Interspeech 2022
18-22 September 2022, Incheon, Korea

Dynamic Vertical Larynx Actions Under Prosodic Focus

Miran Oh[1], *Yoonjeong* Lee[1,2]

[1]Department of Linguistics, University of Southern California, USA
[2]Department of Linguistics, University of Michigan, USA
miranoh@usc.edu, yoonjeol@umich.edu

Abstract

It is well known that there is a positive correlation between fundamental frequency and vertical larynx position. Recently, Lee (2018) observes that one vertical larynx movement (VLM) is associated with an Accentual Phrase (AP) in Seoul Korean. The current study builds on these findings by investigating the effect of prosodic focus on vertical larynx actions. Target sentences were designed to produce four APs (e.g., *Joohyun sold six yards of shabby garden field;* $_{\text{AP}}$[Joohyun-SUBJ] $_{\text{AP}}$[shabby garden field] $_{\text{AP}}$[six yards-OBJ] $_{\text{AP}}$[sold-DECL], presented in Korean) and were used to elicit focus on the initial word of the object phrase (e.g., *six*). Articulatory data on VLM is obtained from five Seoul Korean speakers using real-time MRI. Results indicate that quantifiable VLMs observed for each sentence range from 3 to 6 movements, with 4 movements per sentence being the most frequent. Sentences with focus have more instances of VLM per sentence than those without. Focused sentences exhibit significantly greater vertical larynx displacement around the region of focus than the control. Our findings have implications for prosodic planning and pitch resetting, and ongoing analyses examine how VLMs align with Accentual Phrases in Seoul Korean and correlate with fundamental frequency.

Index Terms: speech articulation, speech imaging, real-time MRI, articulatory phonology, vertical larynx actions, prosody

1. Introduction

Vertical larynx height has been investigated with a direct comparison to fundamental frequency (f0), relating f0 values to the absolute vertical larynx positions in the vocal tract [1, 2, 3, 4, 5, 6, 7]. However, most work was limited to the static larynx data, and only a few studies have examined dynamic actions of vertical larynx positions, let alone vertical larynx activities in the context of *prosodic* variation.

In our recent studies on the effect of prosody on vertical larynx actions, we have quantified dynamic actions of vertical larynx movements observed at varying phrasal positions [8, 9], and found that one vertical laryngeal movement roughly corresponds to a single prosodic constituent [10, 11]. Building on these findings, the current study examines kinematic characteristics of vertical larynx (LX) gestures in sentences with or without prosodic focus, extending our preliminary analysis in [12].

Previous studies suggest that vertical larynx actions are controlled in order to manipulate pitch or tonal sequences. However, spatial aspects of articulatory information are often overlooked when identifying phrase boundaries, while changes in pitch range or boundary-adjacent lengthening are considered as indicators of prosodic phrase edges. The current study aims to find laryangeal correlates of prosodic grouping in spoken utterance.

An Accentual Phrases (AP) in Seoul Korean is a prosodic phrasal unit that is defined by a regular tonal sequence of LH(LH) [13, 14]. In addition to the tonal target in the AP-initial syllable depending on the phrase-initial consonant type (tense consonants associated with a high f0 of the following vowel, and lax consonants with a low f0), the phrasal tonal shape of (L/H)HLH surfaces.

However, no quantification method is currently available for labeling Accentual Phrases. The most common labeling convention is to listen to the production and determine AP boundaries with the reference to f0, but the perception of the boundaries vary across and within speakers even with slight changes in speech rate.

The current study examines articulatory correlates of prosodic constituents with a focus on vertical larynx actions in Korean sentences with multiple APs. Real-time Magnetic Resonance Imaging (rtMRI) which images the entire vocal tract during speech production enables a quantitative assessment of non-oral gestures such as larynx and velum. Using state-of-the-art rtMRI data from native Seoul Korean speakers, this study presents dynamic changes of vertical larynx actions as an effect of prosodic focus. This study examines how prosodic focus modulates vertical larynx motions in a language with a linguistic constituent that is prosodically defined. The number of vertical larynx gestures in a sentence are measured to see if this LX count correlates with predicted number of Accentual Phrases. We hypothesize that vertical larynx gestures contribute to prosodic groupings, and that the number of quantifiable LX gestures is equivalent to the number of predicted APs in a sentence.

Lastly, the spatial properties of vertical larynx actions are analyzed to examine the effect of prosodic modulations (via focus) on laryngeal articulatory dynamics. The presence of focus may reshape the AP structures affecting vertical larynx dynamics. We predict that sentences with focus will be produced with potentially greater LX magnitudes than the sentences without such focal prominence.

2. Methods

Articulatory data were obtained by five native speakers of Seoul Korean (2 female, 3 male), with their age ranging from 25 to 31 at the time of the data collection. Target stimuli included sentences with or without a phrase-initial focus with a fixed syntactic structure. All target sentences consisted of five words, in the order of subject, adjective, noun, object, and verb. Given the identical syntactic structure of the utterances, focus was elicited by instructing the participants to produce the sentence with an emphasis on the initial word of the object, placing a corrective focus on the number classifier ("*six*"). An example

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10661

target sentence (with the utterance-medial focused word in bold) is shown in (1).

(1) $_{\text{AP}}$[Joohyun] $_{\text{AP}}$[shabby garden field] $_{\text{AP}}$[**six yards**] $_{\text{AP}}$[sold]
SUBJ ADJ NOUN OBJ VERB
'Joohyun sold six yards of shabby garden field.'

As indicated in (1), the target sentences are predicted to have roughly four APs, while some variations can be observed due to speech rate. At a slower speech rate, each word can be produced as one AP, resulting in five APs per sentence instead of four. At a faster rate, there can be fewer APs per sentence. We observed cases where the object and the verb at the end of the sentence form a single AP, resulting in a total of three APs per sentence. There were nine different tokens with variations in the content words. Each sentence was repeated 8 times per speaker (7 repetitions for one speaker) for both focus and non-focus conditions. A total of 702 sentence tokens were collected and analyzed (2 conditions x 9 items x 8 repetitions x 4 speakers + 2 conditions x 9 items x 7 reps x 1 speaker).

2.1. Data acquisition

Real-time MRI (rtMRI) data of the midsagittal vocal tract and simultaneous audio recordings were acquired using an rtMRI protocol developed for speech production research [15, 16]. Speech imaging data were acquired with a field of view of 200 x 200 mm, a spatial resolution of 84 x 84 pixels, and a temporal reconstruction rate of 83 frames per second [17, 18]. The imaging data included the entire midsagittal plane of the human vocal tract, allowing dynamic visualization of both the lingual and larynx articulation. Data collection followed the approved IRB protocol at the University of Southern California.

2.2. Data analysis

Articulatory landmarks of the vertical larynx (LX) gesture included i) the onset position (the start of the raising movement) and ii) its maximum raising position. Kinematic trajectories of the vertical larynx "centroids," i.e., the pixel intensity-weighted center positions of an object that is closest from the selected point within a defined vocal tract region in the image, are computed using an automated centroid tracking tool, ACT [see 19 for details]. A narrow rectangular vocal tract region was placed near the larynx mass in the image to track the vertical larynx movement to minimize the intrusion of the tongue root and the epiglottis into the region. Temporal landmarks of the LX gestures are captured using the MVIEW find_gest algorithm [20], with landmarks determined by the 20% threshold of the peak velocity of the movement.

From the temporal landmarks of the quantified LX raising gestures, *vertical displacement* values from the movement onset to its maximum and the absolute movement *onset positions* of the LX gestures were obtained for analysis, as displayed in Figure 1. *Vertical displacement* indicates the relative magnitude of each LX gesture, and *onset position* is an absolute position of the LX in the vocal tract. Patterns of LX raising movements between sentences with prosodic focus and those without are evaluated using generalized additive mixed models (GAMMs) [21] with the significance level for statistical testing set at $p < .05$.

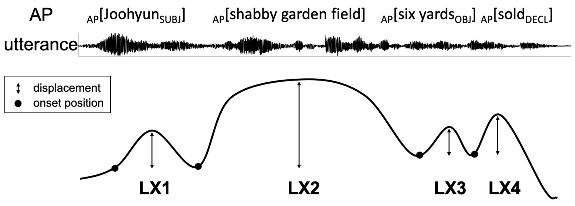

Figure 1: *LX spatial measurements.*

3. Results

The quantifiable vertical larynx gestures identified via the movement velocity profile per sentence are presented in Table 1. As predicted, the most frequent LX counts are four per sentence, which corresponds with the anticipated number of Accentual Phrases for a given sentence. For both focus and non-focus conditions, each sentence has four LX gestures 60 percent of the time on average. Additionally, sentences with focus have more counts of LX gestures per sentence than the control; for example, there are twice as many sentences with 5 LX counts in the focus condition compared to the non-focus condition (24.5% vs. 13.7%, respectively).

Table 1: *LX raising movement counts per sentence.*

LX count	**Count (%)**	
per sentence	Focus	Non-focus
6	11 (3.1)	2 (0.6)
5	86 (24.5)	48 (13.7)
4	208 (59.3)	202 (57.5)
3	46 (13.1)	99 (28.2)
Total	351 (100)	351 (100)

To connect the articulatory findings on LX patterns with the actual realization of APs, the perceived AP counts are determined and cross-checked by the two authors, who are ToBI-trained phoneticians and native speakers of Seoul Korean. For about 90 percent of the data, the produced sentences are perceived to be produced with four APs for both focus and non-focus conditions, as illustrated in Table 2. This confirms that the current stimuli set contains four APs per sentence regardless of the presence of the instructed focus. Both the articulatory LX counts and the perceived AP counts show that target sentences dominantly constitute four prosodic phrases (N.A. represents sentences that are omitted from identifying the number of APs due to speech disfluency).

Table 2: *Perceived AP counts per sentence.*

AP count	**Count (%)**	
per sentence	Focus	Non-focus
5	18 (5.1)	21 (6)
4	323 (92)	315 (89.7)
3	7 (2)	13 (3.7)
(N.A.)	3 (0.9)	2 (0.6)
Total	351 (100)	351 (100)

3.1. Vertical larynx trajectory

Figure 2 presents sample vertical larynx trajectories of sentences with the identical segmental makeup for focus versus non-focus conditions. As shown in the figure, both prosodic conditions have four LX raising gestures. When under focus, we observe a larger magnitude of the gesture contributed by both the onset position and the maximum position. In the domain of focus on the third LX gesture (indicated by a red arrow in the left panel of Figure 2), the vertical displacement of the larynx gesture in pixel (px) is larger, accompanied by a lower absolute position of the movement onset followed by a higher peak position, compared to the third LX movement in the non-focus condition (indicated by a blue arrow in the right panel of Figure 2).

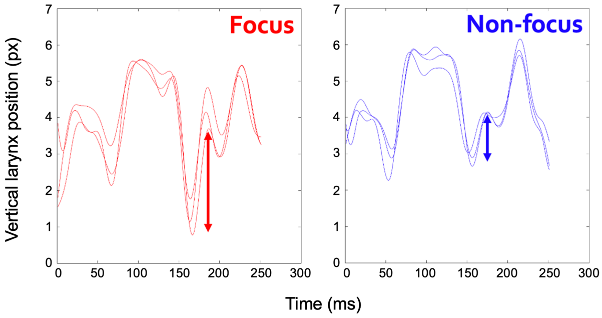

Figure 2: *Sample vertical larynx trajectories with 4 LX gestures (left: focus; right: non-focus).*

3.2. Vertical displacement

The larynx displacement data from five speakers is presented in Figure 3. The LX displacement values visualized over normalized time using GAMMs indicate that there is a significant difference in displacement between focus and non-focus conditions at around two thirds of the way through the sentence, corresponding to the domain of focal prominence (the range of the statistical significance indicated by the interval with dotted lines in Figure 3). Additionally, the LX displacement decreases over the course of the utterance.

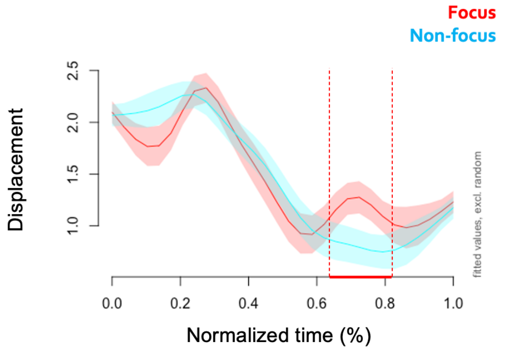

Figure 3: *Overall LX displacement.*

3.3. Onset position

Figure 4 displays the larynx movement onset positions obtained from five speakers. In the domain of the focus manipulation (i.e., 60% to 80% through the sentence), LX raising onset positions are significantly lower under prosodic focus than the control sentence with no focus. In addition to the displacement patterns reported in the previous section, this finding confirms that focus is realized both by lowering larynx at the initiation of focus and by increasing magnitude of larynx raising.

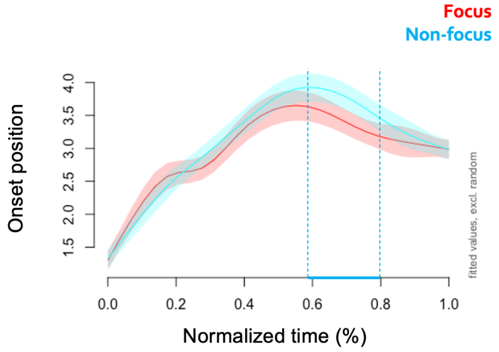

Figure 4: *Overall LX onset position.*

3.4. Individual speaker analysis

Individual speaker larynx behaviors are presented in Figure 5. Overall, phrases under focus have greater vertical larynx displacement and lower onset positions, compared to the non-focus condition (except for Speaker A). Moreover, across conditions, the LX displacement decreases over time through the sentence, indicating the diminished gestural magnitude near the end of an utterance. Despite this trend, the focused phrase, which is near the end of the utterance, still exhibits a larger displacement than that of its unfocused counterpart. As a whole, the results show that vertical larynx actions are spatially conditioned by the prosodic planning of focus implementation.

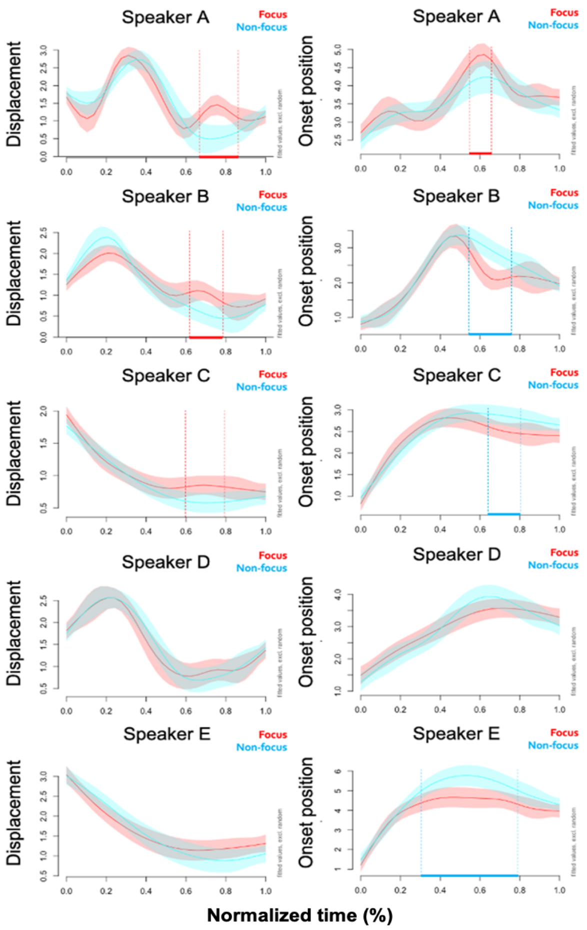

Figure 5: *LX displacement and LX onset position by each speaker.*

4. Discussion

This study presents spatio-temporal characteristics of vertical larynx actions related to the prosodic structure of Seoul Korean. Using real-time MRI data allowing for the detailed examination of non-oral articulators such as vertical larynx movements, we reveal the role of laryngeal gestures on realizing prosodic constituents and information highlighting. We devise an automated tool that quantifies kinematic characteristics of the LX, which is used to identify phrasal structures in a methodologically sounding way. Our findings via the methodological advance are a novel contribution for developing the models of prosodic systems, which, to date, has relied heavily only on the acoustic signal.

First, we observe that each vertical larynx movement roughly corresponds to the linguistically meaningful phonological unit, in this case, an Accentual Phrase in Seoul Korean. Although LX height can be controlled and further modulated by many other factors (e.g., the tense and lax distinction for consonants, global tonal patterns, laryngeal consonants), our findings provide evidence that LX gestures are governed by the language's prosodic system.

Next, the patterns of LX gestures are modulated by prosodic variations such as focal prominence. The LX travels up further under the influence of focus than no focus. The realization of focus is also accompanied by the lowered absolute position of larynx raising onset. Focal prominence has been difficult to quantify in that it is sometimes realized with lowered pitch or raised pitch, in addition to potential increase in intensity or in the magnitude and/or duration of constrictions. The kinematic details of LX, i.e., the spatial characteristics of vertical larynx dynamics, provided in this study reveal the articulatory representation of focus implementation.

Our findings further suggest that tonal adjustments can be modulated by vertical larynx dynamics. For example, there is a sudden increase in vertical larynx displacement and lowered LX onset position occurring in the domain of focus. This indicates that focal prominence may induce pitch resetting at the start of a new phrase. The LX height that affects the tonal register may cue the beginning of a new (focused) phrase, facilitating the perception of phrase boundaries.

Lastly, the gradual decay in gestural magnitude of the LX raising over time may explain the gradual pitch declination over the course of an utterance an utterance, which is observed universally [22, 23, 24]. This implies how physical realization of laryngeal movements necessarily results in tonal downdrift.

In most cases, we observe a single vertical larynx gesture associated with a single prosodic phrase, supporting the proposal of an articulatory reality of a phrasal constitute presented in our earlier study [10]. However, the number of observed LX gestures and the perceived number of APs do not always match (potentially due to the other factors that control LX dynamics, such as consonant type or tonal system). The results presented here are drawn from the articulatory data only, and additional analysis on comparing the f0 contour with the articulatory trajectories available from the current speech imaging data is currently underway. Measuring the number of APs per sentence (currently presented with the articulatory data) will be further supplemented by the f0 trajectory analysis. For instance, Figure 6 presents an example of f0 trajectories along with the LX trajectories of a sentence with 3 APs as well as 3 LX gestures (each phrase separated by dotted lines). This additional analysis step may clarify the mismatch between the production and perception. By investigating the relation between f0 and vertical larynx action, the question of whether there is a more direct association between a single vertical larynx gesture and a single prosodic phrase will be further examined.

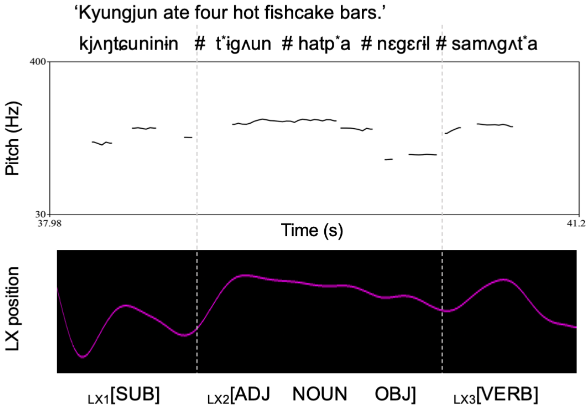

Figure 6: *F0 contour aligned with the LX trajectory with 3 LX gestures ('#' indicates a word boundary)*

5. Conclusions

This study provides evidence for articulatory realities of prosodic planning, by investigating the role of vertical laryngeal actions in prosodic organization and information. Specifically, the study examines the effect of focal prominence on vertical larynx actions to investigate articulatory contributions to prosodic phrasing. With the instrumental tool of real-time MRI speech imaging data and the observed vertical laryngeal control, our findings illuminate how prosodic information is manifested in the articulation of vertical larynx gestures. This work advances our understanding of the vertical laryngeal control in prosodic modeling and speech production.

6. Acknowledgements

This work was supported by NSF 2116376 [PI: Byrd, Co-PI: Oh], NIH DC03172 [Byrd], NIH DC007124 [Narayanan], and NSF SPRF-2105410 [Lee].

7. References

[1] S. L. Hamlet, "Ultrasonic measurement of larynx height and vocal fold vibratory pattern," *Journal of the Acoustical Society of America,* vol. 68, no. 1, pp. 121–126, 1980.

[2] H. Hirai, K. Honda, I. Fujimoto, and Y. Shimada, "Analysis of magnetic resonance images on the physiological mechanisms of fundamental frequency control," *Journal of Acoustical Society of Japan,* vol. 50, pp. 296–304, 1994. (In Japanese)

[3] K. Honda, H. Hirai, S. Masaki, and Y. Shimada, "Role of vertical larynx movement and cervical lordosis in F0 control," *Language and Speech,* vol. 42, no. 4, pp. 401–411, 1999.

[4] K. F. Andersen and A. Sonninen, "The function of the extrinsic laryngeal muscles at different pitch," *Acta Oto-Laryngologica,* vol. 51, no.1-2, pp. 89–93, 1960.

[5] J. Lindqvist, M. Sawashima, and H. Hirose, "An investigation of the vertical movement of the larynx in a Swedish speaker," *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics,* vol. 7, pp. 27–34, 1973.

[6] Y. Kakita and S. Hiki, "Investigation of laryngeal control in speech by use of thyrometer," *Journal of the Acoustical Society of America,* vol. 59, no. 3, pp. 669–674, 1976.

[7] A. M. Laukkanen, R. Takalo, E. Vilkman, J. Nummenranta, and T. Lipponen, "Simultaneous videofluorographic and dual-channel electroglottographic registration of the vertical laryngeal position in various phonatory tasks," *Journal of Voice,* vol. 13, no. 1, pp. 60–71, 1999.

[8] M. Oh, D. Byrd, L. Goldstein, and S. S. Narayanan, "Vertical larynx actions and larynx-oral timing in ejectives and implosives," *3rd Phonetics and Phonology in Europe (PaPE),* Lecce, Italy, Jun. 2019.

[9] M. Oh, *Articulatory Dynamics and Stability in Multi-Gesture Complexes.* Doctoral Dissertation, University of Southern California, 2021.

[10] Y. Lee, *The Prosodic Substrate of Consonant and Tone Dynamics.* Doctoral Dissertation, University of Southern California, 2018.

[11] Y. Lee, L. Goldstein, and D. Byrd, "Laryngeal consonant and tone dynamics in Seoul Korean," *Linguistic Society of America (LSA) 2020 Annual Meeting,* New Orleans, USA, Jan. 2020.

[12] M. Oh and Y. Lee, "Focusing on vertical larynx action dynamics," *Acoustical Society of America (ASA) 179th Meeting,* Acoustics Virtually Everywhere, Dec. 2020.

[13] S. A. Jun, *The Phonetics and Phonology of Korean Prosody.* Doctoral Dissertation, Ohio State University, 1993.

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

[15] S. S. Narayanan, K. S. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America,* vol. 115, no. 4, pp. 1771–1776, 2004.

[16] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Computer Speech & Language,* vol. 52, pp. 1–22, 2018.

[17] S. G. Lingala, A. Toutios, J. Töger, Y. Lim, Y. Zhu, Y. C. Kim, C. Vaz, S. Narayanan, and K. S. Nayak, "State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function," in *Proceedings INTERSPEECH 2016,* San Francisco, Sep. 2016, pp. 475–479.

[18] S. G. Lingala, Y. Zhu, Y. C. Kim, A. Toutios, S. S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," *Magnetic Resonance in Medicine,* vol. 77, no. 1, pp. 112–125, 2017.

[19] M. Oh and Y. Lee, "ACT: An Automatic Centroid Tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data," *Journal of the Acoustical Society of America,* vol. 144, no. 4, EL290–EL296, 2018.

[20] M. Tiede, *MVIEW: Multi-channel Visualization Application for Displaying Dynamic Sensor Movements.* New Haven, CT: Haskins Laboratories, 2010.

[21] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics,* vol. 70, pp. 86–116, 2018.

[22] A. Cohen and R. Collier, "Declination: Construct or intrinsic feature of speech pitch?" *Phonetica,* vol. 39, no. 4-5, pp. 254–273, 1982.

[23] D. R. Ladd, "Declination: A review and some hypotheses," *Phonology,* vol. 1, pp. 53–74, 1984.

[24] B. Connell, "Downdrift, downstep, and declination," in *Proceedings of Typology of African Prosodic Systems Workshop,* Bielefeld, Germany, May 2001, pp. 3–12.