Psychological Research (2020) 84:547–557
https://doi.org/10.1007/s00426-018-1073-9

ORIGINAL ARTICLE

Sumacku or Smack? The value of analyzing acoustic signals when investigating the fundamental phonological unit of language production

Rinus G. Verdonschot[1] · Hinako Masuda[2]

Received: 23 March 2018 / Accepted: 6 August 2018 / Published online: 3 September 2018

Abstract
An ongoing debate in the speech production literature suggests that the initial building block to build up speech sounds differs between languages. That is, Germanic languages are suggested to use the phoneme, but Japanese and Chinese are proposed to use the mora or syllable, respectively. Several studies investigated this matter from a chronometric perspective (i.e., RTs and accuracy). However, a less attention has been paid to the actual acoustic utterances. The current study investigated the verbal responses of two Japanese–English bilingual groups of different proficiency levels (i.e., high and low) when naming English words and found that the presence or absence of vowel epenthesis depended on proficiency. The results indicate that: (1) English word pronunciation by low-proficient Japanese English bilinguals is likely based on their L1 (Japanese) building block and (2) that future studies would benefit from analyzing the acoustic data as well when making inferences from chronometric data.

Introduction

In psycholinguistic theories on language production (e.g., Caramazza, 1997; Dell, 1986; Levelt, Roelofs, & Meyer, 1999), it is typically assumed that the phoneme is the fundamental phonological unit (FPU) underlying speech production. In these theories, the FPU is meant to reflect the unit used in the segment-to-frame (or prosodification) process to fill the metrical frame. That is, according to Levelt et al., (1999) to create the pronunciation for a word, specific phonemes must be assembled in their correct order and then must be put into their respective place in the syllabic structure of the utterance. For example, a word like "guitar" would be spelled out as /g/ /ɪ/ /t/ /ɑ:/ /r/ and these phonemes would be put at their correct position in the metrical structure of the word [i.e., /gɪ.ˈtɑ:r/; see Levelt et al., (1999) and Roelofs (2015)]. Several studies (e.g., Meyer, 1990, 1991) as well as linguistic phenomena, such as re-syllabification (e.g., "I read it again", usually spoken as: /aɪ/ /rɛ/ /dɪ/ /tə.gen/ instead of /aɪ/ /red/ /ɪt/ /ə.gen/), have been used to support this claim. However, recently, others have indicated that the FPU may not be similar between languages. For example, in Chinese, it is likely to be the syllable, and in Japanese, it is likely to be the mora (e.g., O'Seaghdha et al., 2010; Verdonschot et al., 2011, see also Roelofs, 2015).

In addition, a valid question would be: what is the FPU then with respect to bilinguals for which the L1 and L2 are proposed to differ in size (e.g., the syllable in Chinese but the phoneme in English)? This has been investigated for Chinese–English high-proficient bilinguals (Li, Wang & Davis, 2015; Verdonschot et al., 2013) and for Japanese–English high- and low-proficient bilinguals (Ida et al., 2015; Nakayama et al. 2016). For Chinese–English high-proficient bilinguals, it was found that the initial construction of speech sounds in these bilinguals' L2 is prepared like native speakers of English (Verdonschot et al., 2013). Interestingly, under certain conditions, they even showed that the smaller FPU for the L2 (English) could influence L1-naming (Chinese), though this finding awaits replication. Similarly, Li et al., (2017) using a picture naming task found that Chinese–English bilinguals could process sub-syllabic units, but only after several repetitions of the L2 (English) stimuli, no sub-syllabic effects were found within

✉ Rinus G. Verdonschot
rinusverdonschot@gmail.com

[1] Department of Oral and Maxillofacial Radiology, Institute of Biomedical and Health Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan

[2] Faculty of Science and Technology, Seikei University, Musashino, Japan

the L1 (Chinese). In addition, here, L2 proficiency was not manipulated as a factor.

For Japanese–English bilingual speakers, using masked primes and to-be-read-aloud target words (i.e., a production task with written words), it has been shown that L2 (English) onset effects (i.e., faster naming latencies through experimental manipulation of the overlapping onset between stimuli) can be obtained if bilinguals' proficiency level is sufficiently high (Ida et al. 2015, Exp. 1; Nakayama et al. 2016, Exp. 2). However, importantly, in Japanese, it has also been shown that, for low-proficient bilinguals, onset effects do not appear in their L2 (Nakayama et al., 2016, Exp.1). More specifically, in Nakayama et al. (2016), employing the masked priming paradigm, English primes and targets were administered to low- and high-proficient Japanese–English bilinguals (as well as monolingual native speakers). In this paradigm, participants needed to name a target (e.g., BENCH) which was preceded by either an (1) onset-related prime, e.g., bark, (2) an onset control, e.g., dark, (3) a mora/CV-related prime, e.g., bell, or (4) a mora/CV control, e.g., cell. In addition, an identity condition was added using a different set of targets (e.g., target: SOFT with identity prime soft vs. control). It was found that the identity and CV (mora)-related conditions gave rise to equal priming effects for all groups; however, only native monolingual English speakers and high—(but not low) proficient Japanese–English bilinguals showed onset priming effects. Nakayama et al. (2016) concluded that their high-proficient bilinguals used an FPU similar to monolingual native speakers. Furthermore, the size of the onset priming effect was correlated with the length of time spent in English-speaking countries, which suggests that extensive exposure to L2 phonology may play a key role in the emergence of a language-specific phonological unit in L2 word production. Interestingly, these authors also ran limited acoustic analyses on some of their participant's verbal responses and found evidence of the insertion of vowels (i.e., epenthesis) for low-proficient but not high-proficient groups. However, Nakayama et al. (2016) did not employ high-quality audio recordings, nor were their materials and experimental design created with such analyses in mind. Therefore, they were only able to run limited analyses on a small subset of their stimuli.

The current paper aims to extend upon Nakayama et al. (2016) by furthering the case for recording and analyzing acoustic information besides solely relying on chronometric measures. That is, we believe that it is sensible to not only focus on the time that it takes to initiate an utterance but also to investigate the actual utterance itself, especially when predictions are made about its structure and quality (e.g., how the FPU shapes the utterance). In addition, Nakayama et al. (2016) stated that their stimuli were not specifically designed to investigate the epenthesis issue as, for example, the likelihood of vowel insertion varies between consonant clusters [which is most pronounced for consonant clusters which contain voiced stops; see Nakayama et al. (2016; p8)] and that additional acoustic analyses are necessary to corroborate their findings. We believe that future studies contributing to the ongoing debate on the underlying fundamental unit of speech production would benefit from investigating the acoustic signal. We propose that this extra source of information will provide additional insights into the inner workings of the language production system. Here, we specifically focus on the case of vowel insertion (epenthesis) in Japanese bilinguals when speaking English.

In the perception of speech literature, it has been found that Japanese native speakers are very likely to perceive and produce epenthetic vowels between consonants. For instance, Dupoux, Kakehi, Hirose, Pallier, and Mehler (1999; see also Masuda & Arai, 2010) found that, in contrast to French participants, Japanese participants heard illusory vowels perceived between consonants. For example, in one experiment, the non-word "ebzo" was perceived as "ebuzo", and in another experiment, Japanese participants could not distinguish between "ebzo and ebuzo" contrasts (though French participants could). Similarly, Masuda and Arai (2010) conducted perception and production experiments using Dupoux et al.'s non-word list (target: word-medial consonant clusters) and found that Japanese participants were generally likely to perceive vowels between consonants, and those with lower English proficiency had a higher tendency of doing so. In their production experiment, participants' epenthesis was categorized into three epenthesis degrees: full epenthesis, partial epenthesis, and no epenthesis. Their acoustical measurements revealed that participants with lower English proficiency produced a higher degree of epenthesis compared to those with higher English proficiency. Furthermore, they also found that consonant voicing had a stronger effect in the lower proficiency group than the higher group (i.e., voiced consonants evoked more epenthesis in lower proficiency group). In addition, Dehaene-Lambertz, Dupoux, and Gout (2000) have shown that, in an auditory oddball task (a task in which a frequent stimulus, e.g., "ebzo" is occasionally replaced by a deviant stimulus, e.g., "ebuzo"), the underlying mismatch negativity brain potential indexing the neuronal detection of the deviant is absent for Japanese in "ebuzo" but present for French participants. These authors speculated that fast uncovering of a phonological representation might help to reconstruct the actual/intended utterance in case of a noisy environment (or mispronunciation). That is, when Japanese would detect the phototactically implausible cluster "ebzo", they would reconstruct it into "ebuzo" (i.e., more likely in Japanese).

On the production side, native speakers of Japanese are well known for producing epenthetic vowels in non-native words to repair the syllabic structure to match that of Japanese (e.g., producing vowel epenthesis to avoid syllables

to end with a consonant, i.e., breaking consonant clusters into CV sequences, as well as after word-final consonants, again to create a CV sequence). Concerning the epenthesis for Japanese speaking English, one usually observes that /u/ is inserted after consonants (as it has the shortest intrinsic duration among the five Japanese vowels with low sonority; Kubozono, 1999). For example, "brave" likely becomes /bureɪbu/, though notable exceptions are alveolar stops (t, d) and palato-alveolar affricates (tʃ, dʒ) in which the ensuing vowels would be /o/ (as t and d become different allophones before high vowels) and /i/ (because its similar place of articulation), respectively (see Masuda & Arai, 2010). This substantiates Nakayama et al.'s (2016) claim that their low-proficient group likely has been adhering to their L1 unit (mora) when speaking in their L2 (English). However, as Nakayama et al. (2016) did not have their stimuli specifically designed for the detection of vowel epenthesis (and only analyzed a limited acoustic set), they proposed that a more detailed study would be necessary to investigate this phenomenon.

The current study has undertaken this task. We predict that, if Nakayama et al. (2016) were, indeed, correct, then low-proficient Japanese–English bilinguals are more likely to exhibit vowel epenthesis for L2 English words to adhere to their L1 phonotactic structure (compared to high-proficient bilinguals) to repair a non-native sequence (i.e., the focus is on vowel epenthesis, including but not limited to consonant clusters). Furthermore, it is unclear whether the position affects the frequency of epenthesis, an answer which could possibly benefit the understanding of mechanism in non-native speech production as well as for English education. There are three reasons why investigating epenthesis position is warranted: (1) although epenthesis can occur in multiple positions, few studies in the area of non-native speech production have carried out thorough analyses on its effect on production, (2) to verify which position most likely elicits epenthesis leads to better understanding of the difficulties Japanese speakers face when producing illegal sequences, and, finally, (3) if there are, in fact, differences among frequencies of epenthesis depending on position, this will be useful information that likely benefits English education. This study, therefore, took the following three positions into consideration: word-initial consonant cluster (e.g., strike), word-medial consonant cluster (e.g., instead), and word-final singleton consonant (e.g., sip). Furthermore, the current experiment also carried out the task of using real words, in contrast to Dupoux et al. and Masuda & Arai's study which used non-words.

We have carried out an acoustic analysis on the responses to these stimuli given both by high- and low-proficient bilinguals which are drawn from the same student population as Nakayama et al. (2016) and expected to see larger effects of epenthesis for the low-proficient compared to the high-proficient bilingual group irrespective of epenthesis position.

Experiment

Method

Participants Twenty low- and twenty high-proficient JP-ENG bilingual speakers studying at Waseda University (i.e., the same student population as in Nakayama et al., 2016) participated in the experiment in return for 1000 yen (9 USD). None had taken part in the Nakayama et al. (2016) study. Table 1 shows their bilingual aptitudes as assessed using a questionnaire. We found that all variables except "TotYrStudEng" (i.e., Total Years Studied English) and AoAEng (Age of Acquisition for English) were significantly different between LPB and HPBs.

Materials We selected 75 target words (see Appendix A, Table 4) which were divided into three potential epenthesis sites, i.e., the onset (e.g., "smack"), medial (e.g., "increase"), and final (e.g., "sob") positions. Word frequencies taken from Brysbaert & New (2009) indicated no differences between groups ($F[3,144] = 1.4$, $p = .24$). We used a wide range of consonant clusters and consonant coda endings to allow any outcome not to be generalized to any specific set. We avoided obvious loanwords (e.g., 'drug') as they might trigger the Japanese, mora-based, utterance [e.g.,ドラッグ;

Table 1 Participant (group) information

Variable	LPB Mean (SD)	HPB Mean (SD)
Age (year)	21 (2)	22 (5)
Gender (F/M)	5/15	8/12
TOEIC	463 (65)	874 (56)
AoAEng (year)	11.0 (2.0)	10.8 (2.5)
TotYrStudEng (year)	9.7 (2.8)	9.6 (4)
SchYrAbroadEng (month)	0 (0)	11 (11)
LiveAbroadMonth	0 (0)	17.8 (2.5)
PropJapEng (in %)	96/4	80/20
EngRead (1–10)	5.0 (1.8)	7.0 (1.8)
EngWrite (1–10)	4.3 (1.9)	7.1 (1.0)
EngSpeak (1–10)	3.5 (2.1)	6.8 (1.2)
EngList (1–10)	4.0 (2.1)	7.3 (1.1)

LPB low-proficient bilinguals, *HPB* high-proficient bilinguals, *TOEIC* Test of English for International Communication, *AoAEng* Age of Acquisition English, *TotYrStudEng* total years studying English in any capacity, *SchYrAbroadEng* studied English at school abroad, *SchYrAbroadEng* studied abroad in an English-speaking country, *LiveAbroadMonth* lived abroad in an English-speaking country, *PropJapEng* (proportion Japanese/English used daily, *EngRead/Write/Speak/List* self-assessed English, reading, writing, speaking, and listening ability

/doraggɯ/]. Stimuli were checked in a Japanese dictionary to ensure that they were not loanwords, except for "signature"; which is a loanword, but, since the word is rarely used in daily life, we included this in the stimuli list. In addition, we administered 75 filler words which did not have any obvious epenthesis site.

Apparatus, design, and procedure Participants were instructed to embed the target words in a carrier sentence "I say [target]". If a participant did not know a word, she/he had to still try naming it to the best of his/her abilities. After practicing with four target words, the experiment started. Stimuli were presented using PowerPoint and the timing was self-paced, though they had to name the words without delay and as accurately possible. Participants were placed in a soundproof booth and recorded using a digital sound recorder (Marantz Professional Solid-State Recorder PMD661) with a microphone (AudioTechnica AT4022) at a sampling frequency of 48 kHz. Two pseudo-randomized counterbalanced lists containing all target and filler words were created (each list totaled 150 to-be-named words) and equally distributed. There were two breaks at 1/3 and 2/3 of each list; after each break, the naming was always continued using two filler stimuli.

Analysis and results Each utterance from each participant was evaluated by the second author (HM: rater was blinded to the experimental group) using 'Praat' version 6.0.23 (Boersma & Weenink, 2016). The epenthesis judgment was binary (present or absent). An utterance was judged as "epenthesis present" if there were (1) clear first and second formants in the spectrogram, (2) a clear voice bar in the spectrogram, and (3) it was longer than 45 milliseconds (ms) in terms of duration. Although 45 ms seems to be a rather conservative threshold (as short /u/ in Japanese can be shorter; e.g., Beckman, 1982) it has been shown that short, vowel-like segments can appear "epenthesized" between (voiced) consonant clusters due to gestural mistiming (e.g., Davidson & Stone, 2003). In addition, in Masuda & Arai (2010), no epenthesis was reported to be shorter than 45 ms. The results of the analyses were later randomly double checked for confirmation.

Utterances with partially devoiced vowels were judged as being epenthesized, if the epenthesis was partially voiced with visible first and second formant traces and the duration was longer than 45 ms. An utterance was judged as "no epenthesis" if the duration of the vowel, if any, was under 45 ms, and if the vowel was devoiced with a little evidence of first and second formants (see Figs. 1 and 2 for examples).

While performing the analysis, we found that we accidentally used one word (i.e., "trash") in both onset + final conditions. We kept this word in the analysis and added any epenthesis count to its relevant position (i.e., "trash" epenthesis at the end was marked as "trash2").

First, all clearly mispronounced words were excluded from the analysis (84 out of 3000 responses or 2.8%; e.g., "*avertise" instead of "advertise").

Table 2 shows that for the total counts for words per condition, LPBs show much higher rates of epenthesis than HPBs. We subjected the data to a logistic mixed-effects analysis (Baayen, 2008; Jaeger, 2008; Janda, Nesset, & Baayen, 2010). This analysis allows for a likelihood estimation of the presence of epenthesis by transforming the "yes" and "no" counts (see Table 2 and Appendix B Table 5) into a log-odds ratio (see Appendix C Fig. 3 for the log-odds ratios for all 75 target words; see: Baayen, 2008, Janda et al. 2010). We obtained the values in Appendix C Fig. 3 by calculating the natural log for (Epenthesis: Yes + 1)/(Epenthesis: No + 1).

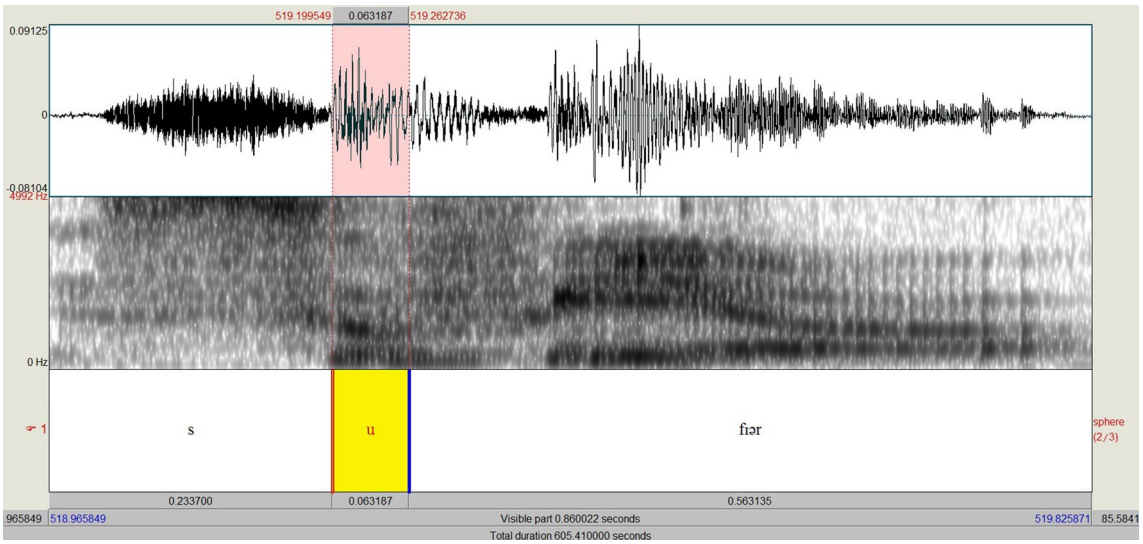

Fig. 1 Example of the utterance "sphere" with an epenthesis of 63 ms (/sufiər/)

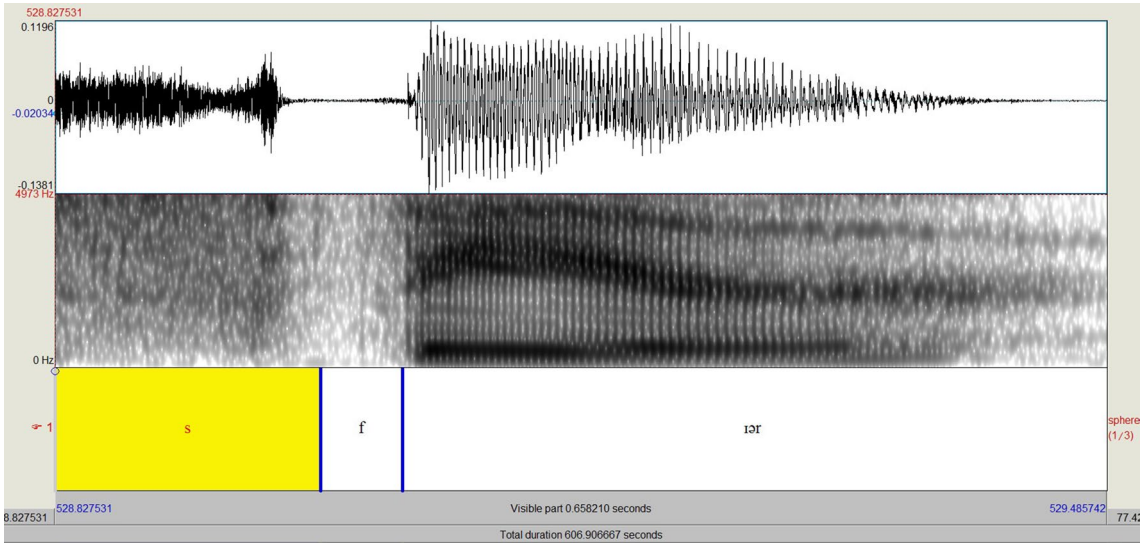

Fig. 2 Example of the utterance "sphere" without epenthesis (0 ms) (/sfɪər/)

Table 2 Total epenthesis counts for high-proficient bilinguals (HPB) and low-proficient bilinguals (LPB) per position in the word (onset, medial, and final)

Group	Epenthesis	Position in word			Total
		Onset	Medial	Final	
HPB	No	386	394	485	1265
	Yes	103	97	9	209
LPB	No	133	150	418	701
	Yes	348	329	64	741

Total represents respective yes or no counts per group across positions

Please note that we added + 1 to both our counts to avoid dividing by zero. For example: for smack (low-proficient group), this would be $ln((15+1)/(5+1)) = 0.99$ indicating a higher inclination for no epenthesis.

Initial position First, the data in Appendix C Fig. 3 show that high-proficient JP-ENG bilinguals were not immune to inserting epenthetic vowels. For example, for /dw/ as in 'dwarf', about 60% of high-proficient bilingual responses showed epenthesis (e.g., /dowarf/); this pattern also holds for /gr/ as in 'grab', though slightly less (e.g., /gurab/). We will return to this issue in the General Discussion. However, low-proficient bilinguals did show a much higher rate of epenthesis than high-proficient bilinguals. In this group, except for three specific consonant clusters (i.e., /sk/ 'scar', /sp/ 'spat' and /st/ 'stab'), all other clusters experienced a significant vowel epenthesis such as /fr/ in 'fridge'. The absence of epenthesis for /sp/, /st/, /sk/ clusters is most likely due to high vowel devoicing. Fricative–stop combinations such as these are devoiced categorically more so than fricative–fricative combinations (e.g., Fujimoto 2015), so it is not surprising that a few people epenthesized their vowels in this environment. Note that other /s/ + consonant stimuli (such as /sl/ 'slave', /sm/ 'smack', and /sn/ 'snail') *did* observe a significant epenthesis.

Medial position Although high-proficient bilinguals did, for particular consonant clusters, show signs of epenthesis; for example, /dv/ in 'advertise' might become /ædəvətaɪz/ and /gn/ in 'signature' might become /sɪgənətʃər/, the frequency was much less, compared to low-proficient bilinguals (see Table 2). In this group, nearly all stimuli (except those which were prone to devoicing such as /sk/ 'reschedule', /sp/ 'inspire', /st/ 'instead') showed large effects of epenthesis. There were even combinations for which there was no correct instance (i.e., all showed epenthesis) such as /gl/ 'burglar' and /gn/ 'signature'.

Final position As can be seen in Table 2, whereas high-proficient bilinguals hardly showed epenthesis, low-proficient bilinguals showed substantial vowel epenthesis with final consonants like /g/ and /l/, that is: words like 'fog' and 'fell' are more likely to become /foggu/ and /felu/ for the low-proficient bilingual group.

Overall It seems to be clear that when modeling the presence of epenthesis; we should not only consider the individual's differences (i.e., proficiency level) but also the differences between the target words themselves. Accordingly, when applying a logistic mixed-effects model analysis, we used random intercepts for participants as well as target words to allow for modeling the overall patterns more accurately by allowing for exact intercept adjustments for each participant and each target word.

Logistic mixed-effects model analysis In all, 2609 data points were submitted to the logistic mixed-effects analysis

which excluded the mispronunciations as well as tokens which contained devoicing environments (i.e., the /sp/, /st/ and /sk/ clusters). In line with the recommendation to keep the random effect structure maximal (Barr, Levy, Scheepers, & Tily, 2013), the initial model included random slopes on participants and stimuli; the final model which we report was selected using a backward stepwise model selection procedure. Somewhat unexpected, none of the variables from the post hoc questionnaire (i.e., TOEIC, AoAeng, TotYrStudEng, SchYrAbroadEng, LiveAbroadMonth, PropEng, EngRead, EngWrite, EngSpeak, and EngList) when entered (as fixed factors) in the model contributed significantly. The optimal (converging) model was obtained by fitting a mixed model (binomial family) using the following formula: Epenthesis ~ group × position + (1|participant) + (1|word) in which the log-odds was modeled as a function of group (LBP, HPB) and position (onset, medial, and final) as well as *LiveAbroadMonthCentered in* combination with random intercepts for participants and words. Table 3 shows the coefficient estimates along with their *z* values, associated *p* values, and standard errors.

As can be seen, there is a clear difference between HPBs and LPBs with the estimate for HPBs, indicating that less epenthesis occurred for this group. In addition, there are differences between the three positions as indicated by the significant interaction, signifying that the final position exhibited significantly less epenthesis than the onset and medial positions.

General discussion

Recent studies have revealed that the fundamental unit underlying the construction of phonology is the phoneme in most European languages (e.g., English, Dutch; Roelofs, 2015), the syllable in Mandarin Chinese (Chen et al., 2002; O'Seaghdha et al., 2010), and the mora in Japanese (Kureta et al., 2006; Verdonschot et al., 2011). However, what then of bilinguals learning a second language which differs in the fundamental phonological unit? Nakayama et al. (2016) using a masked priming task found that proficiency seems to matter and low-proficient Japanese–English bilinguals showed no onset phoneme priming but only mora and identity priming consistent with their L1 (Japanese) language. High-proficient bilinguals, however, also showed onset/phoneme priming showing a similar pattern to native speakers of English. Nakayama et al. (2016) proposed that this represented a change in unit due to increased proficiency level.

The current study examined a consequence of this proposition. That is, if low-proficient Japanese–English bilinguals really adhere to an L1 Japanese moraic system, then even when speaking in their L2, there should be a strong presence of epenthetic vowels. To this end, we administered 75 to-be-named words having consonant clusters (e.g., 'blast') which are likely to elicit epenthesis to participants (from the same student population as Nakayama et al., 2016). We also manipulated epenthesis location using three distinct positions, i.e., onset, medial, and final, as Japanese can have consonant clusters in medial positions (but not onset and final). We found that, consistent with Nakayama et al. (2016), low-proficient bilinguals, indeed, inserted far more epenthetic vowels than high-proficient bilinguals, likely because of their mora-based L1 (Japanese). Our data fit well with findings by Masuda and Arai (2010) who also found the differential effects of highly-proficient JP-ENG bilinguals vs. JP monolinguals on the production of consonant clusters. They also found higher occurrence of epenthesis, showing a profound influence of L1 on L2, but, interestingly, also that consonant voicing had a much stronger influence on production in monolinguals compared to bilinguals. This trend seems to also hold for the current paper.

The combination of voiced consonants seems generally more likely to exhibit epenthesis compared to voiceless consonants/clusters which is shown by the fact that even our high-proficient bilinguals were not immune. As we stated earlier, for 'dwarf', about 60% of high-proficient bilingual responses showed epenthesis (e.g., /dowarf/) and this pattern also held for /gr/ as in 'grab' (e.g., /gurab/), though slightly less. However, the combination of voiced consonants per se does not constitute a sufficient condition to exhibit a higher rate of epenthesis (as we had other words that also had that combination with high-proficient bilinguals not inserting vowels). We speculate that "dwarf" is perhaps less encountered by Japanese in general, considering that this combination (i.e., dw) does not frequently

Table 3 Mixed-effect model coefficients, standard errors, *z* values, and *p* values

	Estimate	Standard error	*z* value	*p* value
Intercept	2.2131	0.4177	5.298	$<0.001^{***}$
HPB (group)	−3.8478	0.4275	−9.000	$<0.001^{***}$
Medial (position)	−0.4489	0.4931	−0.979	0.328
Final (position)	−4.9963	0.4584	−10.132	$<0.001^{***}$
HPB:medial (interaction)	0.1797	0.4656	0.586	0.558
HPB:final (interaction)	1.0372	0.3066	2.228	0.05^{*}

appear in English. This could mean that less frequent clusters trigger L1 phonotactics, even for high-proficient bilinguals. However, that is not the whole story as "grab" (i.e., gr) also yielded a high rate of epenthesis, though it has a high-frequency cluster combination in English (e.g., grab, gray, green, grip, grow, etc.). This combination is likely to be familiar to Japanese; therefore, this observation needs a different explanation compared to "dwarf". Let us then compare the case of "grab" with "embrace", which has a similar consonant combination to "grab" (i.e., voiced stop + /r/). Here, high-proficient bilinguals are not inserting vowels. Looking at other br- or bl-clusters in the initial and medial positions, some words are epenthesized (e.g., blast, reliably), while others are not (e.g., embrace). In the case of /g/, all cluster combinations, although only three (i.e., glide, grab, and signature), showed a tendency of being epenthesized, compared to other types of clusters. These observations suggest that the consonant /g/ as the first consonant of the cluster may tend to exhibit epenthesis by Japanese speakers. Admittingly, this explanation is not complete and further investigation on these speculations including more native speakers' data is necessary to confirm its validity. Finding such patterns using acoustic analysis is important as it will provide insight on what specific patterns elicit L1 phonotactic adherence and which do not, and this will eventually enhance our understanding what is going on and how speech production processes in the (bilingual) brain take place.

Another angle to examine epenthesis is through sonority. According to Gouskova (2002), rising sonority in consonant clusters elicits internal epenthesis (e.g., English "fruit" becomes /fɪrut/ in Hindi, "glass" becomes /gelaʃ/ in Bengali), whereas falling sonority elicits external epenthesis (e.g., English "school" becomes /ɪskul/ in Hindi and /ɪʃkul/ in Bengali), and such patterns are proposed to exist across unrelated languages. This phenomenon is also discussed in Fleischhacker (2005) using the example of Hawai'ian Creole (English "plenty" becomes /puranti/; Nagara 1972, cited in Fleischhacker 2005). To confirm whether this pattern is also the case in Japanese learners of English, we examined the likelihood of epenthesis within consonant clusters according to the sonority scale (Carr & Montreuil 2013) and we found (though our design was unbalanced as we had many more rising sonority words than falling sonority words) that, indeed, words with rising sonority gave more rise to internal epenthesis ($z = 9.4$, $p < 0.001$).

In conclusion, we would like to recommend that future chronometric investigations which consider how bilinguals construct speech sounds on a fundamental level (besides focusing on reaction times and error data) should also take the acoustic signal itself into account. This source of information is extremely valuable and can reveal some of the underlying processes such as L1- to L2 unit transfer causing epenthesis in Japanese, and likely different phonological effects in other language combinations.

Funding This study was funded by a Grant-In-Aid (C) from the Japan Society for the promotion of Science (No. 17K02748) to Rinus G. Verdonschot.

Compliance with ethical standards

Conflict of interest Rinus G. Verdonschot declares that he has no conflict of interest. In addition, Hinako Masuda declares that she has no conflict of interest.

Research involving human and/or animal participants All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Appendix 1

See Table 4.

Table 4 Stimuli

Position	CC	Word	Position	CC	Word	Position	CC/C	Word
Onset	sm	Smack	Medial	kr	Increase	Final	b	Sob
Onset	sn	Snail	Medial	fl	Reflex	Final	d	Cod
Onset	st	Stab	Medial	tr	Introduce	Final	f	Laugh
Onset	sw	Swam	Medial	pr	Approve	Final	g	Fog
Onset	sk	Scar	Medial	sp	Inspire	Final	j	Huge
Onset	sl	Slave	Medial	st	Instead	Final	k	Sock
Onset	sp	Spat	Medial	br	Embrace	Final	l	Fell
Onset	sf	Sphere	Medial	pl	Replace	Final	m	Fame
Onset	dw	Dwarf	Medial	kl	Include	Final	p	Nap
Onset	tw	Twice	Medial	fr	Refrain	Final	s	Loss
Onset	thr	Throat	Medial	sk	Reschedule	Final	t	Got
Onset	dr	Drain	Medial	sl	Enslave	Final	v	Cave
Onset	tr	Trash	Medial	shr	Enshrine	Final	z	Goes
Onset	kw	Quit	Medial	bl	Reliably	Final	th	Math
Onset	kr	Crime	Medial	gl	Burglar	Final	sh	Trash
Onset	kl	Clap	Medial	dv	Advertise	Final	b	Rub
Onset	pr	Praise	Medial	gn	Signature	Final	g	Slug
Onset	fr	Fridge	Medial	lt	Difficulty	Final	v	Starve
Onset	br	Brag	Medial	tl	Honestly	Final	f	Enough
Onset	gr	Grab	Medial	dn	Madness	Final	k	Bark
Onset	pl	Plead	Medial	pn	Deepness	Final	p	Sip
Onset	fl	Flood	Medial	sd	Wisdom	Final	s	Mess
Onset	bl	Blast	Medial	thl	Northland	Final	t	Bought
Onset	gl	Glide	Medial	tf	Doubtful	Final	sh	Posh
Onset	shr	Shrine	Medial	bd	Abduct	Final	z	Those

C consonant, *CC* consonant cluster

Appendix 2

See Table 5.

Table 5 Epenthesis counts for each presented item

Target	HPB		LPB		Target	HPB		LPB		Target	HPB		LPB	
	Stand	Epen	Stand	Epen		Stand	Epen	Stand	Epen		Stand	Epen	Stand	Epen
Onset (total)	386	103	133	348	Medial (total)	394	97	150	329	Final (total)	485	9	418	64
Blast	14	6	2	18	Abduct	14	3	3	15	Bark	20	0	19	0
Brag	18	1	5	15	Advertise	7	12	0	20	Bought	20	0	19	0
Clap	16	4	2	18	Approve	17	3	4	16	Cave	17	2	12	4
Crime	17	3	5	15	Burglar	11	7	0	15	Cod	19	1	17	2
Drain	14	6	2	18	Deepness	15	5	3	17	Enough	20	0	20	0
Dwarf	7	13	0	20	Difficulty	20	0	6	13	Fame	19	1	15	4
Flood	13	7	3	17	Doubtful	19	1	17	2	Fell	19	1	9	9
Fridge	19	1	3	15	Embrace	17	3	3	16	Fog	20	0	15	5
Glide	10	9	0	20	Enshrine	19	1	8	11	Goes	19	0	18	2
Grab	9	10	0	19	Enslave	18	2	3	16	Got	20	0	19	1
Plead	17	3	6	13	Honestly	17	3	4	14	Huge	20	0	13	5
Praise	15	2	9	11	Include	11	9	1	19	Laugh	20	0	17	0
Quit	18	1	6	7	Increase	14	6	1	19	Loss	20	0	20	0
Scar	19	0	19	0	Inspire	20	0	20	0	Math	20	0	20	0
Shrine	13	7	1	16	Instead	20	0	19	1	Mess	20	0	19	1
Slave	18	2	0	20	Introduce	15	5	5	15	Nap	19	0	20	0
Smack	15	5	4	16	Madness	18	2	4	16	Posh	20	0	20	0
Snail	20	0	1	19	Northland	19	1	5	15	Rub	19	1	16	3
Spat	20	0	20	0	Reflex	16	4	3	17	Sip	17	0	20	0
Sphere	16	1	9	8	Refrain	17	3	7	13	Slug	19	1	5	15
Stab	20	0	20	0	Reliably	8	9	3	15	Sob	19	1	15	5
Swam	14	6	5	15	Replace	14	6	5	15	Sock	20	0	18	1
Throat	12	8	3	16	Reschedule	20	0	15	1	Starve	20	0	18	1
Trash	19	1	4	16	Signature	8	12	0	19	Those	20	0	18	2
Twice	13	7	4	16	Wisdom	20	0	11	9	Trash2	19	1	16	4

HPB high-proficient bilinguals, *LPB* low-proficient bilinguals, *Stand* standard pronunciation, *Epen* epenthesized pronunciation

Appendix 3

See Fig. 3.

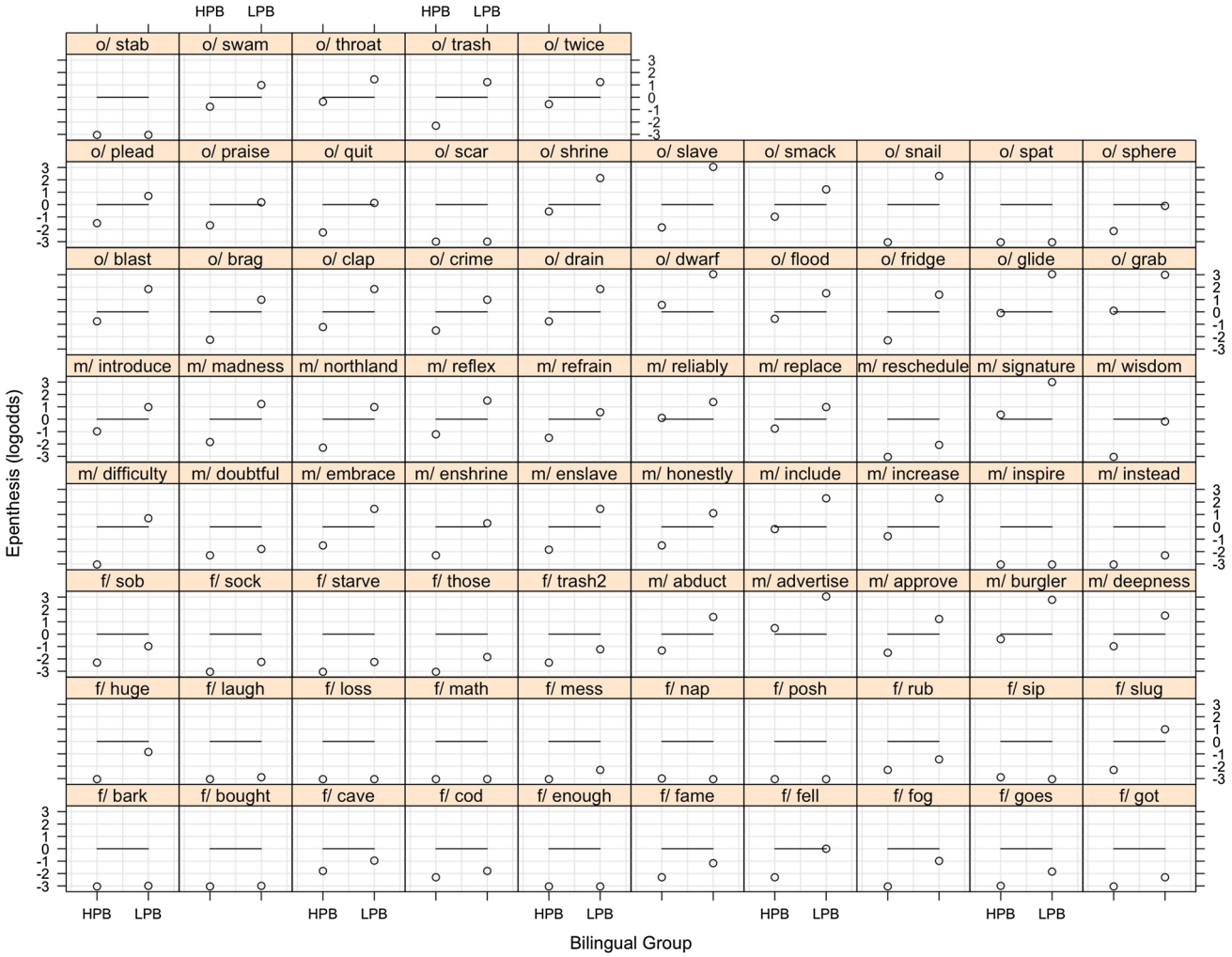

Fig. 3 Trellis plot depicting the log-odds ratios (see Baayen, 2008; Janda et al., 2010) for all 75 words divided by Bilingual Group (*o* onset, *m* medial, *f* final, *HPB* high-proficient bilingual, *LPB* low-proficient bilingual). A log-odds ratio lower than zero (i.e., the middle line) indicates an inclination to no-epenthesis, while a ratio higher than zero indicates an inclination to epenthesis. The zero line indicates equal inclination for both

References

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Beckman, M. (1982). Segment duration and the 'mora' in Japanese. *Phonetica, 39*(2–3), 113–135.

Boersma, P., & Weenink, D. (2016). PRAAT: doing phonetics by computer [Computer program]. Version 6.0.23. Retrieved December 12, 2016 from http://www.praat.org/

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Carr, P., & Montreuil, J. P. (2013). *Phonology* (2nd edn.). London: Palgrave MacMillan.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Wordform encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language, 46*, 751–781.

Davidson, L., & Stone, M. (2003). Epenthesis versus gestural mistiming in consonant cluster production: an ultrasound study. In *Proc. of the West Coast Conference on Formal Linguistics* (Vol. 22, pp. 165–178).

Dell, G. S. (1986). A speading activation theory of retrieval in language production. *Psychological Review*, *93*, 226–234.

Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: A cross-linguistic study. *Journal of Cognitive Neuroscience, 12*(4), 635–647.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1568.

Fleischhacker, H. (2005). Similarity in phonology: Evidence from reduplication and loan adaptation. UCLA Ph.D. dissertation.

Fujimoto, M. (2015). *Vowel devoicing. The handbook of Japanese language and linguistics: Phonetics and phonology*. Berlin: Mouton de Gruyter.

Gouskova, M. (2002). Falling sonority, loanwords, and Syllable Contact. Papers from the 37th Meeting of the Chicago Linguistic Society, 1, 175–186.

Ida, K., Nakayama, M., & Lupker, S. J. (2015). The functional phonological unit of Japanese–English bilinguals is language dependent: Evidence from masked onset and mora priming effects. *Japanese Psychological Research, 57*, 38–49.

Ito, J., & Mester, A. (1995). Japanese phonology. In J. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 817–838). Oxford: Blackwell.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Janda, L. A., Nesset, T., & Baayen, R. H. (2010). Capturing correlational structure in Russian paradigms: a case study in logistic mixed-effects modeling. *Corpus Linguistics and Linguistic Theory, 6*(1), 29–48.

Kubozono, H. (1999). *Nihongo no Onsei*. Tokyo: Iwanami Shoten.

Kureta, Y., Fushimi, T., & Tatsumi, I. F. (2006). The functional unit of phonological encoding: Evidence for moraic representation in native Japanese speakers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1102–1119.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.

Li, C., Wang, M., & Davis, J. (2017). The phonological preparation unit in spoken word production in a second language. *Bilingualism: Language and Cognition, 20*, 351–366. https://doi.org/10.1017/S1366728915000711.

Masuda, H., & Arai, T. (2010). Processing of consonant clusters by Japanese native speakers: Influence of English learning backgrounds. *Acoustical Science and Technology, 31*, 320–327.

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29*, 524–545.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30*, 69–89.

Nagara, S. (1972). *JapanesePidgin English in Hawaii: a bilingual description*. Honolulu: University of Hawaii Press.

Nakayama, M., Kinoshita, S., & Verdonschot, R. G. (2016). The emergence of a Phoneme-sized unit of speech planning in Japanese–English bilinguals. *Frontiers in Psychology, 7*, 175.

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115*, 282–302. https://doi.org/10.1016/j.cognition.2010.01.001.

Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: from germanic languages to mandarin Chinese and Japanese. *Japanese Psychological Research, 57*, 22–37. https://doi.org/10.1111/jpr.12050.

Verdonschot, R. G., Kiyama, S., Tamaoka, K., Kinoshita, S., La Heij, W., & Schiller, N. O. (2011). The functional unit of Japanese word naming: evidence from masked priming. *Journal of Experimental Psychology: Learning, Memory and Cognition., 37*, 1458–1473.

Verdonschot, R. G., Nakayama, M., Zhang, Q.-F., Tamaoka, K., & Schiller, N. O. (2013). The proximate phonological unit of Chinese–English bilinguals: proficiency matters. PLoS One, *8*, e61454.