THE ROLE OF INITIAL F0 RISE IN SPEECH SEGMENTATION: A CROSS-LINGUISTIC STUDY

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 paper investigates the cross-linguistic use of initial F0 rise in speech segmentation. Previous studies using real word detection experiments have provided suggestive evidence that F0 rise serves as a universal cue to word beginnings. We examined if there would be more direct support for this claim by conducting an experiment with listeners of four typologically different languages: English, French, Japanese, and Taiwanese Southern Min. Participants learned an artificial language by listening to speech streams in which the language's words were concatenated without pauses in between and then identified the words in a test. Analysis of identification accuracy indicated that the words with an F0 rise on the initial syllables were identified more accurately than those without any F0 cue, suggesting that initial F0 rise facilitated segmentation. This was found for all listener groups, adding support to the view that F0 rise is a universally accessible cue to word beginnings.

Keywords: F0 rise, speech segmentation, universal cues, artificial language learning

1. INTRODUCTION

The ability to segment physically continuous speech into discrete chunks is in part modulated by the native language. There is accumulated evidence that listeners use segmentation strategies promoted by experience with language-specific phonotactics ([16, 17]), allophonic variation ([11]), prosodic structures ([7]), distribution of prominence ([22]), and so on. However, certain segmentation solutions seem to be cross-linguistically employed. For example, it has been shown that Japanese listeners exploit F0 rise to locate word beginnings ([23]) and the same is also reported for listeners of French ([24]) and Korean ([13]). The cross-linguistic parallel has led [23] to assume that F0 rise may be a universally salient cue to word onsets. This paper aims to investigate this possibility by examining the performance of native listeners of four typologically distinct languages in an artificial language segmentation task.

Presently, the aforementioned evidence for the cross-linguistic use of F0 rise is limited due to methodological restrictions of studies that lend support for such use. One of the methods adopted by these studies is to ask listeners to detect real words (of their language) embedded among nonsense syllables (e.g., [13, 23]). Detection accuracy or speed then serves as a measure of segmentation success. Yet, as segmentation is driven primarily by lexical knowledge ([15]), this method may be subject to confounds such as word frequency. Importantly, none of the studies performed direct cross-linguistic comparisons, which are difficult in designs using real-word materials.

One experimental paradigm that overcomes the above limitations and is widely adopted in recent segmentation literature is artificial language learning (ALL). In an ALL experiment, subjects first learn the words of an artificial lexicon just by listening to speech streams in which tokens of the words are concatenated together with no pauses in between. Next, they identify the words in a test, with higher identification accuracy indicating more successful segmentation during the learning. With the ALL technique, research has provided much insight into the cross-linguistic use of prosodic information in segmentation. For instance, it is suggested that final lengthening is a universal perceptual cue to finality as it is exploited by listeners of several different languages, including English, French, Dutch, and Korean ([14, 22]). Yet, recent ALL experiments have also found that the effects of such a putatively universal cue may be overridden by the influence of language-specific phonology (e.g., [18]). Due its potential to reveal both cross-linguistic and language-specific patterns, the ALL is suitable for more conclusively establishing whether F0 rise is indeed a universally salient cue to word onsets.

The current study explores this question by testing native listeners of English, French, Japanese, and Taiwanese Southern Min (TSM). These languages are meant to represent four different groups in the typology of lexical prosody. Japanese is a lexical pitch accent language in which one syllable in a word is accentuated and associated with some F0 pattern that is perceived as pitch accent. Lexical stress languages like English accentuate one syllable in a word with a combination of acoustic correlates that include but are not limited to F0 cues (e.g., [8, 9]). TSM is a lexical tone language, which uses F0 patterns over individual syllables to contrast word meanings. Finally, French is generally thought to have no lexical prosody. An ALL experiment was carried out to examine whether the listener groups would all exploit F0 rise in word-initial positions despite the different functions of F0 in their native languages.

2. METHOD

2.1 Hypothesis

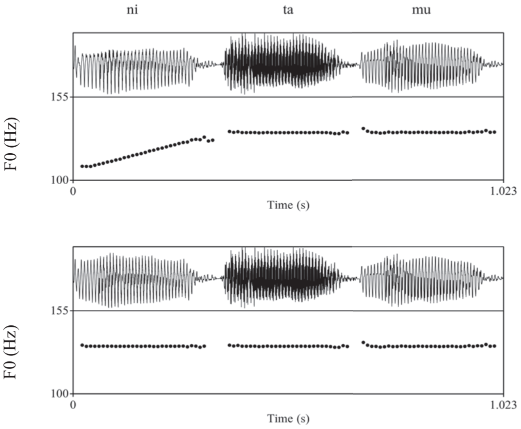
It was hypothesized that listeners of English, French, Japanese, and TSM all exploit F0 rise as a word-initial position to support segmentation if it serves as a universal cue. To test the hypothesis, we created an artificial language in which half of the words received an F0 rise on the word-initial syllable whereas the other half had a monotonous F0 contour (i.e., no F0 cue). Evidence for the hypothesis would be that listeners of the four languages generally identified the words with initial F0 rise more accurately than those with no F0 cue in the test of the ALL experiment.

2.2 Materials

Six consonants ([p, t, k, m, n, l]) and four vowels ([a, i, e, u]) were combined to create 18 consonant-vowel syllables, which were then concatenated to form six trisyllabic sequences: [pakime], [tulepi], [kemina], [mapeku], [nitamu], and [linute]. The sequences were the words of the artificial language. Their component syllables were individually recorded by a male sequential bilingual speaker of TSM and Taiwan Mandarin, who inserted each syllable into the same carrier sentence ([gua kɔŋ ____ tsit pian] 'I said ____ once') and read the result out. The syllables were excised from the recorded items and subjected to prosodic manipulations using Praat ([2]). First, their F0 contours were all flattened at 132 Hz, which was the average F0 of the original syllables. Next, their durations were normalized such that each syllable had a length of 341 milliseconds (ms), the mean duration of the syllables.

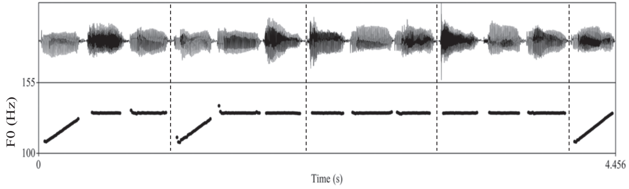
Among the six words, half of them would have initial F0 rise while the other half would have no F0 cue. Thus, after their durations were normalized and their F0 contours were flattened, the syllables were directly concatenated to construct the words without any F0 cue. As for the cue-bearing words, further manipulations were performed. Following [3], we used a 3.5-semitone change and created the initial F0 rise for each of these words by lowering the onset F0 value of its first syllable by 3.5 semitones and then linearly increasing F0 value throughout the syllable. Two versions of the artificial language were constructed to counterbalance potential effects of placing the initial F0 rise on a fixed set of words. In one version, the cue-bearing words were [nitamu], [linute], and [mapeku]; in the other, they were [pakime], [tulepi], and [kemina]. Shown in Figure 1 are the waveforms and F0 profiles for the two versions of an example word—one had the initial F0 rise and the other had no F0 cue.

Figure 1: Waveforms and F0 contours of [nitamu] with initial F0 rise and without any F0 cue.



The experiment comprised of a learning phase and a test phase. The stimuli used in the learning phase were five speech streams generated by randomly concatenating the six words of the artificial language with the restriction that no word immediately followed itself. No pause was placed between any two words, and each word occurred 20 times in each stream. As a result, there were 100 tokens in total (20 times × 5 speech streams) for each word. Following [22], we applied five-second fade-in and fade-out effects to each stream, gradually increasing the volume at the beginning and gradually decreasing it at the end. This prevented subjects from finding out the very first and last syllables and using them as a segmentation cue. Each stream was around two minutes long, and the total duration of the five streams (and thus the exposure to the artificial language) was approximately 10 minutes. Shown in Figure 2 is an excerpt taken from one speech stream in one version of the artificial language.

Figure 2: Waveform and F0 contour of a sample speech stream.



The test phase consisted of a two-alternative forced-choice test. In each trial, two stimuli were presented with an interval of 500 ms: one was a word of the artificial language and the other was a nonword. Serving as a distractor, the nonword was a trisyllabic sequence that never occurred during the learning phase. There were six nonwords: [pamiku], [tupena], [kenupi], [makite], [nilemu], and [litame]. They were constructed using the same 18 syllables of the words of the artificial language. Consequently, the nonwords might also bear initial F0 rise. For example, if the [tu] in the word [tulepi] was marked with F0 rise, then so was the [tu] in the nonword [tupena]. The test contained 36 trials, resulting from all possible pairings of the words with the nonwords. The presentation orders of the words and nonwords were counterbalanced. E-Prime 2.0 ([19]) was used to control stimulus presentation and record responses.

2.3 Procedure

Participants were told that their task was to learn an artificial language by listening to five speech streams in which the words of the language were concatenated together. They were not given any cues to word boundaries or to the length or number of the words. All they needed to do was to pay as much attention to the speech streams as they could. They were made aware that they would be tested later. After the learning phase, participants took the two-alternative forced-choice test. In each trial, they heard two successive stimuli and then selected the one which sounded more like a word from the artificial language by pressing a button ‘1’ or ‘2’ on a response device. They had 10 seconds to respond and failures to respond within the allotted time were treated as omissions. All participants completed the experiment individually.

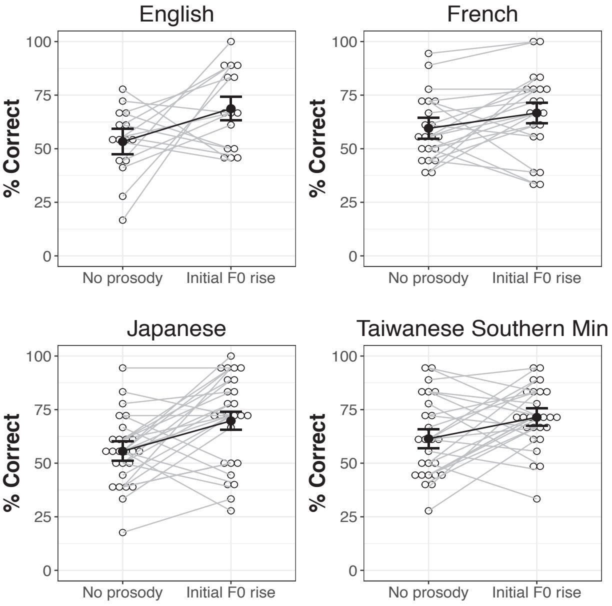
2.4 Participants

A total of 88 subjects participated in the study. For English, French, Japanese, and TSM, the numbers of native listeners recruited were 15, 21, 26, and 26, respectively. They were randomly assigned to either version of the artificial language. None reported hearing impairments or history of auditory disorders.

3. RESULTS

Subjects’ responses in the test were analyzed. We first excluded invalid responses such as omissions, which accounted for 0.25% of the data. For the remaining data, we computed the accuracy rates of individual subjects of each group and compared the group means against the 50% chance level using one-sample *t*-tests. The results indicated that the four groups all performed significantly better than would be expected by chance (English: $t(14) = 4.301$, French: $t(20) = 4.104$, Japanese: $t(25) = 4.217$, TSM: $t(25) = 6.444$, all *p*s < 0.001), suggesting that the subjects were generally able to formulate some hypotheses regarding the words of the artificial language during the 10-minute learning phase. Presented in Figure 3 are the mean percentages of correct responses of the four groups, calculated for the words with initial F0 rise and those with no F0 cue.

Figure 3: Individual participants’ accuracy rates (empty dots) and mean accuracy rates of the four groups (solid dots), calculated for the words with initial F0 rise and those with no F0 cue. The grey lines represent participants’ accuracy across these two types of words. The error bars indicate 95% confidence intervals.



To examine the effect of initial F0 rise on segmentation, we fitted linear mixed-effects logistic regression models to the valid responses separately for each group, using the GLMER() function from the lme4 package ([1]) of R ([20]). Included as fixed effects in the models were *Prosody* and *Trial*. The former represented the contrast between the words with initial F0 rise and those with no F0 cue (dummy-coded as 1 and 0, respectively) and was of most interest to the current research. The latter was the ordinal number of a trial, included to factor out the potential influence of practice or fatigue. The random-effects structure consisted of random intercepts for subject, word, and nonword as well as a by-subject random slope for *Prosody*.

The results of the mixed-effect model of each group are given in Table 1. The *Trial* terms in all the models indicated that there was a decline in identification accuracy over trials, although this effect reached significance only for the Japanese group. Importantly, it was found that the effect of *Prosody* was the same across the four groups: overall, identification of the words with initial F0 rise was significantly more accurate than that of the words without an F0 cue, as can be seen in Figure 3. This is consistent with the hypothesis that initial F0 rise can be exploited to support speech segmentation, regardless of the listener's native language.

Table 1: Results of mixed-effects models

English:

Fixed effect	β	SE(β)	z	p
Intercept	0.389	0.241	1.615	0.106
Trial	-0.013	0.009	-1.473	0.141
Prosody	0.766	0.368	2.081	0.038

French:

Fixed effect	β	SE(β)	z	p
Intercept	0.679	0.208	3.252	0.001
Trial	-0.014	0.008	-1.892	0.059
Prosody	0.411	0.202	2.031	0.042

Japanese:

Fixed effect	β	SE(β)	z	p
Intercept	0.508	0.238	2.139	0.032
Trial	-0.014	0.007	-1.996	0.046
Prosody	0.795	0.234	3.390	< 0.001

Taiwanese Southern Min:

Fixed effect	β	SE(β)	z	p
Intercept	0.739	0.255	2.892	0.004
Trial	-0.011	0.007	-1.563	0.118
Prosody	0.455	0.200	2.273	0.023

4. DISCUSSION AND CONCLUSION

The present study investigates the use of F0 rise in speech segmentation by listeners of four typologically different languages: English, French, Japanese, and TSM. Our ALL task revealed that the segmentation performance of all the listener groups was enhanced by a rise in F0 placed on the initial syllables of the words of the artificial language. This finding is compatible with previous studies examining word segmentation by French and Japanese listeners with different experimental paradigms (e.g., [23, 24]), thus providing support for the hypothesis that an increase in F0 can be a universally available cue for locating word onsets.

Our experiment furnishes additional evidence for this hypothesis by showing that listeners of a lexical stress language (i.e., English) and those of a lexical tone language (i.e., TSM) also seem to exploit the initial F0 rise cue. Yet, given that lexically stressed syllables in English are predominantly word-initial ([6]) and one correlate of lexical stress is F0 rise ([21]), it might be speculated that the English listeners in the experiment interpreted syllables with F0 rise as stressed syllables and therefore as word beginnings. In other words, they were simply applying a stress-based segmentation procedure developed from experience with the stress patterns of their language ([5, 22]). However, such a language-specific explanation fails to account for all the obtained findings. The results of TSM listeners bear on this point. TSM is a lexical tone language in which the only rising tone is strongly associated with word-finality because the tone is restricted by the language's tone sandhi process to the final position of the tone sandhi domain. There is no reason to assume that experience with TSM would encourage the use of F0 rise as a cue to word onsets. Still, the initial F0 rise significantly improved their segmentation performance. The cross-linguistic utility of initial F0 rise as observed in the current experiment suggests that speech segmentation is possibly guided by certain universally accessible mechanisms.

Currently, it is not clear what exactly these mechanisms could be. Previous studies on rhythmic grouping (e.g., [10]) have demonstrated that listeners (of English and French) apply the so-called trochaic law to syllables alternating in intensity in continuous speech. That is, they tend to parse the speech into units beginning with syllables with relatively higher intensity. It may be the case that just as with greater intensity, a rise in F0 generates a percept of prominence that triggers trochaic grouping. The use of initial F0 rise may also be linked to domain-initial strengthening (e.g., [12]), which enhances the acoustic salience of the beginnings of prosodic units and can guide listeners to locate word onsets ([4]). Experimentation with a wider range of salience or prominence cues and languages would be needed to provide further insight into what may underlie the cross-linguistic use of initial F0 rise in segmentation.

5. ACKNOWLEDGEMENT

The current research is supported by a grant from the Ministry of Science and Technology to the first author (MOST 107-2410-H-110-036).

6. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48.

[2] Boersma, P., Weenink, D. 2016. Praat: doing phonetics by computer (Version 6.0.21) [Computer program]. Retrieved from http://www.praat.org/.

[3] 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. *Studies in Second Language Acquisition* 37, 335–357.

[4] Cho, T., McQueen, J. M., Cox, E. A. 2007. Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics* 35, 210–243.

[5] Cutler, A. 1990. Exploiting prosodic probabilities in speech segmentation. In: Altmann, G., (ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press, 105–121.

[6] Cutler, A., Carter, D. M. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language* 2, 133–142.

[7] Cutler, A., Mehler, J., Norris, D., Segui, J. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25, 385–400.

[8] Fry, D. B. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* 27, 765–768.

[9] Gay, T. 1978. Physiological and acoustic correlates of perceived stress. *Language and Speech* 21, 347–353.

[10] Hay, J. S., Diehl, R. L. 2007. Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception & Psychophysics* 69, 113–122.

[11] Jusczyk, P. W., Hohne, E. A., Bauman, A. 1999. Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics* 61, 1465–1476.

[12] Keating, P., Cho, T., Fougeron, C., Hsu, C. S. 2004. Domain-initial articulatory strengthening in four languages. In Odgen, R., Local, J., & Temple, R. (eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press, 143–161.

[13] Kim, S. 2003. The role of post-lexical tonal contours in word segmentation. *Proc. 15th ICPhS* Barcelona, 495–498.

[14] Kim, S., Broersma, M., Cho, T. 2012. The use of prosodic cues in learning new words in an unfamiliar language. *Studies in Second Language Acquisition* 34, 415–444.

[15] Mattys, S. L., Bortfeld, H. 2017. Speech segmentation. In: Gaskell, M. G. & Mirković, J. (eds.), *Speech Perception and Spoken Word Recognition*. London: Routledge, 55–75.

[16] Mattys, S. L., Jusczyk, P. W. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78, 91–121.

[17] McQueen, J. M. 1998. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language* 39, 21–46.

[18] 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. *Memory & Cognition* 45, 863–876.

[19] Psychology Software Tools. 2012. E-Prime 2.0. Pittsburgh, PA: Author.

[20] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[21] Spitzer, S. M., Liss, J. M., Mattys, S. L. 2007. Acoustic cues to lexical segmentation: A study of resynthesized speech. *J. Acoust. Soc. Am.* 122, 3678–3687.

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

[23] Warner, N., Otake, T., Arai, T. 2010. Intonational structure as a word-boundary cue in Tokyo Japanese. *Language and Speech* 53, 107–131.

[24] Welby, P. 2007. The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication* 49, 28–48.