JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Long-term within-speaker consistency of filled pauses in native and non-native speech

Meike M. de Boer,[1,a] Hugo Quené,[2,b] and Willemijn F. L. Heeren[1,c]
[1]Leiden University Centre for Linguistics, Leiden University, Reuvensplaats 3-4, 2311 BE Leiden, The Netherlands
[2]Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands
m.m.de.boer@hum.leidenuniv.nl, h.quene@uu.nl, w.f.l.heeren@hum.leidenuniv.nl

Abstract: Filled pauses are widely considered as a relatively consistent feature of an individual's speech. However, acoustic consistency has only been observed within single-session recordings. By comparing filled pauses in two recordings made >2.5 years apart, this study investigates within-speaker consistency of the vowels in the filled pauses *uh* and *um*, in both first language (L1) Dutch and second language (L2) English, produced by student speakers who are known to converge in other speech features. Results show that despite minor within-speaker differences between languages, the spectral characteristics of filled pauses in L1 and L2 remained stable over time.

[Editor: Anders Lofqvist] https://doi.org/10.1121/10.0009598

Received: 27 November 2021 **Accepted:** 28 January 2022 **Published Online:** 4 March 2022

1. Introduction

The filled pauses *uh* and *um* are considered useful features in forensic speaker comparisons because their (spectral) characteristics are highly speaker specific (e.g., Hughes *et al.*, 2016). Speaker specificity means not only that there is variation between different speakers but also that a feature is rather consistent in an individual's speech. Within the forensic context, where one or more anonymous recordings are compared with reference material made during a separate occasion, it is important to establish within-speaker consistency across speech tasks or recording sessions. So far, for filled pauses, this has been demonstrated in terms of their frequency of occurrence (Goldman-Eisler, 1961, p. 24), the proportions filled to silent pauses and *uh* to *um* (Künzel, 1997), the extent to which a speaker uses alternative pausing strategies such as word-final lengthening (McDougall and Duckworth, 2017), and the filled pauses' fundamental frequency (Braun and Rosin, 2015). In addition, the vowels used in filled pauses have been described as acoustically consistent within speakers in a single recording (e.g., Hughes *et al.*, 2016; Künzel, 1997). But are speakers' filled pause vowels also acoustically consistent across recordings made weeks or months apart? In this paper, we aim to answer that question, using recordings from the same speakers made >2.5 years apart.

Despite the lack of empirical evidence across recordings, several explanations have been offered for the presumed within-speaker consistency of filled pause acoustics (see also Hughes *et al.*, 2016). First, during the production of the mid-central filled pause vowel (McDougall and Duckworth, 2017), the articulators are configurated in a neutral position that is different from speaker to speaker (e.g., Swerts, 1998). Second, there is limited coarticulation because a silent pause typically precedes and/or follows a filled pause (Hughes *et al.*, 2016). Third, being produced with a prolonged duration compared with lexical vowels, filled pause vowels have a relatively stable part in the middle (Shriberg, 2001, p. 165), further reducing coarticulation effects. Finally, filled pauses are produced with relatively little conscious cognitive control (Jessen, 2008, p. 690), which is used as an argument for the prediction that the within-speaker consistency of filled pauses may even hold during voice disguise (Hughes *et al.*, 2016; McDougall and Duckworth, 2017).

According to Clark and Fox Tree (2002), multilingual speakers may also be consistent in their filled pause realizations across their languages. This assumption has been tested in prior work (de Boer and Heeren, 2020) among 58 speakers (a subset of whom is used in the current paper) speaking in their first language (L1) Dutch and second language (L2) English. A control experiment presented in this study showed that these languages both have *uh* and *um* with a vowel described as schwa, but it is realized with a 30–40-Hz higher first formant (F1), i.e., more open, by L1 English speakers than L1 Dutch speakers (see de Boer and Heeren, 2020). The sequential bilinguals from our dataset had somewhat higher F1 and lower second formants (F2) in their L2 English-filled pauses than in their L1 Dutch counterparts (de Boer and

[a]Author to whom correspondence should be addressed, ORCID: 0000-0003-0161-9115.
[b]ORCID: 0000-0001-7988-1346.
[c]ORCID: 0000-0001-7124-027X.

Heeren, 2020). Other characteristics, i.e., duration, fundamental frequency, and formant 3 (F3), did not differ between L1 and L2. Thus, filled pauses seem to be partly language dependent within speakers (see also Lo, 2020, on simultaneous bilinguals). According to Rose (2017), the extent to which L1 Japanese speakers adapted their filled pauses in their L2 English compared with their L1 depended on their L2 proficiency: The higher the proficiency, the more language dependent the filled pauses were. Thus, proficiency and ongoing practice in speaking in L2 may affect filled pause realization in that language.

To test whether filled pauses are as consistent over time within speakers as is implied in the phonetic literature, we investigated in the current study filled pause realizations in speakers' L1 Dutch and L2 English in two noncontemporaneous recordings made >2.5 years apart. In a previous study, we analyzed filled pause acoustics across languages of 58 speakers in one moment in time (de Boer and Heeren, 2020). For the current study, we used recordings from a subset of 25 of the same speakers who had been recorded at two moments in time. The speakers were students at University College Utrecht (UCU), a Dutch liberal arts and science college (Orr and Quené, 2017). UCU is a tight, relatively closed community of students from different language backgrounds who use English as a *lingua franca*. Only a small minority of the students speaks English as a native language; L1 Dutch speakers are the majority (Orr and Quené, 2017). Previous studies have shown that the speakers in this multilingual community seem to converge toward a shared accent of English. Whereas most of these highly proficient L2 English speakers showed cross-linguistic contrasts already in year 1, e.g., with a higher center of gravity (CoG) for /s/ in English than in Dutch, after >2.5 years on campus, students' English /s/ pronunciations have become more similar to one another (Quené *et al.*, 2017), as has their speech rhythm (Quené and Orr, 2014). In the current study, we aimed to investigate to what extent this particular linguistic environment may affect the within-speaker consistency of filled pauses in L1 Dutch and L2 English. If these speakers, who show convergence in other phonetic features, stay consistent in their filled pause realizations over 2.5 years, this would suggest that filled pauses are indeed highly stable within speakers.

Because of the convergence that seems to be taking place in the speakers' L2 English, we expect changes over time to be more likely in English than in Dutch. In addition, if the filled pause realizations in the L2 have changed over time, we expect the direction of this shift to be further away from the L1 because the L2 proficiency is expected to have increased after a period of >2.5 years.

2. Materials and methods

2.1 Speaker characteristics

To investigate the within-speaker consistency of filled pauses in the context of convergence, 25 native Dutch female UCU students were selected from the Database of the Longitudinal Utrecht Collection of English Accents (D-LUCEA; see Orr and Quené, 2017). To be admitted into UCU, students are selected partly on the basis of their above-average English-language proficiency. The first recordings were made within 6 weeks of arrival at UCU, when the speakers were a mean age of 18.4 years (s.d., 0.8). The same students were recorded again for several times over a period of 2 years and 8 months until the end of their third year at UCU. For the current study, to allow maximum convergence to have taken place, the first and final recordings were used, recorded >2.5 years apart.

2.2 Recordings

For this study, we used semispontaneous informal monologues in L1 Dutch and L2 English of 2 minutes per language. The recordings were made in a quiet, furnished office using a close-talking microphone (Sennheiser HSP 2-EW) attached to a headset. The speech was recorded digitally (44.1 kHz, 16 bits) using a Focusrite Saffire Pro 40 multichannel preamplifier and analog-to-digital (A/D) converter (Quené *et al.*, 2017). During the monologue, although not involved in an interactional conversation, speakers directed their speech toward an interlocutor who could understand them in both Dutch and English. Filled pauses *uh* and *um* occur in monologues (Clark and Fox Tree, 2002; Swerts, 1998). Clark and Fox Tree (2002) found no systematic differences in numbers of occurrences in monologues relative to dialogues (p. 93, Table 2). In addition, we see no reason to assume that the acoustic realization of filled pauses will differ by discourse type, although their position in an utterance is likely to affect their acoustics (Swerts, 1998; de Boer and Heeren, 2020).

2.3 Segmentation and acoustic measurements

In total, the speakers produced 1656 filled pauses (see Table 1),[1] with 10–121 tokens per speaker (mean, 66; s.d., 27). Because of the focus on phonetic measurements, only the most common filled pause types *uh* and *um* were included, and related, but sparser phenomena, such as vowel lengthening, lexical fillers, and nasal-only filled pauses, were excluded. In Dutch, as expected on the basis of previous studies (e.g., de Leeuw, 2007), speakers used *uh* more often than *um*; in L2 English, their *uh*:*um* proportions were more equal (see for a discussion de Boer and Heeren, 2020). The filled pauses *uh* and *um* were segmented manually by two coders using Praat speech analysis software (Boersma and Weenink, 2016), and the vowels of *um* tokens were segmented from the nasal. Segmentation was based on oscillogram and spectrogram, and was supported by repeated listening using a Focusrite Scarlett 2i4 audio interface and beyerdynamic DT 770 PRO headphones.

Table 1. Overview of the number of filled pause tokens (*uh*, *um*) in the dataset and their distribution over the languages (L1 Dutch, L2 English) and recording sessions (year 1, year 3).

	Year 1		Year 3		Total	
	uh	*um*	*uh*	*um*	*uh*	*um*
L1 Dutch	320	161	260	156	580	317
L2 English	212	212	156	179	368	391
Total	532	373	416	335	948	708
Total *uh* + *um*	905		751		1,656	

The filled pause vowels were analyzed in Praat on fundamental frequency (F0) and the first three vowel formants (F1, F2, F3). The mean F0 was measured in hertz over the full duration of the vowel within a 100–350-Hz range, using an autocorrelation method implemented in Praat. Previous studies have shown the F0 to be highly stable across a filled pause's duration (e.g., Swerts, 1998). All values <150 Hz were checked auditorily to see whether any octave errors were made in pitch estimation. If it was not heard as an extremely low F0, then the measurement was considered an octave error, and the F0 value was doubled (n = 23). Vowel formants were measured in hertz over the central 50% of each vowel's duration, using the Burg method (window length, 25 ms; ceiling, 3500 Hz). For the analysis, measurements were transformed to psychoacoustical scales, thus modeling auditory evaluation of the data. Formants were converted from hertz to bark using equation 6 from Traunmüller (1990), and F0 to semitones related to the grand mean. The data can be found in Supplementary Material 1.[2]

2.4 Statistical analysis

The data were analyzed with Bayesian mixed-effects modeling (linear mixed model [LMM]), using the *bmrs* package (Bürkner, 2017; Bürkner, 2018) in R (R Core Team, 2020). Bayesian LMMs allow evaluation of the strength of evidence in favor of or against the null hypothesis (H0) (see, e.g., Nicenboim and Vasishth, 2016). Prior distributions were sampled

Table 2. Summary of Bayesian mixed-effects models for *uh* and *um*, with means and highest density intervals (HDIs) of posterior distributions of significant coefficients (for F0, in semitones relative to overall mean; for formants, in bark), and with Bayes factors (BFs).

		F0	F1	F2	F3
uh					
Intercept	Mean	0.08	5.9	11.7	15.1
	HDI	−0.55, 0.73	5.65, 6.15	11.57, 11.92	14.92, 15.24
Year 3	Mean	⋯	⋯	⋯	⋯
	HDI	⋯	⋯	⋯	⋯
	BF	0.188	0.022	0.015	0.025
English language	Mean	⋯	0.2	−0.1	⋯
	HDI	⋯	0.04, 0.32	−0.25, −0.02	⋯
	BF	0.144	0.323	0.167	0.012
Year 3 English language	Mean	⋯	0.2	⋯	⋯
	HDI	⋯	0.02, 0.39	⋯	⋯
	BF	0.255	0.184	0.022	0.029
um					
Intercept	Mean	−0.07	6.1	11.6	15.0
	HDI	−0.76, 0.62	5.84, 6.39	11.42, 11.80	14.81, 15.18
Year 3	Mean	⋯	⋯	⋯	⋯
	HDI	⋯	⋯	⋯	⋯
	BF	0.036	0.030	0.016	0.031
English language	Mean	⋯	0.2	⋯	⋯
	HDI	⋯	0.11, 0.39	⋯	⋯
	BF	0.031	2.35	0.029	0.078
Year 3 English language	Mean	⋯	⋯	⋯	⋯
	HDI	⋯	⋯	⋯	⋯
	BF	0.053	0.019	0.053	0.098

44 000 times (in four independent chains of 11 000 samples each, of which the first 1000 warm-up samples were discarded so that 40 000 samples remained), using No-U-Turn sampling (NUTS) (Bürkner, 2018) to further avoid dependencies within each chain. The prior distributions were chosen on the basis of acoustic–phonetic theory: For schwa-like vowels spoken by females, *a priori*, one would expect the F1, F2, and F3 midformant frequencies to be ~5.3, 11.7, and 15.1 bark, respectively (these values correspond with 550, 1650, and 2750 Hz, respectively). We assume a gaussian distribution of formant frequencies in the bark space, with the center frequencies given above and with large standard deviations of 2, 1.5, and 1 bark, respectively. For F0, we assumed a gaussian distribution in the semitone space (centered at 0 semitones relative to the grand mean F0); a standard deviation of 4 semitones was assumed (Hincks, 2004). Using these weakly informative priors, 1000 prior predictive distributions were sampled for each formant and F0 and plotted in a scattergram (using the mean and standard deviation of the prior predictive distribution as coordinates). For each of the three formants and F0, the observed mean and standard deviation fell in the most closely filled area of these scattergrams (Supplementary Material 2).[2] This confirms the validity of the priors used. Analyses of formant frequencies converted to bark units (as reported in Sec. 3) and of the same measurements in hertz (reported in the Supplementary Material only) showed highly similar patterns.

We built Bayesian LMMs that included random intercepts and random slopes for speakers. Although only the vowel segments were analyzed, separate models were built for *uh* and *um* vowels because filled pause type may affect spectral characteristics (de Boer and Heeren, 2020). The fixed parts of the Bayesian LLMs (F0, F1, F2, F3) × (*uh*, *um*) were summarized in two ways. First, the 95% highest density interval (i.e., the smallest interval containing 95% of the posterior distribution) was calculated using the *logspline* package (Kooperberg, 2020). This was preferred over the default credibility interval (CrI) (i.e., over the Q2.5–Q97.5 interval of the posterior distribution). Second, we calculated the Bayes factor (BF) using the *bayestestR* package (Makowski *et al.*, 2019). In interpreting the BF, values of ~1 indicate no preference in favor of or against H0, large values (>10) provide strong evidence against H0, and small values (i.e., <1/10) indicate that there is strong evidence in favor of H0 over the alternative hypothesis (Kass and Raftery, 1995).

3. Results

Results showed that F0 and F3 of the speakers' filled pause vowels have not changed after 3 years of being immersed in an English-speaking environment either in L1 or in L2. This is illustrated by the summaries of the Bayesian LMMs in Table 2 (see also speakers' average values in Supplementary Material 2). The very small Bayes factors for F0 and F3 measurements indicate very strong support in favor of H0, without any effects of time (year 3) and without an interaction between language and time; this is also indicated by the ellipses in Table 2.

F1 and F2 are summarized in Fig. 1. The panels show no systematic differences between languages within the same recording (points are scattered around the diagonal) or between subsequent recordings within languages (trajectories vary in direction and length). This is confirmed by the Bayesian LMMs in Table 2. For F1, BFs show that there were no changes over time (with BF = 0.02–0.2 in favor of H0). For *uh* tokens, there was some evidence that F1 was the same for L1 and L2 in year 1 (with BF = 0.323 in favor of H0), and remained so in year 3 (with BF = 0.184 in favor of H0). For *um* tokens, however, there was some evidence that F1 was higher for L2 than for L1 in year 1 (with BF = 2.35 in favor of H1), with this difference remaining the same in year 3 (with BF = 0.019 in favor of H0); this is visible in Fig. 1 (lower left panel) by most points being to the right of the diagonal. Although the evidence is weak at best, this may suggest a somewhat more open articulation of the *um* vowel in L2 English than in L1 Dutch, within speakers and across years. For *uh* tokens, the posterior distribution suggests that F2 may be slightly lower for L2 than for L1 in year 1, whereas the BF suggests no difference.

Crucially, all these results indicate that over time, the filled pauses remained stable both in L1 and in L2. Full results are provided in Supplementary Material 3.[2]

4. Discussion and conclusion

Apart from a very small language effect on F1, spectral characteristics of filled pauses were remarkably stable across the speakers' languages of Dutch and English and across time. Only for *um*, speakers had a slightly more open pronunciation in L2 English than L1 Dutch. The absence of an effect of time in L1 shows that individual speakers realize their filled pauses quite consistently, even in noncontemporaneous sessions recorded >2.5 years apart. This confirms the idea that the within-speaker consistency of filled pauses is high, as was already widely assumed for L1 speech (e.g., Hughes *et al.*, 2016; Künzel, 1997). These findings are promising for forensic speaker comparisons, where noncontemporaneous recordings are inherent.

Even in L2, in which this speaker population has converged toward a shared English accent on /s/ and in speech rhythm after 2.5 years (Quené and Orr, 2014; Quené *et al.*, 2017), filled pauses remained fairly stable over time. Despite ongoing learning from perceiving and producing L2 English extensively over a period of >2.5 years, the speakers did not show significant changes in their filled pause realizations in L2 over time. This finding suggests that speaker-specific characteristics of filled pause realizations may be largely robust against convergence processes among a group of speakers. A possible explanation for this may be found in the relatively high level of inconsistency in language input for filled pauses. Articulatory freedom for filled pauses seems relatively high, so the between-speaker variation is higher than that of lexical vowels (e.g., Hughes *et al.*, 2016). This variation in filled pause language input may delay the acquisition of language-specific realizations in L2 English, especially in an environment with only a few native speakers.

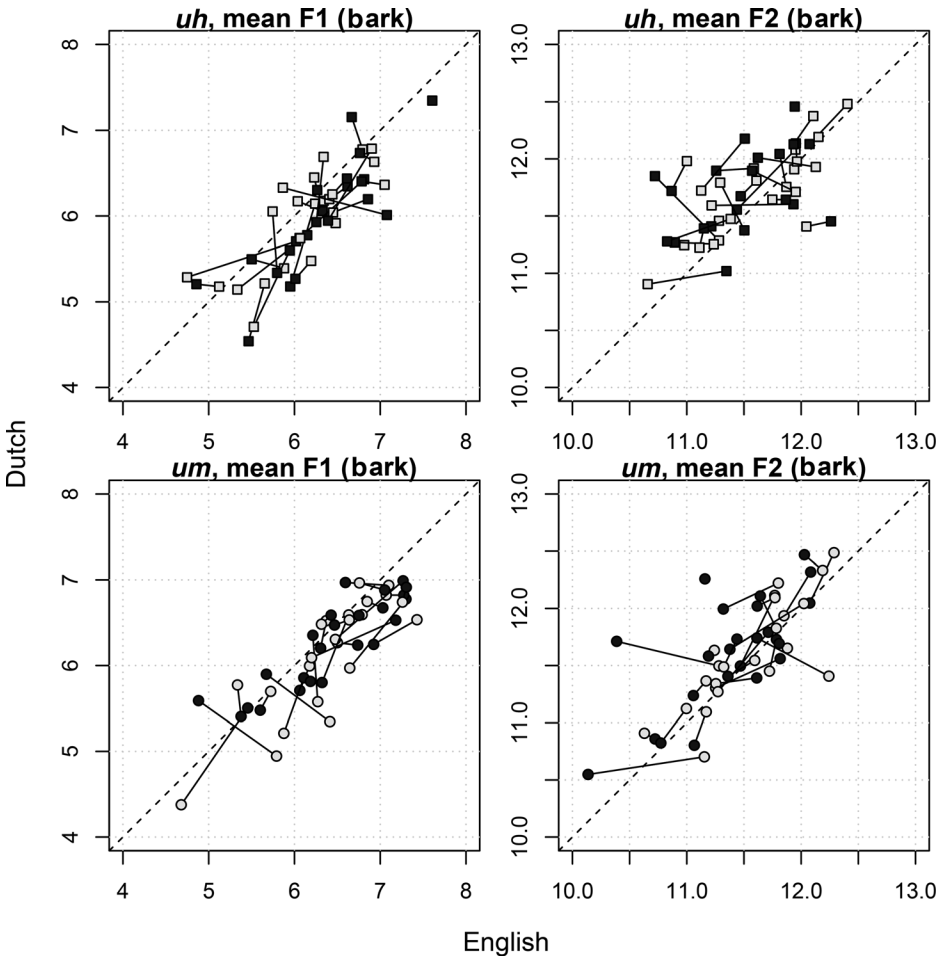

Fig. 1. Average F1 (left) and F2 (right) frequency in bark, for filled pauses *uh* (top) and *um* (bottom) in English and in Dutch, for each speaker and recording separately. Light symbols are for recordings in year 1 and dark symbols for recordings in year 3. The dotted diagonal indicates equal values in both languages.

Despite filled pauses being language specific (e.g., Candea *et al.*, 2005; de Boer and Heeren, 2020), most spectral characteristics were consistent within speakers even across languages, especially for *uh*. This again shows the value of filled pauses for forensic case work, where speech materials may be found in more than one language (van der Vloed *et al.*, 2014). This finding is discussed more elaborately in previous work (de Boer and Heeren, 2020).

The finding that filled pauses remain stable over a period of >2.5 years in a tight, relatively closed, and phonetically converging language community corroborates the claim that *uh* and *um* are a consistent feature of speakers' speech and, hence, that these filled pauses may be very useful for forensic speaker comparisons.

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 and Rosemary Orr for inspiration.

References and links

[1]The critical reader may notice that the speakers used fewer filled pauses in their L2 than L1, even though in both languages they spoke for approximately 2 min. Although this may seem counterintuitive, it may be partly explained by the order of the speech tasks in which the L2 English monologue almost always followed the L1 Dutch one (see de Boer and Heeren, 2020).

[2]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0009598 for the acoustic measurements of *uh* and *um*, the data preparation, scatter plots of speaker averages, the definition and validity check of the priors, and the Bayesian multilevel analyses.

Boersma, P., and Weenink, D. (**2016**). "Praat: Doing phonetics by computer (version 6.1.10) [computer program]," http://www.praat.org (Last viewed 2/17/2022).

Braun, A., and Rosin, A. (**2015**). "On the speaker-specificity of hesitation markers," in *Proceedings of the 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK, pp. 10—14, https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0731.pdf.

Bürkner, P. C. (**2017**). "brms: An R package for Bayesian multilevel models using Stan," J. Stat. Softw. **80**, 1–28.

Bürkner, P. C. (**2018**). "Advanced Bayesian multilevel modeling with the R package brms," R Journal **10**, 395–411.

Candea, M., Vasilescu, I., and Adda-Decker, M. (**2005**). "Inter- and intra-language acoustic analysis of autonomous fillers," in *Proceedings of the 5th Disfluency in Spontaneous Speech Workshop*, September 10–12, Aix-en-Provence, France, pp. 47—51. http://disfluency.org/app/download/29670926/DiSS2005_Proceedings.pdf.

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

de Boer, M. M., and Heeren, W. F. L. (**2020**). "Cross-linguistic filled pause realization: The acoustics of *uh* and *um* in native Dutch and non-native English," J. Acoust. Soc. Am. **148**, 3612–3622.

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

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

Hincks, R. (**2004**). "Standard deviation of F0 in student monologue," in *Proceedings, FONETIK 2004*, May 26–28, Stockholm, Sweden, https://www.speech.kth.se/prod/publications/files/1038.pdf.

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

Jessen, M. (**2008**). "Forensic phonetics," Lang. Ling. Compass **2**, 671–711.

Kass, R. E., and Raftery, A. E. (**1995**). "Bayes factors," J. Am. Statistical Ass. **90**, 773–795.

Kooperberg, C. (**2020**). "logspline: Routines for logspline density estimation, R package version 2.1.16 [computer program]," https://CRAN.R-project.org/package=logspline (Last viewed 1/21/2022).

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

Lo, J. J. H. (**2020**). "Between *Äh(m)* and *Euh(m)*: The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals," Lang. Speech **63**, 746–768.

Makowski, D., Ben-Shachar, M., and Lüdecke, D. (**2019**). "bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework," J. Open Source Softw. **4**, 1541–1549.

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

Nicenboim, B., and Vasishth, S. (**2016**). "Statistical methods for linguistic research: Foundational ideas—Part II," Lang. Ling. Compass **10**, 591–613.

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, London), pp. 177–190.

Quené, H., and Orr, R. (**2014**). "Long-term convergence of speech rhythm in L1 and L2 English," Proc. Speech Prosody **7**, 342—345, see https://www.isca-speech.org/archive_v0/SpeechProsody_2014/pdfs/58.pdf.

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. Am. **142**, EL519–EL524.

R Core Team (**2020**). "R: A language and environment for statistical computing [computer program]," https://www.R-project.org (Last viewed 1/21/2022).

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

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

Traunmüller, H. (**1990**). "Analytical expressions for the tonotopic sensory scale," J. Acoust. Soc. Am. **88**, 97–100.

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 Odyssey: The Speaker and Language Recognition Workshop*, June 16–19, Joensuu, Finland, pp. 6–13.