Underlying Targets of Initial Glides — Evidence from Focus-Related F_0 Alignments in English

Fang Liu and Yi Xu

The University of Chicago, USA

E-mail: liufang@uchicago.edu, xuyi@uchicago.edu

ABSTRACT

In connected speech, initial glides such as /j/ and /w/ often manifest hardly any steady-state formant patterns. In this paper, however, we report data suggesting that, despite their rapid formant movements, glides might have underlyingly static targets. The evidence was revealed when we attempted to determine syllable boundaries involving initial glides as well as the retroflex, using focus-related F_0 alignment in syllables with initial nasals as reference. In addition to finding that the likely onset of a glide or retroflex is much earlier than what is acoustically the most obvious, i.e., the point where formants reach their extremes, we also found that the location of those extreme points was in fact sequentially analogous to the point of nasal release. With this alignment pattern, the entire interval of a glide seemed to consist of a formant trajectory toward a horizontal asymptote, suggesting a static rather than dynamic target.

1. INTRODUCTION

Glides such as /j/ and /w/, as their collective name may suggest, are among the most dynamic-looking elements in speech. In connected speech, they often exhibit hardly any steady-state formant patterns. In the spectrogram of the English word "New York" in Fig. 1, both F2 and F3 are in constant movement, including around the time where the glide [j] might occur. Although F2 may appear to have reached a steady state at the highest value around the time of the arrow in the figure, F3 continues to rise and never reaches a steady state before turning downward, indicating that the vocal track never stops changing its shape. This characteristic of glides presents us with a number of problems, some theoretical and some practical. Theoretically, there are questions as to whether glides are inherently dynamic or static, and whether it has a targeted duration. Practically, there are questions as to how glides should be segmented so that they can be used properly in concatenated synthesis, how they should be simulated in rule-based synthesis, and how they can be recognized in automatic speech recognition.

To address these theoretical and practical issues, it would help to know where a glide begins and where it ends. One may of course point out that with their continuous nature, it is simply impossible to determine the boundaries of glides even if there are any, or that it is actually misguided to attempt to find definitive boundaries for any sound, not to mention glides. Nonetheless, a recent attempt to segment glides in Mandarin has yielded interesting results [13]. The idea is to use the finding that F_0 critical points such as peaks and valleys are often consistently aligned to segmental points such as onset of nasal murmur [1, 7, 9, 10, 11, 12]. Although the interpretation of the onset of nasal murmur as the onset of a syllable itself may be called into question, it offers at least a consistent reference point. In other words, one may start with a quite naïve question: what would be the points in a glide that are somewhat analogous to the onset and offset of the nasal murmur? In the Mandarin study, we attempted to answer this question using tone-related F_0 alignment patterns in syllables with initial nasals as reference [13]. What we found was quite interesting. First, the likely onset of a glide seemed much earlier than what was acoustically the most obvious, i.e., the point where formants reached their extremes (e.g., where the arrow is in Fig. 1). Second, the location of those extreme points seemed in fact temporally comparable to the point of nasal release. Third, with this alignment pattern, the entire interval of a glide seemed to consist of a formant transition toward a horizontal asymptote, suggesting a static rather than dynamic target. Naturally, these findings could be something unique to Mandarin. So, they need to be verified in other languages. The present paper reports an experiment on American English using similar methods as in [13]. Because English does not have lexical tones, we exploited an intonation pattern — focus, that had been found in previous studies to be rather consistent [4, 15]. The experiment used focus-related F_0 peak alignments in syllables with initial nasals as reference to determine the alignment patterns in syllables with initial glides. Because the initial retroflex [ɹ] in English is acoustically somewhat similar to glides, it is also included in the study.

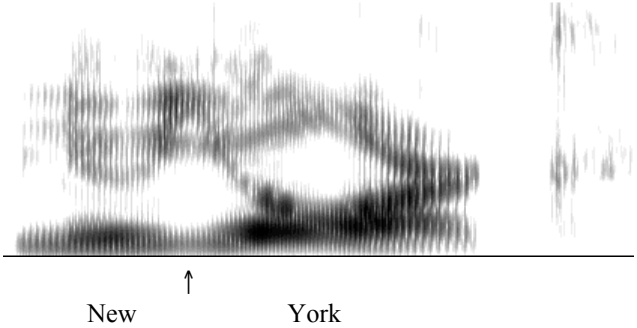

Fig. 1. Spectrogram of the English word "New York". (General American dialect)

ISBN 1-876346-48-5

2. METHOD

Material

Fifteen phrases are used as testing material. They are divided into six comparison sets, each consisting of a nasal phrase, a glide phrase, and, in the first three sets only, a retroflex phrase, as shown below.

1. my meal / my wheel / my reel
2. my mail / my whale / my rail
3. my mike / my wife / my right
4. you knew it / you use it
5. new name / new Yale
6. new novel / new Yahoo

Each set shares the following characteristics:

1. In the nasal phrase, the second word starts with a nasal, while in the other phrase(s), the second word starts with [j], [w] or [ɹ].
2. The initial consonants of the second word have similar F2 values — low: [m], [w], [ɹ], or high: [n], [j].
3. The rhyme of the first word is a diphthong whose F2 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 word.
4. The rhyme of the second word 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 word.

To control the pitch patterns, each phrase is paired with two alternate leading questions. Some examples are given below. The capitalization was shown also to the subject during the recording to further reduce any potential uncertainty as to which word should be emphasized.

What's that?	My WHEEL.
Whose wheel?	MY wheel.
New novel or new movie?	New NOVEL.
Old novel or New novel?	NEW novel.

Subjects

Three female and two male speakers of General American English served as subjects. They were graduate or undergraduate students at Northwestern University. Their age ranged from 22 to 28. None of them reported having any speech disorders.

Recording

Recording was done in a sound-treated booth. A program written in JavaScript was run on a web browser to control the flow of the experiment. The subject was seated comfortably in front of a computer monitor in the booth, wearing a head-mounted microphone, which was placed approximately one inch away from the left side of the subject's mouth. For each trial, the subject read aloud the leading question as well as the target phrase displayed together on the computer screen. They were instructed to say the sentences at a normal rate. The phrases were presented in random order, and a different order was used for each subject.

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 display the spectrogram of each phrase together with the F_0 tracking and a TextGrid for manually adding event labels. The labels are shown in Fig. 2 and they are defined as follows.

b — beginning of word 1
f — turning point of F2 near the end of word 1
n — onset of nasal murmur
p — F_0 peak in word 1 (early focus) or word 2 (late focus)
m — offset of nasal murmur
x — F2 or F3 maximum or minimum in glide or retroflex
e — end of final (or first for one subject) vowel in word 2

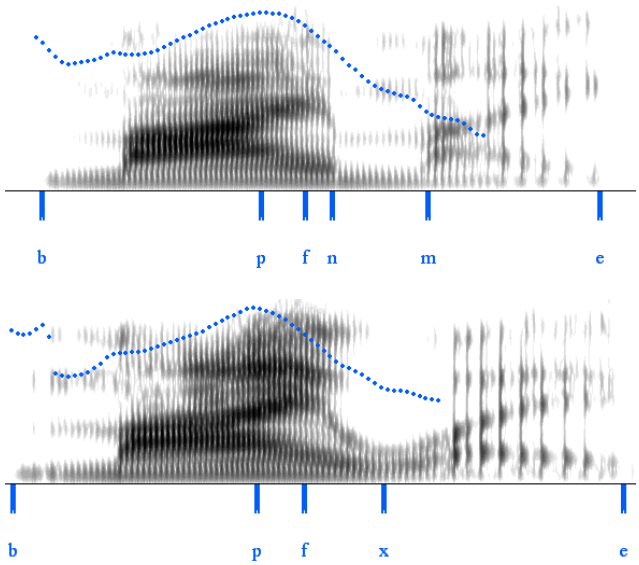

Fig. 2. Spectrograms of "MY mike" (top) and "MY wife," pitch tracking (dotted curves), and event labels.

A C program computed the following values from the event labels.

p-to-f — time lapse from *p* to *f*
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*

Analysis and Results

Reported in this section are analysis results of only part of the data collected in the study. They include all phrases with early focus. The remaining data are currently under analysis and will be reported in future publications. The goal of the analysis is to determine the likely equivalents of the nasal murmur onset and offset in glides and the retroflex using F_0 turning points as reference. Fig. 3 is a summary plot of the mean values of *p-to-f*, *p-to-n*, *p-to-m* and *p-to-x* of all phrases with early focus. In the figure, the F_0 turning point (*p*) is plotted at time 0 and other measurements are plotted relative to it. Displayed this

ISBN 1-876346-48-5

way, the time relation among the measurements provides useful information for determining the equivalents of nasal onset and offset in the glides and the retroflex. First, the mean values of *p-to-f* are similar in the three consonant groups, with that of the nasal phrases only 2 ms longer than the other two. A repeated measure ANOVA did not find this difference to be significant. This similarity means that the formant transition toward the initial consonant of the following syllable started at about the same time after the F_0 peak. This suggests that we may use the F_0 peak as a reasonable indicator for the equivalent of the nasal murmur interval in the glides and the retroflex. In the nasal phrases, the mean value of *p-to-n* is about 63 ms. This means that, on average, the onset of the glides and retroflex should also have occurred about 63 ms after the F_0 peak. Looking at Fig. 3, we can see that this inferred location is well ahead of point *x* where the formants have the most extreme values: 52 ms earlier in the glides, and 24 ms earlier in the retroflex. Looking at Fig. 2 once again, we can see that at this inferred point, F2 is still quite high, and is nowhere near its lowest level.

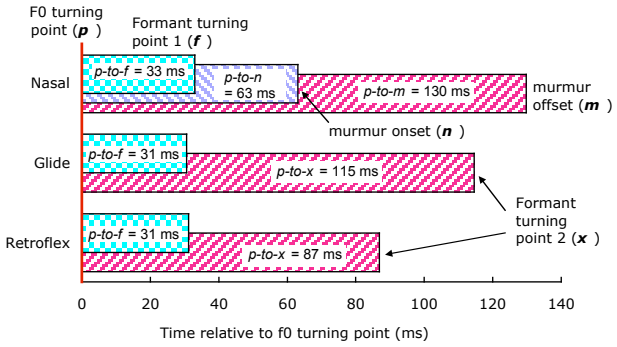

Fig. 3. Mean values of p-to-f, p-to-n, p-to-m and p-to-x, averaged across all five subjects. The F_0 peak (p) is plotted at time 0, which serves as the reference point for all other measurements.

If the point of formant extremes is not equivalent to the nasal murmur onset, what, then, is its nature? In the case of nasals, the next critical event after the onset of nasal murmur is its offset, which is the point of release of the oral occlusion for the nasal consonant. Would it be possible, then, that the point of formant extremes is equivalent or analogous to the point of nasal release? This is plausible at least in terms of the sequence of events. But that itself would not constitute the ultimate evidence. In fact, when comparing the values of *p-to-mx* , which is the time lapse from *p* to either *m*, i.e., nasal release, or *x*, i.e., formant extremes, using a single factor repeated measure ANOVA, we found significant difference due to consonant type, $F(1,2) = 81.2$, $p < 0.0001$. This difference could mean at least two things. Either equating *m* and *x* is totally misguided, or it mostly reflects the different intrinsic durations in these consonant groups. The latter 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. So we defer the question about the equivalent of nasal murmur offset in glides and retroflex to the Discussion.

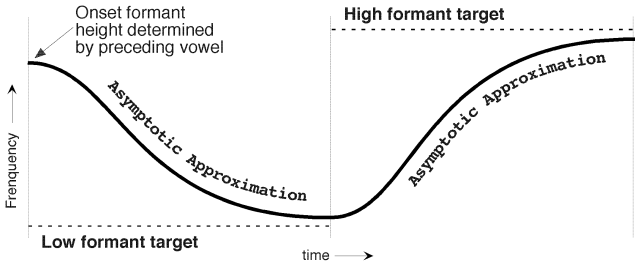

Figure 4. A hypothetical simulation of the F2 trajectory in a glide-vowel sequence. The first and last vertical lines mark the onset of the glide and offset of the vowel. The second vertical line marks the offset of the glide and onset of the vowel. The dashed lines represent the F2 targets of the glide and the vowel. The thick curve represents the formant trajectory resulting from articulatory implementation of the targets.

3. Discussion

The preliminary results of the present study seem to have two direct indications. First, it is highly likely that in a glide or retroflex the event point equivalent to the nasal murmur onset in a nasal consonant is much earlier than the point where formants of the glide or retroflex reach their extremes. Second, at least in terms of the order of the events, the point where formants of the glide or retroflex reach their extremes is analogous to the point of nasal murmur release. The alignment data alone, however, cannot tell us the exact meaning of these indications. Further observation of the formant movements in spectrograms like those in Fig. 2 may help. First, let us refer to the equivalent of *n*, i.e., nasal murmur onset, in a glide as *n′*. In the lower panel of Fig. 2 we note that during the interval between *n′* and *x* F2 continues to drop till it reaches the lowest value. However, we note also that *n′* is not the onset of this continuous drop. Rather, the drop actually starts at *f*, i.e., the start of F2 transition toward [w]. In other words, it is during the entire interval of *f-to-x* that F2 moves continually toward its lowest level. Furthermore, we note also that the transition is not at an even speed during the interval *f-to-x*. Rather, it is asymptotic: starting fast at first but slowing down as the lowest level is being approached. This observation reminds us of a recent development in the understanding of lexical tones. That is, the production of tones seems to be a process of continually approximating its underlying pitch target during the allocated time interval, i.e., its host syllable. And, this approximation seems to terminate when the host syllable is over [12, 14]. Could it be the case, then, that the interval of *f-to-x* is actually the time period during which a glide or retroflex is implemented articulatorily, just as a tone is during the syllable? Fig. 4 sketches a hypothetical simulation of the movement of a single formant asymptotically approaching two consecutive static heights. The simulation is modeled after the pitch target model for lexical tones [14] which is based on findings of tonal realization in continuous speech [8, 10]. We can see

ISBN 1-876346-48-5

in the figure that the shape of the curve bears some resemblance to the F2 trajectory in the lower panel of Fig. 2. In particular, the curve in the left half of Fig. 4 looks quite similar to the shape of F2 during the interval *f-to-x*. This resemblance is quite interesting, because it appears to make the process of articulatorily implementing the glides or the retroflex seem more straightforward than before. That is, they could be implemented simply by asymptotically approximating their most ideal *static* formant values during the allocated time intervals.

Note, however, the foregoing interpretation is inconsistent with the traditional understanding of the acoustic segmentation of adjacent consonant and vowel. This is because, if we take the point *f*, i.e., the onset of the formant transition toward the following consonant, as the starting point of the glide or retroflex, we should also take *f* , rather than *n* — the nasal murmur onset, as the onset of the nasal consonant. One may argue, of course, that this seeming inconsistency can be readily explained by the notion that articulatory gestures of adjacent consonants and vowels are always overlapped [3, 5]. But the understanding outlined in our discussion actually suggests a somewhat different mechanism. That is, whenever a single articulator (e.g., the larynx in the case of tones, and the tip or back of the tongue or the lips in the present study) is involved in making conflicting movements for adjacent sounds, be it consonants or vowels, there is actually no articulatory overlap. The articulator simply finishes one movement before starting the next. Wouldn't this in fact make the speaker's task somewhat simpler? Certainly, we are not even the first one to suggest this understanding. Similar evidence has been reported about the movement of the velum [2]. And, of course, whenever more than one articulator is involved, there is plenty of concrete evidence for actual gestural overlap ([3] and many others), and we would not argue otherwise.

4. CONCLUSION

To summarize, we started in the present study with an attempt to find for the initial glides [j] and [w] and the retroflex [ɹ] in American English the equivalents of the onset and offset of the nasal murmur in initial nasals. The results of our preliminary data analysis suggest that, (a) the equivalent of the nasal murmur onset in the glides and the retroflex is very likely to be much earlier than the points where the formants are closest to the extreme values, and (b) the equivalent of the nasal murmur offset is at least sequentially analogous to the point of formant extremes in the glides and the retroflex. We interpret these results as suggesting that the production of the glides and the retroflex is a process of articulatorily approximating their most characteristic formant patterns during their allocated time intervals. This interpretation seems to have implications for the understanding of segmentation of consonants and vowels in general. Further studies along similar lines therefore need to be conducted to fully explore the implications suggested by our data.

ACKNOWLEGEMENT

This work is supported in part by NIH Grant DC03902.

REFERENCES

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

[2] Bell-Berti, F., "Understanding velic motor control: studies of segmental context"*, Nasals, Nasalization, and the Velum*. M. K. Huffman and R. A. Krakow. Academic Press, San Diego**,** pp.63-85, 1993.

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

[4] Cooper, W. E., Eady, S. J. and Mueller, P. R., "Acoustical aspects of contrastive stress in question-answer contexts", *J. Acoust. Soc. Amer.,* **77**, pp.2142-2156, 1985.

[5] Daniloff, R. G. and Hammarberg, R. E., "On defining coarticulation", *J. Phonetics,* **1**, pp.239-248, 1973.

[6] 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**, pp.1543-1554, 1999.

[7] Ladd, D. R., Mennen, I. and Schepman, A., "Phonological conditioning of peak alignment in rising pitch accents in Dutch", *J. Acoust. Soc. Amer.,* **107**, pp.2685-2696, 2000.

[8] Xu, Y., "Contextual tonal variations in Mandarin", *J. Phonetics,* **25**, pp.61-83, 1997.

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

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

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

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

[13] Xu, Y. and Liu, F., "Segmentation of glides with tonal alignment as reference", *Proceedings of 7th International Conference On Spoken Language Processing*, Denver, Colorado, pp.1093-1096, 2002.

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

[15] Xu, Y. and Xu, C. X., "Exploring underlying pitch targets in English statements." *J. Acoust. Soc. Amer.,* **110, Pt. 2.**, pp.2736-2737, 2002.