THE EFFECT OF SHORTENING ONSET CONSONANTS ON SPEECH SEGMENTATION BY TAIWANESE SOUTHERN MIN LISTENERS

Shu-chen Ou[1], Zhe-chen Guo[2]

[1]National Sun Yat-sen University, [2]The University of Texas at Austin
sherryou@mail.nsysu.edu.tw, zcadamguo@utexas.edu

ABSTRACT

This study examined whether and how shortening voiced syllable-onset consonants impacts listeners' speech segmentation. Cross-linguistically, listeners tend to use longer onset consonants to locate word beginnings, possibly because of the greater phonetic richness and hence auditory-perceptual salience of these consonants. An alternative account, however, holds that it is the increased consonantality that guides the use of longer onsets as word beginning cues. Since shortening voiced stops and nasals enhances their consonantality, we tested the competing explanations in an artificial language learning experiment where shorter voiced onsets provided a potential segmentation cue. Taiwanese Southern Min listeners segmented the words of an artificial language through listening to continuous repetitions of the words. Shortening the voiced onsets in word-initial syllables did not significantly improve segmentation. The results thus did not support the consonantality-based explanation but was in line with the phonetic richness account. Possible mechanisms underlying the findings are discussed.

Keywords: speech segmentation, prosody, consonant shortening, artificial language learning.

1. INTRODUCTION

Speech is continuous without reliable pauses between words [1]. Yet, listeners can still parse it into discrete words as if they were beads on a string. One well-established finding from research into this ability is that listeners across languages use longer vowels to locate word-final positions [2]–[4]. More recently, studies have begun to explore whether and how consonant duration is also used in segmentation. In this paper, we aim to add further insight into this issue by testing whether shortened voiced syllable-onset consonants are useful for speech segmentation.

Like longer vowels, longer syllable-onset consonants can guide speech segmentation, but in a different way. Specifically, they are interpreted as signalling word beginnings rather than word-finality. This was demonstrated by White et al. [5] in a study with English listeners using the artificial language (AL) learning paradigm [2]. The listeners learned to segment trisyllabic nonsense words (e.g., /pabiku/) by listening to speech streams in which the words were concatenated continuously. The results showed that their segmentation improved when the word-initial onset consonants were lengthened. Similar findings were replicated in Italian, Hungarian, and Taiwanese Southern Min (TSM) [6], [7], even though in languages like Italian, consonant duration contrast is phonemically relevant. Using longer onsets to locate word-initial positions seems to be a cross-linguistic tendency superseding language-specific functions or patterns of consonant duration.

Kim et al. [8] observed a similar tendency in an AL learning study with Dutch and Korean listeners. They found that when the AL words began with a voiceless stop onset, both groups segmented speech better when the onset was longer with a longer voice onset time (VOT) than when it was shorter with a shorter VOT. This ran contrary to the prediction that due to the different phonetic realizations of domain-initial strengthening in the two languages [9], [10], Dutch listeners would exploit stops with shorter VOTs better than those with longer VOTs to locate word beginnings, while Korean listeners would show the opposite pattern. Kim et al. attribute their findings to the richer phonetic information in stops with longer VOTs, which confer perceptual salience on the stops to make them so useful for segmentation as to override language-specific strengthening patterns.

However, while phonetic richness can explain the cross-linguistic parallel in [8], there is an alternative possibility. Domain-initially, voiceless onsets such as voiceless stops can be produced with longer VOTs to become more consonant-like and enhance the syntagmatic contrast with the following vowel [11]. Yet, in the same environment, voiced onsets become more consonant-like through reduced sonority such that nasals, for example, are shorter with decreased nasal airflow [11]–[13]. Since [8] tested only voiceless stops, it is possible that the cross-linguistic preference for stops with longer VOTs is not driven by phonetic richness, but by a universal tendency to interpret heightened consonantality as signalling word beginnings. Voiced nasals and stops can provide a test of this alternative account as their consonantality increases when they are shortened.

Here, we examined the effect of shortening voiced onset consonants on speech segmentation by TSM

listeners. Ou and Guo [7] have shown that TSM listeners exploit lengthening of voiced onsets to locate word-initial positions. Currently, it has not been explicitly tested whether and how they would use shortening of these consonants. To assess the consonantality-based account, we conducted an AL learning experiment with TSM listeners which was modelled after [7] but with the cue voiced onset lengthening replaced by voiced onset shortening.

2. EXPERIMENT

2.1. Design and hypotheses

As in previous studies (e.g., [2]), our experiment began with a learning phase in which listeners were exposed to continuous repetitions of the (nonsense) words of an AL. Next, they recognized the words in a forced-choice test, with higher accuracy suggesting more successful segmentation during the learning.

If there is a tendency to interpret increased consonantality as cueing word beginnings, TSM listeners' segmentation should improve in an "initial shortening (IS)" condition where voiced onsets in the initial syllables of the AL words were shortened, compared with a "no shortening" (NS) condition where segment duration did not vary. Also, for the consonantality-based explanation to be supported, it had to be shown that improved segmentation was not due to shortening in any boundary-adjacent syllables. That is, the listeners were not expected to benefit from a "final shortening (FS)" condition where onsets of word-final syllables were shortened. Yet, given the evidence for longer onsets as cues to word-initiality [5]–[7] and the phonetic richness hypothesis [8], it may be that only onset lengthening is used to identify word beginnings. This predicted that TSM listeners would not benefit from onset shortening at all.

2.2. Stimuli

Following [7] and [14]–[16], we created an AL consisting of six trisyllabic CVCVCV words: /banume/, /bimɔna/, /genigɔ/, mimabu/, /nebɔgi/, and /nɔgamu/. These words were constructed from four voiced consonants (/b, g, m, n/) and five vowels (/a, i, e, u, ɔ/), all occurring in TSM. Each CV syllable was produced in a sound-treated room by a native male TSM speaker with phonetic training, who was instructed to read the syllable aloud in monotone in a carrier sentence (/gua kɔŋ __ tsit e dʑi/ "I said the word __"). His speech was recorded with Audacity and a Zoom H4n Handy Recorder at a 44.1 kHz sampling rate and digitized as WAV files.

The syllables were extracted from the carrier sentence and prosodically neutralized using Praat [17]: the F0 contours were flattened, the amplitudes of the syllables were equalized, and the durations of all consonants and vowels were set to a base duration of 150 ms. The resulting syllables were concatenated to form the AL words for the NS condition. In the IS and FS conditions, the onsets of the first syllables and third syllables were shortened by a factor of 1.5 (i.e., to 100 ms, or 0.67 of the base duration), respectively. The shortening factor was based on the magnitudes of lengthening used in [6], [7]. Fig. 1 shows the waveforms of an AL word under the three conditions.

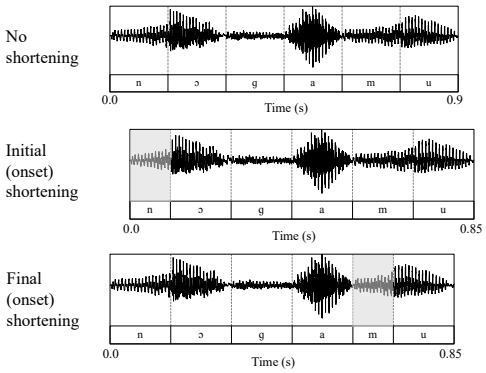

Figure 1: Waveforms of /nɔgamu/ in the three conditions

For each condition, repetitions of the AL words were concatenated without pauses to produce six long speech streams, which were presented as stimuli in the learning phase. Each stream contained 20 repetitions of each word in a pseudo-random order such that the same word never followed itself. The total duration of the six streams was 10-11 minutes. Five-second fade-in and fade-out effects with logarithmic ramps were applied to each stream to prevent listeners from identifying the very first and last syllables and using them as word boundary cues.

The test phase was a two-alternative forced-choice task which presented two stimuli, separated by a 500-ms silence, in each trial: a word of the AL and a "partword." The partword was a trisyllabic CVCVCV sequence formed by combining substrings of two AL words (e.g., /mabuge/, from /mimabu/ and /genigɔ/). There were six partwords. The average between-syllables transitional probability, computed based on the learning-phase speech streams following [2], was around 0.6 for the partwords and one for the AL words. The test contained 36 trials, resulting from all word-partword pairings. The orders of the word and partword were balanced and no stimuli in the test phase had the onset shortening cue. The experiment was compiled using E-prime 2.0 [18].

2.3. Procedure

Seated individually in front of a desktop computer and randomly allocated to either condition, participants first learned the AL by listening to the six

speech streams passively via headphones. They were encouraged to pay as much attention to the streams as possible and made aware that they would be tested on how much they knew the words of the AL. No cues to the words (e.g., their lengths) were provided. After listening to the speech streams, participants completed the forced-choice test, in which they heard two stimuli in each trial and selected, in five seconds, the one that they thought was a word of the AL. The experiment took about 30 minutes.

2.4. Participants

Ninety native speakers of TSM were recruited (45 males, 45 females; $N = 30$ for each condition). Their mean age in years was 21.8 (range: 20–35). All had also acquired Taiwan Mandarin as a native language and learned English as an academic subject. None reported speech or hearing impairments.

3. RESULTS

Participants' test response accuracy was analyzed. Timeout trials (0.25% of all trials) were discarded. The remaining responses ($N = 3{,}232$) were coded as correct (1) if the AL word was selected and incorrect (0) if not. Fig. 2 shows the response accuracy of each participant and the mean of each condition.

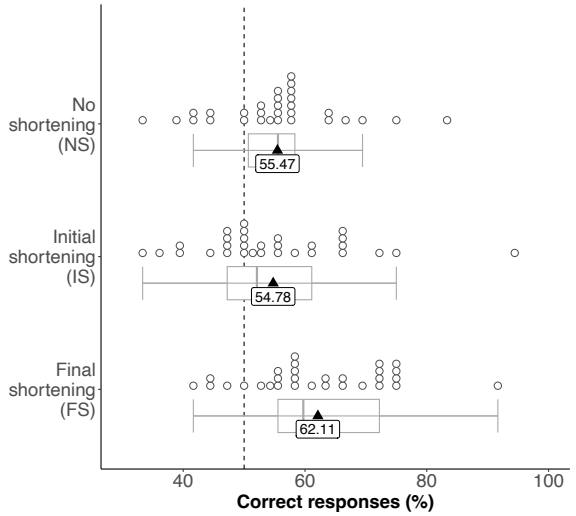

Figure 2: Response accuracy of individual participants (circles) and mean accuracy of each condition (triangles). The gray bar in each box marks the median and the dashed vertical line represents the chance accuracy (50%).

A Bayesian hierarchical logistic regression model was fitted to the responses using the *brms* [19] package of R [20]. The population-level (fixed) effect of main interest was Condition, which contrasted the IS and FS conditions with NS (baseline). As in [21], the model also contained the order number of each trial in the test (Trial) and log-transformed response time (LogRT) to capture any effects of practice or fatigue and response speed-accuracy relationship, respectively. These two continuous predictors were scaled and entered as population-level covariates. The group-level (random) effects included intercepts varying by participant, AL word, and partword and slopes for Condition varying by AL word. Weakly informative priors (e.g., Normal(0, 10)) were assigned to all population- and group-level effects. Following [22], we used a weakly informative LKJ(2) prior [23] on the intercept-slope correlation in the group-level effects. Posterior distributions of model parameters were approximated using four 2,000-iteration Monte Carlo Markov Chains with 1,000 warm-ups. $\hat{R}$ values [24] reached one for all parameters, confirming chain convergence.

Fig. 3 shows the marginal posterior distributions of estimates for the population-level effects. We determined statistical significances based on the 95% highest density interval (HDI) criterion [25]. That is, the effect of a predictor was significant if the 95% HDI of the relevant posterior distribution excluded zero. Thus, LogRT had a significant effect suggesting that slower responses tended to be incorrect, 95% HDI = [–0.261, –0.094], possibly due to listeners' response certainty or confidence [7], [21]. Trial had no significant effect, 95% HDI = [–0.082, 0.069].

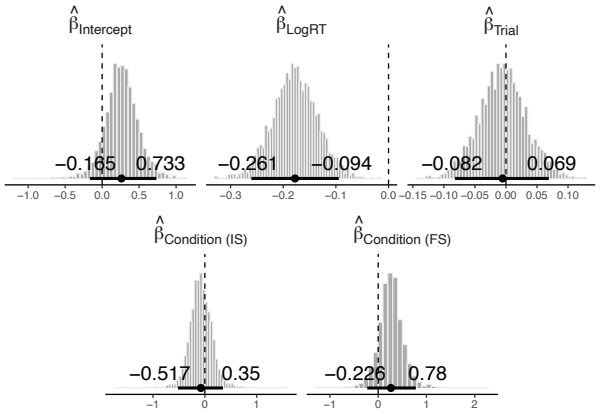

Figure 3: Marginal posterior distributions of population-level effect estimates. The bar under each distribution represents the 95% HDI with the circle marking the mean. The dashed vertical line indicates zero.

Of main interest was the Condition predictor contrasting the NS condition with the other two conditions with shortening cues. The model revealed that neither IS, 95% HDI = [–0.517, 0.350], nor FS, 95% HDI = [–0.226, –0.780], differed significantly from NS. There was thus no evidence that shortening voiced onset consonants, either word-initially or word-finally, affected speech segmentation.

As no shortening conditions differed significantly from NS, we further explored the listeners'

performance by comparing it against the chance level. To do so, we computed the posterior distributions of correct response percentages for the conditions as shown in Fig. 4. No condition was significantly better than chance (as none had a 95% HDI that excluded 50%). However, the posterior probability that the accuracy was better than chance was high, especially for the FS condition (NS: 90%; IS: 77%; FS: 96%). Thus, it was likely that the listeners might still be able to segment out some of the AL words simply by listening to them repeated continuously.

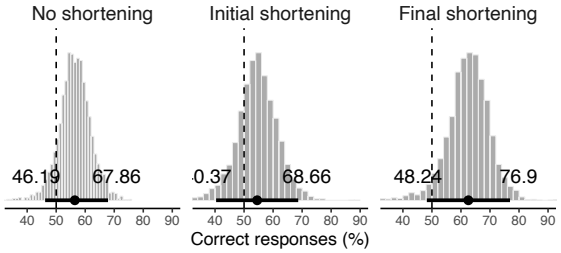

Figure 4: Posterior distributions of probability of correct responses (%) for the three conditions. The bar represents the 95% HDI with the circle marking the mean. The dashed line indicates the chance level (i.e., 50%).

4. DISCUSSION AND CONCLUSION

This study examined whether shorter onset consonants are exploited in speech segmentation. One explanation for the cross-linguistic use of longer onsets in segmentation is that listeners interpret increased onset consonantality as signalling word beginnings. We tested it in an AL learning experiment in which the voiced stops or nasals in syllable onset were shortened to enhance their consonantality. The results showed that onset shortening, either in the initial or final syllables of the AL words, did not significantly improve segmentation. Thus, at least for TSM listeners, there is no evidence for the above explanation. Our findings indirectly align with the phonetic richness hypothesis [8], which holds that the greater auditory-perceptual salience of longer onsets underlies their cross-linguistic use in segmentation.

A question then arises as to the specific mechanisms leading listeners to use onset lengthening as reported in previous studies [5]–[7] but not onset shortening. We assume that the listeners' behavior may be explained from a functional-informational perspective. According to the Effort Code account of intonational variation [26], listeners associate what speakers produce with high pitch or greater pitch range (or more generally, with greater articulatory effort) with greater informational load. When the onsets of word-initial syllables in an AL learning task are made longer and perceptually more salient, they are interpreted as the speaker's effort to signal important information, which directs listeners' attention to word beginnings. In contrast, though word-initial onsets were consistently shortened in our IS condition to provide a potentially useful segmentation cue, listeners would perceive them as informationally trivial and ignore them.

The patterns of findings here and in previous AL learning studies may also be accounted for based on the timing of the perceptual center (p-center) of a syllable, or the perceptual moment of its occurrence [27]. White et al. [5] suggest that lengthening the onset of the initial syllable of an AL word delays its p-center relative to the p-center of the last syllable in the previous word, and such a delay is interpreted as a marker of prosodic boundary. On the other hand, shortening a word-initial onset brings its p-center closer to that of the final syllable in the previous word, producing no boundary percept. Moreover, since the shortening reduces the temporal distance between the p-centers of the syllables across a word boundary, it could lead listeners to incorrectly analyze the two syllables as belonging to the same word.

In addition to the lack of an effect of onset shortening, it was found that although no conditions showed segmentation better than chance, the posterior probability of above-chance performance was relatively higher in the FS condition. We speculate that shortening the onsets of word-final syllables might cause the following vowels to be perceived as longer and hence word-final, reflecting the well-documented use of longer vowels to locate word-finality [2], [3]. However, further confirmatory evidence would be needed to verify this speculation.

To conclude, the current study did not find evidence that shortened voiced onsets with increased consonantality are used to identify word beginnings. Word-initial onset consonants, whether they are voiced or voiceless, facilitate segmentation when they are lengthened. Yet, it remains to be tested in future work if onset shortening can become useful when combined with other cues. For instance, while a nasal onset in the domain-initial position is produced with a shorter duration to enhance consonantality, there is a concomitant reduction in coarticulatory nasalization in the following vowel [12], [28]. It is possible that shorter onsets are exploited when accompanied by a coarticulatorily appropriate reduction of their influence on the vowel. Additionally, our AL stimuli were shortened artificially and their F0 and amplitude were neutralized. It is then important to examine onset shortening effects on segmentation of meaningful, naturalistic speech in which other segmental and prosodic phenomena co-vary with onset duration. Further research is also needed to test if the lack of onset shortening effects is specific to TSM or can generalize to other languages.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the grant support to the first author from National Science and Technology Council, Taiwan (MOST109-2410-H-110-069-MY3).

6. REFERENCES

[1] Cole, R. A., Jakimik, J., Cooper, W. E. 1980. Segmenting speech into words. *J Acoust Soc Am* 67, 1323–1332.

[2] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: The role of distributional cues. *J Mem Lang* 35, 606–621.

[3] Tyler, M. D., Cutler, A. 2009. Cross-language differences in cue use for speech segmentation. *J Acoust Soc Am* 126, 367–376.

[4] Ordin, M., Polyanskaya, L., Laka, I., Nespor, M. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Mem Cognit* 45, 863–876, 2017.

[5] White L., Mattys, S. L., Stefansdottir, L., Jones, V., 2015. Beating the bounds: Localized timing cues to word segmentation. *J Acoust Soc Am* 138, 1214–1220.

[6] White, L., Benavides-Varela, S., Mády, K. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J Phon* 81, 100982.

[7] Ou, S.-C., Guo, Z.-C. 2021. The differential effects of vowel and onset consonant lengthening on speech segmentation: Evidence from Taiwanese Southern Min. *J Acoust Soc Am* 149, 1866–1877.

[8] Kim, S., Cho, T., McQueen, J. M. 2012. Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *J Phon* 40, 443–452.

[9] Cho T., McQueen, J. M. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *J Phon* 33, 121–157.

[10] Cho T., Keating, P. A. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *J Phon* 29, 155–190.

[11] Cho T., Keating, P. 2009. Effects of initial position versus prominence in English. *J Phon* 37, 466–485.

[12] Cho, T., Kim, D., Kim, S. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *J Phon* 64, 71–89.

[13] Hsu, C.-S. K., Jun, S.-A. 1998. Prosodic strengthening in Taiwanese: Syntagmatic or paradigmatic? *UCLA Working Papers in Phonetics* 96, 69–89.

[14] Kim, S., Cho, T., McQueen, J. M. 2012. Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *J Phon* 40, 443–452.

[15] Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., Christiansen, M. H. 2015. Factors influencing sensitivity to lexical tone in an artificial language: Implications for second language learning. *Stud Second Lang Acquis* 37, 335–357.

[16] Saffran, J. R., Newport, E. L., Aslin, R. N. 1996. Word segmentation: The role of distributional cues. *J Mem Lang* 35, 606–621.

[17] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/.

[18] Psychology Software Tools. 2012. E-Prime 2.0. Pittsburgh, PA.

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

[20] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[21] Ou, S.-C., Guo, Z.-C. 2021. The language-specific use of fundamental frequency rise in segmentation of an artificial language: Evidence from listeners of Taiwanese Southern Min. *Lang Speech* 64, 437–466.

[22] Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., Kong, E. J. 2018. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *J Phon* 71, 147–161.

[23] Lewandowski, D., Kurowicka, D., Joe, H. 2009. Generating random correlation matrices based on vines and extended onion method. *J Multivar Anal* 100, 1989–2001.

[24] Gelman, A., Hill, J. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

[25] Kruschke, J. K., Aguinis, H., Joo, H. 2012. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organ Res Methods* 15, 722–752.

[26] Gussenhoven, C. Intonation and interpretation: phonetics and phonology. *Proc. Speech Prosody 2002*, 47–57.

[27] Morton, J., Marcus, S., Frankish, C. 1976. Perceptual centers (P-centers). *Psychol Rev* 83, 405–408.

[28] Jang, J., Kim, S., Cho, T. 2018. Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences. *J Acoust Soc Am* 144, EL33–EL39.