Listeners are sensitive to prosody in segmental categorization

Jeremy Steffman

University of California Los Angeles, USA
jsteffman@g.ucla.edu

Abstract

Two experiments were designed to test if and how listeners' awareness of prosodic structure might modulate categorization of speech segments. This possibility has been challenged by recent experiments showing an effect originally analyzed as originating from awareness of prosodic structure might simply be due to speech rate normalization. The current studies test for listener awareness of prosodic structure in a way that is not confounded with rate normalization. Experiment 1 shows that tonal melodies influence categorization, where an IP boundary tone appears to give the percept of increased speech rate when compressed onto a short vowel, suggesting listener awareness of intonationally defined prosodic structures and their temporal manifestation. Experiment 2 shows that expectations about phrase final lengthening modulate segment categorization in a directionality that is not predicted by rate normalization. Taken together, the experiments suggest that prosodic structure is relevant for listeners in their categorization of speech segments.

Index Terms: speech perception, prosodic structure, intonation, speech rate normalization, segment categorization.

1. Introduction

Listeners have been shown to modulate categorization of segments to accommodate contextual variation in segmental neighbors, formant frequency distributions, etc. [1]-[4]. Recent research [5] [6] has explored how this perceptual accommodation might extend to prosodically conditioned segmental variation, given that the realization of a segment varies systematically based its prosodic position [7]-[10].

Kim & Cho [5] investigated how initial strengthening of voice onset time (VOT) in English might affect categorization of a VOT continuum (categorized as /p/ or /b/). Because VOT is systematically longer at the beginning of an Intonational Phrase (IP) in English [11] [12], the authors predicted that listeners would require longer VOT to categorize a sound as voiceless when the target stop was IP-initial in a carrier phrase. Using the carrier phrase "Let's hear *x* again", the authors found that placing a phrase boundary (produced with phrase-final lengthening and falling F0 (L-L%)) before the target ("Let's hear % *x* again"), shifted listeners' categorization to higher VOT values for a /p/ response. The authors interpreted this as indicating listeners' awareness of initial strengthening and the segmental encoding of prosodic structure more generally.

This view has been challenged more recently by Mitterer et al. [6] who performed several experiments suggesting that the effect documented by Kim & Cho is explainable as speech rate normalization. Listeners have been shown to shift categorization of VOT (and other temporal cues) on the basis of contextual information about speech rate, where longer durations preceding *or following* the target shift categorization to higher VOT vales for a voiceless stop response, e.g. [13] [14]. Because the phrase boundary used by Kim & Cho was cued by phrase-final lengthening, and because rate normalization has been shown to occur on the basis of *local* slowdowns in speech rate [14], the shift observed by Kim & Cho may have originated only from normalization for preceding length. Mitterer et al. show that global slowdowns in the words preceding the target (using the same stimuli as Kim & Cho) shift categorization in the same direction as the phrase boundary, suggesting that the more localized slowing at the boundary could be shifting categorization via rate normalization. Mitterer and colleagues also show that flattening the F0 in the preceding words, effectively removing F0 information, causes no shift in categorization, while the monotonized stimuli still shifted categorization on the basis of duration. Based on these findings, the authors conclude that listeners may not be aware of prosody for the purposes of categorizing speech segments.

The two experiments reported here challenge this view, in providing some evidence that listeners do indeed make reference to prosodic structure in their categorization of speech segments. Experiment 1 independently manipulates F0 as a cue to prosodic boundary, and shows that speech rate percepts appear to integrate tonal distributions, implicating listener knowledge of prosodic structure defined by intonation. Experiment 2 shows that expectations about the segmental encoding of prosodic position modulate categorization in a way that is not explainable as rate normalization.

2. Experiment 1

2.1. Motivation

Experiment 1 is designed to test how intonational cues to prosodic structure might independently influence categorization. A 2x2 design of duration by F0 variables was used in a two-alternative forced choice (2AFC) categorization task. Manipulations were made to the vowel immediately preceding the target in the carrier phrase, "I'll say *pa/ba* again", where participants categorized the target sound as /p/ or /b/. The two length variables are SHORT and LONG where the SHORT condition refers to a non-lengthened IP-medial vowel in "say", and the LONG condition refers to a vowel in "say" that has undergone phrase-final lengthening. The two F0 variables in the experiment are chosen so that in one condition (called the LH condition), F0 cues an IP boundary regardless of the length condition that it co-occurs with, while in the other F0 condition (called the FLAT condition), F0 will cue an IP boundary only in the LONG condition.

In the intonational phonology of English [15] [16], there are four IP boundary tones (excluding downstep): L-L% (low falling F0), L-H% (low rising F0), H-L% (high flat F0), and H-H% (high rising F0). These boundary tones occur on the

10.21437/SpeechProsody.2018-56

last syllable of an IP with substantial lengthening. A single non-pitch-accented syllable (with the L aligned early in the syllable) can have two tonal targets *only* with L-H%. Because it occurs exclusively at phrasal boundaries, it is expected that this low rising contour (L-H%) on unaccented "say" should, in theory, inform listeners of an IP boundary even in the absence of durational cues. This is the LH F0 condition. The contour used for the FLAT condition is high flat F0 (H-L%). In the SHORT condition, this F0 should not cue the presence of a boundary, as it is a natural transition between adjacent H* pitch accents in the carrier phrase (described below). However, in the LONG condition, it is predicted to be interpretable as H-L%. In summary: both low rising (LH) and high flat (FLAT) F0 contours are possible boundary tones in the LONG condition, but in the SHORT condition, LH F0 should cue a boundary, while FLAT F0 should not. The conditions thus present a possible test for listener awareness of prosodic structure, independent of its durational correlates. If listeners are aware of prosodic structure for the purposes of segmental categorization, categorization might be expected to shift in the SHORT condition, where LH F0 cues a boundary. Following the proposal made by Kim & Cho, this boundary-cuing F0 may inform listeners that the syllable over which is it distributed is IP-final, meaning the following stop is initial to an IP, which would cause them shift categorization of the target stop to higher VOT values for a /p/ response (due to initial strengthening). On the other hand, this F0 information may be integrated into listeners' rate percepts. That is, because the contour, being IP-final, is typically distributed over a lengthened syllable in natural speech, when it is compressed on a short vowel (in the SHORT condition), it may sound like that syllable was spoken more quickly. If listeners incorporate tonal information in this way when computing speech rate, it would be predicted that LH F0 would lower the VOT threshold for a /p/ response in the SHORT condition.

2.2. Experimental Design

2.2.1. Stimuli

The stimuli for the four conditions were made by resynthesizing the speech of a ToBI-trained English speaker recorded at 44.1 kHz (32 bit) using an SM10A Shure^{TM} microphone and headset in a sound attenuated room in the UCLA Phonetics Lab. PSOLA resynthesis [17] in Praat [18] was used (preserving the original intensity contour). The utterances from which the stimuli were made are shown with English ToBI notation [19] [20] (*"pa"* is phonetically [p^{h}ɑ]).

I'll say *pa* again (1)
H* H* L-L%

I'll say *pa* again (2)
H* L-H% H* L-L%

The creation of the stimuli proceeded as follows. First, the vowel in "say" was excised from (2) above, and was used as the vowel in the LONG condition. This vowel was produced with IP-final lengthening (duration = 245 ms). The remainder of (2) served as the frame for all stimuli. The duration of each segment was resynthesized to be the mean duration for that segment in (1) and (2), to minimize biases that non-boundary-adjacent segment durations might introduce (as boundaries can affect the durations of non-adjacent segments, e.g. [21]).

The vowel from "say" in (1) was excised and was the vowel in the SHORT condition (duration = 145 ms). This vowel with a naturally produced high flat contour (transitioning between the two adjacent H* pitch accents in (1)) is the FLAT SHORT condition. This vowel with the F0 contour from "say" in (2) (L-H%) overlaid, is the LH SHORT condition. The excised LONG vowel from (2) was overlaid with the high flat F0 contour from *"say"* in (1). This created a LONG vowel with FLAT F0. The LONG vowel naturally produced with L-H% is the LH LONG condition. These four vowels were inserted into the duration-normalized frame. VOT manipulations were made with PSOLA resynthesis as well. VOT was set to 0 to 45 milliseconds in 5 millisecond steps, for 10 steps total. These manipulations in combination created 40 unique stimuli (2 length conditions x 2 F0 conditions x 10 VOT steps).

2.2.2. Participants

55 monolingual American English speaking students at UCLA participated in the experiment for course credit. 4 participants were excluded because their proportion of /p/ responses at the endpoints of the continuum fell more than two standard deviations outside the mean proportion of /p/ responses for either endpoint for the group. One participant was excluded because of failure to perform the experimental procedure. Reported results are for the remaining 50 participants.

2.2.3. Procedure

Participants completed testing in a sound attenuated room in the UCLA Phonetics Lab seated in front of a desktop computer. Stimuli were presented binaurally via a Peltor^{TM} 3M^{TM} listen only headset, on the platform Appsobabble [22]. The participants heard a stimulus and saw "p" on one side of the screen and "b" on the other. They indicated their choice via keypress where 'f' indicated the sound on the left side of the screen and 'j' indicated the sound on the right side of the screen. For 25 participants /p/ was on the left and for 25 it was on the right. Stimuli were grouped by the set of 40 unique stimuli, and randomized within this block. Participants heard 10 blocks with a short break halfway through. The experimental procedure took approximately 25 minutes to complete.

2.3. Results

Results were assessed using a linear mixed effects model with a logistic linking function. The fixed effects in the model were VOT (centered at zero), F0 and length. Categorical fixed effects were contrast coded. LONG was mapped to 1 and SHORT was mapped to -1. For F0, FLAT was mapped to 1 and LH was mapped to -1. The random effect structure in the model was by-subject random intercepts with maximal random slopes.

Table 1: *Model output: values are rounded. A colon indicates an interaction.*

	B **(SE)**	**z value**	**p value**
(Intercept)	1.51(0.11)	13.18	< 0.001
VOT	2.77(0.16)	17.42	< 0.001
F0	-0.06(0.03)	-2.24	0.03
length	-0.24(0.05)	-4.55	< 0.001
VOT:F0	0.022(0.05)	0.48	0.63
VOT:length	0.35(0.05)	6.55	< 0.001
F0:length	0.04(0.03)	1.39	0.16
VOT:F0:length	-0.08(0.03)	-2.44	0.015

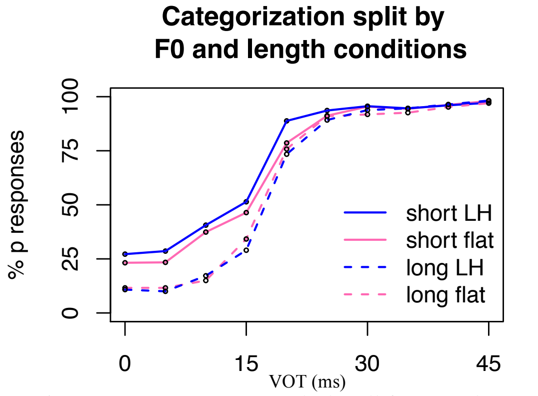

Figure 1: *Categorization split by all four conditions.*

Length had a significant effect in the model, whereby a LONG preceding syllable significantly decreased /p/ responses ($p < 0.001$). Length also significantly interacted with VOT in the model ($p < 0.001$) where the effect of length diminished as VOT values increased, indicating that lower VOT values are more ambiguous to listeners and thus more susceptible to shifts based on context. These effects are visible in Figure 1. The effect of length replicates the same effect found by Kim & Cho [5] and Mitterer et al. [6]. The crucial prediction of an asymmetry in F0's effect across length conditions is borne out in the three-way interaction ($p = 0.015$). This interaction between VOT, F0, and length was assessed first by using the lsmeans post hoc test, testing for the effect of F0 within each length condition. The test found a significant effect of F0 in the SHORT condition (Estimate = -0.187, z-ratio = -2.53, p = 0.01), but not in the LONG condition ($p = 0.52$). Because the interaction is also crucially linked with VOT, it was further investigated by collapsing F0 as a variable. This was done by calculating the proportion of /p/ responses for each participant for each VOT value in each F0 condition and then subtracting the proportion of /p/ responses in the FLAT condition from those in the LH condition. This was done *within* each length condition, thus providing information about how F0 is shifting categorization in each length condition. These values are named delta LONG and delta SHORT. Averaging these values across participants provides a visual assessment of the effect of F0, where a positive value indicates that LH F0 increases /p/ responses, and a negative value indicates that FLAT F0 increases /p/ responses. The larger the absolute delta value, the larger the magnitude of the effect. This is shown in Figure 2.

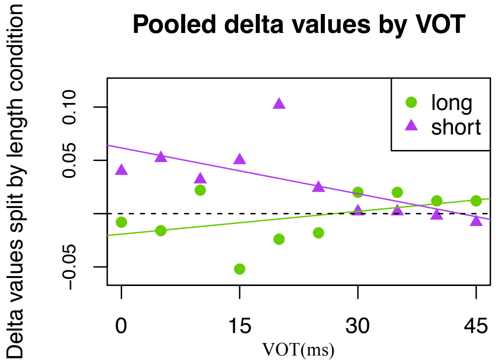

Figure 2: *Delta values split by length condition.*

Figure 2 shows that in the SHORT condition only, the delta values are higher at the lower (more ambiguous) end of the VOT continuum. No such trend exists in the LONG condition. A significant negative correlation (Kendall's rank) was found between the pooled delta SHORT values and VOT ($r_\tau = -0.68$, $p < 0.01$), while no correlation was found in the LONG condition. Overall, the results show that F0 only exerted an influence in the SHORT condition, where LH F0 significantly increased /p/ responses: visually, the SHORT lines in the categorization function in Figure 1 are separate while the LONG lines are not. This suggests that listeners interpret the compressed boundary tone (L-H%) over a short vowel as being spoken more quickly, providing evidence that tonal distributions are crucially linked to listener's computation of speech rate. More broadly, this suggests awareness of prosodic structure as cued by intonation, which modulates categorization of VOT.

3. Experiment 2

3.1. Motivation

Experiment 2 tests if listeners' expectations about phrase final lengthening affect their categorization of another temporal contrast. Listeners categorized a vowel length continuum as a word ending in either a voiced or voiceless obstruent. In English, vowels preceding voiced obstruents are longer than those preceding voiceless ones [23] [24], and listeners use this information to categorize an obstruent as voiced or voiceless, e.g. [25]. Given that segments are lengthened phrase-finally e.g. [21], Experiment 2 tested if the target word being phrase final shifted listeners' categorization of the vowel length continuum. Because listeners are sensitive to segments' lengths, rating longer sounds as sounding more natural when phrase final [26], if listeners incorporate their expectations of phrase final lengthening in categorization of the vowel length continuum, they would be expected to require longer segment durations for a voiced obstruent response when the target is phrase final. The experiment tested this by placing the target vowel in either phrase-FINAL position ("I'll say *x*") or in the phrase-MEDIAL position ("I'll say *x* now"). A shift in categorization in this experiment would suggest that listeners are sensitive to phrasal position, and crucially that they accommodate expectations about phrase final lengthening, which are separate from normalizing for speech rate. This point will be discussed further below.

3.2. Experimental Design

The experiment was a 2AFC task, with the same testing location and platform used in Experiment 1. Participants categorized a vowel length continuum as one of two lexical items: "coat" or "code". These two words were chosen because they are closely matched for frequency (from the SUBTEX corpus [27]; "coat" $\text{Log10}_{\text{WF}} = 3.33$, "code" $\text{Log10}_{\text{WF}} = 3.43$), meaning frequency effects should have a minimal influence on categorization.

3.2.1. Stimuli

Stimuli were recorded and manipulated by the same method as in Experiment 1. The starting point for the stimuli was a production of "I'll say code now", with the target being "code/coat". This served as the frame for the MEDIAL condition. The FINAL condition frame was made by removing "now" so that the target was phrase final. To create the target sound, the natural production of "code" was excised, and the audible stop voicing was edited out to render it more ambiguous. The vowel duration was then manipulated via resynthesis, to create a vowel length continuum ranging from 80 ms to 240 ms in 20 ms steps (9 steps total). The steps from the continuum were then inserted into both frames to create a total of 18 unique stimuli (9 vowel durations x 2 positions).

3.2.2. Participants

32 monolingual American English speaking students at UCLA participated in the experiment for course credit. 2 participants were excluded by the same criteria used in Experiment 1.

3.2.3. Procedure

The experimental procedure was identical to that in Experiment 1, with the difference being participants saw "coat" and "code" on the computer screen. Trials were divided into small blocks, blocked by condition. Each block consisted of four repetitions of the 9 steps on the continuum in a given position condition (randomized within block). In each half of the experiment, participants heard two FINAL blocks and two MEDIAL blocks, with presentation of blocks randomized as well. They were given a short break and then heard another two FINAL blocks and two MEDIAL blocks. The experimental procedure took about 20 minutes to complete.

3.3. Results

Results were assessed by a linear mixed effects model with a logistic linking function, with vowel length (centered at zero) and position (contrast coded where FINAL was mapped to 1 and MEDIAL was mapped to -1) as fixed effects. Random effects were by-subject random intercepts with maximal random slopes. Categorization split by condition is shown in Figure 3.

Table 2: *Model output.*

	***B*(SE)**	**z value**	**p value**
(Intercept)	0.54(0.12)	4.34	< 0.001
position	-0.25(0.09)	-2.62	< 0.01
duration	2.45(0.19)	12.99	< 0.001
position:duraiton	-0.30(0.08)	-3.62	< 0.01

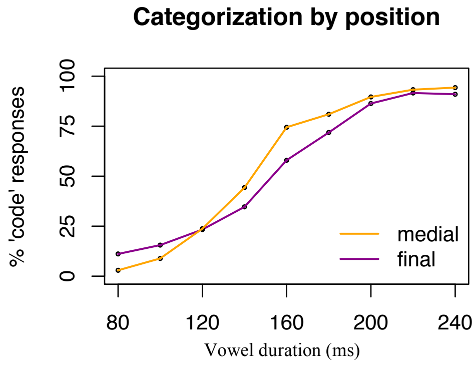

Figure 3: *Categorization split by condition.*

The model found a significant effect of vowel length ($p < 0.001$) as would be expected from any such continuum. Crucially position also showed a significant effect ($p < 0.01$), where the FINAL position decreased "code" responses. In other words, longer durations were required for a "code" response in the FINAL condition, suggesting listener adjustment for expectations about phrase final lengthening in categorization. The significant interaction ($p < 0.01$) was assessed using lsmeans, testing for the effect of position at each vowel duration. The test found that at the three shortest durations of the vowel length continuum, there was no significant effect of position, but at all higher lengths position showed a significant effect ($p < 0.05$). This interaction suggests that the influence of position increases as the vowel becomes longer.

The effect observed here is crucially separable from speech rate normalization. Consider that in the MEDIAL condition the target is followed by "now", which is lengthened by virtue of being phrase-final. Given that following sounds have been shown to shift the categorization of a preceding segment [28] [29], it might be expected that the lengthened "now" (with a duration of 325 ms) following the target in the MEDIAL condition would make it sound relatively short, thus shifting categorization to longer required durations for a "code" response (as compared to the FINAL condition). This is clearly not the case: longer durations are required for a "code" response in the FINAL, not the MEDIAL, condition. This suggests that listeners are sensitive to phrase-final lengthening, and are not simply normalizing for speech rate.

4. Discussion

The two experiments presented above provide some evidence that listeners are sensitive prosodic structure in their categorization of speech segments.

Experiment 1 showed that listeners appear to be sensitive to intonational categories or intonationally defined prosodic structure for the purposes of computing speech rate, where a boundary tone compressed onto a short vowel gives the percept of increased speech rate. This implicates listener awareness of tonal melodies as encoding phrasal boundaries, which mediate rate percepts. Exploring how F0 is incorporated into rate percepts across languages might be insightful, as languages with different intonational systems or different inventories of intonational categories would be expected to show sensitivity to F0 as signaling speech rate in different ways. For example, because the LH pattern can occur at the end of an AP (Accentual Phrase) in Seoul Korean [30], without substantial lengthening as in an IP, Seoul Korean speakers may not shift their categorization on the basis of this tonal contour. Further exploring how more global rate normalization (in longer utterances) interfaces with the current results may also be informative.

Experiment 2 showed that expectations about phrase final lengthening, independently of normalization for speech rate, shifted the required duration for a voiced obstruent response in the categorization of a vowel length continuum. This result shows that listeners are sensitive to the prosodic position of a given target sound, and use this information to mediate categorization of a temporal cue to segment identity.

Taken together, both experiments provide some evidence that listeners do indeed incorporate expectations from prosodic structure in segmental categorization. Extending these results will be crucial to better understanding the central question addressed here. For example, investigating how these factors function cross-linguistically and seeking to better understand their interaction with (non-local) rate normalization in perception will prove insightful.

5. Acknowledgements

Many thanks to Sun-Ah Jun, Megha Sundara, Pat Keating and Jesse Harris for valuable advice. Thanks to Adam Royer for recording for the stimuli, to Henri Tehrani for technical assistance and to Yang Wang for work as an RA.

6. References

[1] Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

[2] Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

[3] Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America*, *69*(2), 548–558.

[4] Sjerps, M. J., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics*, *41*(3), 145–155.

[5] Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19–EL25.

[6] Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics*, *54*, 68–79.

[7] Cho, T. (2002). The Effects of Prosody on Articulation in English. New York, NY: Routledge.

[8] Jun, S.-A. (1993). *The Phonetics and Phonology of Korean Prosody*. The Ohio State University.

[9] Keating, P. (2006). Phonetic Encoding of Prosodic Structure. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 167–186). New York and Hove: Macquarie Monographs in Cognitive Science, Psychology Press.

[10] Keating, P., Fougeron, C., Hsu, C., & Cho, T. (2003). Domain initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press.

[11] Pierrehumbert, J., & Talkin, D. (1992). Lenition of / h / and glottal stop. In G. Doherty & D. R. Ladd (Eds.), *Papers in laboratory phonology II : gesture segment prosody* (pp. 90–116). Cambridge University Press.

[12] Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485.

[13] Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, *37*(1), 46–65.

[14] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance*, *7*(5), 1074–1095.

[15] Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*(01), 255–309.

[16] Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (PhD). Massachusetts Institute of Technology.

[17] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Commun.*, *9*(5-6), 453–467.

[18] Boersma, P. & Weenik, D. (2017). Praat: doing phonetics by computer [Computer Program]. Version 6.0.36.

[19] Beckman, M. E. and Ayers, G. M. (1994), Guidelines for ToBI Labelling. Online MS and accompanying files. Available at http://www.ling.ohio-state.edu/phonetics/E_ToBI.

[20] Beckman, M. E. and Hirschberg, J. (1994), The ToBI Annotation Conventions. Online MS. Available at http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html.

[21] Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

[22] Tehrani, H. (2015). Appsobabble [online applications platform] http://www.appsobabble.com.

[23] Chen, M. (1970). Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica*, *22*(3), 129–159.

[24] Walsh, T., & Parker, F. (1981). Vowel length and voicing in a following consonant. *Journal of Phonetics*, *9*(3), 305–308.

[25] Raphael, L. J. (1972). Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.

[26] Klatt, D. H., & Cooper, W. E. (1975). Perception of Segment Duration in Sentence Contexts. In *Structure and Process in Speech Perception* (pp. 69–89). Springer, Berlin, Heidelberg.

[27] Brysbaert, M. & New, B. (2009) Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. In *Behavior Research Methods*, 41 (4), 977-990.

[28] Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512.

[29] Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560.

[30] Jun, S.-A. (2005). Korean Intonational Phonology and Prosodic Transcription. In S.-A. Jun (Ed.), *Prosodic Typology* (pp. 201–229). New York: Oxford University Press.