DETECTING ACCENTUAL PHRASE BOUNDARIES IN SEOUL KOREAN USING TONAL AND SEGMENTAL CUES

Seung Suk Lee

University of Massachusetts, Amherst, USA
seungsuklee@umass.edu

ABSTRACT

This paper investigates the informativity of tonal and segmental cues for detecting prosodic constituents boundaries in a prosodically unlabeled spontaneous speech corpus of Seoul Korean. The focus of this paper is Accentual Phrases (AP) that start with a Lenis obstruent. Previous work has found that listeners detect AP boundaries by attending to tonal cues (change in pitch from the previous syllable) and segmental cues (allophonic variation of Lenis). This paper reports that in spontaneous speech, while the segmental cues are more informative than the tonal cue for detecting AP boundaries, the tonal cue is informative for detecting the boundary of prosodic constituents higher than the AP.

Keywords: Seoul Korean Accentual Phrase, K-ToBI, Spontaneous speech corpus, Phonetics-prosody interface

1. INTRODUCTION

This paper investigates the informativity of tonal and segmental cues to the Accentual Phrases (AP) of Seoul Korean in a spontaneous speech corpus [1].

In the ToBI framework, prosodic constituent boundaries are associated with Break Indices (BI) [2]. In the autosegmental-metrical model of Seoul Korean intonational phonology, K-ToBI, there are 3 levels of BIs associated with 3 levels of prosodic units: the Phonological Word (PWd), the Accentual Phrase (AP), and the Intonational Phrase (IP) [3, 4, 5], see Table 1.[1] Previous work has shown that the listeners pay attention to both segmental and tonal cues to determine the presence or absence of these prosodic constituent boundaries [6, 7, 8, 9].

One way to assess the informativity of acoustic cues in perceiving these different boundary types is by formalizing perception as a categorization task where the boundary types are phonological categories [10]. This approach typically proceeds by first getting the data labeled with prosodic transcription to mark each syllable in the data with a BI value. Next, the acoustic properties of the labeled syllables are parameterized with relevant phonetic measurements, such as f0 or duration. Then, the effect of such cues in accounting for the differences between the BI levels is measured e.g., with a regression analysis.

BI level	Prosodic Position
3	IP-initial
2	AP-initial
1	PWd-initial ($[_{\text{PWd}}\sigma$)
0	PWd non-intial ($[_{\text{PWd}}...\sigma$)

Table 1: The BI levels and the associated prosodic positions. PWd non-initial includes both PWd medial and PWd final syllables.

Crucially, though, this approach requires the prerequisite step of first prosodically labeling the data, which becomes more expensive as the size gets larger, especially since it is known that the labelers also differ in their judgment of the type and the presence/absence of the prosodic constituent boundaries [11], which means it is ideal to have multiple labelers. This sets the entry barrier quite high to conduct such research and only a handful of datasets for only a few languages are publicly available (e.g., [12] for Japanese). In the case of Korean, there are few publicly available prosodically transcribed speech corpora, and none of them contains spontaneous speech (see [13] for the importance of investigating spontaneous speech).

For example, the data used in this study come from a spontaneous speech corpus of 40 native Seoul Korean speakers [1], but it is not prosodically transcribed. Consequently, while there are a few phonetic studies conducted on this corpus (e.g., [14]), none are on prosody.

On the other hand, one could also assess the informativity of cues without the prerequisite step of prosodic labels by investigating the distributions of the categories along a cue dimension and inspecting how separable the categories are along that dimension. The informativity of cues can be further explored with a clustering analysis, where each syllable in the dataset is considered a token in a phonetic space, constructed with acoustic cues hypothesized to be primary cues for AP detection as the dimensions. The clustering analysis can assess the number of clusters that can best account for the variation in the data. If the clusters that emerge from the data can be inferred to be the boundary type categories based on where they lie in the phonetic space, it provides evidence that the cues that construct the space are informative in separating the categories.

This paper reports an initial exploratory analysis of the informativity of the acoustic cues to the AP boundaries of Seoul Korean in spontaneous speech, in this second sense—in the absence of prosodically labeled data.

2. BACKGROUND

Unlike typical prosodic constituents found in other languages, the AP in Korean is argued to not have final lengthening, but instead to be 'tonally marked' [3, 4]. The K-ToBI model suggests that the tonal marking is distinct at the AP junctures so that the listener can pay attention to tonal targets to notice whether or not there is an AP juncture in between the two syllables. According to [4], the initial tone of the AP can be 'L' or 'H'[2] but in this paper, we limit our discussion to the APs that start with a Lenis obstruent and therefore with an initial 'L' tone. APs typically end with an 'Ha' tone [4, 15]. This suggests that in a typical AP juncture, F0 falls from the pre-boundary syllable to the post-boundary syllable. However, F0 also falls from one syllable to another within an AP in longer APs [4]. Previous work has suggested that the gradient nature of the fall matters, as the fall across an AP juncture seems to be larger than the fall within an AP [16], and that listeners are sensitive to the gradient difference in the fall in perceiving AP boundaries [6, 7, 8].

In addition, [3] showed that the AP also serves as a domain that conditions the realization of segments. In particular, AP-initially, Lenis obstruents are acoustically realized as 'strong' (via 'domain-initial strengthening', [17, 18] among others); but optionally realized as voiced and lenited ('acoustically weak') AP-medially ('Lenis medial voicing', [19, 20] among others). The voicing of Lenis can be seen as a type of cross-linguistically attested domain-medial lenition [21], which can bear a demarcative function, as the lenited realization can signal the absence of a boundary or the continuation of a constituent [22]. In Korean, [9] showed that listeners are sensitive to this allophonic variation of Lenis stops and use it in determining the presence/absence of AP boundaries. However, [9] did not test how this segmental cue interacts with the tonal cue as her stimuli were artificially monotonized.

While experimental work has shown listeners are sensitive to these cues, it is yet to be investigated to what extent each is informative in separating the boundary type categories. Also, it is not yet known how the two types of cue interact since existing experimental work has tested the tonal [6, 7, 8] and segmental cues [9] independently, but not together. For listeners to be able to use these cues in spontaneous speech, these categories should be distinguishable in a phonetic space parameterized with these cues. Also, given the emphasis on the importance of tonal marking in the K-ToBI literature [4], I hypothesize that the tonal cue alone might be sufficient for separating the boundary type categories.

3. METHOD

To take a first step towards filling these gaps in the literature, this paper investigated a spontaneous speech corpus of Seoul Korean [1], choosing two speakers as a first pilot: one teenage female and a male in his forties since they were maximally different in terms of their age and gender. The data were also filtered to only include syllables with a Lenis stop onset (7393 out of 27607 syllables, 27%).

Following [16], it was assumed that the tonal marking for the AP is realized as a change in F0 over a disyllabic window. In order to parameterize this tonal cue, F0 was extracted for every 5 ms, from syllables with a Lenis onset and the syllable that came before it. The mean F0 over the preceding syllable was compared with the syllable with the Lenis onset, to determine whether the F0 was falling or rising over the two syllables. Then, the difference between the minimum and the maximum F0 values was taken so as to maximize the fall/rise. The difference was standardized as a fraction of range [16] for each given utterance[3]. This variable, named '$\text{M}\textsc{ax}\Delta_{F0}$', was negative when the F0 change was a fall, and positive when the change was a rise. The 'utterance initial' syllables were excluded since they did not have a previous syllable in an utterance, by definition. The utterance is segmented by a pause in the corpus [1], which indicates that the utterance initial syllables are always IP-initial syllables. However, as the IP boundary is not always preceded by a pause, this meant that some PWd-initial syllables could be both IP-initial and AP-initial. Considering all the possible tonal markings of AP junctures, including AP junctures that are also IP junctures, both rising and falling contours could occur at an AP juncture, as the pitch resets when an IP starts [5]. As mentioned earlier, the size of the tonal fall is larger across AP boundaries, than within AP boundaries, we expect some syllables to have a more extreme negative value of $\text{MAX}\Delta_{F0}$ (AP initial) than others (AP non-initial).

The segmental realization of Lenis stops was parameterized using three measurements taken over the Lenis stop closure interval that were reported to be correlated with the lenition of stops: percentage of voiced interval [23], the difference between the maximum and minimum rate of change in intensity as in [21], and the closure duration, which was speech-rate normalized following [9]. Principal Components Analysis was used to combine them into a single variable for comparison with the tonal cue, following [24]. The first component, which accounted for 77% of the variance in the data, was named STRENGTH. A Lenis token that was strong had a positive value of STRENGTH, whereas a Lenis token that was realized as weak had a negative value.

The corpus [1] provided three levels of segmentation: phone, 'Eojeol', and utterance. An 'Eojeol' is an orthographically defined sequence of syllables, that is often loosely equated with the PWd in previous work including the K-ToBI labeling conventions [4], and I follow this assumption, i.e., Eojeol non-initial is equated with PWd non-initial ($[_{\mathrm{PWd}}...\sigma$) and Eojeol-initial is assumed to be PWd-initial ($[_{\mathrm{PWd}}\sigma$). While the corpus does not have AP labels, we can still make inferences about the distribution of AP-initial (some of which may also be IP-initial) and AP non-initial syllables in the $\langle \text{MAX}\Delta_{F0}, \text{STRENGTH} \rangle$ acoustic space via the PWd labels with the following logic.

Assuming the prosodic constituents are strictly layered, we can assume that PWd non-initial syllables must be AP non-initial. But we do expect the PWd initial syllables to be a mix of AP initial and AP non-initial syllables, since there can be multiple PWds inside an AP. The PWd initial ones that are AP-non-initial should have an overlapping distribution with PWd non-initial syllables (which must be AP non-initial), but the PWd initial syllables that are AP-initial should be separable from the PWd non-initial ones along a cue dimension. If such separability is observed along a dimension, I call that dimension informative. We can also infer that the subset of PWd initial syllables not overlapping with PWd non-initial syllables is AP-initial.

In addition to the exploratory analysis presented above, a clustering analysis was performed on $[_{\mathrm{PWd}}\sigma$ to further explore the systematicity of variation in the distribution [25, 26]. By the logic above, $[_{\mathrm{PWd}}\sigma$ in the corpus can be partitioned into three sub-categories: IP-initial, AP-initial (but not IP-initial), and AP non-initial (see Table 1). To test whether these three sub-categories emerge from the distribution of $[_{\mathrm{PWd}}\sigma$, a K-means analysis was performed by setting the value of K from 2 to an arbitrary number, where K is the number of clusters that the algorithm is supposed to find. Then for each K value, a silhouette analysis was performed, which is a commonly used metric to evaluate clustering analysis. The output of a silhouette analysis ranges from 1 to -1; where 1 indicates perfect separation, 0 indicates a complete overlap, and -1 indicates complete misrepresentation of the data. I hypothesized that the number of clusters that best accounts for the variation in the data to be 3 and expected the found clusters to match the three sub-categories listed above.

4. RESULTS

The marginal plot on the right edge of Fig. 1 shows that some syllables with Lenis onsets have a rising F0 from the previous syllable (positive $\text{MAX}\Delta_{F0}$) and some have a falling F0 (negative $\text{MAX}\Delta_{F0}$). However, it is clear that $\text{MAX}\Delta_{F0}$alone does not distinguish AP-initial syllables from AP non-initial syllables as the $[_{\mathrm{PWd}}\sigma$ (light gray) and $[_{\mathrm{PWd}}...\sigma$ (dark gray) distributions almost completely overlap.

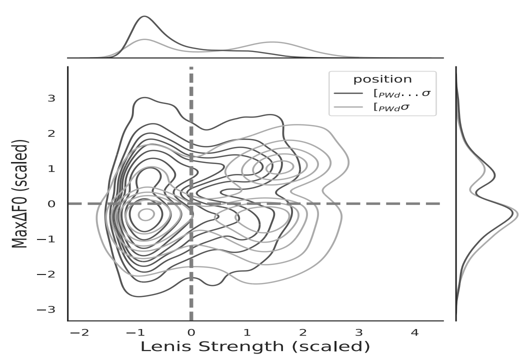

Figure 1: 2D density contour plot with marginal plots showing the distributions of STRENGTH and $\text{MAX}\Delta_{F0}$ on the top and right

On the contrary, the distributions of $[_{\mathrm{PWd}}\sigma$ and $[_{\mathrm{PWd}}...\sigma$ over STRENGTH only partially overlapped (Fig. 1, top). In particular, while the $[_{\mathrm{PWd}}\sigma$ showed two modes in the distribution, one in the positive region and the other in the negative region, $[_{\mathrm{PWd}}...\sigma$ had a unimodal distribution. It is also noteworthy that $[_{\mathrm{PWd}}\sigma$s are distributed to the right

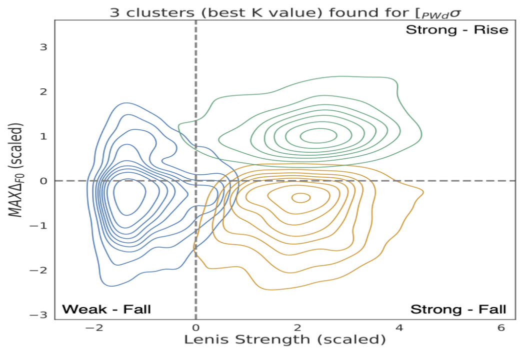

Figure 2: The three found sub-categories of $[_{\mathrm{PWd}}\sigma$, from the K-means clustering analysis

of $[_{\mathrm{PWd}}\ldots\sigma$s in the positive region of the STRENGTH dimension. This indicates that while $[_{\mathrm{PWd}}\ldots\sigma$ could be realized as voiceless (as the Lenis medial voicing is not obligatory, as previously reported [20]), they were still 'acoustically weaker' than $[_{\mathrm{PWd}}\sigma$, as reported in [18].

Overall, contrary to expectation, these findings suggest that the tonal cue is not in fact informative (in the sense operationalized in this paper) for separating the AP-initial syllables from the AP non-initial ones in a large and spontaneous dataset—since the distributions of $[_{\mathrm{PWd}}\sigma$ and $[_{\mathrm{PWd}}\ldots\sigma$ almost completely overlapped. Instead, the results support the informativity of the segmental cue, indexing the allophony of the Lenis stop, for AP boundaries.

In addition to the exploratory analysis presented above, a clustering analysis was performed on $[_{\mathrm{PWd}}\sigma$ to further explore the systematicity of variation in the distribution [25, 26]. As hypothesized, it was found that the number of clusters that best explain the variability in the data was 3 (silhouette score: 0.49). The three clusters found roughly occupied the first, third and fourth quadrants in Fig. 2, which would correspond to 'Strong-Rise', 'Weak-Fall' and 'Strong-Fall', respectively. Among the three, it can be inferred that the Weak-Fall cluster corresponds to $[_{\mathrm{PWd}}\sigma$ that are AP non-initial, since it was expected that voiced/lenited Lenis must be in the AP non-initial position.

There is also evidence suggesting that the Strong-Rise cluster corresponds to IP initial syllables and the Strong-Fall cluster to (non IP initial) AP initial syllables. For Strong-Rise syllables, the top ten most frequent preceding syllables included sentence-final particles; for Strong-Fall, they included case markers, which are indicative of the right edge of an AP [4]. The two Strong clusters were separated along the MAXΔ_{F0} dimension but not along the STRENGTH dimension, which suggests that $[_{\mathrm{PWd}}\sigma$ can be first distinguished into AP-initial or AP non-initial syllables along the STRENGTH dimension, but the Strong $[_{\mathrm{PWd}}\sigma$ can be further separated into AP-initial and the higher constituent initial (e.g., IP) syllables, using the tonal cue. More analysis is required to investigate how AP initial and IP initial syllables can be further separated.

5. DISCUSSION AND CONCLUSION

This paper investigated the informativity of previously proposed tonal and segmental cues to AP boundaries in a corpus of Seoul Korean spontaneous speech. Even though the corpus was not prosodically labeled, I showed how the informativity of the acoustic cues for AP boundaries could nevertheless be investigated, via the distributions of $[_{\mathrm{PWd}}\sigma$ and $[_{\mathrm{PWd}}\ldots\sigma$, and with a clustering analysis. I found that the informativity differs between the tonal and the segmental cues in separating boundary type categories in Seoul Korean spontaneous speech. A clear limitation of this study is that only the syllables with Lenis onset were investigated. It remains to be tested whether other AP-conditioned segmental realizations provide similar informativity as shown in this study (e.g., AP-conditioned denasalization [9]).

Exploring the distributions of the categories in unlabeled data serves as an initial step toward understanding the learnability of the categories, as it resembles the challenge that human learners face, in the sense that they are not given the labels in the data either, cf. the distributional learning literature [27, 28]. Being able to find the expected clusters in the $\langle$MAXΔ_{F0}, STRENGTH$\rangle$ acoustic space could indicate that the AP boundary phonological contrast could be learned from the cues investigated.

6. ACKNOWLEDGEMENTS

I thank Kristine Yu and John Kingston for their valuable feedback.

7. REFERENCES

[1] W. Yun, K. Yoon, S. Park, J. Lee, S. Cho, D. Kang, K. Byun, H. Hahn, and J. Kim, "The Korean corpus of spontaneous speech," *Phonetics and Speech Sciences*, vol. 7, no. 2, pp. 103–109, 2015.

[2] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *ICSLP*, vol. 2, 1992, pp. 867–870.

[3] S.-A. Jun, "The accentual phrase in the Korean prosodic hierarchy," *Phonology*, vol. 15, no. 2, pp. 189–226, 1998.

[4] "K-ToBI (Korean ToBI) Labelling Conventions (version 3.1)." [Online]. Available: https://linguistics.ucla.edu/people/jun/ktobi/k-tobi.html

[5] S.-A. Jun, "Intonational phonology of Seoul Korean revisited," *Japanese-Korean Linguistics*, vol. 14, pp. 15–26, 2006.

[6] S. Kim and T. Cho, "The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean," *JASA*, vol. 125, no. 5, pp. 3373–3386, 2009.

[7] S. Kim, T. Cho, and J. M. McQueen, "Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language," *J. Phon.*, vol. 40, no. 3, pp. 443–452, 2012.

[8] A. Tremblay, T. Cho, S. Kim, and S. Shin, "Phonetic and phonological effects of tonal information in the segmentation of Korean speech: An artificial-language segmentation study," *Appl. Psycholinguist.*, vol. 40, no. 5, pp. 1221–1240, 2019.

[9] K. Yoo, "The production and perception of domain-initial strengthening in Seoul, Busan, and Ulsan Korean," Ph.D. dissertation, University of Cambridge, 2020.

[10] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *JASA*, vol. 91, no. 3, pp. 1707–1717, 1992.

[11] S.-A. Jun, S.-H. Lee, K. Kim, and Y.-J. Lee, "Labeler agreement in transcribing korean intonation with K-toBI," in *INTERSPEECH*, 2000, pp. 211–214.

[12] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. J. Venditti, "X-JToBI: an extended j-toBI for spontaneous speech," in *INTERSPEECH*, 2002.

[13] B. V. Tucker and M. Ernestus, "Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," *The Mental Lexicon*, vol. 11, no. 3, pp. 375–400, 2016.

[14] K. Yoon and S. Kim, "A comparative study on the male and female vowel formants of the Korean corpus of spontaneous speech," *Phonetics and Speech Sciences*, vol. 7, no. 2, pp. 131–138, 2015.

[15] S. Kim, "The role of prosodic phrasing in Korean word segmentation," Ph.D. dissertation, University of California, Los Angeles, 2004.

[16] J. Lee and H. Lee, "Korean Intonation Patterns from the Viewpoint of f0 Percentage Change," *Language and Speech Sciences*, vol. 5, no. 1, pp. 123–130, 2013.

[17] T. Cho and S.-A. Jun, "Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean," *UCLA working papers in phonetics*, pp. 57–70, 2000.

[18] T. Cho and P. A. Keating, "Articulatory and acoustic studies on domain-initial strengthening in Korean," *J. Phon.*, vol. 29, no. 2, pp. 155–190, 2001.

[19] D. J. Silva, "The phonetics and phonology of stop lenition in Korean," Ph.D. dissertation, Cornell University, 1992.

[20] S.-A. Jun, "Lenis Stop Voicing Rule," *Theoretical Issues in Korean Linguistics*, p. 101, 1994.

[21] J. Kingston, "Lenition," in *3rd Conference on Laboratory Approaches to Spanish Phonology*. Cascadilla Proceedings Project, 2008, pp. 1–31.

[22] J. Katz and M. Fricke, "Auditory disruption improves word segmentation: A functional basis for lenition phenomena," *Glossa: a journal of general linguistics*, vol. 3, no. 1, p. 38, 2018.

[23] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *J. Phon.*, vol. 54, pp. 35–50, 2016.

[24] C. Dalcher, "Statistical methods for quantitative analysis of multiple lenition components," in *ICPhS*, 2007.

[25] J. Krivokapic, "The planning, production, and perception of prosodic structure," Ph.D. dissertation, University of Southern California, 2007.

[26] E. Chodroff and J. Cole, "Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English," in *Proceedings of INTERSPEECH 2019*. International Speech Communication Association, 2019, pp. 1966–1970.

[27] N. Feldman, T. Griffiths, and J. Morgan, "Learning phonetic categories by learning a lexicon," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 31, no. 31, 2009.

[28] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, pp. B101–B111, 2002.

[1] There is also an intermediate phrase in the most recent version of K-ToBI [5], but we limit our discussion to the AP in this paper.

[2] The initial tone of an AP is 'H' if the initial segment of an AP is an Aspirated or a Fortis obstruent [3]

[3] Different methods of standardization and measuring not the size of the change but the slope of the fall/rise were tried, but all other measurements patterned similarly.