Cross-linguistic filled pause realization: The acoustics of *uh* and *um* in native Dutch and non-native English

Meike M. de Boer, and Willemijn F. L. Heeren

Citation: The Journal of the Acoustical Society of America **148**, 3612 (2020); doi: 10.1121/10.0002871

View online: https://doi.org/10.1121/10.0002871

View Table of Contents: https://asa.scitation.org/toc/jas/148/6

Published by the Acoustical Society of America

JASA ARTICLE

Cross-linguistic filled pause realization: The acoustics of *uh* and *um* in native Dutch and non-native English

Meike M. de Boer[a] and Willemijn F. L. Heeren[b]
Leiden University Centre for Linguistics, Leiden University, Reuvensplaats 3-4, 2311 BE Leiden, The Netherlands

ABSTRACT:
It has been claimed that filled pauses are transferred from the first (L1) into the second language (L2), suggesting that they are not directly learned by L2 speakers. This would make them usable for cross-linguistic forensic speaker comparisons. However, under the alternative hypothesis that vowels in the L2 are learnable, L2 speakers adapt their pronunciation. This study investigated whether individuals remain consistent in their filled pause realization across languages, by comparing filled pauses (*uh*, *um*) in L1 Dutch and L2 English by 58 females. Next to the effect of language, effects of the filled pauses' position in the utterance were considered, as these are expected to affect acoustics and also relate to fluency. Mixed-effects models showed that, whereas duration and fundamental frequency remained similar across languages, vowel realization was language-dependent. Speakers used *um* relatively more often in English than Dutch, whereas previous research described speakers to be consistent in their *um:uh* ratio across languages. Results furthermore showed that filled-pause acoustics in the L1 and L2 depend on the position in the utterance. Because filled pause realization is partially adapted to the L2, their use as a feature for cross-linguistic forensic speaker comparisons may be restricted. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0002871

(Received 24 March 2020; revised 10 November 2020; accepted 13 November 2020; published online 11 December 2020)
[Editor: Ewa Jacewicz] Pages: 3612–3622

I. INTRODUCTION

While the majority of the world's population speaks more than one language (Bhatia and Ritchie, 2012), most research in the domain of forensic speech science has been carried out in a monolingual context (cf. Mok *et al.*, 2015). At the same time, in criminal cases, it is not uncommon to find speech samples in multiple languages—sometimes even within one recording (van der Vloed *et al.*, 2014). Therefore, there is a need to explore the existence of acoustic-phonetic features that are stable across languages and thus may be used in the future to perform cross-linguistic forensic speaker comparisons.

There are indications that filled pauses, such as *uh* and *um*, may be such a feature. In the native language, filled pauses have been found to be highly speaker-specific, with large variation between speakers and low variation within speakers (e.g., Hughes *et al.*, 2016; Künzel, 1997). Moreover, filled pauses are produced relatively unconsciously and are therefore claimed to be less subject to deliberate disguise (Hughes *et al.*, 2016; McDougall and Duckworth, 2017). Also, earlier work claims that they are transferred from the first language (L1) into the second language (L2; Clark and Fox Tree, 2002; de Leeuw, 2007). If filled pauses are indeed consistent across languages within speakers, this would make them suitable for cross-linguistic speaker discrimination (cf. Wong and Papp, 2018). However, empirical support for the L1-transfer hypothesis is very limited, whereas L2 acquisition theories predict phonetic-acoustic shifts in a speaker's productions when speaking in an L2 (e.g., Flege, 1995). Therefore, as a first step toward assessing the contribution of filled pauses to cross-linguistic speaker comparisons, the current study tested the transfer versus learning hypotheses by comparing filled pause productions in proficient L2 English speakers with their own Dutch L1 realizations.

A. Filled pauses in the second language

Numerous studies on L1 speech found that speakers are reasonably consistent in their filled pause production. First, speakers are consistent in the number of filled pauses they produce (Künzel, 1997), even across different speech tasks (Goldman-Eisler, 1961). Second, speakers are consistent in their choice of either *uh* or *um*, in the proportion between filled and silent pauses, and in the extent to which they use alternative pausing strategies such as word-final lengthening (Künzel, 1997; McDougall and Duckworth, 2017). This consistency in disfluency preferences holds across non-contemporaneous sessions (Braun and Rosin, 2015). Third, speakers are rather consistent in the spectral realization of the vowels in *uh* and *um* (Hughes *et al.*, 2016; Künzel, 1997), partly because filled pauses are often surrounded by silences and thereby less susceptible to coarticulation (Swerts, 1998). Finally, because of their relatively unconscious nature, filled pauses are thought to remain consistent even when a speaker disguises their voice (Hughes *et al.*, 2016; McDougall and Duckworth, 2017).

[a] Electronic mail: m.m.de.boer@hum.leidenuniv.nl, ORCID: 0000-0003-0161-9115.
[b] ORCID: 0000-0001-7124-027X.

According to Clark and Fox Tree (2002), within-speaker consistency also holds across languages: based on their experience with L2 speakers of English, they claim that L2 speakers can often be identified as such because they transfer their filled pauses from the L1 (see also de Leeuw, 2007). However, there is little acoustic evidence for this claim, because cross-linguistic studies examining filled pauses are rare. One study with 15 German-French simultaneous bilinguals showed that they developed language-specific filled pauses (Lo, 2019). In today's society, many speakers are sequential bilinguals. They for instance acquire English as a second language from an early age at school. In addition, in Western European countries such as The Netherlands, children receive a high amount of English language input through pop culture, e.g., music and television (Smakman and De France, 2014). The question remains whether these sequential bilinguals adapt their filled pause realizations to language-specific norms similarly to simultaneous bilinguals.

Studies measuring characteristics of filled pauses in L2 speech are scarce, and so far, small-scale and heterogeneous in terms of age and gender. Vasilescu and Adda-Decker (2006) studied filled pauses in the L2 English of eight native French speakers and found that their vowel realizations—especially F1—showed intermediate values between those of L1 English speakers and other L1 French speakers. There was high variability between the speakers, with some producing native-like English filled pauses. However, the French participants' L1 was not recorded, thus the study did not allow for cross-linguistic, within-speakers comparisons. A recent study analyzed filled pause use of 14 speakers with L1 Afrikaans and L2 Spanish, living in a Spanish-speaking community (García-Amaya and Lang, 2020). These speakers showed intermediate vowels when compared to L1 control groups, and had separate F1 and F2 values in their two languages. In addition, they used the appropriate filled pause types in both languages, using nasal-only filled pauses only when speaking Spanish. Wong and Papp (2018) compared the cross-linguistic use of *uh* and *um* of 21 individuals from New Zealand speaking English and te reo Māori. Most speakers acquired the latter as an L2 and had English as their L1. The findings showed that speakers used *um* relatively more often when speaking English than when speaking te reo Māori, adapting to the language-specific pattern. Regarding spectral characteristics, the first two vowel formants (F1, F2) were slightly higher when speaking English, showing cross-linguistic shifts. Rose (2017) found that 16 Japanese learners of English with low proficiency did not adapt the F1 and F2 of their filled pauses when speaking English, while 16 speakers with higher proficiency realized their filled pauses more similarly to native speakers. Based on these studies, it seems that advanced L2 speakers adapt their filled pauses in the direction of the target language. However, this question has not yet been investigated with a substantial speaker set, homogeneous in terms of gender and age.

According to Clark and Fox Tree (2002), filled pauses have meaning and are planned and produced like any other conventional word, which means that L2 acquisition of vowels in filled pauses should be similar to that of other lexical vowels. According to Flege's Speech Learning Model (SLM; e.g., Flege, 1995), new L2 sounds that are similar but not identical to sounds in the L1 are most difficult to learn. Although empirical evidence is limited, we assume that filled pauses in native Dutch and native English are quite similar, but not identical (see below). This level of similarity implies that the perception of differences between Dutch and English filled pauses is difficult for Dutch learners of English. According to the SLM, L2 learners first need to acquire the ability to perceive the difference between the two sounds before they will be able to produce the new sound. Since filled pauses are perceived relatively unconsciously (Martin and Strange, 1968; Shriberg and Lickley, 1993), this may delay the perception of the L2 sound as being different from the Dutch one.

Moreover, one could argue that the ease with which subtle L1-L2 differences are acquired depends on the level of consistency in the language input. Filled pause vowels show high between-speaker variation when compared to lexical vowels (e.g., Hughes *et al.*, 2016). Because of their distinct characteristics (e.g., prolonged syllable, lower F0), articulatory freedom seems to be relatively high. Also, speakers differ in their preference for *uh* or *um* (e.g., Clark and Fox Tree, 2002). This variation in filled pause language input may delay the adaptation of filled pause realizations by the L2 speaker. Finally, interjections such as filled pauses, and especially their phonetic realization, are not explicitly taught in the L2 classroom, which may further hinder their acquisition (cf. Chen, 2009).

Even when filled pause realizations in the L2 are not learned, they could be different in the L2 than the L1 due to decreased fluency. Speech production in the L2 tends to be less automatic (Guz, 2015) and increases cognitive load (Fehringer and Fry, 2007); it thus may involve more speech planning difficulties. Hence, the number, position, and duration of filled pauses may be different in the L2. Guz (2015) found that L1 Polish speakers who were highly proficient learners of English used more and longer filled pauses when speaking in their L2. De Jong *et al.* (2015) found a similar increase in filled pauses in the L2 speech of intermediate to advanced Dutch learners (L1: Turkish or English). At the same time, they noted that the number of filled pauses in the L2 was highly correlated to that in one's L1 and is partly a feature of individual speaking style (cf. Fehringer and Fry, 2007). Furthermore, de Jong (2016) found that these same L2 Dutch speakers only used more filled pauses than L1 Dutch speakers *within* utterances and not *between* them, possibly due to increased problems with lexical retrieval. The ratio between different hesitation categories—including lexical fillers, repetitions, and filled pauses—was found to remain remarkably similar among sequential French-German bilinguals with high L2 proficiency (Fehringer and Fry, 2007), indicating that hesitation profiles may transfer from the L1 (cf. Wiese, 1984).

Accumulating evidence shows that despite claims that filled pauses are transferred from the L1, they may be different in the L2 due to language-specific adaptations and decreased fluency. The goal of this paper is to investigate whether filled pauses in L2 English by L1 Dutch speakers are realized differently than in the L1, and if so, why: through adaptations or disfluencies. Alternatively, speakers' filled pauses may remain consistent cross-linguistically, which would support the L1-transfer hypothesis.

B. Characteristics of native Dutch and English filled pauses

Filled pauses are also called hesitation markers because they often indicate delays in speech production and are used for planning the next word or utterance (Maclay and Osgood, 1959). According to some, they are mere symptoms of planning problems, whereas others believe they serve a signaling function to announce expected delays or to hold or cede the floor (see Clark and Fox Tree, 2002). Even though filled pauses seem to share some universal characteristics across languages (Clark and Fox Tree, 2002) they are language-specific in terms of distributional and spectral characteristics (e.g., Candea *et al.*, 2005; Clark and Fox Tree, 2002; Swerts, 1998). This section describes characteristics of Dutch and English filled pauses in terms of type, number, position, and phonetic realization.

As for filled pause type, *uh* and *um* are the most common types in both Dutch and English (e.g., de Leeuw, 2007; Wieling *et al.*, 2016).[1] *Uh* is an open syllable consisting of a neutral vowel, whereas in *um*, the neutral vowel is followed by a bilabial nasal (Hughes *et al.*, 2016; Wieling *et al.*, 2016). In general, *um* signals longer delays and greater difficulty in speech production (Clark and Fox Tree, 2002), and *uh* is more typically used in cases of local lexical planning difficulties (Shriberg, 1994). Despite these general tendencies, L1 speakers of Dutch and English differ in the relative occurrences of *uh* and *um*. Overall, British English speakers show a preference for *um*, using it in 81% of the cases, whereas Dutch speakers use *um* in only 27% of the instances (de Leeuw, 2007). Speakers of American English seem to use *um* relatively less often than speakers of British English (Shriberg, 1994), but in both varieties of English, and across different age groups and genders, *um* is used relatively more often than in Dutch (Wieling *et al.*, 2016).

Regarding the number of filled pauses, some studies found that speakers of Dutch use more filled pauses than speakers of English (de Leeuw, 2007; Wieling *et al.*, 2016). However, the number of filled pauses depends on the nature of the discourse, such as speech style and speech task difficulty (Goldman-Eisler, 1968). Moreover, large between-speaker differences have been found in the frequency of occurrence of filled pauses. In a study on British English, this ranged from 1.2 to 88.5 filled pauses per 1000 words (Clark and Fox Tree, 2002). Other studies on several Germanic languages describe similar results (Braun and Rosin, 2015; Künzel, 1997; Maclay and Osgood, 1959; McDougall and Duckworth, 2017). Thus, differences found between Dutch and English speakers could in fact be explained by task or speaker selection rather than language.

Regarding position, i.e., occurrence within or between phrases, Dutch and English filled pauses seem to differ. In general, silent pauses are considered more typical within phrases than filled pauses, which tend to occur at phrase boundaries (Maclay and Osgood, 1959). Indeed, de Leeuw (2007) found that the majority of filled pauses in British English occurred in combination with at least one silent pause, whereas filled pauses without any silent pauses—typical in mid-phrase position—occurred in only 15% of the cases. In Dutch, however, filled pauses without one or two adjacent silent pauses were more common and occurred in 36% of the cases. When de Leeuw (2007) considered the type of filled pauses, she found that in both languages, filled pauses surrounded by silent pauses are likely to be *um*. This corresponds to previous findings that *um* is more likely in major delays and *uh* more likely in minor delays (e.g., Clark and Fox Tree, 2002). The difference in filled pause placement between English and Dutch, however, might have implications for not only their type but also their phonetic realization because of position-dependent variation in, e.g., F0 (declination, boundary effects) and duration (final lengthening).

The phonetic realization of English and Dutch filled pauses seems similar in terms of duration and fundamental frequency (F0). De Leeuw (2007) reported that *um* tokens were somewhat longer in Dutch than in English, although this may also be attributed to other aspects of the speech corpora. In general, filled pauses are considerably longer in duration than lexical vowels (Hughes *et al.*, 2016; Shriberg, 2001) and can be prolonged extensively if speakers anticipate a longer delay (Clark and Fox Tree, 2002). In British English filled pauses, Hughes *et al.* (2016) found that the duration of *uh* was longer than the vowel in *um* and had a larger range, with *uh* tokens up to 1.5 s. In addition, filled pause duration is affected by the position in the utterance: utterance-initial filled pauses were found to be longer than utterance-medial ones (Swerts, 1998). Position also affects the F0 of filled pauses. For Dutch, filled pauses using mid-utterances have been described as having a lower F0 than those at the start of an utterance (Swerts, 1998). American English filled pauses in mid-utterance position were found to have a lowered F0 relative to the rest of the utterance (e.g., Shriberg, 2001; Shriberg and Lickley, 1993). As a low F0 is associated with a creaky voice quality (Keating *et al.*, 2015), filled pauses are spoken with the same creaky quality relatively often (Candea *et al.*, 2005; Shriberg, 2001).

Considering the spectral vowel realization in *uh* and *um*, Dutch and English filled pauses are similar but not identical. Filled pauses tend to be built around a central vowel in the language, requiring minimal articulatory effort and enabling quick production of the filled pause in combination with any following speech sound (Clark and Fox Tree, 2002). The British English filled pause *uh* has been described as a mid-central (McDougall and Duckworth, 2017) or schwa-like vowel, but longer in duration (Hughes *et al.*, 2016; Shriberg, 2001). Similarly, Dutch *uh* has been described as a lengthened schwa (Stouten and Martens, 2003). In spite of these "schwa" descriptions for both languages, formant measurements are rare. Hughes *et al.* (2016) reported that filled pauses of 60 Standard Southern British English (SSBE) male speakers had mean F1 values in the range of 450–700 Hz, and mean F2 values of 1250–1550 Hz, and were relatively stable throughout the vowel. Bon *et al.* (2018) measured formants in the filled pauses of 30 Dutch male speakers and reported mean F1 values per speaker of 476–644 Hz, and mean F2 values of 1268–1634 Hz. The differences between the ranges seem limited, but language-specific tendencies are present: we confirmed this in a control experiment (see Sec. III A) with data made available by Hughes *et al.* (2016) for English and van der Vloed *et al.* (2020) for Dutch (including the speakers from Bon *et al.*, 2018). Whereas British English is the educational target for Dutch L2 learners, the majority of pop culture input is in American English (AE). In AE, filled pauses have been described as a mid-open vowel between /ʌ/ and /æ/ (Candea *et al.*, 2005): this would mean that especially their F1 is higher than in British English and thus further away from the Dutch realization.

Hence, the literature shows that English and Dutch differ most clearly on *um*:*uh* ratios, with *um* being more common in English and *uh* in Dutch. Other characteristics of filled pauses, i.e., number, duration, F0, and F3, may be similar in the two languages or more dependent on speaker than on language-specific norms. Finally, phonetic differences between Dutch and English vowel realizations (i.e., F1, F2) of *uh* and *um* are present, but subtle. This offers an interesting case to assess cross-linguistic filled pause realization: if learnable, L2 English filled pauses may be among the most difficult elements to learn for L1 Dutch speakers. They contain subtle acoustic differences with variability in the language input, are perceived relatively unconsciously, and are not explicitly taught.

C. Research question and hypotheses

The current study investigates whether Dutch L1 speakers adapt their filled pause realizations when speaking in their L2 English. To allow learning of the English filled pause vowels to have potentially taken place, advanced learners were used. Overall, we hypothesize finding large differences between speakers in the extent to which they adapt their filled pauses when speaking in their L2 (cf. Quené *et al.*, 2017). In addition, some aspects of filled pauses are more likely to be adapted across a speaker's languages than others (see below). In contrast with earlier studies, filled pause position was included as a factor because it is expected to correlate with phonetic realization (i.e., duration, F0, and possibly vowel formants). Because of the predicted differences between realizations of *uh* versus *um* and their expected dependency on position (e.g., Clark and Fox Tree, 2002), the two filled pause types were analyzed separately. Thus, no language by position interaction is expected.

Within a homogeneous group of L1 speakers with above-average proficiency in L2 English, the percentage of *um* realizations is most likely to be adapted. There seem to be substantial differences in *um*:*uh* proportions between L1 Dutch and English, where English uses *um* in places where in Dutch *uh* would be more likely (e.g., de Leeuw, 2007). This difference is noticeable and therefore expected to be perceived and thus incorporated into L2 production relatively easily (cf. Flege, 1995). In addition, number and duration of filled pauses are expected to increase in the L2 due to decreased fluency (e.g., de Jong *et al.*, 2015; Guz, 2015). Although the speakers in this study are relatively proficient in their L2, they are not expected to be as fluent as in their native language. In terms of vowel realization (F1 and F2), adaptations are somewhat likely to occur, and individual variation therein is predicted to be relatively high. Although differences between Dutch and English filled pause vowels are subtle, more proficient learners may have gained sufficient input to incorporate separate L2 English filled pause realizations in their speech production (cf. Flege, 1995). Finally, F3 and F0 are predicted to remain similar between L1 Dutch and L2 English. F3 is mostly a feature of a speaker's vocal tract length and is not as salient and adaptable as F1 and F2 (Rose, 2002).

II. METHODS

A. Speaker characteristics

Spontaneous speech recordings were extracted from the Database of the Longitudinal Utrecht Collection of English Accents (D-LUCEA; Orr and Quené, 2017). D-LUCEA contains recordings of students from the liberal arts and sciences college University College Utrecht. The students live on campus, where English is the official language in class and the lingua franca for communication with fellow students (Orr and Quené, 2017). We selected a homogeneous speaker group consisting of female speakers whose single L1 was Dutch, who spoke Standard Dutch without an audible accent, and who were recorded within one month after arrival at campus. By selecting these early recordings, the multilingual environment is not yet expected to have had an effect on the speakers' filled pause realization. One speaker was removed from the dataset, because her speech sounded prepared rather than spontaneous.[2] The selection resulted in 58 speakers ($M_{age} = 18.4$ yr, SD = 0.8 yr).

In order to be accepted at UCU, students are required to have scored 8 out of 10 for English in high school (Quené *et al.*, 2017). They are estimated to have at least B2 level according to the Common European Framework of Reference for Languages (see Council of Europe, 2019), which is above average when compared to the rest of The Netherlands (Quené *et al.*, 2017). Almost all speakers reported that they were raised in a monolingual Dutch environment. Two speakers indicated that they grew up in an environment where English was spoken, but neither regarded English as their native language. The majority of speakers (48) learned English as an L2 in a Dutch school, the other ten in an English-language school (daycare, elementary, or secondary). Regarding the variety of English, 22 of the students reported speaking American English and 12 British English. Another 14 students said to speak both, and the final ten students could not indicate which variety they spoke. This variability shows that there is not a clear target variety of English in The Netherlands. As mentioned before, this could be explained by British English (i.e., Received Pronunciation) being the model in L2 English education, which is often taught by L1 Dutch speakers (Chen, 2009), whereas American pop culture has a high prominence in the personal lives of the Dutch (see also Smakman and De France, 2014). In the current study, the lack of uniformity in target variety is not considered to be problematic because any adaptation from Dutch toward British or American English is in the same direction.

B. Materials and procedure

Speakers in D-LUCEA performed several tasks with varying degrees of spontaneity and formality (see Orr and Quené, 2017). For the current study, a task was selected in which speakers talked for two minutes about an informal topic of their choice, such as their hobbies or vacation. After the speakers did this in their native language, here Dutch, they repeated the task in English. Two of the selected students performed the tasks in reversed order. The language order of the selected tasks was not counterbalanced because the purpose of D-LUCEA was to study the development of the students' English (Orr and Quené, 2017). This may have led to a reduced number of filled pauses in the second monologue, in general L2 English, because of repetition (cf. Goldman-Eisler, 1968). We expect repetition to potentially influence the number of hesitations but not their phonetic realization. Therefore, for number of filled pauses, we compared results for speakers who did (N = 28) and did not repeat their monologue (see below). The students were informed beforehand about the nature of the recordings and tasks. This led to different levels of preparation, ranging from students selecting their topic during the monologue itself to speakers who clearly prepared the outline of their monologue. Overall, the monologues can be considered as semi-prepared spontaneous speech. According to Clark and Fox Tree (2002), while monologues do not require filled pauses to hold the floor as in dialogues, filled pauses are used in the same way.

The sessions took place in a quiet, furnished office and were recorded by eight microphones. We selected the recordings made by a close-talking microphone (Sennheiser HSP 2ew) attached to a headset, to keep the distance to the speaker's mouth consistent. The speech was recorded digitally (44.1 kHz, 16 bits) using a FocusRite Saffire Pro 40 multichannel preamplifier and an A/D converter (Quené *et al.*, 2017). All recording sessions were led by one of five interlocutors, who introduced the tasks in English and monitored the timing. During the monologue, while not involved in an interactional conversation, speakers directed their speech toward the interlocutors, who could understand them in both Dutch and English.

C. Segmentation and measurements

To investigate filled pauses phonetically, only the most common filled pause types *uh* and *um* were included. Much sparser but related phenomena were excluded, such as vowel lengthening (*aaand*), lexical fillers (*like*), and nasal-only filled pauses. The onset and offset of all *uh* and *um* tokens were segmented manually in Praat (Boersma and Weenink, 2016) by at least two coders. In cases of disagreement on the inclusion of a potential filled pause, a panel of three additional phoneticians was consulted, which was required for a handful of tokens. In total, 2,101 *uh* and *um* tokens remained in 257 min of speech (see Table I). For both *uh* and *um*, boundaries were placed at the onset of the vocalic part of the segments, where the signal showed voicing, and at the offset of the vowel and/or nasal part. Segmentation was based on the waveform and spectrogram, and was substantiated by repeated listening using a FocusRite Scarlett 2i4 audio interface and Beyerdynamic DT 770 PRO headphones. The presence/absence of creaky voice quality on filled pauses was hand-coded.

For each token, temporal and spectral measurements were performed in Praat (Boersma and Weenink, 2016). Duration was measured using the manually set onsets and offsets. For *um*, vowel duration was measured separately from nasal duration. Temporal measurements were log-transformed after inspection of model residuals. Over the full duration of the vowel, the mean value for F0 was measured in Hertz. This was done within a 100–350 Hz range, using an autocorrelation method implemented in Praat. Over the mid 50% of each vowel, formants (F1, F2, F3) were measured using the Burg method (window length = 25 ms). For F3, 97 values (4.6%) were missing. To remove erroneous estimates, formant measurements that were over ±2.5 standard deviations from the mean over all speakers were excluded, leaving the other measurements for those tokens. This led to the exclusion of 1.5% to 1.9% of the measurements per formant.

The automatic F0 measurements showed some rather low values for female voices, with minimum values around 100 Hz. All values below 150 Hz were auditorily checked.

TABLE I. Overview of the distribution of the filled pauses *uh* and *um* across language and position (N = 59).

		Position				
Language	Type	Single	Start	Mid	End	Total
Dutch (L1)	uh	34	67	501	40	642
	um	140	69	231	38	478
	uh + um	*174*	*138*	*734*	*78*	*1120*
English (L2)	uh	11	62	322	12	407
	um	132	94	304	44	574
	uh + um	*143*	*156*	*627*	*56*	*981*
Total		***317***	***292***	***1358***	***134***	***2101***

Only if the filled pause was not considered to be extreme by ear, the low F0 value was judged to be an octave error and doubled. In this way, 29 values from ten different speakers were corrected (1.6% of the data). In addition, 307 (14.6%) of the F0 measurements were excluded because of the presence of creaky voice quality.

Per filled pause, its position in the utterance was coded manually by considering grammatical phrases. Filled pauses that occurred in between two grammatical phrases were considered as *single* when they were surrounded by silent pauses of $\geq$150 ms. If they were adjacent to a silent pause of $\geq$150 ms on one side only and to a grammatical phrase at the other side, they were considered to be at the *start* or *end* of an utterance. If a filled pause interrupted a grammatical phrase, no matter how long the silences before or after the filled pause, the position was considered to be *mid* utterance. In order to be *mid*, the surrounding utterance had to proceed as if the filled pause was not there. Restarts, repairs, and repetitions were considered to be new utterances.

In addition to acoustic measurements, the absolute and relative occurrence of *uh* and *um* were counted per speaker. Per language, we calculated the number of filled pauses, time-normalized per minute,[3] and the percentage of the total number of filled pauses (*uh* plus *um* tokens) that was *um*. To control for repetition effects, we determined per speaker whether they repeated the same content in both languages by manually coding whether the speaker spoke about (1) entirely different topics (N = 28), (2) partly the same topics (N = 5), or (3) only the same topics (N = 25) in their second monologue. Table I gives an overview of the total amount of filled pause tokens in our analysis and their distribution across the languages, filled pause types, and positions in the utterance.

D. Statistical analysis

To analyze whether the same speakers differ in their filled pauses in L1 Dutch versus L2 English, the data were analyzed using mixed-effects models in R (R Core Team, 2018), using the package lme4 (Bates *et al.*, 2015). In total, we built 13 different models on the data. Because of the many predicted differences between *uh* and *um* realizations, and their expected correlation to Position, separate sets of models were built for *uh* and *um*. Per filled pause type, we modelled vowel and/or full duration, F0, F1, F2, and F3. In addition, we modelled the number of filled pauses (time-normalized per minute) and the percentage of *um*. A Bonferroni correction was used, setting our α at 0.004 ($\approx$ 0.05/13). For each of the temporal and spectral measurements and number of filled pauses, linear mixed-effects models were built using *lmer*(), and for the percentage of *um* we used generalized linear mixed-effects models [*glmer*() function]. For models predicting spectral and temporal features of filled pauses, fixed factors were Language (L1 Dutch, L2 English) and Position (single, start, mid, end). Factor levels were treatment-coded. For the factor Language, the L1 (Dutch) was the reference level. For the factor Position, the reference level varied between acoustic parameters and will be explained per parameter in the results section.

Significance was evaluated through likelihood ratio testing with stepwise inclusion of predictors. First, effects in the fixed part of the model were evaluated, and interaction effects were tested regardless of the significance of main effects. Subsequently, in the random part of the model, the same procedure was followed by adding random slopes over speaker. If an optimal model includes random slopes, this indicates that speakers varied in the extent to which the overall effect was found in their filled pause realizations.

For the mean formant frequencies, as a control step, a model with the fixed factor of Creak (present, absent) was tested against an empty model to see whether creaky vowels behaved differently than modal vowels.[4] This was only the case for the F1 of *um* vowels, which led to the decision to build these models without using the measurements of creaky vowels. For other models, they were left in. For the count variables, which were calculated per speaker, we again included Language in the fixed part of the model. In addition, we included Repetition (1: no repeated elements, 2: some repeated elements, 3: only repeated elements) to assess whether this had any effect on the frequency of occurrence of filled pauses or the selection of *uh* or *um*.

Before turning to the main analysis, we present a control experiment using native filled pause data in British English (from Hughes *et al.*, 2016) and Dutch (from van der Vloed *et al.*, 2020) to establish that the vowel realization of L1 English and L1 Dutch filled pauses is different. Hughes *et al.* (2016) used high quality studio recordings of conversations with male speakers, and extracted the first 20 *uh* and 20 *um* tokens per speaker (N = 75, n = 3000).[5] Van der Vloed *et al.* (2020) provided us with filled pauses in spontaneous speech of 60 male speakers (including the 30 speakers of Bon *et al.*, 2018), from which we selected *uh* and *um* tokens that were recorded indoors and in a quiet environment. This led to 2104 *uh* tokens and 208 *um* tokens. Both corpora were built to serve as realistic background data for forensic purposes and were comparable in terms of speaker demographics (i.e., young, adult, male speakers). We compared both L1 corpora on the vowel realization (F1, F2) of the filled pauses using linear mixed-effects models. Although the control experiment is performed on male speakers, we assume that any differences found between Dutch and English will also be present for female speakers.

III. RESULTS

A. Control experiment: Filled pause realization in L1 Dutch and L1 English

Comparing the filled pauses of L1 Dutch and L1 English male speakers from forensic reference corpora (reference level = Dutch; α = 0.025), we found that Language was included in the optimal F1 and F2 models (see the supplementary material for modeling details).[6] The models predicting F1 showed intercepts of 534 Hz for Dutch *uh* and

568 Hz for Dutch *um*, with filled pauses in English being 41 and 33 Hz higher, respectively. This difference may be noticeable for listeners ('t Hart, 1981). Hence, in English, pronunciation is more open than in Dutch. The F2 intercepts were 1411 Hz for Dutch *uh* and 1366 Hz for Dutch *um*. Although Language improved the model fit [$\chi^2(1) \geq 6.7$, $p \leq 0.01$], it was not a significant predictor in the optimal F2 models ($t < 2.0$; see supplementary material). Taken together, these results indicate that Dutch and English filled pauses differ on F1. Although part of this difference could be explained by a difference in sampling or measurements, we consider it to be relevant for the interpretation of potential differences in our L1-L2 experiment.

Hence, the control experiment showed that there are language-specific realizations, with a more open pronunciation of filled pause vowels in British English than in Dutch. For American English, we expect even higher F1 values (see Candea *et al.*, 2005). The F1 difference between Dutch and English could have been picked up by the L2 speakers in our main analysis. The remainder of the results section presents the main analysis, with filled pauses from 58 female speakers in L1 Dutch and L2 English.

B. Count features in L1 Dutch and L2 English

The optimal model for the number of filled pauses (time-normalized per minute) was an empty model, that is, without the fixed factors of Language or Repetition. On average, speakers did not use more filled pauses in their L2 than their L1; when looking at the count data, of the 58 speakers, 31 used more filled pauses in L1 Dutch than in L2 English, whereas only 19 used more filled pauses in English. A language effect was also absent for speakers who did not repeat the content from the Dutch monologue in their second monologue (in English). In addition, the two speakers who performed the tasks in reversed order, with their L2 first, also did not use an increased number of filled pauses in their L2 when compared to their L1. The intercept of the empty model was 8.1 filled pauses per minute, with large variation between speakers; speakers used between 1.5 and 18.4 filled pauses per minute in Dutch, and between 0 and 16.9 filled pauses per minute in English. Indeed, some speakers can talk for up to 60 s without using a single filled pause, whereas other speakers may use a filled pause every three seconds.

For *um*:*uh* ratios, Language was included in the optimal model [$\chi^2(1) = 22.8$, $p < 0.001$]. In Dutch, 44.8% of all filled pauses was *um* (and 55.2% *uh*), whereas in English, speakers used *um* in 59.8% of the cases. Repetition was not included in the optimal model; whether part of the monologue was repeated did not seem to be relevant for the choice between *um* or *uh*. Again, speakers varied highly from each other, with ranges between 0 and 100% in both languages: some speakers used only *uh*, others used only *um*. The difference between Dutch and English was quite consistent: for 42 out of 58 speakers, the *um* percentage was higher in English than in Dutch. For example, none of the 22 filled pauses uttered by speaker 45 in Dutch was *um*, whereas she used it in 14 of her 21 hesitations in English.

C. Acoustics of filled pauses in native Dutch and non-native English

For duration, the reference level was set to Dutch, single filled pause realizations. For none of the temporal measurements, Language or a Language × Position interaction were included in the optimal model (see Table II); the duration of filled pauses was similar in L2 English and L1 Dutch, regardless of their position. Variation in filled pause duration was high, with *uh* tokens ranging 42–1113 ms and *um* tokens ranging 148–1312 ms (and their vowel durations 53–863 ms). This variation was also visible within speakers, with durations for instance varying 150–1113 ms.

For *uh* duration, the optimal model included Position [$\chi^2(3) = 60.6$, $p < 0.001$]. Single *uh*s (back-transformed intercept: 381 ms) were significantly longer than *uh*s at the start (261 ms), mid (246 ms), and end (312 ms) of an utterance. Random slopes for Position and Language over Speaker were not included in the optimal model; speakers did not differ from each other in the extent to which they changed their *uh* duration by position or language.

For the vowel duration of *um*, Position was not included in the optimal model [$\chi^2(3) = 12.7$, $p = 0.005$]. The optimal model included random slopes for Position over Speaker [$\chi^2(9) = 25.0$, $p = 0.003$], indicating that some speakers differed in positional adaptations to vowel duration. The intercept was –0.583 (SD = 0.013, t = –46.2), i.e., 261 ms. For the full duration of *um*, so including the nasal (back-transformed intercept: 506 ms), Position was part of the optimal model [$\chi^2(3) = 68.1$, $p < 0.001$]. Single *um*s were significantly longer than *um*s in other positions. Random slopes for Position and Language over Speaker were included in the optimal model [$\chi^2(14) = 33.2$, $p = 0.003$], showing that speakers differed in the extent to which they varied their *um* duration across positions and across languages.

For F0, Position was part of the optimal model for *uh* [$\chi^2(3) = 87.4$, $p < 0.001$] and *um* [$\chi^2(3) = 170.6$, $p < 0.001$]. Filled pauses at the start of an utterance, the reference level of all spectral measurements, had a higher F0 than filled pauses in other positions. Language was included in neither of the models. The optimal F0 model of *uh* included random slopes for Position over Speaker [$\chi^2(9) = 33.5$, $p < 0.001$]. For *um*, random slopes for both Position and Language over Speaker were included in the optimal

TABLE II. Optimal linear mixed-effects models predicting log-transformed duration of filled pauses in Dutch (L1) and English (L2).

	Duration *uh*		(Full) duration *um*	
	Coefficient (SE)	t	Coefficient (SE)	t
intercept	–0.419 (0.029)	–14.6	–0.296 (0.012)	–25.4
Position: start	–0.165 (0.030)	–5.4	–0.079 (0.011)	–3.4
Position: mid	–0.190 (0.027)	–6.9	–0.063 (0.009)	–7.5
Position: end	–0.087 (0.036)	–2.4	–0.047 (0.014)	–7.1

TABLE III. Optimal linear mixed-effects models predicting fundamental frequency (in Hz) of filled pauses in Dutch (L1) and English (L2).

	F0 *uh*		F0 *um*	
	Coefficient (SE)	t	Coefficient (SE)	t
intercept	202 (3.4)	59.5	208 (2.9)	–71.1
Position: mid	–13 (2.2)	–6.0	–20 (2.3)	–8.5
Position: end	–17 (3.2)	–5.2	–23 (2.7)	–8.6
Position: single	–8 (3.1)	–2.7	–6 (2.1)	–2.7

TABLE IV. Optimal linear mixed-effects models predicting the F1 of filled pauses in Dutch (L1) and English (L2).

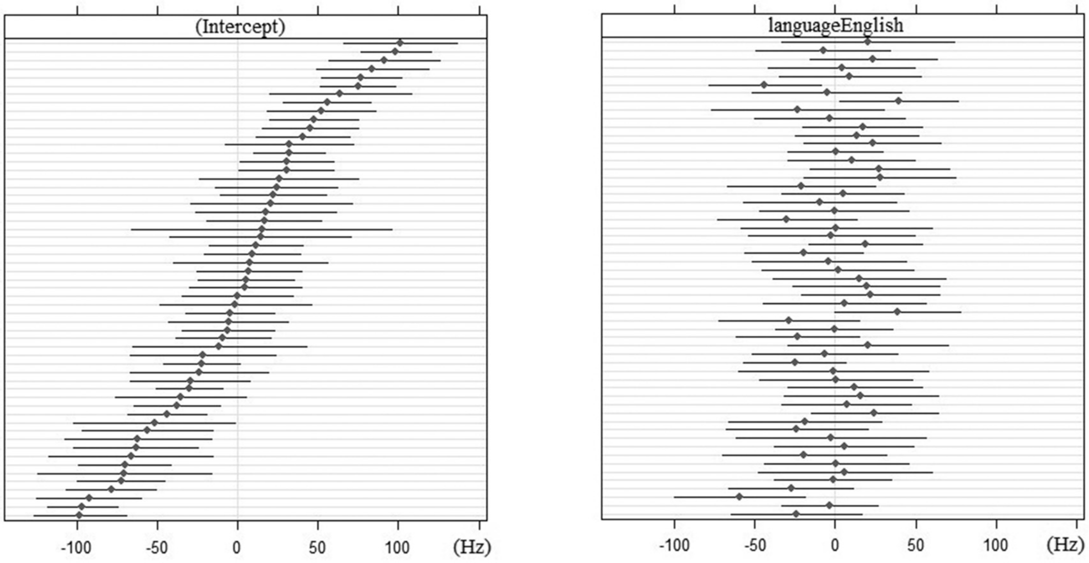
	F1 *uh*		F1 *um*	
	Coefficient (SE)	t	Coefficient (SE)	t
intercept	656 (9.6)	68.5	673 (11.4)	59.1
Language: English	26 (6.2)	4.1	37 (4.1)	9.1
Position: mid	–25 (6.3)	–3.9	–29 (8.1)	–3.6
Position: end	–36 (11.1)	–3.2	–22 (10.2)	–2.1
Position: single	–28 (11.8)	–2.4		

model [$\chi^2(5) = 20.0$, p = 0.001], so F0 adaptations across positions and languages differed per speaker. Table III gives an overview of the optimal F0 models. Variation in F0 was relatively high (cross-speaker SD = 24.5 Hz), with means per speaker ranging 160–236 Hz.

F1 and F2 showed shifts from the L1 to the L2. In both optimal F1 models (see Table IV), Language and Position were included [*uh*: $\chi^2(3) = 18.4$, $p < 0.001$; *um*: $\chi^2(3) = 32.2$, $p < 0.001$]. F1 was about 30 Hz higher in English than in Dutch, and higher for filled pauses at the start of an utterance than in other positions. In the random part, the F1 model for *uh* included random slopes for Language over Speaker [$\chi^2(2) = 15.6$, $p < 0.001$], showing between-speaker differences in cross-linguistic F1 adaptations (see Fig. 1). For *um*, random slopes for Position over Speaker were included in the optimal model [$\chi^2(9) = 25.4$, p = 0.003]; speakers differed in the extent to which they changed their F1 across positions. This difference of about one semi-tone between L1 Dutch and L2 English F1 values may be noticeable to listeners (cf. 't Hart, 1981).

Optimal models for F2 included Language [*uh*: $\chi^2(1) = 15.0$, $p < 0.001$; *um*: $\chi^2(1) = 16.8$, $p < 0.001$], but not Position. For *uh*, the 1661 Hz Dutch intercept (SE = 13.6, t = 122.5) was lowered by 28 Hz (SE = 7.2, t = –3.9) for English, and for *um*, the 1614 Hz intercept (SE = 14.7, t = 109.8) was lowered by 29 Hz (SE = 8.3, t = –3.6) for English. These differences, although significant, are unlikely to be perceptible to listeners (cf. 't Hart, 1981). Random slopes were not included in the optimal model for *uh*. The optimal model for *um* included random slopes for Language and Position over Speaker [$\chi^2(12) = 31.2$, p = 0.002], which means speakers differed in the extent to which they varied their F2 values across positions and across their languages. Across filled pause types, individual F2 means varied 1456–1993 Hz for Dutch (cross-speaker SD = 140 Hz), and 1364–1924 Hz for English (cross-speaker SD = 156 Hz).

The optimal F3 model for *uh* was an empty model (β = 2742 Hz, SE = 18.6, t = 147.5) without Language or Position. Random slopes were not included in the optimal model for *uh*, indicating that the F3 remained stable across conditions. In the optimal model for *um*, Position was included [$\chi^2(3) = 20.7$, $p < 0.001$]. The F3 intercept of *um* tokens at the start of an utterance (β = 2764 Hz, SE = 24.5, t = 112.7) was higher than the F3 of *um* tokens in the middle (β = –53 Hz, SE = 13.3, t = –4.0) or at the end



(a)

(b)

FIG. 1. Caterpillar plots showing the by-speaker (a) intercepts and (b) L2 adaptations for the F1 model of *uh* (in Hertz). Horizontal lines represent the speakers, sorted from lowest to highest mean F1 in L1 Dutch. In (a), x = 0 represents the intercept over all speakers, (b) shows by-speakers adaptations relative to their intercepts when speaking in the L2, English, (0 means no adaptation, –100 Hz is the most extreme adaptation).

(β = −67 Hz, SE = 19.8, t = –3.4) of an utterance. This model did not include Language, so speakers' F3 of *um* was language-independent. Random slopes were also not included, so speakers were similar in the extent to which they changed their F3 across positions. Individuals' F3 means ranged 2301–3053 Hz (cross-speaker SD = 187 Hz).

IV. DISCUSSION

We investigated whether L1 Dutch speakers of L2 English incorporate language-specific features when producing filled pauses in their L2, or whether they transfer their filled pauses from the L1 without making any adaptations—as claimed by Clark and Fox Tree (2002) and de Leeuw (2007). In the latter case, filled pauses could be useful in cross-linguistic forensic speaker comparisons (see also Wong and Papp, 2018), whereas when filled pauses are language-specific within speakers, this may be inadvisable. As predicted, we found that speakers adapt their vowel realizations and *um*:*uh* proportions across languages. Like L1 speakers of British and American English, the L2 speakers realized their filled pauses with a higher (i.e., more open) F1 and used *um* relatively more often in English than in Dutch. In addition, they realized their filled pauses with a somewhat lower (i.e., more back) F2, although this difference was minimal. The finding that filled pause realizations of proficient L2 speakers are language-specific is in line with prior work. Similar results have been described for L1 Japanese-L2 English speakers (Rose, 2017), L1 Afrikaans-L2 Spanish speakers (García-Amaya and Lang, 2020), and L1 English-L2 te reo Māori speakers (Wong and Papp, 2018). Using highly proficient L2 English speakers with L1 Dutch, we showed that L2 speakers may make language-specific adaptations in their filled pause realizations also for a closely related language combination for which filled pauses are described as being similar (see García-Amaya and Lang, 2020).

The fact that the speakers pronounced their filled pauses with a more open vowel in English, in combination with the absence of more and longer pauses in the L2, indicates that they learned the difference between Dutch and English filled pauses rather than showed more disfluencies. Although we do not know the target variety for the speakers, the shift is in the direction of both British and American English. Note that this does not mean that the speakers produced native-like filled pauses in English. The spectral difference between L1 Dutch and L2 English was smaller than between the L1 male control groups, whereas for females—with higher formants—one would expect a larger difference between the languages. In addition, there was much variability between speakers in their L1-L2 differences, with some speakers also producing similar F1 values in both languages, and some speakers shifting toward the opposite direction in English. Still, our findings suggest that despite the relatively unconscious processing of filled pauses (e.g., Martin and Strange, 1968), the relatively small differences in native filled pauses between English and Dutch, and the relatively high amount of variation in language input (e.g., Hughes *et al.*, 2016), L2 speakers seem capable of learning different realizations of filled pauses in their non-native language.

As do L1 speakers of British and American English, the Dutch speakers used relatively more *um* when speaking L2 English, indicating that they learned this feature as well. An alternative explanation for the increased use of *um* could be that L2 speech induces the use of *um*. The addition of the nasal has been related to longer delays (e.g., Clark and Fox Tree, 2002), which may be expected to occur more often in L2 speech. However, as mentioned, we found that the speakers did not increase the number and duration of their filled pauses (see also below). In addition, several studies reported consistency of *um*:*uh* ratios across speakers' languages (e.g., Fehringer and Fry, 2007; Wiese, 1984). This seems to indicate that the increased use of *um* in L2 English is learned from L1 English speech input.

The absence of more and longer pauses in the L2 contradicted our expectations. Both features have been related to L2 fluency and are usually higher in the L2 than the L1. De Jong (2016) found that a cross-linguistic difference in number of filled pauses may be attributed to an increased number of filled pauses *within* utterances, while those between utterances are similar to L1 use. We checked in our data whether the number of filled pauses within utterances (i.e., mid-utterance) increased in the L2; this was not the case. The absence of increased hesitation in the L2 could have been caused by the order of the speech tasks, which was not counterbalanced, although number of filled pauses in the L2 did not lower with repetition of the content (cf. Goldman-Eisler, 1968). Moreover, the two speakers who did the monologue first in L2 English and then in L1 Dutch did not use more filled pauses in their L2 either, despite the lack of practice. According to Vasilescu and Adda-Decker (2006), increased hesitation in the L2 may be more related to increased stress than to decreased proficiency. By the time speakers started their second monologue, they may have experienced less stress because they had gotten used to the task. In addition, the speakers in this study may feel as comfortable when speaking English as when speaking Dutch, since they applied to an English-speaking study program, were selected based on their English proficiency, and had often spent time abroad—an experience shown to lead to a decreased number of filled pauses in the L2 (Wiese, 1984).

Unlike most prior studies, we considered the effect of position on filled pause acoustics, since this could affect or explain certain differences between the L1 and L2. For example, one might expect L2 speakers to have different production problems than L1 speakers (e.g., more lexical retrieval problems), which could affect filled pause positions (i.e., more within-utterance) and acoustics. However, we did not find such effects. Contrary to de Jong's (2016) claim that filled pauses occurring within an utterance are less common in L1 speech, this was the most common position for filled pauses to occur in our corpus, both in the L1 (65.3%) and the L2 (63.8%). De Leeuw (2007), comparing L1 Dutch to L1 English, reported that filled pauses without any surrounding silent pauses were the least common form in English, whereas they were the most common form in Dutch. If filled pauses within utterances or without adjacent silences are more acceptable in Dutch than English, this may indicate that the current speakers transferred this feature from their L1 to their L2, showing some within-speaker consistency in the cross-linguistic use of filled pauses. However, this requires further research including the analysis of silent pauses.

As expected, filled pause characteristics depended on their position in the utterance. For example, filled pauses occurring as an independent utterance were longer than those that were part of an utterance, and filled pauses at the start of an utterance had a higher F0 (cf. Swerts, 1998). While these overall effects are theoretically predictable, random slopes indicated that speakers differed in the extent to which they adapted their filled pauses across positions. This shows that when analyzing filled pauses, whether for theoretical or forensic purposes, the position and linguistic context of the token may affect phonetic properties and must be taken into account (cf. Smorenburg and Heeren, 2020).

Despite language-specific tendencies, there was variation between speakers within one language, which is in line with prior findings (e.g., Hughes *et al.*, 2016). In addition, speakers differed from each other in the extent to which they shifted their acoustics in the L2, and some speakers shifted in the opposite direction. Language effects thus depend on the speaker, even in a relatively homogeneous speaker group with speakers who have been selected based on their L2 proficiency. In part, this variation may still have been introduced by differences in L2 proficiency or the English target varieties of the speakers, for which there was no detailed, objective information available. However, in real-life forensic cases, such details regarding the speakers involved are also lacking. Previous studies suggested that the transferability of filled pauses could be used in forensic phonetics to perform cross-linguistic speaker comparisons (de Leeuw, 2007; Wong and Papp, 2018). Based on the current study, we cannot fully support this claim; most of the advanced L2 speakers changed their filled pause realizations and *um*:*uh* ratios from the L1 to the L2. At the same time, spectral differences were relatively small, and most filled pause characteristics stayed consistent (i.e., number, duration, F0, and F3). A remaining question, therefore, is how much speaker-specific information is still retained in filled pauses to use them in cross-linguistic forensic speaker comparisons.

V. CONCLUSION

The speaker-dependent nature of filled pauses does not seem to overwrite language-specific tendencies. Advanced L2 English speakers with L1 Dutch adapt their filled pause realization in the direction of the target language, showing sensitivity to subtle acoustic differences even in filled pauses, which are perceived relatively unconsciously and are often neglected in second language teaching.

ACKNOWLEDGMENTS

This research was supported by the Dutch Research Council (NWO VIDI Grant No. 276-75-010). We thank Jade van der Graaf, Yara Sleeboom, and Thomas Haga for their assistance in the annotations. We also thank Vincent Hughes and David van der Vloed (assisted by Marit Bon and Tessi Marks) for kindly sharing their databases. Finally, we thank the anonymous reviewers and the associate editor Ewa Jacewicz for their thorough and helpful comments on previous versions of this manuscript.

[1] In addition to *uh* and *um*, some speakers use nasal-only filled pauses (*mmm*), although this type is much sparser (e.g., de Leeuw, 2007; McDougall and Duckworth, 2017).

[2] This was speaker 262 from D-LUCEA.

[3] The number of filled pauses was time-normalized by dividing the total number of *uh* and *um* tokens by the duration of the monologue (in seconds), multiplied by 60 to get to the number per minute. Building models on the total number of filled pauses (i.e., not time-normalized) led to the same effects as the time-normalized models presented in this paper.

[4] The reliability and validity of formant measurements may be affected by creakiness (Ladefoged *et al.*, 1988; Moosmüller, 2007).

[5] We received the data files from Dr. Vincent Hughes on 5 December 2019 and from David van der Vloed on 30 July 2020.

[6] See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0002871 for the model comparisons and optimal models for the F1 and F2 of *uh* and *um* in L1 Dutch and L1 English.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). “Fitting linear mixed-effects models using lme4,” J. Stat. Softw. **67**, 1–48.

Bhatia, T. K., and Ritchie, W. C. (**2012**). *The Handbook of Bilingualism and Multilingualism* (Wiley-Blackwell, West-Sussex, UK), pp. xxi–xxiii.

Boersma, P., and Weenink, D. (**2016**). “Praat: Doing phonetics by computer [computer program],” http://www.praat.org/ (Last viewed 3 July 2018).

Bon, M., Heeren, W., and van der Vloed, D. (**2018**). “The speaker-dependency of features of hesitation markers in Dutch spontaneous phone conversations,” in *Proceedings of the 26th IAFPA*, Huddersfield, UK, pp. 71–72.

Braun, A., and Rosin, A. (**2015**). “On the speaker-specificity of hesitation markers,” in *Proceedings of the 18th ICPhS*, Glasgow, UK, pp. 731–735.

Candea, M., Vasilescu, I., and Adda-Decker, M. (**2005**). “Inter- and intra-language acoustic analysis of autonomous fillers,” in *Proceedings of the 5th Worksh. Disfl. Spont. Speech*, Aix-en-Provence, France, pp. 47–51.

Chen, A. (**2009**). “Perception of paralinguistic intonational meaning in a second language,” Lang. Learn. **59**, 367–409.

Clark, H. H., and Fox Tree, J. E. (**2002**). “Using *uh* and *um* in spontaneous speaking,” Cogn. **84**, 73–111.

Council of Europe (**2019**). “Common European framework of reference for languages: Learning, teaching, assessment (CEFR),” Retrieved 13 March 2020 from https://www.coe.int/en/web/common-european-framework-reference-languages.

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. Teach. **54**, 113–132.

de Jong, N. H., Groenhout, R., Schoonen, R., and Hulstijn, J. H. (**2015**). “Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior,” Appl. Psycholing. **36**, 223–243.

de Leeuw, E. (**2007**). “Hesitation markers in English, German, and Dutch,” J. Germanic Ling. **19**, 85–114.

Fehringer, C., and Fry, C. (**2007**). “Hesitation phenomena in the language production of bilingual speakers: The role of working memory,” Folia Linguistica **41**, 37–72.

Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York Press, York), pp. 233–277.

García-Amaya, L., and Lang, S. "Filled pauses are susceptible to cross-language phonetic influence: Evidence from Afrikaans-Spanish bilinguals," Studies Sec. Lang. Acq. (published online, 2020).

Goldman-Eisler, F. (**1961**). "A comparative study of two hesitation phenomena," Lang. Speech **4**, 18–26.

Goldman-Eisler, F. (**1968**). *Psycholinguistics: Experiments in Spontaneous Speech* (Academic Press, New York), pp. 11–31.

Guz, E. (**2015**). "Establishing the fluency gap between native and non-native-speech," Research Lang. **13**, 230–247.

Hughes, V., Foulkes, P., and Wood, S. (**2016**). "Strength of forensic voice comparison evidence from the acoustics of filled pauses," Int. J. Speech, Lang. Law **23**, 99–132.

Keating, P. A., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th ICPhS*, Glasgow, UK, paper no. 0821.

Künzel, H. J. (**1997**). "Some general phonetic and forensic aspects of speaking tempo," Int. J. Speech, Lang. Law **4**, 48–83.

Ladefoged, P., Maddieson, I., and Jackson, M. (**1988**). "Investigating phonation types in different languages," in *Vocal Physiology: Voice Production, Mechanisms and Functions*, edited by O. Fujimura (Raven Press, New York), pp. 297–317.

Lo, J. J. H. (**2019**). "Between *Äh(m)* and *Euh(m)*: The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals," Lang. and Speech **63**, 746–768.

Maclay, H., and Osgood, C. E. (**1959**). "Hesitation phenomena in spontaneous English speech," Word **15**, 19–44.

Martin, J. G., and Strange, W. (**1968**). "The perception of hesitation in spontaneous speech," Percept. Psychophys. **3**, 427–438.

McDougall, K., and Duckworth, M. (**2017**). "Profiling fluency: An analysis of individual variation in disfluencies in adult males," Speech Commun. **95**, 16–27.

Mok, P. P., Xu, R. B., and Zuo, D. (**2015**). "Bilingual speaker identification: Chinese and English," Int. J. Speech Lang. Law **22**, 57–77.

Moosmüller, S. (**2007**). "The influence of creaky voice on formant frequency changes," Int. J. Speech Lang. Law **8**, 100–112.

Orr, R., and Quené, H. (**2017**). "D-LUCEA: Curation of the UCU Accent Project data," in *CLARIN in the Low Countries*, edited by J. Odijk and A. van Hessen (Ubiquity Press, Berkeley), pp. 177–190.

Quené, H., Orr, R., and van Leeuwen, D. (**2017**). "Phonetic similarity of /s/ in native and second language: Individual differences in learning curves," J. Acoust. Soc. of Am. **142**, EL519–EL524.

R Core Team (**2018**). *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria), http://www.R-project.org/ (Last viewed 13 March 2020).

Rose, P. J. (**2002**). *Forensic Speaker Identification* (Taylor & Francis, London), pp. 195–268.

Rose, R. L. (**2017**). "A comparison of form and temporal characteristics of filled pauses in L1 Japanese and L2 English," J. Phonetic Soc. Jpn. **21**, 33–40.

Shriberg, E. E. (**1994**). "Preliminaries to a theory of speech disfluencies," Ph.D. thesis, University of California.

Shriberg, E. E. (**2001**). "To 'errrr' is human: Ecology and acoustics of speech disfluencies," J. Int. Phonetic Assoc. **31**, 153–169.

Shriberg, E. E., and Lickley, R. J. (**1993**). "Intonation of clause-internal filled pauses," Phonetica **50**, 172–179.

Smakman, D., and De France, T. (**2014**). "The acoustics of English vowels in the speech of Dutch learners before and after pronunciation training," in *Above and Beyond the Segments: Experimental Linguistics and Phonetics*, edited by J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller, and E. van Zanten (John Benjamins Publishing Company, Amsterdam), pp. 288–301.

Smorenburg, L., and Heeren, W. (**2020**). "The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues," J. Acoust. Soc. Am. **147**, 949–960.

Stouten, F., and Martens, J. P. (**2003**). "A feature-based filled pause detection system for Dutch," in *Proceedings of the 2003 IEEE Worksh. Autom. Speech Recogn. and Underst.*, St. Thomas, Virgin Islands, IEEE Cat. No. 03EX721, pp. 309–314.

Swerts, M. (**1998**). "Filled pauses as markers of discourse structure," J. Pragm. **30**, 485–496.

't Hart, J. (**1981**). "Differential sensitivity to pitch distance, particularly in speech," J. Acoust. Soc. Am. **69**, 811–821.

van der Vloed, D. L., Bouten, J. S., and Van Leeuwen, D. A. (**2014**). "NFI-FRITS: A forensic speaker recognition database and some first experiments," in *Proceedings of the Odyssey Speaker and Lang. Recogn. Workshop*, Joensuu, Finland, pp. 6–13.

van der Vloed, D., Kelly, F., and Alexander, A. (**2020**). "Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA: A forensically realistic database," in *Proceedings of the Odyssey Speaker and Lang. Recogn. Workshop*, Tokyo, Japan, pp. 402–407.

Vasilescu, I., and Adda-Decker, M. (**2006**). "Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech," in *Proceedings of the 9th Int. Conf. on Spoken Lang. Proc. Interspeech*, Pittsburgh, PA, pp. 1850–1853.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M. (**2016**). "Variation and change in the use of hesitation markers in Germanic languages," Lang. Dyn. Change **6**, 199–234.

Wiese, R. (**1984**). "Language production in foreign and native languages: Same or different?," in *Second Language Productions*, edited by H. W. Dechert, D. Möhle, and M. Raupach (Gunter Narr Verlag, Tübingen), pp. 11–25.

Wong, S. G. J., and Papp, V. (**2018**). "Transferability of non-lexical hesitation markers across languages: Evidence from te reo Māori-English Bilinguals," in *Proceedings of the 26th IAFPA*, Huddersfield, UK, pp. 35–36.