis of autonomous fillers, In *DiSS-2005*, 173-176, 2005.