Prosodic boundary information modulates phonetic categorization

Sahyang Kim
*Department of English Education, Hongik University, Sangsu-dong, Mapo-gu,
Seoul 121-791, Korea*
sahyang@hongik.ac.kr

Taehong Cho[a]
*Hanyang Phonetics and Psycholinguistics Laboratory, Department of English Language
and Literature, Hanyang University, Haengdang-dong 17, Seongdong-gu,
Seoul 133-791, Korea*
tcho@hanyang.ac.kr

Abstract: Categorical perception experiments were performed on an English /b-p/ voice onset time (VOT) continuum with native (American English) and non-native (Korean) listeners to examine whether and how phonetic categorization is modulated by prosodic boundary and language experience. Results demonstrated perceptual shifting according to prosodic boundary strength: A longer VOT was required to identify a sound as /p/ after an intonational phrase than a word boundary, regardless of the listeners' language experience. This suggests that segmental perception is modulated by the listeners' computation of an abstract prosodic structure reflected in phonetic cues of phrase-final lengthening and domain-initial strengthening, which are common across languages.

PACS numbers: 43.71.Es, 43.71.Hw, 43.71.Sy [SGS]
Date Received: February 1, 2013 **Date Accepted:** May 7, 2013

1. Introduction

It is well known that a listeners' phonetic categorization is modulated by numerous contextual factors. Some perceptual modulation results from higher-level linguistic structure (e.g., lexical knowledge, Ganong, 1980; sentential contexts, Connine, 1987). It also comes from factors that bring about systematic subphonemic variation in speech production, such as coarticulation (Mann and Repp, 1980) and speech rate (Miller *et al.*, 1984). Another important source of subphonemic variation that has received increasing attention is prosodic structure of an utterance. Phrase-final lengthening and domain-initial strengthening are two well-known examples of such variation, lengthening segments in the vicinity of the prosodic boundary in proportion to boundary strength. For example, segments, whether consonantal or vocalic, are longer in the phrase-final position than in the phrase-medial position (phrase-final lengthening; e.g., Turk and Shattuck-Hufnagel, 2007); and consonants are produced with longer constriction duration [and longer voice onset times (VOTs), if they are voiceless aspirated stops] in the phrase-initial position than in the phrase-medial position (domain-initial strengthening; e.g., Fougeron and Keating, 1997; Cho and Keating, 2009).

The boundary-related effects are assumed to signal the prosodic structure of a given utterance, providing testable hypotheses regarding how the boundary information is exploited by listeners at different levels of speech comprehension [see Chap. 7 of Cutler (2012) for a review]. Some recent studies have indeed shown that such

[a] Author to whom correspondence should be addressed.

information plays a role in lexical segmentation and recognition (e.g., Cho *et al.*, 2007; Tyler and Cutler, 2009; Kim and Cho, 2009) and in syntactic parsing (e.g., Schafer *et al.*, 2000; Carlson *et al.*, 2001). This multi-level contribution of the perceived prosodic boundary in speech comprehension, along with the listeners' general sensitivity to segmental variations in phonetic categorization, leads to another testable hypothesis: Listeners use the boundary information in categorizing upcoming (post-boundary) segments. To the best of our knowledge, no study has tested the hypothesis directly. The present study therefore takes the initiative to test the effect of perceived boundary information on phonetic categorization of an upcoming stop consonant along a /b/-/p/ continuum in American English.

The /b/-/p/ continuum was created by manipulating VOT in different prosodic boundary contexts. A target-bearing syllable was inserted in a carrier sentence, "Let's hear # /Xa/ again" (X = a sound in a /b-p/ continuum), in which the prosodic boundary "#" varied with an intonational phrase (=IP) vs a prosodic word (=Wd) boundary. The critical fact is that a voiceless stop is produced with a longer VOT after an IP than a Wd boundary (Pierrehumbert and Talkin, 1992). If the perceived boundary information modulates phonetic categorization of an upcoming (post-boundary) segment, listeners may take into account the boundary-induced VOT lengthening pattern, expecting a relatively longer VOT for a voiceless /p/ percept after an IP vs a Wd boundary. If this is the case, the categorization function may show a rightward shift when the preceding context is perceived as signaling an IP boundary.

The present study tests both native and non-native (Korean) listeners of American English in order to determine whether the hypothesized effect can be observed across the listener groups. Korean aspirated stops are generally known to be produced with longer VOTs than their English counterparts by some 20 ms (e.g., Lisker and Abramson, 1964). More recent studies, however, suggest that the VOT difference between the two languages has been reduced over time to less than a 10 ms difference in utterance-initial position (e.g., 73 ms vs 68 ms; Kang and Guion, 2006; cf. Silva, 2006) as well as in phrase-initial position inside an utterance (e.g., ranging from 30 to 40 ms in both languages; Cho and Keating, 2001, 2009). As the listeners' native language is known to influence phonetic categorization of non-native speech sounds (e.g., Cutler, 2012, Chap. 2), Korean listeners are expected to show a different categorical perception pattern as compared to native listeners of English. What is crucial for the purpose of the present study, however, is whether non-native (Korean) listeners shift a /b/-/p/ categorical boundary in English according to perceived boundary strength, similar to native listeners. Given that both Korean and English have phrase-final (pre-boundary) lengthening as a robust cue for prosodic boundary used in speech comprehension (e.g., Tyler and Cutler, 2009; Kim *et al.*, 2012), and given that Korean shows domain-initial strengthening by lengthening VOTs of post-boundary aspirated stops as in English (Cho and Keating, 2001), both Korean and English listeners may show a perceptual shift in a comparable way. However, because non-native listeners often attend to different perceptual cues than native listeners (e.g., Tremblay *et al.*, 2012) it may be possible that Korean listeners do not make use of the boundary information in the same way as native English listeners. Moreover, in order to evaluate whether any perceptual shift observed with Korean listeners is due to their experience with English or to their knowledge of their own language, two subgroups of Korean listeners were tested: An advanced learner group and a beginner group. If Korean listeners' ability to use boundary information in English as a function of their English experience, then advanced Korean learners of English would behave more like the English native listeners than novice Korean learners of English. If, however, the boundary-dependent categorical perception has less to do with English experience, but more to do with Korean listeners' knowledge of their first language ($L1$), then no difference between the learner groups will be observed.

The speech materials were further manipulated in two other prosodic aspects. First, the stop closure duration (CD) was manipulated as it also varies as a function of prosodic boundary strength: Like VOT, CD is longer after an IP boundary than after a word boundary in both English and Korean (Cho and Keating, 2001, 2009). It is therefore possible that a longer CD may help signal an IP boundary whereas a shorter CD (mismatched with an IP boundary) may weaken the perceived strength of an IP boundary. Alternatively, given that a voiceless stop is produced with a longer VOT as well as with a longer CD (than a voiced stop), there may be a trade-off between CD and VOT, so that upon hearing a shortened CD, listeners may require a longer VOT to compensate for it. Thus, with the possibility that different CDs may influence the way that a prosodic boundary is perceived, two values of the stop CD (long vs short, appropriate for the IP and the Wd boundary conditions, respectively) were used to test how the boundary-dependent perceptual shift is further conditioned by CD.

Second, the target-bearing syllable varied in terms of pitch accent or phrase-level stress (accented vs unaccented) as pitch accent in English is also known to affect speech comprehension at various linguistic levels (Cutler *et al.*, 1997). Since a voiceless stop in English is produced with a longer VOT in pitch-accented than in unaccented syllables (Cole *et al.*, 2007; Cho and Keating, 2009), the presence or absence of pitch accent may also influence VOT-based categorical perception of the stop. Furthermore, given that the boundary effect is often more robust when there is no conflicting effect coming from pitch accent in speech production (Cho and Keating, 2009), a more robust perceptual effect of prosodic boundary may also arise when the target-bearing syllable receives no pitch accent. Testing these two additional factors along with the boundary and the listener group factors will therefore allow us to examine the hypothesized boundary-induced perceptual shift in various contexts, which would illuminate its nature in a more informed way.

2. Methods

2.1 Participants

For the native listener group, 20 native listeners of American English (NE) in their twenties participated. They were temporary residents in Korea, and only one could speak Korean fluently. For the native Korean listener (NK) group, 40 Korean undergraduates living in Korea in their twenties participated: Half of them were advanced English learners (TOEIC score 940 to 990, average percentile rank = 98), and the other half were beginners (TOEIC score 315 to 600, average percentile rank = 25). (TOEIC is a standardized English proficiency test offered by the Educational Testing Service on receptive listening and reading proficiency.) All were paid for their participation and none reported any known hearing problems.

2.2 Stimuli

A male native speaker of American English recorded the target syllables *pah/bah* multiple times in a sentence, *Let's hear #/Xa/ again*. The speaker was trained to produce the sentence with four different prosodic patterns (Table 1). The target-bearing sentences were varied with the boundary (#) before the target syllable (IP vs Wd). In these sentences, a high pitch accent (H*) was placed on the target syllable in the accented

Table 1. The carrier sentence produced with different phrasing and accent. Accented words are marked in bold. Square brackets indicate Intonational Phrases. The prosodic transcription provided is based on AE-ToBI (Beckman *et al.*, 2005).

	Pitch accent	No pitch accent
IP-initial	[**Let's** hear]$_{\mathrm{IP}}$ [**pah** again]$_{\mathrm{IP}}$ H* L-L% H* L-L%	[**Let's** hear]$_{\mathrm{IP}}$ [pah **again**]$_{\mathrm{IP}}$ H* L-L% H* L-L%
Word (IP-medial)	[**Let's** hear **pah** again]$_{\mathrm{IP}}$ H* H* L-L%	[**Let's** hear pah **again**]$_{\mathrm{IP}}$ H* H* L-L%

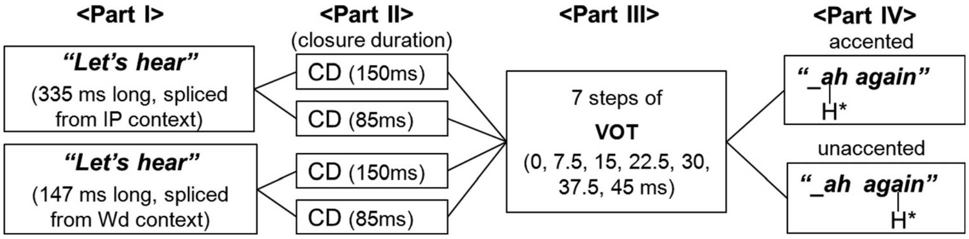

Fig. 1. Stimuli concatenation for the phonetic categorization experiment.

condition, and on *again* in the unaccented condition. The sentence-initial word *Let's* was pitch-accented with H* in all conditions. Using these speech materials, the stimuli to be used for experiments were created by concatenating four parts taken from the recorded sentences as illustrated in Fig. 1.

For Part I, two *Let's hear* tokens were selected, one spliced from an IP context sentence and one spliced from a Wd (IP-medial) context sentence as given in Table 1. The selected *Let's hear* token from the IP context was 355 ms long and the *Let's hear* token from the Wd context was 147 ms long, showing a substantial durational difference. The similarity in *F*0 distribution was also considered in selecting the two tokens, in order to ensure that the *F*0 transition from *Let's hear* to the rest of the sentence did not noticeably vary between IP and Wd contexts. For both conditions, *F*0 rises during the *Let's* portion (with H*) and falls on the *hear* portion. For Part II (CD), 150 ms (for IP) and 85 ms (for Wd) CDs were employed. The latter was the average CD for IP-medial /p/ from the recorded tokens. Since IP-initial CD cannot be measured from the acoustic data, its value was obtained from data in Cho and Keating's (2009) electropalatographic study. Other than CD, no additional pause was added to the stimuli. For Part III, the VOT for the target consonant was manipulated, using PSOLA resynthesis in Praat. There were seven 7.5 ms VOT steps, ranging from 0 to 45 ms. Note that more VOT steps above 45 ms could have been used, but results of our recent phonetic categorization task with Korean listeners (Kim *et al.*, 2012) indicated that the 50% crossover points for aspirated stops centered around 30 ms. We therefore decided to use steps up to 45 ms in order to reduce the experiment time. Finally, for Part IV (the remaining string), two -*ah again portions* were selected, one from an accented context and one from an unaccented context. The accent difference on the target-bearing syllable was confirmed both numerically and perceptually. In the accented compared to the unaccented condition, the vowel /a/ had a higher *F*0, longer duration, and higher intensity; and it was perceived as accented (by the authors). The four parts were concatenated with all possible combinations (as shown in Fig. 1), creating 56 stimuli (=2 × 2 × 7 × 2).

2.3 Procedure

Each stimulus was repeated 10 times, yielding 560 stimuli. The stimuli were distributed across four blocks separated by Accent and CD conditions: accented/long CD, accented/short CD, unaccented/long CD, and unaccented/short CD. Stimuli were presented in two 30-min sessions on different days. The stimuli and blocks were presented in random order. The Nijmegen Experiment Set Up, developed at the Max Planck Institute for Psycholinguistics, was used for stimulus presentation and data collection. Subjects were seated in front of a PC with a button box in the sound treated perception booth at the Hanyang Phonetics and Psycholinguistics Lab., and heard the stimuli through Sennheiser PC-161 headphones at a comfortable listening level—i.e., its peak was initially set at about 65 dB sound pressure level, but during the practice session, subjects were allowed to adjust the volume to their comfortable level. A 2AFC procedure was used. Half of the subjects were presented with the written "bah" on the left and "pah" on the right side of the screen, and the other half with the opposite order. They were asked to press a button as fast and as accurately as possible.

3. Results and discussion

A logistic regression line was fitted to the boundary region curve of each subject's raw responses, and repeated measures analysis of variances (ANOVAs) were performed on the estimated 50% crossover points with one between-subject factor, Listener group (NE, advanced NK, beginning NK) and three within-subject factors, Boundary (IP vs Wd), CD (85 ms vs 150 ms), and Accent (accented vs unaccented).

ANOVAs revealed a significant main effect of the Listener group ($F[2,57] = 11.57$, $p < 0.001$). Results of *post hoc* tests (Bonferroni/Dunn) revealed that the crossover points were significantly higher for NK listeners (21.2 ms for both advanced and beginner groups) than for NE listeners (16 ms) [$p < 0.01$; see Fig. 2(a)], perhaps reflecting the fact that Korean voiceless aspirated stops are produced with longer VOTs than their English counterparts. ANOVAs also showed a significant main effect of Boundary: The crossover point was higher in the IP than in the Wd condition (18 ms vs 14.3 ms; $F[1,57] = 47.82$, $p < 0.001$), showing a boundary-induced perceptual shift. Crucially, Boundary did not interact with the Listener group ($F[2,57] = 1.03$, $p > 0.1$), indicating that the Boundary effect was independent of listeners' language background and experience as can be seen in Fig. 2(b). Boundary did not interact with CD, either ($F[1,57] = 3.01$, $p > 0.05$), indicating that the effect was also independent of closure duration.

Boundary, however, interacted with Accent, which was reflected in both a two-way interaction between Boundary and Accent ($F[1,57] = 4.39$, $p < 0.05$) and a three-way interaction among Boundary, Accent, and CD ($F[1,57] = 5.35$, $p < 0.05$). Results of *post hoc* tests and eta statistics indicated that the Boundary by Accent interaction stemmed at least in part from the fact that the boundary effect was more robust when the target-bearing syllable was unaccented (mean diff., 6.5 ms, $t^{59} = 51.19$, $p < 0.001$, eta$^2 = 0.47$) than when it was accented (mean diff., 3.7 ms, $t^{59} = 42.19$, $p < 0.001$, eta$^2 = 0.42$). Results of *post hoc* tests for the three-way interaction, however, revealed that the robust boundary effect in the unaccented condition was largely due to an even more robust boundary effect when CD was short (unaccented/short CD: mean diff., 7 ms, $t^{59} = 63.91$, $p < 0.001$, eta$^2 = 0.52$; unaccented/long CD: mean diff., 4.5 ms, $t^{59} = 24.49$, $p < 0.001$, eta$^2 = 0.29$). This can be seen as coming from a trading relation between CD and VOT: When CD was not sufficiently long for an IP-initial stop (in the short CD condition), listeners compensated for it by requiring a relatively longer VOT for a voiceless percept, but only when the target-bearing syllable was not prominent (i.e., when unaccented). While we do not have a clear explanation for why a similar trading did not occur in the accented (prominent) condition, it may be related to accent-induced perceptual saliency of the target-bearing syllable. Given that accent

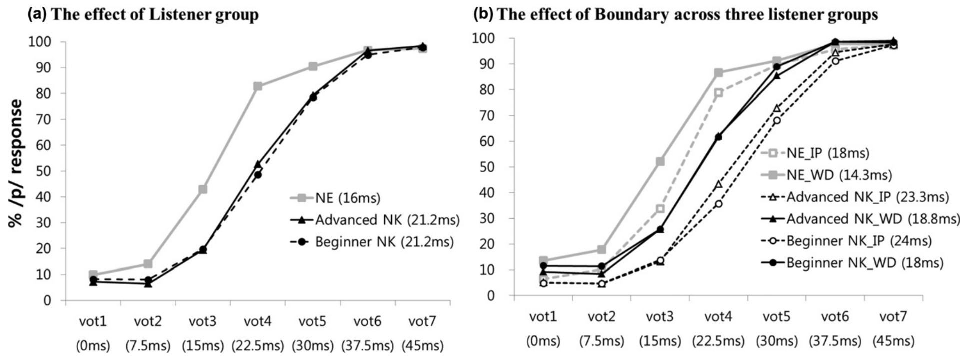

Fig. 2. Phonetic categorization (/ba-pa/) as a function of (a) the Listener group (NE, Advanced NK, Beginner NK) and (b) Boundary (IP vs Wd) across three listener groups. The numbers in parentheses indicate the estimated 50% crossover point (in ms) in each condition. The categorization curves were drawn based on the data pooled across conditions of Pitch Accent and Closure Duration.

is phonetically realized primarily on the vowel, listeners may pay more attention to the vowel when accented, thus being less sensitive to the consonantal variation in CD and VOT. On the other hand, when the syllable is unaccented, the locus of attention may be shifted to the consonantal variation that is known to be modulated by boundary strength, particularly for unaccented stimuli (Cho and Keating, 2001, 2009).

Accent did not produce a significant main effect ($F[1,57] = 1.48$, $p > 0.1$). As discussed above, there were significant interactions involving Accent, but *post hoc* comparisons revealed no single case in which its effect reached significance. Accent did not interact with the Listener group, either, suggesting that its null effect was consistent across listener groups. CD, on the other hand, yielded a significant main effect ($F[1,57] = 7.25$, $p < 0.01$), but interacted with Boundary and Accent as discussed above. Further *post hoc* comparisons revealed that the effect of CD was significant only in one condition: When the target-bearing syllable was unaccented after an IP boundary ($p < 0.001$). This three-way interaction can also be accounted for by the same trading relation that was discussed with respect to the Boundary effect: The relatively weakened voiceless percept (caused by a shortened CD) is compensated for by a lengthened VOT. This effect did not interact further with the Listener group, suggesting that the observed trading effect was independent of the listeners' language background.

4. General discussion

The present phonetic categorization study revealed three major results. First, both native English and non-native (Korean) listeners showed a robust perceptual shift in phonetic categorization as a function of a perceived prosodic boundary. Categorization of an ambiguous sound as /p/ required a longer VOT in the IP boundary than in the Wd boundary contexts. This implies that when the preceding context provided IP boundary cues, listeners took into account the IP-boundary induced (domain-initial) lengthening of VOT, and therefore required a corresponding, longer VOT for an upcoming post-boundary stop. Second, non-native (Korean) listeners showed a prosodic boundary-induced perceptual shift similar to native listeners, and thus the effect was independent of their experience with English as a second language. This implies that non-native listeners' knowledge of their native language can be easily carried over in processing the prosodic structure of a non-native language. When both the source and the target languages employ similar patterns in phrase-final and domain-initial lengthening, listeners' perceptual adjustments to such prosodically conditioned speech variations appear to come about in processing a non-native language without substantial learning. This is also consistent with the suggestion that prosodic cues are more available than segmental cues to listeners of different languages because many prosodic cues are common across languages (Cutler, 2012, Chap. 10). Finally, results showed a kind of perceptual trading relation between CD and VOT when the target-bearing syllable was unaccented. Specifically, listeners required an even longer VOT for an IP-initial voiceless stop when CD was short, but only in the unaccented condition where the locus of perceptual attention may be shifted to the consonantal variation. This implies that categorical speech perception is modulated not only by boundary information that precedes the target sound but also by its interaction with Accent that is realized on the target-bearing syllable.

Taken together, the results of the present study demonstrate a case in which both native and non-native listeners behave in comparable ways in terms of boundary-induced adjustment of phonetic categorization, presumably driven by cross-linguistically applicable prosodic information of boundary finality and domain-initial strengthening. Although the data presented in the present study are limited to one particular stop voicing contrast in English, it is hoped that this study sparks future research exploring how phonetic manifestation of the abstract prosodic structure is indeed perceptually relevant at different levels of speech processing and how such prosodic modulation can be incorporated into current models of native and non-native speech perception.

Acknowledgments

We thank Anne Cutler for inspiring us with her initial input on this research project.

References and links

Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. (**2005**). “The original ToBI system and the evolution of the ToBI framework,” in *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford), pp. 9–54.

Carlson, K., Clifton, C., Jr., and Frazier, L. (**2001**). “Prosodic boundaries in adjunct attachment,” J. Mem. Lang. **45**, 58–81.

Cho, T., and Keating, P. A. (**2001**). “Articulatory and acoustic studies of domain-initial strengthening in Korean,” J. Phonetics **29**, 155–190.

Cho, T., and Keating, P. A. (**2009**). “Effects of initial position versus prominence in English,” J. Phonetics **37**, 466–485.

Cho, T., McQueen, J. M., and Cox, E. (**2007**). “Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English,” J. Phonetics **35**, 210–243.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (**2007**). “Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News Speech,” J. Phonetics **35**, 180–209.

Connine, C. M. (**1987**). “Constraints on interactive processes in auditory word recognition: The role of sentence context,” J. Mem. Lang. **26**, 527–538.

Cutler, A. (**2012**). *Native Listening: Language Experience and the Recognition of Spoken Words* (The MIT Press, Cambridge, MA), pp. 1–576.

Cutler, A., Dahan, D., and van Donselaar, W. (**1997**). “Prosody in the comprehension of spoken language: A literature review,” Lang. Speech **40**, 141–201.

Fougeron, C., and Keating, P. A. (**1997**). “Articulatory strengthening at edges of prosodic domains,” J. Acoust. Soc. Am. **101**, 3728–3740.

Ganong, W. F. (**1980**). “Phonetic categorization in auditory word perception,” J. Exp. Psychol. Hum. Percept. Perform. **6**, 110–125.

Kang, K.-H., and Guion, S. (**2006**). “Phonological systems in bilinguals: Age of learning effects on the stop consonant systems of Korean-English bilinguals,” J. Acoust. Soc. Am. **119**, 1672–1683.

Kim, S., Broersma, M., and Cho, T. (**2012**). “The use of prosodic cues in processing an unfamiliar language,” Stud. Second Lang. Acquis. **34**, 415–444.

Kim, S., and Cho, T. (**2009**). “The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean,” J. Acoust. Soc. Am. **125**, 3373–3386.

Lisker, L., and Abramson, A. S. (**1964**). “A cross-language study of voicing in initial stops: Acoustical measurements,” Word **20**, 384–422.

Mann, V. A., and Repp, B. H. (**1980**). “Influence of vocalic context on perception of the [S]-[s] distinction,” Percept. Psychophys. **28**, 213–228.

Miller, J. L., Grosjean, F., and Lomanto, C. (**1984**). “Articulation rate and its variability in spontaneous speech: A reanalysis and some implications,” Phonetica **41**, 215–225.

Pierrehumbert, J., and Talkin, D. (**1992**). “Lenition of /h/ and glottal stop,” in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by G. Docherty and D. R. Ladd (Cambridge University Press, Cambridge), pp. 90–117.

Schafer, A. J., Speer, S. R., Warren, P., and White, S. D. (**2000**). “Intonational disambiguation in sentence production and comprehension,” J. Psycholinguist. Res. **29**, 169–182.

Silva, D. J. (**2006**). “Acoustic evidence for the emergence of tonal contrast in contemporary Korean,” Phonology **23**, 287–308.

Tremblay, A., Coughlin, C. E., Bahler, C., and Gaillard, S. (**2012**). “Differential contributions of prosodic cues in the native and non-native segmentation of French speech,” Lab. Phonology **3**, 385–423.

Turk, A. E., and Shattuck-Hufnagel, S. (**2007**). “Multiple targets of phrase-final lengthening in American English words,” J. Phonetics **35**, 445–472.

Tyler, M., and Cutler, A. (**2009**). “Cross-language differences in cue use for speech segmentation,” J. Acoust. Soc. Am. **126**, 367–376.