THE SPEAKER-SPECIFICITY OF FILLED PAUSES: A CROSS-LINGUISTIC STUDY

Meike de Boer & Willemijn Heeren

Leiden University Centre for Linguistics, Leiden University, The Netherlands
m.m.de.boer@hum.leidenuniv.nl; w.f.l.heeren@hum.leidenuniv.nl

ABSTRACT

We investigated the speaker-specificity of filled pauses across languages and time. The filled pauses *uh* and *um* contain speaker-specific information in a speaker's native language. Since speakers are relatively unaware of their hesitation behavior, it might transfer from their first (L1) to their second language (L2). We examined filled pauses using several phonetic-acoustic features in spontaneous L1 Dutch and L2 English speech of 20 female speakers, recorded at two times, three years apart.

Using linear mixed-effects models, we found that speakers differ in their first and second formants of *uh* and *um* in L1 versus L2, while duration and fundamental frequency remain stable. Speaker classification models trained on filled pauses in one language perform worse – but still relatively well – on the other language. With the exception of a few speakers, hesitation behavior remained stable over time. In spite of L1-L2 differences, some speaker characteristics of filled pauses remain.

Keywords: forensic phonetics, hesitation markers, speaker-specificity, second language acquisition

1. INTRODUCTION

Individual speakers have distinct and consistent patterns in their usage of the filled pauses *uh* and *um* [e.g. 3, 9, 11, 14]. Clark and Fox Tree [4] mention that "speakers of English as a second language often import fillers from their first language" (p. 93). However, empirical research investigating this claim is limited. Hence, this study addresses the following research questions:

1. Are speakers consistent in their hesitation behavior across languages?
2. How does hesitation behavior change with second language development?

We study these questions in a population of Dutch students living in a multilingual community with English as the lingua franca, whose proficiency in English is relatively high [17]. We used recordings made during the first and last semester of a three-year period. Prior studies show that these students converge towards a shared English accent over time [18, 19].

Based on the literature, different expectations arise. Some findings predict that hesitation behavior could be consistent across languages, as suggested by [4], while other findings predict that speakers adapt their hesitation behavior when speaking in another language. [10] demonstrates that the number and duration of silent pauses remain stable in speakers' first (L1) and second (L2) language. Filled pauses might behave similarly. However, research on language fluency shows that while L2 speakers do not use more silent pauses than L1 speakers, they do use more *filled* pauses [5]. This implies that the number of filled pauses may differ across one's languages. Research on fluency also shows that L2 learners decrease their use of filled pauses over time [12, 16]. This suggests that the speakers in our study might use fewer filled pauses in their L2 than less advanced L2 learners, decreasing the variation in the use of filled pauses across their L1 and L2. Three years in an English-speaking environment might decrease between-language differences even more.

Regarding the phonetic realization of filled pauses, Flege's Speech Learning Model (SLM) [7] predicts that L2 learners only adapt their pronunciation to a more native-like one if they perceive that the sound is different from a sound in their L1. Since the vowels in filled pauses are realized quite similarly in Dutch and English [6, 20] and filled pauses are a relatively unconscious part of language [4], the SLM predicts that Dutch L1 speakers do not adapt their vowel realizations when speaking English. However, the more proficient L2 learners become, the easier they are able to perceive subtle sound differences, and the more likely they are to adopt a more native-like pronunciation [19]. A more noticeable feature to L2 learners is the preference in English to end filled pauses with a bilabial nasal (*um*), while in Dutch, *uh* is preferred [6]. According to the SLM, speakers should therefore be more prone to show between-language differences in their *um:uh* proportions than in their realizations of the vowel. For both *um:uh* proportion and vowel realization, we expect between-language differences to increase over time.

Overall, we expect that speakers are consistent in their use of filled pauses across languages on number per minute and duration, but show changes in their spectral realizations.

2. MATERIALS AND METHOD

2.1. Speakers

We selected 20 female speakers with Standard Dutch as L1 from the Longitudinal Corpus of University College English Accents (LUCEA), collected in 2010-2013 by Orr and Quené [17]. The speakers were students from University College Utrecht, who were recorded on multiple occasions over the course of three years. During this time, the students lived on campus – a multilingual community with English as the lingua franca. University Colleges select their students based on English language proficiency, which has to be at a level similar to B1 in the Common European Framework of Reference. The selected speakers thus form a relatively homogeneous group in terms of age, gender, education level, L1 background, L2 proficiency, and linguistic environment.

2.2. Recordings

The LUCEA corpus consists of multiple speaking tasks [see 17], of which we selected a task where the students were asked to speak about an informal topic for two minutes; first in L1 Dutch and then in L2 English. The order of the languages was not counterbalanced. This could have caused practice effects in the data [8], since some students talked about the same topics in both languages. In such cases, speakers are expected to use fewer filled pauses in the L2 than without practice.

The students were recorded five times over the course of three years [17]. In this study, only the first and fifth recordings of the selected speakers were used, since these recordings are expected to show the largest development over time.

The recordings were made in a quiet furnished room with eight different microphones [see 17]. We used the recordings of the close-talking headset, to keep a consistent distance to the speaker's mouth.

2.3. Segmentation and measurements

The filled pauses *uh* and *um* were segmented manually in Praat [2], separating the vowel and optional nasal part. The total number of filled pauses was 1,472, of which 826 (56%) were *uh*. Each speaker contributed 74 filled pauses on average (ranging from 25 to 121).

The following measurements were taken:

- the duration of the filled pause incl. or excl. the optional nasal (in ms);
- the mean fundamental frequency (F0) of the filled pause over the middle 50% (in Hz);
- the mean first, second, and third formant (F1, F2, F3) of the filled pause over the middle 50% of the vowel (in Hz);
- the number of filled pauses per minute;
- the *um:uh* proportions.

Spectral measurements were performed in Praat. Measurement errors (max. 6% per feature), as well as outliers (max. 3% per feature), were excluded. Outliers were determined by visual inspection of the histograms against a normal distribution. For the analysis of F0, low values indicating creak on the vowel (1%) were also excluded.

2.4. Statistical analyses

Linear mixed-effects models were used to investigate the effects of the fixed factors Language (Dutch, English) and Time (recording 1, 5) on the acoustic measurements: duration, F0, and F1~3. For modeling, the *lmer*() function from the *lme4* package [1] was used. Significance was evaluated through likelihood ratio testing with stepwise inclusion of predictors. In the random part of the model, in addition to by-speaker intercepts, the effect of maximization of the random structure on model fit was evaluated for each final model. Modeling was done separately for *uh* and *um*. As reference levels, Dutch and recording 1 were used (treatment-coding).

Number of filled pauses per minute was calculated using recording duration, and compared between languages and recordings using log-linear regression via the *glmer*() function from the *lme4* package [1]. The *um:uh* proportions were analyzed using logistic regression through the same function.

Finally, linear discriminant analysis (LDA) was used to evaluate speaker classification performance under different conditions, based on the language spoken and the moment of recording. As predictors, the acoustic measurements duration, F0, and F1~3 were entered. For the LDAs, the filled pauses *uh* and *um* were analyzed together to increase the number of instances per speaker. Therefore, instead of duration of the entire filled pause, the more comparable measurement vowel duration was used. Vowel duration was log-transformed to better meet the normality criterion. The classifications were cross-validated using leave-one-out validation, and the model's structure coefficients were examined to find the predictors contributing most to the classification outcome. Speakers with fewer than six filled pauses in a certain language at a certain time were excluded from the analysis. This led to the exclusion of one to three speakers per LDA (5−15%).

3. RESULTS

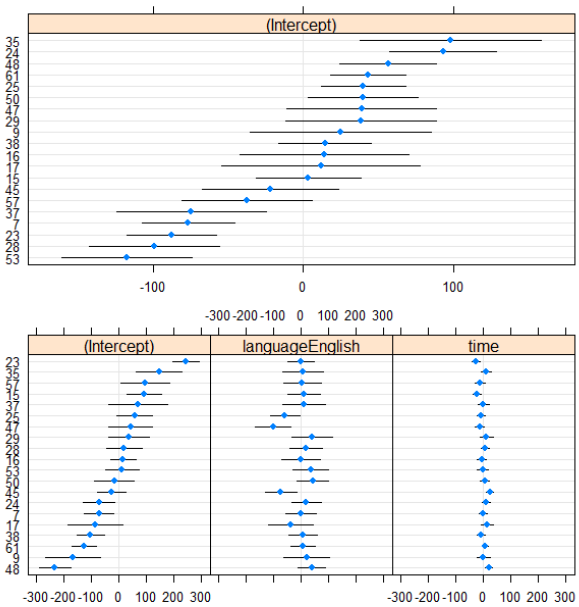
3.1. Modeling of acoustic features

The intercept for *uh* duration was 300 ms and 492 ms for *um*. Adding the predictors Time or Language did not improve the *uh* or *um* models ($\chi^2(1) \leq 1.08$, $p \geq .30$). By-speaker random intercepts showed that speakers varied in their durations of *uh* (range: −57 – 90 ms) and *um* (range: −100 – 209 ms).

The F0 intercept was 187 Hz, both for *uh* and *um*. For both hesitation markers, adding the predictors Time or Language did not improve the models ($\chi^2(1) \leq 0.83$, $p \geq .36$). By-speaker random intercepts showed that speakers differed somewhat in F0 of both *uh* (range: −27 – 18 Hz) and *um* (range: −27 – 17 Hz).

The intercepts for F1 of the *uh* and *um* vowels were 604 Hz and 629 Hz, respectively. The optimal models included the fixed factors Language ($\chi^2(2) \geq 29.0$, $p < .001$) and Time ($\chi^2(2) \geq 4.1$, $p < .05$), by-speaker intercepts and for *um* by-speaker slopes for Time. In English, speakers' F1 of *uh* was on average 37 Hz higher than in Dutch (SE = 6.0, t = 6.2), and of the *um* vowel 39 Hz higher (SE = 6.5, t = 6.0). At time 5, speakers' F1 of *uh* was on average 4 Hz higher than at time 1 (SE = 1.5, t = 2.4). For *um*, the increase in F1 over time was not significant.

Figure 1: Caterpillar plots of by-speaker random intercepts (F1 of *uh*; upper) and by-speaker random intercepts plus by-speaker slopes for Language and Time (F2 of *uh*; lower).



The intercepts for F2 of the vowels in *uh* and *um* were 1,648 Hz and 1,633 Hz, respectively. For both hesitation markers, the optimal models included the fixed factor Language ($\chi^2(2) \geq 14.9$, $p < .001$), by-speaker intercepts and by-speaker slopes for Language and Time. The speakers' F2 was 43 Hz lower in English for the *uh* vowel (SE = 14.7, t = −2.9) and 41 Hz lower for the *um* vowel (SE = 15.6, t = −2.7). Figure 1 illustrates that individual speakers varied in their F1 and F2 random intercepts of *uh*.

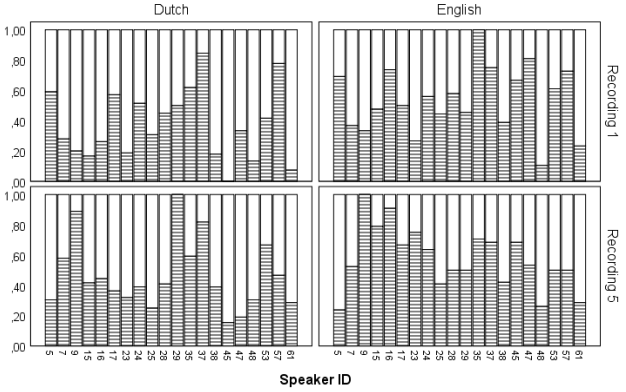
The intercepts for F3 of the *uh* and *um* vowels were 2,717 Hz and 2,725 Hz, respectively. For the *uh* vowel, the optimal model included the fixed factor Time ($\chi^2(1) = 5.8$, $p < .05$), by-speaker random intercepts and by-speaker slopes for Time. After three years, the speakers' F3 was on average 7 Hz higher, but this was not significant (SE = 5.0, t = 1.5). For the *um* vowel, adding the predictors Time or Language did not improve the model ($\chi^2(1) \leq 0.79$, $p \geq .38$). By-speaker random intercepts showed between-speaker variation in F3 of both *uh* (range: −217 – 308 Hz) and *um* (range: −327 – 288 Hz).

3.2. Modeling of count features

The final model's intercept showed that speakers used 8.7 filled pauses per minute. The model included the factor Language ($\chi^2(1) = 4.1$, $p < .05$), reflecting that speakers used less filled pauses in English (7.4/minute). By-speaker random intercepts showed speaker variation between 5.0 and 12.7 filled pauses per minute.

The back-transformed intercept for the *um:uh* proportion was 0.76. The optimal model also included the fixed factor Language ($\chi^2(1) = 40.2$, $p < .001$), altering the *um:uh* proportions to 1.10 for English. In the random part, by-speaker intercepts and by-speaker slopes for Language were included. Figure 2 shows the by-speaker *um:uh* proportions.

Figure 2: Proportions of *um* and *uh* by speaker, per language and per recording time.



3.3. Reflection on mixed-effects modeling

The results of the mixed-effects models show that some features remained stable across languages (i.e. number, duration, F0, F3), whereas others varied by language (i.e. F1, F2, *um:uh* proportions).

The acoustic parameters were used for speaker classification to assess how hesitation markers may contribute to between-language (forensic) speaker comparisons. The vowel formants were expected to contribute most to the models because of their high speaker-specificity [e.g. 13, 15]. Since time differences seemed minimal, excepting results from a few speakers (see section 3.1), we predicted that a model built on data from recording 1 would perform quite well on recording 5. Because language effects were more prominent, we expected a model trained on one language to perform worse on the other.

3.3. Linear Discriminant Analyses

Models built on either Dutch or English at one moment in time performed worse on the other language recorded at the same time (see table 1). When Dutch data from recording time 1 (T1) were used for training, cross-validated speaker classification performance on Dutch data from T1 was 44% correct, whereas performance on T1 English data was 38% (chance level = 5%). When training on T1 English data, performance was 46% correct on English and 31% on Dutch. Comparable results were obtained with data from T5.

Table 1: Speaker classification performance of LDA models (trained on one language at T1 or T5) on the same language and on the other language at that time.

	Trained on Dutch:		Trained on English:	
	(T1)	(T5)	(T1)	(T5)
Dutch	44%	44%	31%	31%
English	38%	29%	46%	45%

Models built on data from either T1 or T5 performed worse on within-language data from the other time (see table 2). For instance, training on Dutch T1 data gave 47% correct classification on Dutch at T1, but 25% at T5. Results on the other within-language comparisons showed the same advantage of training data Time.

Table 2: Speaker classification performance of LDA models (trained on T1 or T5 in one language) on the same time and on the other time in that language.

	Trained on time 1:		Trained on time 5:	
	(NL)	(EN)	(NL)	(EN)
Time 1	47%	48%	26%	29%
Time 5	25%	23%	43%	45%

In both types of models, the formants (F1~3) carried most weight, and to a lesser extent F0. Duration had a minimal contribution.

4. DISCUSSION AND CONLUSION

We assessed whether speakers are consistent in their hesitation behavior across languages, and how this develops over time. Results showed that filled pauses in speakers' L1 and L2 differed in a number of features. Over time, their filled pauses in either language showed only small changes.

When talking English, the Dutch altered the F1 and F2 of their filled pauses. The vowels were pronounced more open and more back in English than in Dutch. Also, the speakers used *um* relatively more often than *uh* in English, as native English speakers do [6]. Apparently, the students' L2 proficiency was high enough to perceive differences between Dutch and English filled pauses and use them in their L2 speech, already at T1.

Other features of hesitation behavior remained consistent in L1 and L2. The consistency in duration of filled pauses is in line with the findings of [10] on silent pauses. The finding that the speakers did not use more filled pauses in their L2, as predicted by [5], can be explained in two ways. Firstly, this can be explained by the speakers' high L2 proficiency [12, 16]. Secondly, practice effects could have caused a lower number of filled pauses in the second speaking task, which was in the L2. Overall, speakers did not alter their F0 and F3 when speaking in the L2. Whereas F1 and F2 are highly dependent on the nature of the target sound, F0 and F3 are less dependent on the target [13, 15]. Filled pauses were mostly consistent over time, but some speakers' F1 and F3 changed within both languages.

The LDAs showed that the formants (F1~3) were the best-performing features in speaker classification models, which is in line with previous findings [9, 13, 15]. When comparing classification performance across languages and across time, we found that the speakers made most adaptations on exactly these features, especially on F1 and F2. The formants' instability over languages and time forms a challenge for speaker-specificity, as it causes lower within-speaker consistency.

Speakers are not fully consistent in their use of filled pauses across languages, but partially adapt to the characteristics of the L2. Still, some speaker-dependent information remains in filled pauses. Future research on a larger speaker set will extend these results using a likelihood ratio approach.

5. ACKNOWLEDGEMENTS

This research was supported by a VIDI grant from The Netherlands Organisation for Scientific research. Thanks to J. van der Graaf and Y. Sleeboom for their help in the annotations.

7. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects: Models Using lme4. *Journal of Statistical Software* 67, 1−48.

[2] Boersma, P. 2001. Praat: doing phonetics by computer [computer program]. Version 6.0.37, retrieved 3 July 2018 from http://www.praat.org/.

[3] Braun, A., Rosin, A. 2015. On the speaker-specificity of hesitation markers. *Proc. 18th ICPhS* Glasgow, 731−736.

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

[5] De Jong, N. H. 2016. Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *Int. Rev. Appl. Ling. Lang. Teaching* 54, 113−132.

[6] De Leeuw, E. 2007. Hesitation markers in English, German, and Dutch. *J. Germ. Ling.* 19, 85–114.

[7] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* York: York Press, 233−277.

[8] Gaito, J. 1961. Repeated measurements designs and counterbalancing. *Psychological Bulletin* 58, 46-54.

[9] Hughes, V., Wood, S., Foulkes, P. 2016. Filled pauses as variables in forensic voice comparison. *Int. J. Speech Lang. Law* 23, 99−132.

[10] Kolly, M. J., Leemann, A., De Mareüil, P. B., Dellwo, V. 2015. Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study. *Proc. 18th ICPhS* Glasgow, 294−299.

[11] Künzel, H. F. 1997. Some general phonetic and forensic aspects of speaking tempo. *For. Linguist.* 4, 48−83.

[12] Lennon, P. 1990. Investigating fluency in EFL: A quantitative approach. *Lang. Learning* 3, 387−417.

[13] McDougall, K. 2006. Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *Int. J. Speech Lang. Law* 13, 89−126.

[14] McDougall, K., Duckworth, M. 2017. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Comm.* 95, 16−27.

[15] Moos, A. 2010. Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician* 101, 7-24.

[16] O'Brien, I., Segalowitz, N., Freed, B., Collentine, J. 2007. Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition* 29, 557−582.

[17] Orr, R., Quené, H. 2017. D-LUCEA: Curation of the UCU Accent Project data. In: Odijk, J., Van Hessen, A. (eds), *CLARIN in the Low Countries.* London: Ubiquity Press, 177–190.

[18] Quené, H., Orr, R. 2014. Long-term convergence of speech rhythm in L1 and L2 English. *Social and Linguistic Speech Prosody* 7, 342–345.

[19] Quené, H., Orr, R., Van Leeuwen, D. 2017. Phonetic similarity of /s/ in native and second language: Individual differences in learning curves. *J. Acoust. Soc. Am.* 142, 519–524.

[20] Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., Liberman, M. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change* 6, 199−234.