Journal of Phonetics 54 (2016) 68–79

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

How does prosody influence speech categorization?

Holger Mitterer [a], Taehong Cho [b,*], Sahyang Kim [c]

[a] Department of Cognitive Science, University of Malta, Msida, Malta
[b] Hanyang Phonetics and Psycholinguistics Lab, Department of English Language and Literature, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Republic of Korea
[c] Department of English Education, Hongik University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 24 December 2014
Received in revised form
8 September 2015
Accepted 10 September 2015
Available online 8 October 2015

Keywords:
Prosody
Phonetic categorization
English stops
Perceptual compensation
Speaking rate normalization
Phrase-final lengthening
English and Korean listeners

ABSTRACT

A recent study (Kim & Cho, 2013, *The Journal of the Acoustical Society of America*) reported that the perception of a prosodic boundary leads to a shift in a stop-identification function in English, so that stops with a relatively long VOT are accepted as voiced if occurring after a major prosodic boundary. Even Korean learners of English showed such a shift. This shift would seem to result from compensation for post-boundary lengthening effects (or domain-initial strengthening) and thereby help to overcome the invariance problem in speech perception. In two experiments, we ask how this effect comes about. The first experiment tested whether a simple adjustment to a change in overall speaking rate would be sufficient to account for the shift. Results showed that while the global speaking-rate change modulates phonetic categorization in a similar way as a change in the prosodic boundary strength, the speaking-rate effect is not sufficient to explain the boundary effect. That is, there was a more robust shift in a stop identification function with localized slowing down of the final syllable due to an intonational phrase (IP) boundary than with global slowing down of speaking rate. The second experiment therefore investigated the contribution of an *F*0 cue to the observed perceptual shift and found that the presence or absence of the *F*0 cue did not mediate the effect of prosodic boundaries on phonetic categorization. This suggests that a perception shift in phonetic categorization stems primarily from the listeners' adjustment to temporal variation, though its source is different from the speaking rate. The results are considered in terms of two possible accounts: one that takes both the boundary-induced and the speaking rate-induced effects as listeners' adjustments to low-level temporal variation, and the other that separates them by taking the boundary-induced effects to arise with computation of higher-level prosodic structure, given that the source of the localized slowing down effect is a prosodic boundary.

1. Introduction

Reductionism may be an unavoidable element of the scientific enterprise. Although a complex whole system is better understood by examining the characteristics of its parts one at a time, the interactions of those parts are then often unwantedly overlooked. In the field of phonetics, for instance, this practice may be found in a division of labor between segmental and suprasegmental (or prosodic) levels. As the term 'supra' indicates, the suprasegmental elements of speech have been taken to be realized independently of segmental elements, as described in most textbooks on phonetics (Ladefoged & Johnson, 2014; Reetz & Jongman, 2011; Rietveld & van Heuven, 1997). Recent years, however, have witnessed a large body of evidence for interactions between the two levels, illuminating an important aspect of the phonetics-prosody interface (for reviews see Cho, 2011; Fletcher, 2010; Mücke, Grice, & Cho, 2014; Shattuck-Hufnagel & Turk, 1996). It is now well established that segmental realization is intricately conditioned by the prosodic structure of a given utterance, so that, for example, the actual phonetic form of a phoneme is determined by the prosodic position in which it occurs (e.g., Cho & Keating, 2009; Cho, 2011; Cho, Lee, & Kim, 2014; Fletcher, 2010; Fougeron & Keating, 1997).

This leads to the question of how listeners can recognize both prosodic boundaries and the intended segment. Two quite different answers are provided independently by two fields that do not seem to communicate enough. On the one hand, from the field of prosodic phonology, it has often been assumed that listeners take into account the higher-order prosodic structure when interpreting

* Corresponding author. Tel.: +82 2 2220 0746; fax: +82 2 2220 0741.
E-mail addresses: holger.mitterer@um.edu.mt (H. Mitterer), tcho@hanyang.ac.kr (T. Cho), sahyang@hongik.ac.kr (S. Kim).

0095-4470/$ - see front matter
http://dx.doi.org/10.1016/j.wocn.2015.09.002

segments (e.g., Cho, McQueen, & Cox, 2007; Christophe, Peperkamp, Pallier, Block, & Mehler, 2004). On the other hand, work on speech perception suggests an alternative interpretation in terms of speech-rate normalization arising from low-level auditory processing (e.g., Newman & Sawusch, 2009; Reinisch & Sjerps, 2013). The present study explores these competing views in order to understand how much leverage each of these approaches may have and to explore the extent to which these could be further disentangled.

Prosodically-conditioned variation of speech adds to the invariance problem—that is, how the perception system copes with variation and extracts underlying linguistic information from the speech signal. This issue can be considered in terms under which variation-adding influences in production are mirrored by variation-subtracting processes in perception that allow the listener to arrive at a context-independent rendition of a given speech category. This perception-production "mirroring" has been well reflected in the way listeners compensate for variation-inducing factors, such as physiological differences in vocal-trace size (Ladefoged & Broadbent, 1957), coarticulation (Mann, 1980), phonological assimilation (Mitterer & Blomert, 2003), and speech rate (Summerfield, 1981). To provide one example of this mirroring, if a speaker speaks quickly, the listener will assume that an interval of medium duration is relatively long, while the same interval would be considered relatively short with a slow background speech rate. It is hence natural to ask whether listeners compensate for prosodic influences on segmental realization, and, if so, how. In this vein, Kuzla, Ernestus, and Mitterer (2010) indeed demonstrated that listeners' compensation for (voicing) assimilation is further modulated by the size of the prosodic boundary that falls between the assimilated and assimilating segments—i.e., compensation for assimilation is more robust when the boundary is smaller (e.g., a prosodic word boundary vs. a phrase boundary). This perceptual compensation again mirrors speech production, as speakers also produce stronger assimilation across smaller boundaries. There are two ways to explain such an effect, either as a higher-level prosodic effect or as a lower-level temporal effect. According to the first possibility, listeners recognize the prosodic boundary as such, and use the information to fine-tune the compensation for assimilation process. This is generally in line with the assumption that speech perception is modulated by the prosodic structure that is computed on-line by the listener (e.g., Cho et al., 2007; Christophe et al., 2004). On the other hand, a lower-level temporal explanation would be that the compensation effect is stronger because the assimilating segments that straddle a smaller prosodic boundary are shorter and less temporally separated than those that occur across a larger prosodic boundary. This is plausible, given that shorter and hence more ambiguous segments tend to be more vulnerable to contextual influences (Massaro, 1998), and that compensation for assimilation is a graded process that is influenced by phonetic detail (Gow, 2003; Mitterer, Csépe, Honbolygo, & Blomert, 2006).

Additional evidence for compensation for prosodic influences, however, has been provided by Kim and Cho (2013). In a phonetic categorization experiment, they employed a VOT continuum from /ba/ to /pa/ in English which was presented to the listener in a carrier phrase, as in *Let's hear* $^{b}/_{p}$*a again.* Crucially, the carrier phrase was recorded with two prosodic renditions: one with an Intonational phrase (IP) boundary and the other with a prosodic word (Wd) boundary before the test syllable (see Shattuck-Hufnagel & Turk, 1996 for a review of prosodic boundaries). The critical condition was therefore whether the test syllable was preceded by *Let's hear* as a full IP (*Let's hear* # $^{b}/_{p}$*a again*, where the prosodic boundary ('#') is an IP boundary) or as a part of an IP (*Let's hear* # $^{b}/_{p}$*a again*, '#' = an Word boundary). These two different prosodic boundaries suggest different phrase structures. The word boundary suggests that *again* is an adverb modifying *hear* ([$hear_{V}$ [ba_{N}]$_{NP}$]$_{VP}$ [*again*]$_{AdvP}$), while the IP boundary suggests that *ba again* is part of the object of the verb *hear* ([$hear_{V}$ [*ba again*]$_{NP}$]$_{VP}$). Given that the VOT of an aspirated stop in English is likely to be longer after an IP than after a Word boundary (Cho & Keating, 2009; Pierrehumbert & Talkin, 1992), Kim and Cho (2013) tested whether listeners would indeed take that into account in phonetic categorization of /b/-/p/—i.e., whether more aspiration (a longer VOT) is required for /p/ percept after an IP than after a Word boundary. The results showed that participants accepted stimuli with relatively longer VOTs as /b/s after an IP boundary compared to when the same stimuli were presented after a Wd boundary. That is, participants gave fewer /p/ responses to the same /ba/-/pa/ continuum in the IP boundary than in the Wd boundary condition, suggesting that listeners indeed adjusted their phonetic categorization of /b/-/p/ as a function of preceding prosodic contexts.

Kim and Cho (2013) ran the same experiment with two groups of Korean listeners with different levels of English proficiency (beginners vs. advanced learners). A basic finding was that both groups of Korean learners required more aspiration for /p/, thus giving more /b/ responses than native English listeners on the same /ba/-/pa/ continuum. This is in line with the fact that Korean aspirated stops are produced with a longer VOT than English aspirated stops (Abramson & Lisker, 1964; see Cho & Keating, 2001, 2009 for Korean and English data collected in similar ways), suggesting that Korean listeners used their native Korean categories for the identification of English stops. Nevertheless, the Korean listeners seemed to be attuned to the prosodic influences, even in English. Like English-speaking listeners, the Korean listeners also showed a shift of their categorization function between the two prosodic contexts. This shift was similar in size as that of the English listeners, and did not also differ between intermediate and advanced learners of English. As Kim and Cho (2013) suggest, it appears that the prosodic boundary in Korean and English is marked by some common (suprasegmental) features, and the Korean learners employ the common prosodic knowledge when processing English as an L2.

But just as in the case of Kuzla et al. (2010) discussed above, it is not clear whether the shift of phonetic categorization observed in Kim and Cho (2013) is attributable solely to the perception of the prosodic boundary *per se*. A closer look at the methods in Kim and Cho (2013) suggests that there may be a simpler explanation for both the effect of prosodic boundary on speech categorization and the native-like performance of the Korean listeners with English stimuli. The precursor (i.e., *Let's hear*) before the test syllable in the carrier sentence had a duration of 710 ms in the IP boundary condition and a duration of 450 ms in the Wd boundary condition, showing that the overall speaking rates were different between the two conditions (i.e., it was about 1.6 times faster in the Wd condition). As reviewed above, listeners tend to accept a relatively long segment as being "short" when the preceding context is longer. Summerfield (1981), for instance, presented continua from voiced to voiceless stops (similar to Kim and Cho, 2013) after

sentences with either a slow or a fast rate. After the slower (i.e., longer) precursor, listeners labeled the stops more often as voiced, parallel to the finding of Kim and Cho (2013). Importantly, recent evidence indicates that this is a 'low-level' effect, arising as a consequence of early auditory processing. Newman and Sawusch (2009) reported that speaking rate normalization occurs even when there is a change in speaker between the precursor and the target syllable (i.e., when the precursor sentence is spoken by one speaker and the target syllables by another) or a change in speaker's spatial location (i.e., the precursor is heard from one direction and the target syllable from another). This suggests that speaking rate information is used by listeners at a relatively early processing stage which precedes adjustments to speaker differences and auditory perceptual grouping. This conclusion is reinforced by Reinisch and Sjerps (2013), who used a time-sensitive, eye-tracking method to show an early effect of precursor speaking rate on the perception of a phonological duration contrast (in their case, a Dutch vowel contrast).

Proponents of the auditory account might therefore propose that speaking rate normalization[1] would be a more parsimonious explanation for the results of Kim and Cho (2013), as it would not be necessary to assume that both the speaking rate and the perception of prosodic boundary directly (and separately) influence speech categorization. That is, the influence of the prosodic boundary may be indirect, via speaking rate normalization. This interpretation also suggests a radical re-interpretation of Kim and Cho's (2013) results regarding non-native listeners. Kim and Cho suggested that Korean listeners, regardless of their English proficiency levels, may perform relatively well in English because they transfer their knowledge of Korean prosodic structure to English. If speaking rate normalization is however the driving force and this normalization is a relatively basic low-level auditory phenomenon that occurs at relatively early processing stages (Newman & Sawusch, 2009; Reinisch & Sjerps, 2013), one does not have to posit transfer of native-language knowledge to explain the results. That is, one cannot rule out the possibility that Kim and Cho's results are caused by a low-level general auditory implementation of speech rate effects that comes from the mammalian auditory system that both English and Korean listeners have in common (see, e.g., Lotto, Kluender, & Holt, 1997; Mann, 1986).

Considering the two alternative views, it becomes necessary for the field of speech perception (often carried out by psychologically informed researchers) and the field of prosodic phonology (often carried by linguistically informed researchers) to communicate with each other and share findings. It has been remarked recently that psychologists and linguists often seem to investigate similar issues with a lack of interaction that impedes scientific progress (Embick & Poeppel, 2014; Wagner & Klassen, 2015). The present study therefore incorporates both phonological and psychological approaches, with the goal to enhance our understanding of whether the perceptual shift of phonetic categorization as observed in Kim and Cho (2013) should be attributed to prosodic boundary perception (from the higher-order prosodic structure) or speaking rate normalization (as a low-level effect).

One way to help disentangle the potential confound between speaking rate and prosodic boundary strength is to run an experiment similar to Kim and Cho's (2013) and to compare the effects of the manipulation of prosodic boundaries with the effects of the manipulation of speaking rate, analogous to other studies. We did so in our first experiment by using a diphone synthesizer to generate stimuli in which temporal adjustments near a prosodic boundary (i.e., localized slowing down) vs. a change in speaking rate (i.e., global slowing down) were manipulated independently. Crucially, we created a stimulus in which the prosodic characteristics of the precursor were typical of a Wd boundary (especially in terms of relative segment duration and pitch) but the overall (total) duration was lengthened to be matched with that of a stimulus with an IP boundary. Listeners' performance in this condition (Wd-LONG) is then compared with that in the original IP boundary condition (IP-FULL). (See below for details.) If the two conditions yield different shifts in phonetic categorization, one might assume that the mechanism involved in the boundary-induced perceptual modulation is not the same mechanism as the one responsible for general speaking rate normalization. But if the effects turn out to be indistinguishable, then it is likely that general low-level speaking normalization underlies the boundary-related perceptual shift.

Before getting into the details of Experiment 1, it is also worth noting that, in keeping with the design of Kim and Cho (2013), we also used three listener groups: native speakers of English and intermediate and advanced learners of English with Korean as a native language. (It should be noted, however, that there is no theoretical ground for choosing Korean listeners as L2 listeners since any L2 learners of English may serve the purpose.) The comparison of native and non-native learners was included again in the present study because it allowed us to compare the results of the two studies in exactly the same conditions. Furthermore, given that both the localized and global slowing down effects are assumed to occur across languages, it is useful to understand whether such universally applicable effects easily permeate into L2 perception, or are constrained by listeners' level of English proficiency. Thus, although addressing these questions was not the main goal of the present study, the comparison of native and non-native listeners enriches our understanding of modulation of speech perception as a function of different types of slowing down effects.

2. Experiment 1

2.1. Method

2.1.1. Participants

Forty-eight volunteers participated in the study in three groups of 16: native speakers of American English (8 females, 8 males), advanced Korean learners of English (with TOEIC scores 920–990, average percentile rank = 96th; 10 females, 6 males), and

[1] Speaking rate normalization may refer to two different types of effects: The effect of the duration of a precursor on the perception of a timing based distinction (e.g., voicing based on VOT) of a target syllable as explored in the present study, or the influence of the overall syllable duration on the perception of this distinction in the syllable's initial stop as previously explored in studies such as Miller and Volaitis (1989) and Nagao and de Jong (2007). While the perceptual shifting in question of the present study can be in principle tested against both types of speaking normalization, our discussion here is limited to the first type.

intermediate learners of English (with TOEIC scores 470–700; average percentile rank = 49th; 3 females, 13 males).[2] Native speakers of American English were exchange students or English instructors temporarily residing in Korea, and their age ranged from 21 to 36 (M_{age} = 28). All Korean participants were undergraduate or graduate students in their 20s (M_{age} = 24). Participant dialects were not controlled. All participants were tested at Hanyang University, Seoul.

2.1.2. Materials

We first resynthesized the stimuli of Kim and Cho using the MBROLA diphone synthesizer, which concatenates pre-recorded diphones, but with full control over the pitch contour. The stimuli were aligned with MBROLIGN and then a resynthesis file was generated for both the IP and the Wd boundary context using the pitch contour of the original stimulus. This method generates a file that contains segment durations and ten pitch points during each segment, based on the original stimulus. With these parameters, an MBROLA resynthesis was generated, using the (male) voice "us2". (Note that the study of Kim and Cho (2013) also made use of a male US American speaker.)

Additionally, a third stimulus was generated from the Wd Boundary ("Wd-FULL") stimulus by multiplying the durations by a factor of 1.6, which leads to an overall duration that is the same as the IP boundary stimulus (i.e., the original IP stimulus was 60% longer than the Wd stimulus). Fig. 1 shows the duration of the three stimuli and their segments as well as the pitch curve (as estimated from the resynthesis). As can be seen, the lengthened Wd boundary ("Wd-LONG") and the IP boundary ("IP-FULL") stimuli have the same overall duration, but differ in their durational structure and intonation contour, appropriate for a Wd boundary and an IP boundary, respectively. When compared to the natural Wd boundary ("Wd-FULL") stimulus, the IP boundary ("IP-FULL") stimulus is longer mainly on the last syllable, in line with the localized temporal expansion before a phrase boundary that has been observed in English (Byrd, Krivokapić, & Lee, 2006; Cho, Kim, & Kim, 2013; Turk & Shattuck-Hufnagel, 2007) and many other languages including Korean (e.g., Cho, Lee, & Kim, 2011; Jun, 2005; see Cho, 2015 for a review). As a consequence, the linear expansion of segment duration of the Wd boundary stimulus generates a pattern in which the first syllable is relatively longer than that in the IP boundary stimulus while the second syllable is relatively shorter. Moreover, as the pitch points are aligned within their segments, we also get different pitch contours for the two cases (see the small circles in Fig. 1, with a scale indicated on the lowest group of bars, which is also valid for the other contours). As Fig. 1 shows, the pitch contours have similar ranges (IP: 110–160 Hz; Wd: 115–167 Hz) but the IP stimulus has, compared to the Wd stimulus, a lower mean (128 Hz vs. 141 Hz) and a smaller standard deviation (14.7 Hz vs. 17.9 Hz), due to its long "low" tail before the boundary. It should be also noted that in the IP boundary condition, the tonal pattern is clearly consistent with a H*L-L% ToBI tone pattern, as confirmed by two of the authors, who are trained English ToBI transcribers.

To increase the naturalness of the stimuli, the resynthesized stimuli were then given the same intensity contour as the original stimulus (using a PSOLA lengthened version of the original Wd boundary stimulus for the Wd-LONG case). This was achieved by first giving the stimuli a flat intensity contour and then multiplying them with the intensity contour of the original stimuli.

To generate a VOT continuum for the target syllable, the aspiration parts from the original stimuli were again spliced between burst and voice onset of a resynthesized stimulus. Seven different target syllables were generated with VOTs ranging from 0 to 45 ms in steps of 7.5 ms (=seven steps).

2.1.3. Procedure

Participants were seated in front of a computer screen and informed that they would hear sentences spoken by a native speaker of American English and that their task was to decide whether the speaker had said either "Let's hear 'ba' again" or "Let's hear 'pa' again". They were asked to respond by pressing the left or the right button of a response box. The allocation of answers to the left and the right button was counterbalanced over subjects. In each group, a half of the participants were asked to press the left button for "ba" and the right button for "pa" and the other half vice versa. Each participant reacted to each of the 21 stimuli ten times after completing ten training trials. After each 60 trials, participants had the opportunity to take a short break. The experiment was performed in sound-attenuated booths at the Hanyang Phonetics and Psycholinguistics Lab.

2.2. Results

Fig. 2 shows mean proportions of /b/ responses for each group. Each panel shows three categorization functions over VOT, one for each context condition. First of all, there is a difference between the IP boundary (IP-FULL) condition (with the largest proportion of /b/ responses) and the Wd Boundary (Wd-FULL) condition (with the lowest proportion of /b/ responses) for all three groups, replicating Kim and Cho (2013). Moreover, we also see more /b/ responses overall by the non-native (Korean) listeners, which is also consistent with Kim and Cho (2013). This reflects the fact that Korean aspirated stops are produced with longer VOTs than English aspirated stops, so that Korean listeners need longer VOTs to categorize a stop as aspirated. Most importantly, the "Wd-LONG" condition (with a Wd boundary but the same overall duration as the IP boundary) is intermediate between the IP-Full and the Wd-Full condition. This suggests that while the global speaking rate change has an effect, the localized speaking rate change (due to prosodic structure) has an effect beyond that of global speaking rate.

To statistically validate these patterns, we ran a linear mixed-effect model with a logistic linking function to account for the categorical nature of the dependent variable (Jaeger, 2008). As fixed effects, we used VOT (centered at zero), Boundary (IP-FULL,

[2] TOEIC stands for the Test of English for International Communication, which is a standardized English proficiency administered by the Educational Testing Service on receptive listening and reading proficiency).

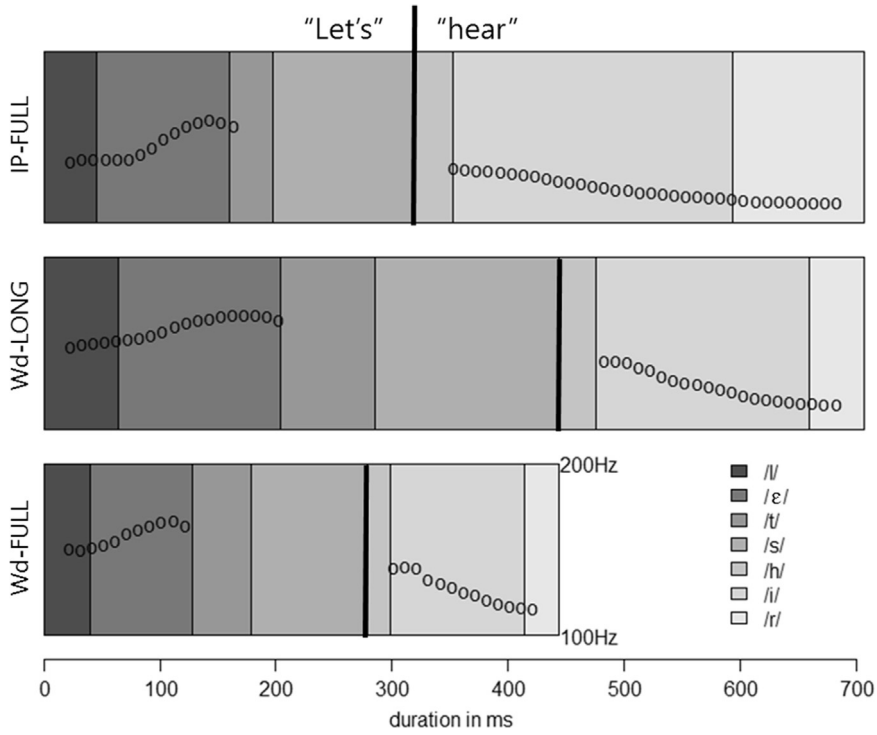

Fig. 1. Duration of the three precursor stimuli (*Let's hear*) in Experiment 1 and the respective segment durations. The small circles indicate the pitch contour as estimated after resynthesis for each stimulus, with a scale indicated on the "WD-Full" stimulus that is valid for all three pitch curves. The "IP-FULL" (IP natural) and the "Wd-FULL" (Wd natural) stimuli were resynthesized based on the original precursor *Let's hear* produced with an IP boundary and a Wd boundary, respectively. The "Wd-LONG" stimulus was matched with the "IP-FULL" stimulus in terms of its overall (total) duration, but with the "Wd-FULL" stimulus in terms of relative durational distribution of segments.

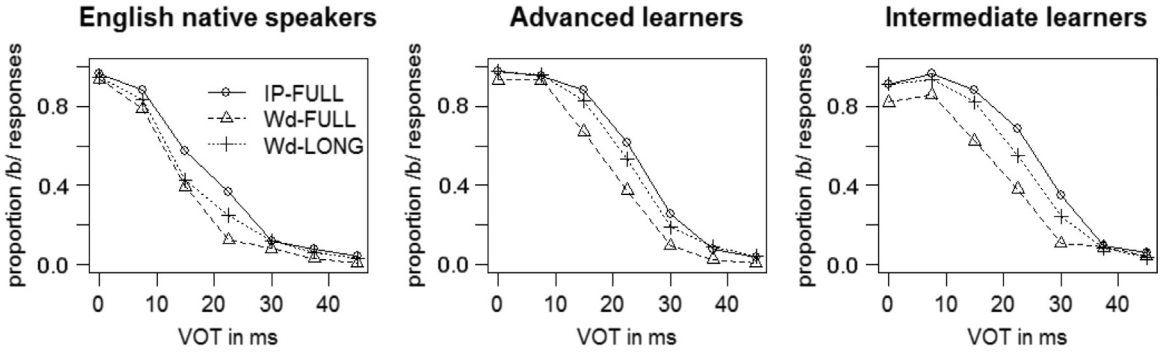

Fig. 2. Proportion of /b/ responses for each of the combination of precursor sentence and VOT, with separate panels for each listener group. The results show that speaking rate has an influence with more /b/ responses for the long Wd boundary stimulus (Wd-LONG) than the short one (Wd-FULL). Overall duration, however, is not the only influence, as there are more /b/ responses in the IP boundary (IP-FULL) condition than in the Wd-LONG condition, despite the identical overall duration of these two conditions.

Table 1
Results from the analysis of Experiment 1, showing similar effects in all three groups. Group differences only arise in the overall proportion of /b/ responses (more both by the advanced and the intermediate learners than by the native listeners) and a reduction is found in the steepness of the categorization function for the intermediate learners in the word-boundary condition.

	English Native listeners		Advanced learners		Intermediate learners		English Native vs.	
	B (SE)	*Z*	*B* (SE)	*Z*	*B* (SE)	*Z*	Advanced	Intermediate
(Intercept)	−1.32 (0.5)	−2.67**	0.56 (0.38)	1.47	0.73 (0.29)	2.5*	3.20**	3.55***
VOT	−2.0 (0.26)	−7.8***	−2.12 (0.23)	−9.39***	−1.6 (0.18)	−9.0***	−0.69	1.06
Boundary = Wd-FULL	−1.3 (0.25)	−5.2***	−1.42 (0.2)	−7.0***	−1.57 (0.25)	−6.3***	−0.69	−1.11
Boundary = Wd-LONG	−0.56 (0.15)	−3.7***	−0.41 (0.19)	−2.1*	−0.62 (0.18)	−3.5***	0.73	−0.39
VOT: Boundary = Wd-FULL	−0.22 (0.14)	−1.6	−0.01 (0.13)	−0.1	0.22 (0.11)	1.96*	0.82	2.3*
VOT: boundary = Wd-LONG	0.04 (0.1)	0.4	−0.04 (0.16)	−0.2	0.01 (0.1)	0.06	0.07	−0.48

*Significance level = 0.05.
**Significance level = 0.01.
***Significance level = 0.001.

Wd-FULL, Wd-LONG), and Group (Native, Advanced, Intermediate). For Boundary, the IP-FULL condition was mapped as the intercept. Two sets of analyses were run. In both analysis sets, subject was added as a random factor, and all possible random slopes were added to the model (Barr, Levy, Scheepers, & Tily, 2013). First, we ran an analysis for each group with VOT and Boundary and their interaction as fixed factors. We then ran a final analysis with a full factorial model using all three fixed factors, in which the native English listeners were mapped on the intercept. This provides two regression weights that show whether any of the two groups of non-native (Korean) listeners behave differently than the native listeners. Table 1 shows the outcome of this procedure. The first three columns show the results for the analyses for each group, showing the regression weights for each predictor. The final two columns indicate whether the differences between the native listeners and the non-native (Korean) listeners of two learner groups are statistically significant in the overall analysis.

The first row of Table 1 provides the intercept, which is the estimated likelihood of a /b/ response in the IP-Full condition (the level mapped on the intercept). The results indicate that both groups of Korean listeners (Advanced and Intermediate) give significantly more /b/ responses than the native English listeners. The effect of VOT indicates the steepness of the categorization function, which is not significantly different between groups. All groups also give significantly less /b/ responses in both Wd-FULL and Wd-LONG conditions than in the IP-FULL condition, which is consistent across groups. Crucially, although the precursor *Let's hear* in the Wd-LONG condition has the same total duration as in the IP-FULL condition, it still triggers less /b/ responses than in the IP-LONG condition for all three groups of listeners (i.e., a longer VOT is needed for /p/ responses in the IP-LONG than in the Wd-LONG condition). Finally, Korean listeners of the intermediate learner group have a shallower categorization function in the Wd-FULL boundary condition than in the other boundary conditions, and in that respect they also differ from the native English listeners.

2.3. Discussion

In Experiment 1, we have found that the phonetic categorization function for /b/-/p/ is shifted depending on the boundary condition, in such a way that listeners (regardless of their language background) gave less /b/ responses in the Wd-FULL condition (when the precursor was relatively shorter) than in both Wd-LONG and IP-FULL conditions (when the precursor was relatively longer). This is largely consistent with speaking rate normalization effect reported in the literature (Newman & Sawusch, 2009; Summerfield, 1981), and suggests that the shift of categorization as a function of boundary condition (IP vs. Wd) reported by Kim and Cho (2013) is in part comparable to the effect that arises with speaking rate normalization. That said, we have also observed a significant difference between the Wd-LONG condition (with global slowing down of speaking rate) and the IP-FULL condition (with boundary-related slowing down) even though the IP precursor had exactly the same duration as the long Wd boundary precursor—i.e., even more /b/ responses were given in the IP-FULL condition than in the Wd-FULL condition. This implies that prosodic boundary perception may still contribute to the modulation of phonetic categorization.

As was discussed in the Introduction, however, this interpretation leads to a hard-to-solve question as to whether the observed effect is indeed attributable to prosodic-boundary perception, or it is merely an augmented speaking rate normalization effect due to the localized temporal expansion associated with an IP boundary. Studies from the psychologically-oriented literature on speaking rate normalization actually have argued that local speaking rate has the most leverage in speaking-rate normalization. Summerfield (1981, Exp. 4), for example, tested the influence of the different syllables in the precursor phrase “Why are you” and found that the duration of the final syllable “you” carries twice the weight of the combined influence of the two other syllables (see also Reinisch, Jesse, & McQueen, 2011). Thus, proponents of the speaking rate normalization account might argue that the difference between the IP-FULL condition and the Wd-LONG condition (with the same overall duration) may not have come as a consequence of the perception of a prosodic boundary per se. Instead, it may be explained by the longer duration of the final syllable in the IP boundary condition, which would more strongly influence low-level auditory processing. From the viewpoint of prosodic phonology, however, Summerfield's (1981, Exp. 4) finding is equally consistent with the prosodic boundary perception hypothesis. Given that a local slowing down is a near-universal feature of prosodic boundaries applicable to both English and Korean (e.g., Cho, 2015), the perceptual weighting on the duration of the final syllable observed by Summerfield (1981) may have stemmed from the heightened perception of a prosodic boundary before the target syllable due to a lengthening of the final syllable.

It would therefore be difficult to determine which of the two has more leverage on speech categorization, especially because the prosodic boundary perception and speaking rate normalization both involve temporal modification of segments. Nevertheless, it may be informative to test whether the observed perceptual shift associated with an IP boundary has anything to do with the intonational (*F*0) cue to the identity of the prosodic boundary—i.e., whether the differences in the intonational (*F*0) cues to the prosodic boundary found in the stimuli could lead to different shifting effects. The rationale is as follows. Given that *F*0 information is an important cue to prosodic boundary perception (e.g., Kim, Broersma, & Cho, 2012; Kim & Cho, 2009; Ladd & Schepman, 2003; Tyler & Cutler, 2009), one might expect that the absence of an *F*0 cue would lead to a less robust prosodic boundary perception. If prosodic boundary perception indeed directly contributes to the perceptual shift of phonetic category, and if the lack of the *F*0 cue weakens the boundary perception, a reduced shifting effect may be observed when the *F*0 cue is not available to the listener. If this turns out to be the case, it may lend some support to the boundary-induced perceptual modulation account. It should be noted, however, that the failure to find such an effect (i.e., no difference between stimuli with and without the *F*0 cue) would not necessarily refute the boundary-induced perceptual modulation hypothesis because listeners may still be able to compute prosodic structure based on the available temporal cues. Rather, it would mean that results are still compatible with both the boundary perception account and the speaking normalization account. In Experiment 2, we explored these possibilities with four conditions: two conditions from Experiment 1 (IP-FULL and Wd-FULL) and two additional conditions with flat *F*0 (IP-FLAT and Wd-FLAT).

3. Experiment 2

3.1. Method

3.1.1. Participants

An additional 48 subjects participated in Experiment 2. As in Experiment 1, they came from three groups: native listeners of American English (6 females, 10 males), advanced Korean learners of English (with TOEIC scores 920–990, average percentile rank = 97th; 4 females, 12 males) and intermediate Korean learners of English (TOEIC scores 470–700; average percentile rank = 49th; 6 females, 10 males). Native speakers of American English were exchange students or English instructors temporarily residing in Korea, and their age ranged from 16 to 49 ($M_{age} = 30$). All Korean participants were undergraduate or graduate students and their age ranged from 19 to 30 ($M_{age} = 23$). Participant dialects were not controlled. (See footnote 1 for an explanation of TOEIC.)

3.1.2. Materials and procedure

The target /ba/-/pa/ continuum was the same as in Experiment 1. The precursors for the full-intonation conditions (IP-FULL and Wd-FULL) were the same as in Experiment 1. Additionally, two precursor stimuli (IP-FLAT and Wd-FLAT) were generated for which the pitch was set to 127 Hz, the overall mean *F*0 of the utterance. These new resynthesized versions were also given the same intensity contour of the original utterances.

The procedure was identical to Experiment 1, except that this experiment had slightly more trials—i.e., 28 stimuli (2 boundary conditions (IP vs. Wd) × 2 *F*0 conditions (Full vs. Flat) × 7 VOT steps) were repeated ten times. (In Experiment 1, 21 stimuli were repeated 10 times.) The experiment lasted about 20 min.

3.2. Results

Fig. 3 shows the proportions of /b/ responses in all three groups. What is immediately apparent is that there is a clear difference that arises with prosodic boundary (IP vs. Wd marked by triangles vs. circles) across groups, showing more /b/ responses after an IP than after a Wd boundary. This effect is generally of the same size, regardless of whether the original *F*0 contour is present (solid lines) or flattened (dotted lines). However, the Korean listeners, especially the intermediate learners seem to show a difference due to *F*0 conditions, generally giving more /b/ responses to stimuli with the flat *F*0 contour than with the original (full) *F*0 contour, especially in the Wd boundary condition (Wd-FLAT vs. Wd-FULL).

For statistical analysis, we used a linear mixed effect model with Group, VOT, Boundary, and *F*0 as fixed factors. The binary variables Boundary and *F*0 were contrast-coded (Boundary: Wd = −0.5, IP = 0.5; *F*0: FLAT = −0.5, FULL = 0.5). Subject was added as a random factor, and all possible random slopes were added to the model (Barr et al., 2013). As in Experiment 1, we first ran models for each group separately and then an overall analysis to see whether the groups differ significantly. Table 2 shows the result of this process.

All three groups show the expected effects of VOT and Boundary, with more /b/ responses with shorter VOTs and at an IP boundary. Regarding the intercept, which reflects the overall proportion of /b/ responses, the Korean listeners of the ‘advanced’ learner group only show a non-significant trend for more /b/ responses than the native English listeners. This effect was significant in Experiment 1, in line with the fact that Koreans require more VOT for a /p/ response. Interestingly, however, the ‘intermediate’ Korean learners do not even show a trend towards a difference with the native English listeners. The intermediate Korean learners instead show an effect of *F*0 on the overall number of /b/ responses (more /b/ responses with the flat *F*0 contour), which differs from the native listeners’ pattern. The advanced Korean learners show a steeper identification function in the IP condition, and in this respect they differ from the native listeners. Most importantly, we find that the effect of Boundary is similar in all three groups (as indicated in the row for the ‘Bound’ parameter in Table 2 which shows a significant main effect of Boundary for each group but no further native vs. learner group interaction), and is not moderated by the presence or absence of the original *F*0 contour (as indicated in the row for

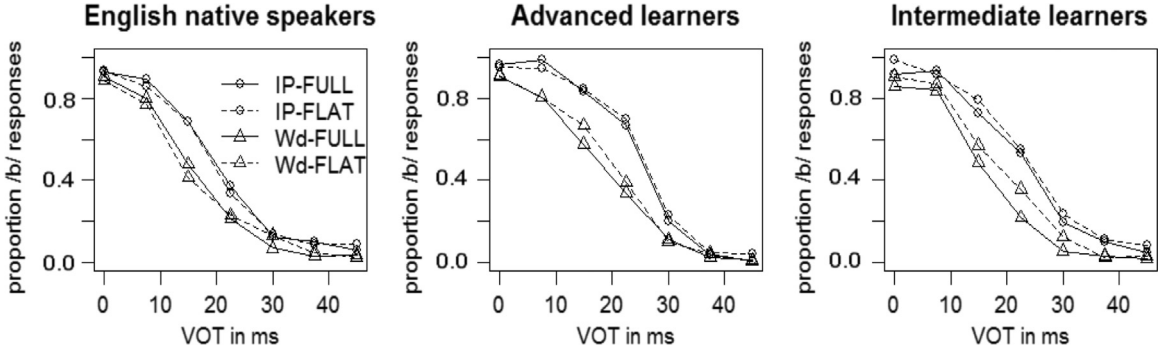

Fig. 3. Proportions of /b/ responses in Experiment 2 for each of the combination of precursor sentence and VOT, with separate panels for each listener group. IP-FULL and Wd-FULL refer to the conditions when the original *F*0 contour is maintained, and IP-FLAT and Wd-FLAT when the *F*0 is flattened. The results show that removing the *F*0 contour has no effect on English native speakers. For Korean speakers, there is a general tendency to give more /b/ responses if the *F*0 contour is flattened, but the size of the boundary effect (IP vs. Wd condition) is roughly similar with the original *F*0 contour and a flat *F*0 contour.

Table 2
Results from the statistical analyses of Experiment 2. Similar effects are found in all three groups. Group differences arise only in three cases: a trend for more /b/ responses by the advanced Korean learners than by the native English listeners), an effect of *F*0 only for the intermediate learners, and steeper categorization functions for advanced learners only in the IP boundary conditions.

	English native listeners		Advanced learners		Intermediate learners		English Native vs.	
	B (SE)	*Z*	*B* (SE)	*Z*	*B* (SE)	*Z*	Advanced	Intermediate
(Intercept)	−1.05 (0.33)	−3.15**	−0.24 (0.26)	−0.91	−0.68 (0.25)	−2.71**	1.67	0.59
VOT	−1.65 (0.2)	−8.06***	−1.91 (0.24)	−7.94***	−1.6 (0.15)	−10.5***	−1.13	−0.07
Bound	0.91 (0.27)	3.34***	1.44 (0.18)	8.08***	1.3 (0.14)	9.22***	1.89	1.45
*F*0	−0.04 (0.11)	−0.39	−0.13 (0.11)	−1.16	−0.41 (0.15)	−2.78**	−0.69	−2.65**
VOT:Bound	−0.12 (0.1)	−1.25	−0.42 (0.12)	−3.45***	0.14 (0.1)	1.48	−2.63**	0.18
VOT:*F*0	−0.18 (0.09)	−2.05*	−0.1 (0.1)	−1.05	−0.04 (0.1)	−0.47	−0.26	1.52
Bound:*F*0	0.32 (0.23)	1.35	0.18 (0.28)	0.64	0.38 (0.22)	1.71	0.05	0.25
VOT:Bound:*F*0	0.25 (0.18)	1.39	−0.26 (0.2)	−1.27	0.1 (0.18)	0.56	−1.99*	0.05

*Significance level = 0.05.
**Significance level = 0.01.
***Significance level = 0.001.

the 'Bound:*F*0' parameter in Table 2 which shows no significant Boundary by *F*0 interaction). (Note that the critical parameters for the question of prosodic modulation of speech perception are the terms Boundary and Boundary by *F*0.)

3.3. Discussion

This experiment examined the extent to which the presence or absence of the *F*0 cue to the prosodic boundary influences phonetic categorization in an effort to identify what aspect of prosodic boundary information contributes to the modulation of phonetic categorization. The results suggest that the *F*0 cue does not moderate the boundary effect when there is a temporal (lengthening) cue to the prosodic boundary. This therefore does not seem particularly to buttress the argument for the boundary-induced perceptual modulation hypothesis, leaving the general speaking normalization account still competitive (see General Discussion for further discussion on this point).

Before we turn to the general discussion of our findings with respect to the perceptual shift as a function of boundary perception vs. speaking rate normalization, it is worthwhile to discuss some unexpected group differences that we found in Experiment 2, although they are tangential to the key research questions of the present study. In Experiment 1, Korean listeners (advanced and intermediate learners alike) gave more /b/ responses than English listeners, indicating that Korean listeners needed relatively longer VOTs for the voiceless /p/ percept. In Experiment 2, however, the language effect was not found to be entirely reliable: There was only a trend effect in the comparison between advanced Korean learners and native English listeners, and the language effect disappeared in the comparison between intermediate Korean learners and native English listeners. We do not have absolute explanations as to why the language effect was heavily attenuated in Experiment 2, and in particular why it disappeared completely with the intermediate learners when the language effect, if it exists, would more likely arise with the intermediate learners rather than with the advanced learners. We only have a rather speculative explanation to offer which is complicated by a possible language-specific use of *F*0 information between English and Korean.

Even though there is a general tendency in the languages of the world for voiced stops to cause lower *F*0s in the following vowel than voiceless stops, the size of this effect varies between languages (Kingston & Diehl, 1994). In fact, it is much larger in Korean than in English, as the effect is phonologized in the intonational phonology (Jun, 1993, 2005). A recent study by Schertz, Cho, Lotto, and Warner (2015) indeed showed that Korean listeners use the *F*0 cue more than English listeners do when processing a non-native English stop contrast in voicing, implying that Korean listeners rely more heavily on a relatively low *F*0 (which is a robust cue to the Korean lenis stop) in identifying English voiced stops.

Our posthoc analyses in fact indicated that the greater number of /b/ responses by the intermediate Korean learners stemmed primarily from the flat *F*0 condition, especially in the Wd contexts (Wd-FLAT vs. Wd-FULL) (as can be seen in the rightmost panel of Fig. 3 where the dotted lines for the flat *F*0 conditions are "above" the solid lines for the full *F*0 conditions). In our stimuli, *F*0 at the acoustic offset of the precursor was relatively higher in the *F*0 flat than in the original *F*0 condition (which had a falling *F*0 contour). So, assuming that Korean learners use *F*0 as a cue to voicing more strongly than English listeners, as has been found in Schertz et al. (2015), and that the perception of *F*0 is context-dependent (e.g., Wong & Diehl, 2003), we can explain why Korean listeners, especially intermediate learners, give more /ba/ responses in the flat *F*0 conditions: the higher *F*0 offset in the precursor leads to a lowering of the perceived *F*0 at the onset of the critical syllable, which in turn leads to more /ba/ responses in conjunction with lower *F*0. This possible perceptual compensation due to the preceding *F*0 context appears to have contributed to more /b/ responses on average by the intermediate learners. On the other hand, given that the advanced Korean learners of English showed no such *F*0-induced compensation effect, an interim interpretation that may be offered here is that Korean listeners' reliance on the *F*0 cue in processing a non-native English stop contrast becomes attenuated as their level of English proficiency becomes more native-like. With this possibility, however, we are still left with a puzzle: why, overall, more /b/ responses were observed with the advanced learners (than with the native English listeners), but not with the intermediate learners, especially given that the effect of *F*0 should

lead to even more /b/ responses for the intermediate learners, all else being equal. On the other hand, just because the intermediate learners gave more /b/ responses in the flat *F*0 condition than in the original *F*0 condition does not necessarily mean that the overall number of /b/ responses by the intermediate learners should be greater than that by the native English listeners. We can only speculate here that the presence of contrastive *F*0 information in Experiment 2 (but not in Experiment 1) might have made Korean learners use this cue somewhat in an unbalanced way, leading to less /b/ responses for the stimuli with the original *F*0 contours, which possibly canceled out or outweighed more /b/ responses for the stimuli in the flat *F*0 condition.

4. General discussion

Two experiments were conducted to elucidate the extent to which speech categorization is directly influenced by the perception of a prosodic boundary. Previous evidence suggesting such an influence (Kim & Cho, 2013) is open to alternative interpretations since prosodic boundary perception was correlated with speaking rate in the preceding context. The present study therefore made an attempt to explore whether the purported boundary-induced modulation of phonetic categorization may be simply seen as an effect of speaking rate normalization at a low level of processing. Although it is not an easy task to completely disentangle the perception of prosodic boundaries and speaking rate normalization, we have made a step further towards understanding prosodically-modulated speech perception. Most importantly, while both global and local slowing down effects (due to a change in general speaking rate vs. due to a major prosodic boundary, respectively) do modulate speech perception, their effects are not the same. The current set of experiments therefore opens up a new way of understanding temporal modulation of speech perception by bridging the gap between the psychologically-informed field in which the effect has previously been attributed to low-level speaking rate normalization and the linguistically-informed field in which the effect is assumed to arise as a consequence of perceiving the higher-order prosodic structure.

Experiment 1 replicated the basic conditions of Kim and Cho (2013, i.e., Wd boundary or IP boundary before the critical syllable) along with an additional condition to compare the boundary effect with the speaking rate effect. Stimuli were created by a diphone synthesizer which allowed for generating a preceding context (the precursor, *Let's hear*) that had the original temporal structure of segments and *F*0 contour associated with a Wd boundary, but its overall duration was elongated to be matched with that of the precursor with an IP boundary. As was expected (Summerfield, 1981), this longer precursor (in the Wd-LONG condition) led to more /b/ responses than the shorter Wd boundary precursor (in the Wd-FULL condition). This means that longer VOTs were still perceived as being relatively "short" after the longer precursor. Importantly, however, the IP boundary stimulus gave rise to an even higher proportion of /ba/ responses, despite the fact that it had the same overall duration as the lengthened Wd precursor.

These results suggest that although the speaking rate change (between Wd-FULL and Wd-LONG conditions) does modulate phonetic categorization, a simplistic conception of speaking rate normalization is not sufficient to explain the effect found in Experiment 1 as well as in Kim and Cho (2013). There remained a difference between the Wd-Long and the IP-full condition, and such local effects may be interpreted as suggesting that listeners adjust their perception in phonetic categorization based on the prosodic boundary that they perceive. In an effort to explore the nature of prosodic boundary perception in relation to speaking rate normalization, a second experiment was carried out. It tested the extent to which the purported boundary-induced modulation of phonetic categorization is attributable to the temporal structure in the absence of another salient prosodic boundary cue, *F*0. It was hypothesized that if the prosodic boundary perception is a direct cause of the modulation of phonetic categorization, the localized slowing down with no *F*0 cue would reduce the effect under the assumption that the absence of *F*0 would weaken the prosodic boundary perception.

The results of Experiment 2 showed that phonetic categorization was influenced by the boundary condition (IP vs. Wd) as was found in Experiment 1, but this effect was not modulated by the presence or absence of the *F*0 cue. That is, there is no *F*0 by Boundary interaction, indicating that the *F*0 information of the preceding context does not contribute to adjusting listeners' perception in speech categorization. (Given that there was no further interaction with Language Group, the lack of *F*0 by Boundary interaction was generalizable to both native English listeners and Korean learners.) Given that the *F*0 information is an important cue to the prosodic boundary (e.g., Ladd & Schepman, 2003; Tyler & Cutler, 2009; Kim & Cho, 2009; Kim et al., 2012), the failure of *F*0's contribution to boundary-induced modulation of phonetic category may not be seen as entirely consistent with the view that speech perception is directly modulated by computation of prosodic structure (see also Cho et al., 2007 for a related argument). However, the fact that *F*0 has no influence on the perceptual shift of phonetic categorization does not necessarily mean that boundary perception has no role. It simply fails to strengthen the boundary perception account. The account is still in line with the results because the local slowing found near the boundary, according to theories of prosodic phonology, results from the boundaries themselves. The boundary-driven local slowing has been computationally modeled under the rubric of the theory of the *pi-gesture* advanced by Byrd and colleagues (e.g., Byrd & Saltzman, 2003). It is therefore likely that the boundary-induced temporal expansion is enough to signal the presence of a stronger prosodic boundary, even in the absence of the *F*0 cue, as to be discussed below.

One of the reasons for the lack of *F*0 contribution may be that, in languages like English, particular *F*0 patterns in a prosodic boundary (or boundary tones) are not predictable, so that there is no fixed *F*0 pattern as an invariant cue to prosodic boundaries in English. That is, multiple types of *F*0 cues for boundary tones (e.g., H%, L%, LH% or HL%, where '%' denotes a tone used for marking an intonational phrase boundary) can be used for the same type of prosodic boundary in English (e.g., Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980). On the other hand, the durational (localized final lengthening) cue is the most robust and consistent prosodic boundary cue (e.g., Shattuck-Hufnagel & Turk, 1996), and is indeed considered as a universal

cue to a prosodic boundary that listeners use across languages (Tyler & Cutler, 2009). Even though languages may have quite different rhythmic (or durational) patterns (Ramus, Nespor, & Mehler, 1999; White, Mattys, & Wiget, 2012), final lengthening seems to be a cue that surpasses those differences, so that languages from different rhythmic classes are still similar in this respect. Thus it appears that a durational cue in line with a major prosodic boundary may suffice to mark a boundary, and therefore an additional (variable) *F*0 cue, regardless of whether it actually enhances the boundary percept or not, does not seem to contribute to a shift in phonetic category boundaries. In a recent artificial language learning study, Kim et al. (2012) found independent evidence for the primacy of durational cues for listeners of Dutch, which is similar to English in terms of use of boundary tones. Dutch listeners' learning of new words in an unfamiliar language was facilitated when the words had a localized final lengthening cue, but did not get any better even when a possible *F*0 cue was superimposed on the final lengthening cue.[3] After all, an *F*0 cue (whether it is H%, L%, LH% or HL%) always co-occurs with phrase final lengthening. Therefore, *F*0 cues appear to be ancillary to the durational cue to a prosodic boundary, especially in languages like English and Dutch in which there is no fixed *F*0 cue for a particular prosodic boundary. It therefore seems to be extremely difficult (if not impossible) to test the unique contribution of *F*0 to the perception of prosodic boundary without an accompanied temporal stretch of the segmental string.

It is, however, worth noting that there is still ample evidence for the role of *F*0 in the perception of prosodic structure. Kim and Cho (2009) demonstrated that lexical segmentation in Korean is facilitated when there is an *F*0 cue consistent with a phrase boundary even in the absence of the preboundary durational cue, and that the performance is no better when the durational cue is added. Similarly, Welby (2007) showed that an alignment of *F*0 which is phonologically specified for an Accentual Phrase in French is an important prosodic cue that also effectively facilitates lexical segmentation. However, Korean and French are different from English and Dutch, in that they employ *F*0 cues more systematically (e.g., both Korean and French demarcate an Accentual Phrase with a rising *F*0 at phrase edges). These studies all together suggest that what is important may not be to examine which prosodic cue takes precedence over another, but to understand their relative roles by considering their universal applicability vs. language-specific prosodic phonology.

In summary, our data indicate that researchers interested in two issues that are usually studied separately (speaking rate normalization and the perception of prosody) may need to take each other's findings into account. The results of the present study do suggest that the boundary-induced modulation of phonetic categorization is not the same as the modulation as a function of a 'global' change in speaking rate. However, they also indicate that the boundary-induced modulation of phonetic categorization may be understood as listeners' adjustment to temporal variation that arises as a consequence of a prosodic boundary. The effect is therefore comparable to a speaking rate normalization effect as well, to the extent that listeners adjust their phonetic categorization to temporal variation (either globally as a function of global speaking rate change or locally as a function of boundary). Thus, we are still left with two possible accounts. On the one hand, a parsimonious account may be to treat both effects as arising with listeners' adjustment to temporal variation that occurs at early auditory processing stages, presumably independent of computation of a higher-order prosodic structure. This account also has an advantage to provide a unified account for similar behaviors across native (American English) and non-native (Korean) listeners. However, this account should not be entirely independent of listeners' parsing of the higher-order prosodic structure, precisely because a local slowing down is very likely to be modulated by prosodic boundary. Furthermore, as the computation of prosodic boundaries does influence speech comprehension at various levels of processing (including lexical and syntactic processing levels; e.g., Brown, Salverda, Dilley, & Tanenhaus, 2011; Cho et al., 2007; Christophe et al., 2004; Jun, 2010), it is still reasonable to assume that the boundary-induced modulation of phonetic categorization arises with computation of a higher-level prosodic structure. Under this account, the similar behaviors of listeners with different language backgrounds are explicable by the role of final lengthening that signals a prosodic boundary across languages. In this context, it is important to note that the role of speaking rate normalization has recently been questioned (Toscano & McMurray, 2012) and that, especially, the VOT distinctions that support the perception of voicing may not be that dependent on speaking rate after all (Kessinger & Blumstein, 1997).

To place one account over the other on a firmer footing, one may suggest that future research is needed. One might test, for example, the effects of non-temporal prosodic cues to prosodic boundaries that may work orthogonally from speaking rate normalization on phonetic categorization. Or one might employ cues that are universally applicable to prosodic boundaries across languages, vs. ones that vary from language to language. (In this regard, it will also be worth extending a study like Experiment 2 of the present study with listeners of languages in which *F*0 cues to prosodic boundaries are comparable (e.g., Korean and French may be such languages) in order to further explore the universality and the language specificity of the perceptual recalibration due to higher-level vs. lower-level effects). But, as discussed above, dissociating the role of non-temporal prosodic cues from that of temporal cues appears to extremely difficult, if not impossible. Alternatively, the effect of a prosodic boundary on a non-temporal, post-boundary speech cue (e.g., a spectral cue) could be studied. However, research indicates that domain-initial strengthening may affect temporal aspects more than spectral ones (Clayards, 2015), so that this is a difficult route as well. In summary, the effects of local speaking rate normalization and boundary perception may not be inseparable, but are integrated into speech perception by virtue of the interface between low-level auditory processing and parsing of the higher-level prosodic structure, a kind of the

[3] Our finding that *F*0 did not have much leverage is in fact consistent with other studies that have pitted pitch and durational properties against each other in the perception of prosody. De Ruiter, Mitterer, and Enfield (2006) found that for the projection of a turn ending during interaction, *F*0 contours had little to contribute in Dutch. Taking away the *F*0 contour did not influence Dutch listener's ability to project the end of a turn. However, durational properties of the utterance turned out to be effective, as listeners were still surprisingly well able to project the end of a turn with just the durational patterns intact (compared to a control condition). Similarly, Michelas and D'Imperio (2015) found that parsing decision in French are more strongly influenced by durational cues than by *F*0 cues to prosodic boundaries.

phonetics-prosody interface in speech comprehension. More interdisciplinary studies are warranted in order to further illuminate the nature of the phonetics-prosody interface in speech perception.

Acknowledgment

We thank our graduate student assistants, Daejin Kim, Miroo Lee and Yuna Baek for assisting us with data acquisition. This work was supported by the research fund of Hanyang University (HY-2013) to the corresponding author (T. Cho).

References

Abramson, A. S., & Lisker, L. (1964). *A cross-language study of voicing in initial stops: Acoustical measurements*. Retrieved from ⟨http://www.researchgate.net/publication/208032881_A_Cross-Language_Study_of_Voicing_in_Initial_Stops_Acoustical_Measurements⟩.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278, http://dx.doi.org/10.1016/j.jml.2012.11.001.

Beckman, M., & Pierrehumbert, J. (1986). Intonational Structure inn Japanese and English. *Phonology Yearbook*, *3*, 255–309.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review*, *18*(6), 1189–1196.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*, 149–180.

Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *The Journal of the Acoustical Society of America*, *120*(3), 1589–1599, http://dx.doi.org/10.1121/1.2217135.

Cho, T. (2011). Laboratory phonology. In N. C. Kula, B. Botma, & K. Nasukawa (Eds.), *Bloomsbury companion to phonology* (pp. 343–368). London/New York: Continuum. Retrieved from ⟨http://books.google.com/books?hl=de&lr=&id=dAPdZWT-RZkC&oi=fnd&pg=PA343&dq=Cho+Laboratory+pHonology+Companion&ots=p6LUDhTMR6&sig=2NwmZVYnhw2U9BWDv_Kyhp4MNjI⟩.

Cho, T. (2015). Language effects on timing at the segmental and suprasegmental levels. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 505–529). Hoboken, NJ: Wiley-Blackwell.

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

Cho, T., Kim, J., & Kim, S. (2013). Preboundary lengthening and preaccentual shortening across syllables in a trisyllabic word in English. *The Journal of the Acoustical Society of America*, *133*(5), EL384–EL390.

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. *Journal of Phonetics*, *39*(3), 344–361.

Cho, T., Lee, Y., & Kim, S. (2014). Prosodic strengthening on the /s/-stop cluster and the phonetic implementation of an allophonic rule in English. *Journal of Phonetics*, *46*, 128–146.

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*(2), 210–243.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, *51*(4), 523–547.

Clayards, M. (2015). Prominence enhances voicelessness and not place distinction in English voiceless sibilants. In *Proceedings of ICPhS 2015*. Glasgow.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515–535, http://dx.doi.org/10.1353/lan.2006.0130.

Embick, D., & Poeppel, D. (2014). Towards a computational (ist) neurobiology of language: Correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 1–10, http://dx.doi.org/10.1080/23273798.2014.980750.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. *The Handbook of Phonetic Sciences, Second Edition*, 521–602.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*(6), 3728–3740.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*, 575–590, http://dx.doi.org/10.3758/BF03194584.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446, http://dx.doi.org/10.1016/j.jml.2007.11.007.

Jun, S.-A. (2005). Korean intonational phonology and prosodic transcription. *Prosodic Typology: The Phonology of Intonation and Phrasing*, *1*, 201.

Jun, S.-A. (2010). The implicit prosody hypothesis and overt prosody in English. *Language and Cognitive Processes*, *25*(7-9), 1201–1233.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168.

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*(03), 415–444.

Kim, S., & Cho, T. (2009). The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean. *The Journal of the Acoustical Society of America*, *125*(5), 3373–3386.

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

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419–454, http://dx.doi.org/10.2307/416481.

Kuzla, C., Ernestus, M., & Mitterer, H. (2010). Compensation for assimilatory devoicing and prosodic structure in German fricative perception. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology*, *10* (pp. 731–758). Berlin: Mouton.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press. Retrieved from ⟨https://books.google.com/books?hl=de&lr=&id=T0aIk3GUJoQC&oi=fnd&pg=PR9&dq=Ladd+intonational+Phonology+2008&ots=B6UsxdL6yg&sig=yjl5aM2rn6s23xZ_nAR984xKW9k⟩.

Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics*, *31*(1), 81–112.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *27*, 98–104.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage Learning. Retrieved from ⟨http://books.google.com/books?hl=de&lr=&id=U4XaAgAAQBAJ&oi=fnd&pg=PP1&dq=ladefoged+Johnson&ots=KZjUj9OE4R&sig=HjHNj45MqSqKc9G9wD7jjw9uu5U⟩.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America*, *102*, 1134–1140, http://dx.doi.org/10.1121/1.419865.

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

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners perception of English "l" and "r". *Cognition*, *24*, 169–196, http://dx.doi.org/10.1016/S0010-0277(86)80001-4.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Michelas, A., & D'Imperio, M. (2015). Prosodic boundary strength guides syntactic parsing of French utterances. *Laboratory Phonology*, *6*(1), 119–146, http://dx.doi.org/10.1515/lp-2015-0003.

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, http://dx.doi.org/10.3758/BF03208147.

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, *65*(6), 956–969, http://dx.doi.org/10.3758/BF03194826.

Mitterer, H., Csépe, V., Honbolygo, F., & Blomert, L. (2006). The recognition of phonologically assimilated words does not depend on specific language experience. *Cognitive Science*, *30*(3), 451–479, http://dx.doi.org/10.1207/s15516709cog0000_57.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment—Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics*, *44*, 1–7.

Nagao, K., & de Jong, K. (2007). Perceptual rate normalization in naturally produced rate-varied speech. *The Journal of the Acoustical Society of America*, *121*(5), 2882, http://dx.doi.org/10.1121/1.2713680.

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, http://dx.doi.org/10.1016/j.wocn.2008.09.001.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (Ph.D. dissertation). Cambridge MA: MIT.

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/and glottal stop. *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 90–117.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292, http://dx.doi.org/10.1016/S0010-0277(99)00058-X.
Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, production, acoustics, and perception* (Vol. 34). John Wiley & Sons. Retrieved from ⟨http://books.google.com/books?hl=de&lr=&id=LpxJL1tJajgC&oi=fnd&pg=PA19&dq=Reetz+Jongman&ots=RG85J95FAg&sig=EZiHNeToUKiUE_ZBFhCOIQAwPil⟩.
Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech*, *54*(2), 147–165.
Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116, http://dx.doi.org/10.1016/j.wocn.2013.01.002.
Rietveld, A. C. M., & van Heuven, V. J. (1997). *Algemene fonetiek*. Bussum, The Netherlands: Coutinho.
Schertz, J., Cho, T., Lotto, A. J., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204.
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247.
Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074.
Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301, http://dx.doi.org/10.3758/s13414-012-0306-z.
Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.
Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*, 367–376, http://dx.doi.org/10.1121/1.3129127.
Wagner, M., & Klassen, J. (2015). Accessibility is no alternative to alternatives. *Language, Cognition and Neuroscience*, *30*(1–2), 212–233, http://dx.doi.org/10.1080/23273798.2014.959532.
Welby, P. (2007). The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, *49*(1), 28–48.
White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665–679.
Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*, 413–421.