Proceedings of the 7th International Conference On Spoken Language Processing, Denver, Colorado, 2002. pp. 1093-1096.

SEGMENTATION OF GLIDES WITH TONAL ALIGNMENT AS REFERENCE

Yi Xu
Department of Communication Sciences and Disorders, Northwestern University
xuyi@northwestern.edu

Fang Liu
Department of Linguistics, University of Chicago

ABSTRACT

This paper reports an attempt to determine the segmentation of glides using existing knowledge about tonal alignment as reference. It is found that the likely onset of a glide is much earlier than what is acoustically the most obvious, i.e., the point where the formants reach their extremes. It is further found that there is indication that the point of formant extremes may in fact be the point of glide release.

1. THE PROBLEM

If we set aside the issue of overlapping gestures [2, 4], determining segmental boundaries can be quite straightforward in some cases, at least for practical purposes. In a sequence of two CV syllables such as /mama/, the boundary between the two syllables can be said to be at the onset of the second /m/. We may refer to this kind of boundary as the "*de facto*" syllable boundary as well as the "*de facto*" segmental boundary. In some other cases, however, the exact locations of even such *de facto* boundaries are not that clear. One of such cases is when the initial consonant is a glide such as /j/ or /w/. Because there is usually no abrupt shift of the formants, it is hard to determine where the glide, and for that matter, the syllable, begins. As can be seen in the Mandarin word "báyá" [to pull a tooth] (where "´ " denotes the Rising tone) in Figure 1, for example, between two regions of steady-state formants corresponding to the first and second [a], all three formants glide quickly but continuously into and out of a set of extreme values appropriate for [i].

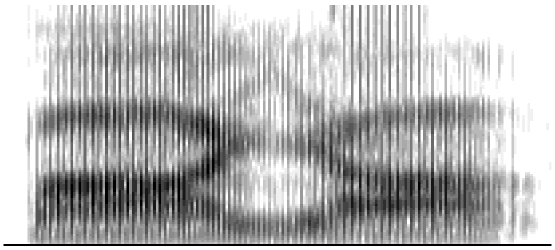

Figure 1. Spectrogram of the Mandarin word "báyá" [paja].

Two questions need to be answered when it comes to the segmentation of glides like the one in Figure 1. First, where is the syllable boundary? Second, what is the interval for the glide, or is there any? Some may argue that, since there is ubiquitous gestural overlap in speech production, it is simply futile to try to look for clear-cut boundaries for any phonetic units, not to mention those as difficult as glides. However, there has been recent evidence that syllable boundaries, at least when it is relatively easy to determine, often function in production as alignment points for prosodic units such as tone and accent [1, 3, 7, 8 9]. This may highlight the importance of recognizing syllable boundaries even if they involve glides. Furthermore, in areas of speech technologies such as synthesis and recognition, the issue of segmentation is often unavoidable. For glides, however, due to lack of empirical data, one can only go with the most obvious acoustic cues such as formant extremes or intensity minima.

So far, we have found only one published paper (by Peterson and Lehiste in 1960 [5]) that provides clear description about how to segment glides from adjacent vowels. To measure the duration of the syllable nuclei, they needed to exclude the duration of initial consonants including [j] and [w]. They claimed that they were able to find regions of relatively steady state for the glides. And they treated certain points where formants values start to change as the boundaries between glides and vowels. As is now well known, however, glides produced in context seldom show apparent steady states, as can be seen in Figure 1.

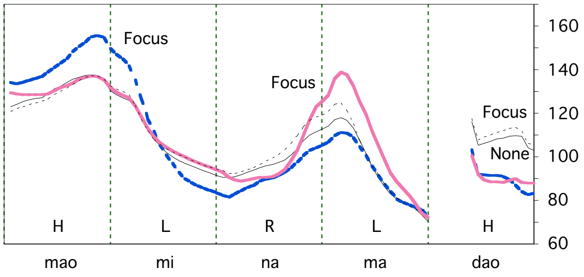

Figure 2. F_0 tracings of the Mandarin sentence "māomǐ ná mǎdāo" [Cat-rice takes the sable]. The four curves differ from each other in terms of focus locations, as indicated by the labels. The vertical dashed lines indicate boundaries of consonant and vowel segments. H, L and R represent the High, Low and Rising tones, respectively.

In a number of recent studies of Mandarin tones, it is found that certain F_0 contours are consistently aligned with certain part of the syllable [7, 8, 9]. For example, in the syllable sequence "ná mă", where "´ " and " ˇ " denote the Rising and Low tones, an F_0 peak was found to usually occur inside the nasal murmur of the initial /m/ of the second syllable. Since it has been found to be rather stable, maybe this alignment pattern can be used to determine the *de facto* boundaries of syllables with initial glides. For example, in the sequence "ná yă", the F_0 peak should also occur within the glide [j] of the second syllable. The *de facto* syllable boundary should then be located somewhere before the F_0 peak. An experiment was therefore designed to determine the *de facto* boundaries in syllables with initial glides using the F_0 alignment patterns in syllables with initial nasals as reference.

2. METHOD

2.1. Material

Eight word pairs were used as testing material, as shown in Table 1.

Table 1. Word pairs used in the experiment. Most of them are nonsense words, although the morphemic meanings of the characters are provided. The tone marks "¯ ´ ˇ `" denote the H (High), R (Rising), L (Low) and F (Falling) tones, respectively.

Pair	Chinese character	Direct English translation	Pinyin	IPA w/o tone	Tone sequence
1	白麻	white hemp	bái má	paɪ ma	RR
	白娃	white child	bái wá	paɪ wa	RR
2	白马	white horse	bái mă	paɪ ma	RL
	白瓦	white roof-tile	bái wă	paɪ wa	RL
3	薄牛	thin ox	báo níu	pɑʊ niou	RR
	薄油	thin oil	báo yóu	pɑʊ jiou	RR
4	薄纽	thin button	báo nǐu	pɑʊ niou	RL
	薄友	cold friend	báo yǒu	pɑʊ jiou	RL
5	败骂	fail scold	bài mà	paɪ ma	FF
	败袜	defeated socks	bài wà	paɪ wa	FF
6	拜妈	worship mother	bài mā	paɪ ma	FH
	拜蛙	worship frog	bài wā	paɪ wa	FH
7	抱拗	hold stubborn	bào nìu	pɑʊ niou	FF
	抱幼	hold baby	bào yòu	pɑʊ jiou	FF
8	抱妞	hold girl	bào nīu	pɑʊ niou	FH
	报忧	report worries	bào yōu	pɑʊ jiou	FH

These disyllabic sequences share the following characteristics:

1. In the first of each pair, the second syllable starts with a nasal, while in the second of each pair, the second syllable starts with a glide. The glide and nasal in each pair share similar places of articulation: [m]:[w], and [n]:[j].
2. The tone of the first syllable is a dynamic one: either R or F.
3. The starting pitch of the second syllable has the opposite value as the ending pitch of the previous syllable: R followed by L or R, and F followed by F or H. This is to guarantee that F_0 makes a sharp turn near the syllable boundary [7, 8, 9].
4. The rhyme of the first syllable is a diphthong whose second formant would end at a very different value from the locus of the following consonant. This is to guarantee a sharp formant turn near the end of the first syllable
5. The rhyme of the second syllable in both words is a vowel or diphthong whose second formant starts at a very different value from the locus of the initial consonant. This is to guarantee a sharp formant turn between the initial consonant and the rhyme of the second syllable

2.2. Subject

Two males and two females who were native speakers of Mandarin served as subjects. Their age ranged from 25 to 46, and none of them reported having any speech disorders.

2.3. Recording

Recording was done in a sound-treated booth in the Speech Acoustics Laboratory in the Department of Communication Sciences and Disorders, Northwestern University. A program was written in JavaScript to control the flow of the recording. The subject was seated comfortably in front of a computer monitor in the booth. The microphone was a head-worn type and was placed approximately one inch away from the left side of subject's mouth.

The subject read aloud the word displayed on the computer screen. In half of the trials, the words were said in isolation while in the other half they were said with a carrier sentence. Subjects were instructed to say the target sentence at a normal rate. The sentences were presented in random order, and a different order was used for each subject.

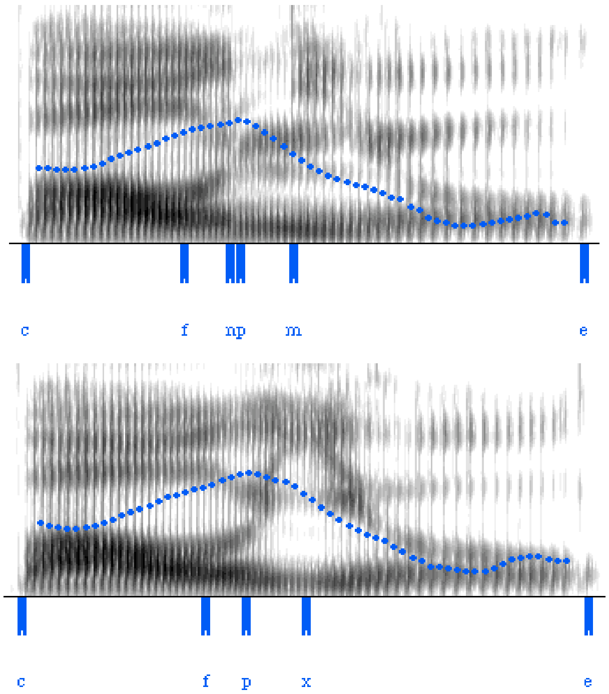

Figure 3. Illustration of markers placed in each word (top: báo nǐu; bottom: báo yǒu). The dotted lines are F_0 tracings generated by Praat. The exact locations of f, p, and x are determined by a computer program

Twelve repetitions of each word were recorded, half with carriers and half without carriers. The first and seventh repetitions were treated as practice trials and were later excluded from the analysis. The utterances were digitized directly into the computer at a sampling rate of 44.5 KHz, but was later down-sampled to 22.05 KHz. The individual trials were then extracted and saved as separate files for later analysis.

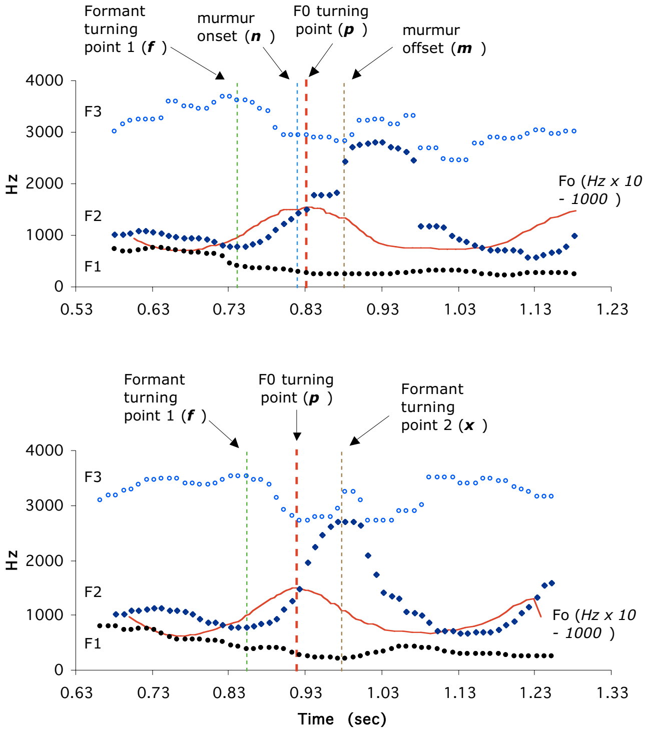

Figure 4. tracings of F_0 and the first three formants for the words "báo níu" (top) and "báo yóu." (bottom) produced by a female subject. The F_0 values are rescaled for clarity of display.

2.4. Measurements

F_0 and formant analyses were done using a procedure that uses Praat (www.praat.org) and a custom-written C program. First, a Praat script was run to perform the following tasks semi-automatically:

1. Load a sound file into Praat.
2. Generate a spectrogram and display it together with the waveform and a one-field TextGrid for manually adding segmental labels. The segmental labels are shown in Figure 3 and their meanings are explained next.
3. Generate markers for individual vocal cycles using the "To PointProcess (periodic)" command and display the markers together with the waveform, which were then manually edited for missing cycles and double markings.
4. Convert the vocal-cycle markers to F_0 values and save them into a file.
5. Generate an LPC formant track using the "To Formant (burg)" command and save it into a file.

The meaning of the labels are as follows:

c — onset of word (starting at stop release)
f — turning point of F2 near the end of syllable 1
n — onset of nasal murmur in the nasal group
p — F_0 turning point near the syllable boundary
m —offset of nasal murmur in the nasal group
x — point of extreme F2 or F3 near the syllable boundary in the glide group
e — end of word

Of these labels, *c*, *n*, *m* and *e* are manually placed, using both spectrogram and waveform as reference. The rest of the markers are placed by a C program which also computed the following measurements.

f-to-p — time lapse from *f* to *p*
p-to-n — time lapse from *p* to *n*
p-to-m —time lapse from *p* to *m*
p-to-x — time lapse from *p* to *x*

Figure 4 shows tracings of the first three formants as well as the F_0 for the words "báo níu" (top) and "báo yóu" (bottom) produced by a female subject. The F_0 values are rescaled for clarity of display.

2.5. Analysis and Results

The goal of the analysis is to determine the likely location of the *de facto* syllable boundaries in the glide group using the F_0 turning point as well as the nasal group as reference. Figure 5 is a summary plot of the mean values of the measurements, including *f-to-p*, *p-to-n*, *p-to-m* and *p-to-x*. In the figure, the F_0 turning point (*p*) is plotted at time 0 and other measurements are plotted relative to it. Plotted this way, the time relation among the measurements provides critical information for determining the *de facto* syllable boundaries of the glide group.

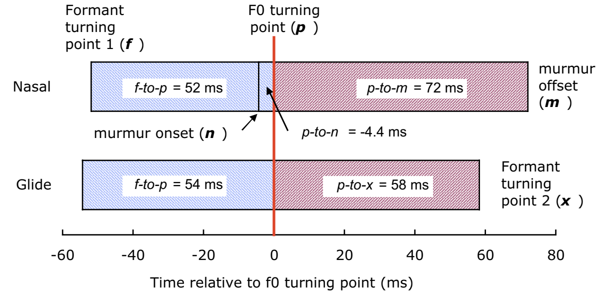

Figure 5. Mean values of f-to-p, p-to-n, p-to-m and p-to-x, averaged across all four subjects. The F_0 turning point (p) is plotted at time 0, which serves as the reference point for all other values.

First, as shown on the upper bar in Figure 5, the mean value of *p-to-n* is negative. This indicates that, on average, the F_0 turning point occurred after the onset of nasal murmur in the nasal group. This agrees with the previous reports about alignment of the dynamic tones in Mandarin [7, 8, 9]. The value of *p-to-n* did vary with the tone of the first syllable, however. It is greater for the F tone than for the R tone (3.6 vs. –12.4 ms). A repeated-measure ANOVA finds this difference marginally significant (F(1,3) = 18.6, *p* = 0.023). This indicates that the turning point occurred earlier in the F tone than in the R tone. This also agrees with previous findings that pitch falls are faster than pitch rises [6, 11].

Second, the values of *f-to-p* seem similar in the nasal group and in the glide group. A repeated measure ANOVA with tone (R/F) and consonant (nasal/glide) as independent variables did not find significant effect of consonant on *f-to-p*. The effect of tone on *f-to-p*, however, is significant (F(1,3) = 38.8, $p < 0.01$), with greater value for R than for F (61 vs. 45 ms). This again confirms faster pitch falls than pitch rises. The similarity in *f-to-p* between the nasal group and the glide group means that the F_0 turning point occurred at about the same time relative to the start of the formant transition toward the initial consonant of the following syllable. This suggests that we may use the F_0 turning point as a reasonable indicator for the location of the *de facto* syllable boundary in the glide group, which is the equivalent of the onset of the nasal murmur in the nasal group. With this indicator, we may conclude that, on average, the onset of the glide also occurs right before the F_0 turning point, just as in the nasal group. To comprehend what exactly this means, one needs to look again at Figure 4. There we can see that this inferred syllable boundary is well ahead of *x* where the formants have the most extreme values for the glide. As shown in Figure 5, on average, the onset of the glide should probably be about 62 ms before *x* (58 + 4 ms).

Finally, and more interestingly, *x* in fact appears to be somewhat comparable to *m*, i.e., the end of the nasal murmur in the second syllable, although the former occurs somewhat earlier than the latter relative to the F_0 turning point. A repeated-measure ANOVA with tone and consonant as independent variables and *p-to-mx* (where *mx* indicate *m* or *x* depending on the consonant) as dependent variable showed marginally significant effect of consonant (F(1,3) = 19.1, $p = 0.022$). This difference, however, does not seem to offset the overall similarity between *p-to-m* and *p-to-x*. The average value of p-to-mx is 65 ms, whereas the difference between *p-to-m* and *p-to-x* is 14 ms, which is a ratio of 4.6:1. It is possible that the 14 ms mostly reflects the difference in the intrinsic duration of the two consonant groups. This is, of course, hard to verify without knowing where exactly the onset and offset of the glides are, which is what the present study is trying to determine. At any rate, the overall similarity between *p-to-m* and *p-to-x* does seem to be a reasonable indication that the point at which formant values are the most extreme for the glide is fairly equivalent to the point of nasal release.

3. DISCUSSION

The present study yields two interesting findings. First, using a known F_0 alignment pattern in words containing initial nasal as reference, and with direct comparisons with words with initial nasals, the *de facto* syllable boundary involving initial glides is found to be much earlier (about 62 ms on average) than the point where the formant extremes are for the glides. This is quite different from what seems to be the obvious acoustic boundary, as can be seen in Figures 1 and 3. Second, to make it even more interesting, comparisons with initial nasals seem to suggest, though more tentatively, that the point where the formants approach the most extreme values for the glides may be in fact equivalent to the point of nasal release. This seems to indicate that by the time formants have reached their most extreme values, the implementation of the glide is over, just as the implementation of the nasal murmur is over by the time the nasal is released.

These findings have interesting implications. Taking yet another look at Figure 4, one may note that the interval between *f* and *x* in a glide is the time during which F2 rises continually from its lowest value to the highest. Suppose we take the onset of the final formant transitions in a vowel toward the following consonant as the starting point for the implementation of the consonant. Then, for a glide, its entire interval would consist of a continuous transition toward its canonical form. This observation reminds us of the recent development in the understanding of the production of lexical tones. That is, the production of the tone has been shown to be a process of continually approximating its underlying target within its allocated time. And, this approximation seems to terminate when the tone's allocated time is over [10, 12]. Should we, then, take the findings of the present study as evidence that this is also the case with the production of glides? If so, should we take the onset of the final formant transitions in the preceding syllable, i.e., *f* in the present study, as the start of the glides? If we do so, how about other segmental sounds? Apparently, the results of the present study seem to raise more questions than answers. Future studies are needed to address these new questions.

4. ACKNOWLEGEMENT

This work is supported in part by NIH Grant DC03902.

5. REFERENCES

[1] Arvaniti, A., Ladd, D. R. and Mennen, I., "Stability of tonal alignment: the case of Greek prenuclear accents", *J. Phon.,* 36, 3-25, 1998.

[2] Browman, C.P. and Goldstein, L., "Articulatory phonology: An overview", *Phonetica, 49*, 155-180, 1992.

[3] Ladd, D. R., Faulkner, D., Faulkner, H. and Schepman, A., "Constant "segmental anchoring" of F0 movements under changes in speech rate", *J. Acoust. Soc. Amer.,* 106, 1543-1554, 1999.

[4] Liberman, A. M. and Mattingly, I. G. "The motor theory of speech perception revised", *Cognition***,** 21, 1-36, 1985.

[5] Peterson, G. E. and Lehiste, I., "Duration of syllable nuclei in English", *J. Acoust. Soc. Amer.,* 32, 693-703, 1960.

[6] Sundberg, J. "Maximum speed of pitch changes in singers and untrained subjects", *J. Phon.,* 7, 71-79, 1979.

[7] Xu, Y., "Consistency of tone-syllable alignment across different syllable structures and speaking rates", *Phonetica*, 55, 179-203, 1998

[8] Xu, Y., "Effects of tone and focus on the formation and alignment of F_0 contours", *J. Phon*. 27, 55-105, 1999.

[9] Xu, Y., "Fundamental frequency peak delay in Mandarin", *Phonetica* 58, 26-52, 2001.

[10] Xu, Y., "Sources of tonal variations in connected speech", *Journal of Chinese Linguistics,* monograph series #17, 1-31, 2001.

[11] Xu, Y. and Sun, X., "Maximum speed of pitch change and how it may relate to speech", *J. Acoust. Soc. of Amer.,* 111, 1399-1413, 2002.

[12] Xu, Y. and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Communication***,** 33, 319-337, 2001.