Mora-based pre-low raising in Japanese pitch accent

Albert Lee[1], Yi Xu[1], Santitham Prom-on[2]

[1] Dept. of Speech, Hearing and Phonetic Sciences, University College London, UK
[2] Dept. of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand.

kwing.lee.10@ucl.ac.uk, yi.xu@ucl.ac.uk, santitham@cpe.kmutt.ac.th

Abstract

This study is an attempt to understand the phonetic properties of pitch accent conditions in Japanese as related to the two observed versions of H tones. We tested the hypothesis that the higher version (accented H) results from pre-low raising (PLR) rather than being inherently higher. Correlation analysis reveals an inverse relation between accent peak and the following low tone, and that the strength of such correlations is affected by both peak-to-word-end distance (categorical effect) and within-mora time pressure (gradient), but the two effects work in opposite directions. We take this as evidence that the former effect is due to mora-level pre-planning while the latter is mechanical. These results suggest that in Japanese a low pitch target raises the preceding high target through anticipatory dissimilation. The findings of this study extend our previous understanding of the mechanisms of pitch production.

Index Terms: pre-low raising, Japanese, pitch accent, extrinsic laryngeal muscles

1. Introduction

In Japanese, an accented word is characterized by an initial rise (unless it bears an initial accent), followed by a sharp fall starting from the accented mora (e.g. LH*L, as in the Autosegmental-Metrical representation [1], shown as the solid curve in Figure 1). In contrast, an unaccented word has an initial rise but no sharp fall (e.g. LH- as the dashed curve in Figure 1). The two distinct surface tones (H* and H) are argued by some to bear the same underlying phonological representation (for a comprehensive review see [2], [3]). Proponents of this hypothesis support their view by the lack of perceptual distinction between unaccented and final-accented words like *hashi* (LH) 'edge' vs. *ha`shi* (LH*) 'bridge' when said in isolation, and the fact that most speakers cannot produce such distinction in isolation [4], [5]. For example, Vance [5] reported in a production study that 3 out of 4 subjects make no reliable distinctions between *hana* and *ha`na*. Meanwhile, acoustic evidence abounds showing clear differences between the two accent conditions (see Figure 1) when produced in context. This has led to proposals to assign different underlying representations (H* vs. H) to accented and unaccented words [1], [3].

Such contradictory findings remain a conundrum to this day. The present study is an attempt to solve this conundrum by testing a new hypothesis: There is only one H pitch accent in Japanese and the acoustic difference between the two versions is due to a well attested phenomenon, namely, pre-low raising of surface F_0, or PLR in short.

Also known as anticipatory dissimilation, regressive H-raising or anticipatory raising [6], pre-low raising (PLR) is a local anticipatory tonal variation where the F_0 of a tone becomes higher when preceding a Low tone. For example, the F_0 of H1 in the sequence H1L1H2 would be higher than in H1H2H3, *ceteris paribus*. PLR has been reported for many languages, including Bimoba [7], Cantonese[8], Mandarin [9], Thai [10], and Yoruba [11]–[13]. Though widely observed, the underlying mechanism of PLR is still unclear despite some speculations [2], [3]. Also, the precise condition that triggers PLR varies from one language to another. For example, in Yoruba PLR is observed when a high tone is followed by a low tone [12], in Cantonese it appears to be only observed in rising tones [8], while in Mandarin both rising tone and low tone can trigger PLR in a preceding tone [14]. What seems common to all cases is that the trigger contains a low pitch point, and the preceding tone has a high pitch point. The Japanese case seems to satisfy this condition, as can be seen in Figure 1. However, because the high F_0 points in the two curves come from two different accent conditions, attributing the higher F_0 of the solid line than the dashed line to PLR could be at least partially circular. One way to reduce the level of circularity is to show that the effect of PLR is gradient rather than all-or-none, because the gradience would be fundamentally incompatible with the two-H-tone hypothesis.

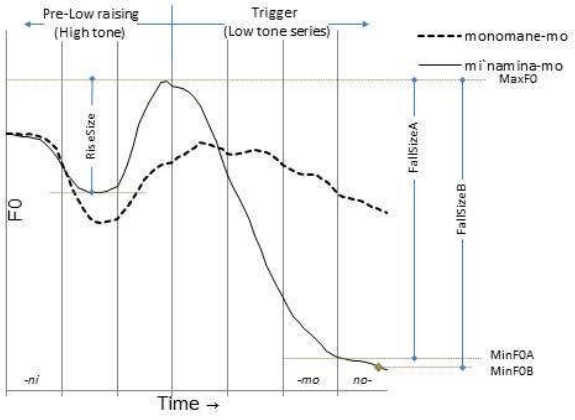

Figure 1: An accented word and an unaccented word.

2. Methodology

A total of 33 Japanese words were chosen as stimuli (see Table 1). The target words varied in length (1-4 morae), accent condition (unaccented and initial/medial/penultimate/final-accent), and syllable structure (CVCV, CVN, CVV). The tokens were presented in the unaccented carrier sentence *Jiten-ni ___-mo nottemasu* 'The word ___ too is found in the dictionary'.

A number of factors were taken into consideration when designing the stimuli. First, past studies of Japanese word prosody often used only bimoraic target words (e.g. [3]) or words that are segmentally incomparable to one another (e.g. [2]), leaving the possibility that the results would be different with longer words and words that have similar segments. Stimuli used in the present study cover a wider range of

phonological contexts (length, accent condition, and syllable structure). And, the use of only nasals as initial consonants avoids most of the distortions from segmental perturbation of F_0. Second, considering that speech rate is generally fast in Japanese and can lead to F_0 target undershoot, we recorded each target sentence at two speech rates to check the effect of speed of articulation. Third, though less directly relevant to the present research question, introducing three types of syllable structure into our stimuli allowed us to gain further insights into the shape of F_0 contours under different conditions, which may be relevant in future studies on prosody modeling.

1-mora	CV		
UA (L-H)	ne		
1 (H*-L)	`ne		
2-mora	CV	CVV	CVN
UA (LH-H)	mane	mai	
1 (H*L-L)	`memo	`mei	`men
2 (LH*-L)	mu`ne		
3-mora	CV	CVV	CVN
UA (LHH-H)	mimono	mimei, neimo	momen
1 (H*LL-L)	`menami	`meimu, `nimei	`ninmu
2 (LH*L-L)	na`name	me`mai	ni`man
3 (LHH*-L)	mimo`no	nui`me	
4-mora	CVCV	CVV	CVN
UA (LHHH-H)	monomane	meimei	nennen
1 (H*LLL-L)		`muumin	`nannen
2 (LH*LL-L)	mi`namina		
3 (LHH*L-L)	nama`nama	mei`mei	men`men
4 (LHHH*-L)	anoma`ma	nimai`me	ninen`me

Table 1. *List of stimuli used in the present study*

Eight subjects (four for each gender, mean age 28.5, s.d. 4.72) were recorded. They were native speakers of Tokyo Japanese from the Greater Tokyo Area (Tokyo, Saitama, Kanagawa, and Chiba) living in London at the time of recording. None of them reported any history of speech, language, or hearing impairment.

Recording took place in a soundproofed chamber in University College London, using a RØDE NT1-A microphone. Subjects were seated in front of a computer screen, on which stimuli were displayed one by one in random order. They produced each sentence first at normal speed, then followed immediately by a slow production. Though speech rate was not stipulated in actual terms, subjects were instructed to speak obviously slower in the second production. When an undesired emphasis was placed on the particle *–mo*, the subject would be asked to repeat the utterance with neutral focus. From each subject a total of 33 stimuli × 5 repetitions × 2 speech rates = 330 tokens were collected. The sampling rate was 44kHz.

Because only nasal stops were used in syllable-initial positions, some low frequency words had to be included. In light of this, subjects were given enough time to rehearse and familiarize themselves with the experiment material before recording commenced. No F_0 patterns peculiar to the less familiar stimuli were observed in subsequent analyses.

Sound files were then annotated using ProsodyPro [15], a Praat [16] script for prosody analysis. Each sound file was labeled, and markings of vocal pulses were manually rectified. Segmentation was done by the "mora", such that a light syllable (CV) counts as one mora while a heavy syllable (CVN or CVV) counts as two. In the latter case, two labeled intervals equal in duration would be assigned. Apart from the target word itself, the mora before (*-ni*) as well as the one after (*no*) were also labeled during annotation, in case any carryover effect extends from or into the target word; other parts of the carrier sentence were not analyzed in the present study. The script then generated all the acoustical measurements from individual files, as well as ensemble files containing data ready for graphical and statistical analysis.

The measurements analyzed in the study include (as illustrated in Table 2): (i) $MaxF_0$—maximum F_0 in the host mora of pitch accent and the following mora, wherever it occurs; (ii) $MinF_0A$—minimum F_0 of the final mora of the target prosodic word, i.e. the final particle –mo in the carrier sentence; (iii) $MinF_0B$—minimum F_0 value of the mora immediately after the target word, i.e. no- in the carrier sentence, with the last 30 ms of the mora excluded; (iv) RiseSize—the difference between $MaxF_0$ and minimum F_0 of all morae that precede the accent peak; (v) FallSizeA—the difference between $MaxF_0$ and $MinF_0A$; (vi) FallSizeB—$MaxF_0$ less $MinF_0B$; (vi) VMaxRise—maximum velocity in initial rise; (vii) VMaxFall—maximum velocity in accentual fall; and (viii) PeakDelay—the difference between accent host onset time and accent peak time divided by accent host duration. Pearson's correlations were calculated between these measurements to examine the relationship between accent peak and the following low tone.

3. Results

Figure 2 displays time-normalized F_0 contours averaged across five repetitions by subjects. We can see that the distance of accent peak away from the end of word is positively related to accent peak height, but inversely related to the F_0 of the right edge of target word. That is, other things being equal, the earlier the pitch accent in a word, the higher its peak F_0 and the lower the F_0 at word end. To verify this observation, we performed Analysis of Variance on the averaged data of accented words. The relative timing of the accent peak (peak-to-word-end distance) significantly affects $MaxF_0$, $F(3,21)=30.684$, $p<0.001$. Post-hoc pairwise comparison confirms ($p<0.05$) that a word has higher $MaxF_0$ when its accent is further away from word end, except that pitch accents that are 3 or 4 morae away do not have significantly different $MaxF_0$. Figure 3 shows how peak-to-end distance affects $MaxF_0$. The first 4 groups from the left on the x-axis are all initial accented words, with the peak of the first group one mora away from word end (i.e. one-mora word), and so on. We can see that **$MaxF_0$** is higher when there is greater distance between the peak and word end (e.g. group 4.1 below).

Our observation about word end F_0 was also confirmed. ANOVA results show that peak to word end distance has a significant main effect on **$MinF_0A$**, $F(3,21)=23.255$, $p<0.001$. Likewise, post-hoc pairwise comparison confirms that groups of accent conditions having different distances from word end have significantly different **$MinF_0A$**, except for those that are 3 morae and 4 morae away from word end, in which case they are not significantly different, as is the case for **$MaxF_0$**.

Then we compared the variables introduced above for possible correlations (N=2640). F_0 data were converted into semitones using the utterance-initial F_0 of each utterance as reference. $MaxF_0$ and $MinF_0A$ were inversely correlated, r=-0.354 (two-tailed, $p<0.001$), suggesting that a lower word-end F_0 is associated with a higher accent peak. However, part of the negative correlation comes from the bimodal distribution as shown in Figure 4, where accented words

generally have a higher MaxF$_0$ and lower MinF$_0$A, and vice versa for unaccented words (blue asterisks).

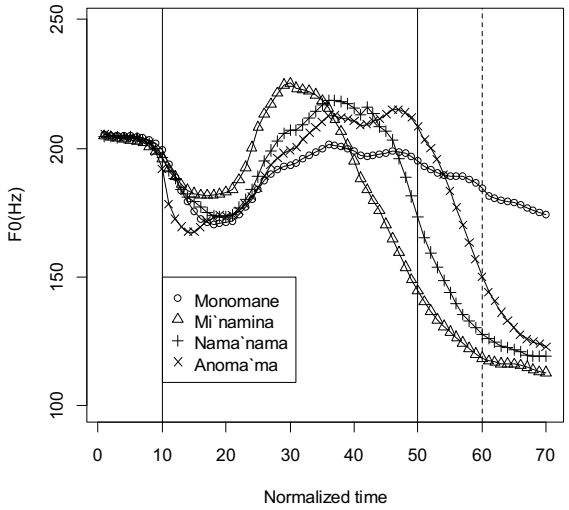

Figure 2. *Time-normalized average F$_0$ contour of four 4-mora words. The solid vertical lines show target word boundaries, while the dashed vertical line marks the end of the particle –mo.*

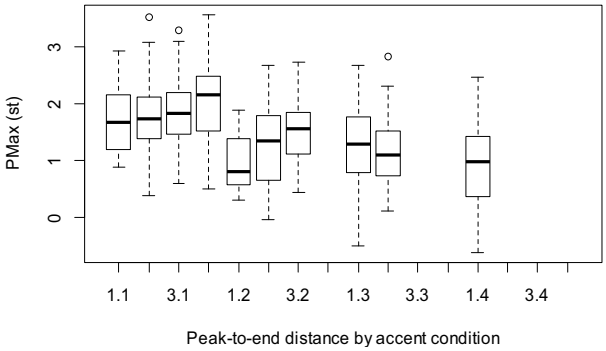

Figure 3. *The effect of peak-to-end distance on MaxF$_0$ in semitones (y-axis). The first digit in each number on the x-axis indicates word length (in morae) while the digit after decimal point indicates the location of pitch accent (the nth mora) in a word.*

Correlations (N=1840)

	MaxF$_0$	MinF$_0$A	MinF$_0$B	VMaxRise	VMaxFall	RiseSize	FallSizeA	FallSizeB
PeakDelay	0.416	-0.176	-0.075			0.207	0.291	0.178
MaxF$_0$						0.694	0.318	0.267
MinF$_0$A			0.767	-0.064	0.057	-0.198	-0.955	-0.745
MinF$_0$B						-0.214	-0.732	-0.967
VMaxRise					-0.125	0.145	0.067	
VMaxFall							-0.060	
RiseSize							0.394	0.382
FallSizeA								0.786

Table 2: *Pearson's correlations of normalized data (converted into semitones using utterance-initial F$_0$ value). Non-significant correlations are not displayed. Data of unaccented words have been removed. The grey intensity in each cell indicates the correlation strength.*

We then repeated the same regression analysis with unaccented words excluded. Table 2 shows that **MaxF$_0$** is positively correlated with **PeakDelay** (r=0.416). That is, in an accented word, when peak occurs later in time, it tends also to be higher. Meanwhile, **PeakDelay** is also inversely related to **MinF$_0$A**, r=-0.176 (or r=-0.235 when normalizing data with word-initial F$_0$ value instead). That is, the lower the word end F$_0$, the later the F$_0$ peak. Similarly, low target also gives rise to a larger initial rise, for **RiseSize~MinF$_0$B** r=-0.214.

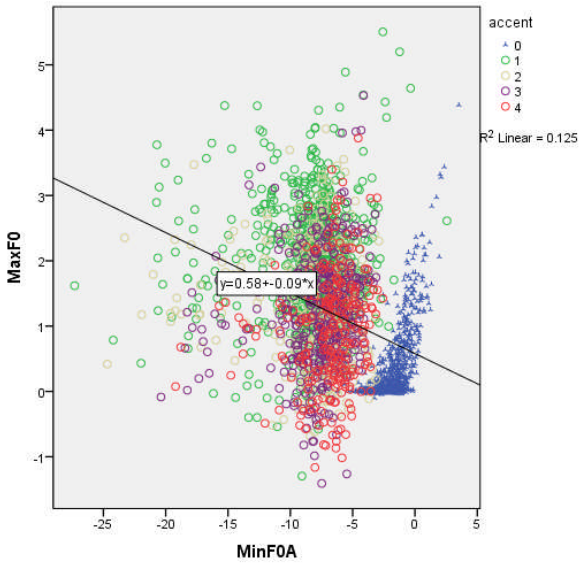

Figure 4. *Scatterplot of MaxF$_0$~MinF$_0$A (N=2640).*

We further divided the data into four subsets according to the distance between accent peak and word end. Here a gradient pattern emerges – **RiseSize~MinF$_0$A** was r=-0.317 when accent was 4 morae away from word end, r=-0.205 when 3 morae away, r=-0.190 when 2 morae away, and r=-0.142 when 1 mora away.

4. Discussion

The above analysis has yielded evidence for PLR in Japanese. First, there is evidence of PLR in the **PeakDelay~MaxF$_0$** and **PeakDelay~MinF$_0$A** correlations. On the one hand, other things being equal, a higher MaxF$_0$ should take longer to achieve, hence a greater peak delay. This is confirmed by the positive **PeakDelay~MaxF$_0$** correlation. On the other hand, a greater **PeakDelay** relative to the accent host mora would leave less time for the movement toward the low F$_0$ at word end, resulting in an undershoot of the low F$_0$. But this is contradicted by the negative **PeakDelay~MinF$_0$A** correlation we have found, which shows that a greater peak delay is associated with a lower F$_0$. Thus a lower **MinF$_0$A** has led to higher **MaxF$_0$**, which in turn led to greater peak delay. This is consistent with previous findings about PLR in other languages, except that no measurement of peak delay was taken in the earlier studies.

The second piece of evidence is that these correlations become stronger as the accent is further away from word end. Given that Japanese is generally spoken very fast (in our data mean mora duration is 117 ms for normal speech and 161 ms for slow speech, respectively), time pressure may have masked part of the PLR effect in the correlations. The fact that the negative correlation in **RiseSize~MinF$_0$A** is the strongest when accent is 4 morae away from word end indicates that a lower F$_0$ indeed gives rise to greater initial rise. Note that the

effect of peak-to-end distance (in terms of number of morae) is not to be confused with gradient measurements like **PeakDelay** and speech rate. In fact, we found that for **PeakDelay~MinF$_0$A,** r=-0.208 at normal speech rate, but r=-0.146 at slow speech rate. This is because, when given more time, the low target is better reached and so less variable, leading to a weaker correlation. Also r is the smallest in **RiseSize~MinF$_0$A** when accent is adjacent to word end, in which case **PeakDelay** is smallest because there is a lack of time to reach the low target, which in turn has led to relatively low **MaxF$_0$**, thus also a smaller negative correlation.

These two pieces of evidence are in support of the hypothesis that variation of F$_0$ height associated with an accent is a function of the height of the following low F$_0$, the lower the following low, the higher the preceding high. But there is also a previously unreported interaction between categorical and gradient effects. At the word level, there seems to be an effect of gross pre-planning, based on the number of morae available to the speaker for achieving the upcoming low target, which the speaker can deduce from lexical knowledge. Within a mora, the exact amount of PLR is dependent on how well the low target is actually achieved, better at the slow speech rate but worse when accent is adjacent to the targeted low. The height of acoustical landmarks in Japanese prosody thus appear to be shaped by both mora-sized planning and mechanistic articulatory effects.

From a physiological perspective, the relationship between **PeakDelay** and **MinF$_0$A** is consistent with our current understanding of the role of laryngeal muscles in phonation. F$_0$ is determined by the tension of the vocal folds, which is antagonistically controlled by the cricothyroids (CT) that lengthen the vocal folds and muscles (thyroarytenoids—TA) that shorten them. But it is also well documented that low F$_0$ involves extrinsic laryngeal muscles like sternohyoid (SH) and thyrohyoid (TH) [17]. This means that an extra force is involved in the production of low F$_0$, which potentially affects the subtle antagonistic balance needed for precise F$_0$ control. There is evidence that this balance is temporarily perturbed after the production of a very low F$_0$, causing F$_0$ to bounce back by an extra amount, a phenomenon known as post-low bouncing [18], [19]. As found in [19], post-low bouncing is highly gradient, and virtually linearly related to the amount of F$_0$ lowering. It has also been shown in [19] that post-low bouncing can be precisely modeled by adding an extra F$_0$ raising force at the onset of the post-low syllable. Unlike post-low bouncing, PLR, as suggested by the present data, is a result of local pre-planning in anticipation of the imminent balance perturbation by the activities of the external laryngeal muscles involved in low F$_0$ production. That is, in anticipating a low target CT muscle activities are increased to pre-balance the antagonistic control of vocal fold tension. This increased activation results in a higher than usual H target, and the amount of increase depends on the predicted amount of forthcoming lowering based on the number of post-accent morae. Interestingly, the present results also indicate that pre-planning can be done only at the level of the smallest unit of individual movement, which is likely the mora in the case of Japanese. That is, speakers seem to anticipate that a low target will be better reached as the amount of time available is increased based on the count of number of morae, as shown by the positive relation between **MaxF$_0$** and distance between the peak and word end. However, as seen above, the within-mora effect (mechanical in nature and more gradient) interacts with pre-planning at the word level (mora-by-mora and more discrete). Thus both post-low bouncing [18], [19] and PLR seem to be the byproducts of maintaining a delicate antagonistic balance in the precise control of vocal fold tension in the production of F$_0$ contours.

On a side note, the distinction between articulatory and mora-by-mora planning effects observed in PLR is also reminiscent of the type of tonal variations in East Asian languages known as tone sandhi. Xu [23] argues that while there is tone sandhi which is postlexical and conditioned by a range of factors (categorical), there is also tonal coarticulation which is phonetically driven and conditioned by tonal context physiological and physical factors (gradient). There are also other more comparable examples in speech where categorical and gradient effects interact, though space forbids us from discussing all of them here.

If pre-low raising is eventually proven to be at work, the current popular accounts of Japanese prosody like [1] will need a fundamental revision. In phonology, where the focus of study is to unearth linguistically meaningful distinctions, the finding of pre-low raising would mean that an apparent surface difference in F$_0$ does not suffice to establish phonemic or tonomic contrasts; neglecting the effect of articulatory mechanisms may reduce phonological analyses to mere annotation of surface forms. Especially, the traditional unidirectional (rightward) approach to word prosody which has been assumed without question will call for serious rethinking. This, of course, will need much more empirical support. Our next step will be to investigate whether the same correlations can be observed in languages where PLR is well established, e.g. Thai, Cantonese and Mandarin. Moreover, although we have found evidence for mora-level categorical pre-planning, it is also possible that the syllable is the real tone bearing unit of the language. We will continue to look at this issue by reanalyzing the data using syllable-based segmentation.

An additional theoretical implication of pre-low raising is its relationship with downstep. In many African languages (see discussion and references cited in [6]), an H1LH2 sequence comes with a raised H1 and a downstepped H2 -- similar to a Japanese pitch accent. In Japanese, a high accent peak is followed by a low target approximated through accentual fall, which is then, if present, followed by a subsequent downstepped word; the downstep is a carryover effect from the preceding low tone. The finding of PLR would allow us to view downstep and pitch accent from a more comprehensive angle, which may lead to models that can better capture contextual tonal variations.

5. Conclusions

In this paper, we have used a quantitative approach to show that in Japanese the F$_0$ peak associated with a pitch accent varies with its following low target. We have found evidence that the variable F$_0$ peak height is the result of pre-low raising. Pearson's r reveals an inverse relation between accent peak and the following low tone, and that such relation becomes more pronounced when the peak is further away from the low target. That the effect of PLR is masked by the proximity between F$_0$ peak and its following low target may explain the absence of similar findings in the literature. These results suggest that in Japanese a low target raises its preceding high tone, which is consistent with our current understanding of the physiology of vocal fold tension control in F$_0$ production.

6. References

[1] Pierrehumbert, J. B., & Beckman, M. E., *Japanese Tone Structure*. Cambridge, MA: Massachusetts Institute of Technology, 1988.

[2] Warner, N., "Japanese final-accented and unaccented phrases," *Journal of Phonetics*, vol. 25, no. 1, pp. 43–60, Jan. 1997.

[3] Sugiyama, Y., *The production and perception of Japanese pitch accent*. Newcastle upon Tyne, England: Cambridge Scholars Publishing, 2012.

[4] Sugito, M., "東京二拍語尾高と平板アクセント考," *Onsei Gakkai Kaihou*, vol. 129, pp. 1–4, 1968.

[5] Vance, T. J., "Final accent vs. no accent: Utterance-final neutralization in Tokyo Japanese," *Journal of Phonetics*, vol. 23, no. 4, pp. 487–499, 1995.

[6] Xu, Y., "Effects of tone and focus on the formation and alignment of F0 contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, Jan. 1999.

[7] Snider, K. L., "Phonetic realisation of downstep in Bimoba," *Phonology*, vol. 15, pp. 77–101, 1998.

[8] Gu, W., & Lee, T., "Effects of tonal context and focus on Cantonese F0," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, 2007, pp. 1033–1036.

[9] Xu, Y., & Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, no. 4, pp. 319–337, Mar. 2001.

[10] Gandour, J. T., Potisuk, S., & Dechongkit, S., "Tonal coarticulation in Thai," *Journal of Phonetics*, vol. 22, pp. 477–492, 1994.

[11] Connell, B., & Ladd, D. R., "Aspects of pitch realisation in Yoruba," *Phonology*, vol. 7, no. 1, pp. 1–29, Oct. 1990.

[12] Laniran, Y. O., & Clements, G. N., "Downstep and high raising: Interacting factors in Yoruba tone production," *Journal of Phonetics*, vol. 31, no. 2, pp. 203–250, Apr. 2003.

[13] Laniran, Y. O., "Intonation in tone languages: The phonetic implementation of tones in Yoruba," Cornell University, 1992.

[14] Xu, Y., "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, no. 1, pp. 61–83, Jan. 1997.

[15] Xu, Y., "ProsodyPro.praat." University College London, London, 2005-2013.

[16] Boersma, P. P. G., & Weenink, D. J. M., "Praat: Doing phonetics by computer." 2012.

[17] Atkinson, J. E., "Correlation analysis of the physiological factors controlling fundamental voice frequency," *Journal of the Acoustical Society of America*, vol. 63, no. 1, pp. 211–222, Jan. 1978.

[18] Chen, Y., & Xu, Y., "Production of weak elements in speech: Evidence from F(0) patterns of neutral tone in Standard Chinese," *Phonetica*, vol. 63, no. 1, pp. 47–75, Jan. 2006.

[19] Prom-On, S., Liu, F., & Xu, Y., "Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling," *Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 421–432, Jul. 2012.