Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 35 (2007) 398–420
Journal of Phonetics
www.elsevier.com/locate/phonetics

Determining the temporal interval of segments with the help of F_0 contours

Yi Xu[a,b,*], Fang Liu[c]

[a]*University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK*
[b]*Haskins Laboratories, New Haven, CT, USA*
[c]*Department of Linguistics, University of Chicago, USA*

Received 21 August 2005; received in revised form 10 June 2006; accepted 14 June 2006

Abstract

The temporal interval of a segment such as a vowel or a consonant, which is essential for understanding coarticulation, is conventionally, though largely implicitly, defined as the time period during which the most characteristic acoustic patterns of the segment are to be found. We report here evidence for a need to reconsider this kind of definition. In two experiments, we compared the relative timing of approximants and nasals by using F_0 turning points as time reference, taking advantage of the recent findings of consistent F_0-segment alignment in various languages. We obtained from Mandarin and English tone- and focus-related F_0 alignments in syllables with initial [j], [w] and [ɹ], and compared them with F_0 alignments in syllables with initial [n] and [m]. The results indicate that (A) the onsets of formant movements toward consonant places of articulation are temporally equivalent in initial approximants and initial nasals, and (B) the offsets of formant movements toward the approximant place of articulation are later than the nasal murmur onset but earlier than the nasal murmur offset. In light of the Target Approximation model (TA) originally developed for tone and intonation [Xu & Wang. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33, 319–337], we interpreted the findings as evidence in support of redefining the temporal interval of a segment as the time period during which the target of the segment is being approached, where the target is the optimal form of the segment in terms of articulatory state and/or acoustic correlates. This new definition may have implications for our understanding of many issues in speech, including, in particular, coarticulation and temporal coordination in speech motor control.

1. Introduction

It is now widely accepted that in speech segmental sounds like consonants and vowels are coarticulated, i.e., overlapped with each other in articulation (Fowler & Saltzman, 1993; Hardcastle & Hewlett, 1999; Joos, 1948; Öhman, 1966, 1967). The articulatory overlap is believed to be so prevalent that a further consensus is that it is simply futile to look for clear-cut boundaries for any phonetic units. What has seldom been pointed out, however, is that the notion of coarticulation itself is in fact heavily rooted in the assumption that we somehow

*Corresponding author. University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK. Tel.: +44 20 7679 5011; fax: +44 20 7679 5107.
E-mail address: yi@phon.ucl.ac.uk (Y. Xu).

0095-4470/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2006.06.002

"know" the temporal whereabouts of segments. That is, a segment is at an interval where its most characteristic acoustic patterns are to be found, namely, patterns that are seen when the sound is produced under optimal conditions. Thus a vowel is at a section of continuous formants with well-defined spectral patterns, delimited by abrupt shifts into adjacent non-vocalic patterns; and a nasal consonant is at a section of continuous formants characterized with broad bandwidth, low overall amplitude and steep spectral tilt, delimited by abrupt shifts into adjacent non-nasal patterns. Abrupt spectral shifts are therefore treated as landmarks that separate one segment from another, and any influence across such landmarks is considered as coarticulation. For example, in the English phrase "a meal", the schwa [ə] is said to be coarticulated with [m] because the formants start to move in the direction of the labial closure "during" the schwa and before the abrupt V-to-C spectral shift. Likewise, because the characteristics of [i] in "meal" also affect the acoustic manifestation of the schwa, as has been found in many studies (Farnetani & Recasens, 1999; Fowler, 1980; Kent & Minifie, 1977; Kühnert & Nolan, 1999; Öhman, 1966), the two vowels are said to be coarticulated across the intervening consonant.

This understanding (i.e., that cross-landmark influences constitute coarticulation) persists even in theoretical approaches that assume articulatory gestures, rather than segments, as the basic phonological units of speech (Browman & Goldstein, 1986, 1992; Fowler, 1986, 1996; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). For example, in the acoustic and perceptual experiments designed to investigate gestural overlap as the mechanism of coarticulation, consonants and vowels are nonetheless divided according to the conventional acoustic landmarks, and acoustic measurements of coarticulatory effects are made with reference to these landmarks (see comprehensive reviews in Hardcastle & Hewlett, 1999). As a result, any cross-landmark acoustic influences, either anticipatory or carryover, are still viewed as due to coarticulation.

Segmentation based on acoustic landmarks runs into problems, however, with approximants such as [j], [w] and [ɹ] that do not involve complete closure of the vocal tract.[1] In connected speech, the acoustic properties most characteristic of these approximants are typically found only in a very brief time interval. For example, in the spectrogram of the English phrase "my wheel" (cf. Fig. 5b), F2 is in constant movement, and it approaches the optimal value of [w] only momentarily, making a sharp turn right away. Although such formant turning points have also been proposed as landmarks that can be used in perception (Stevens, 2002), it is still an open question as to whether they should be viewed as the onset, offset, or center of the approximants. Previous efforts regarding the temporal aspects of approximants have mostly been technical, necessitated by the need to measure their duration. Peterson and Lehiste (1960), for example, tried to measure the duration of utterance-initial /j/,/w/ and /r/ in English, which they said involved steady-state periods. They relied on visual inspection of the spectral patterns but used different criteria to designate the segmental offset. For /w/ they considered the point at which F2 moved up from the steady-state position as its offset, for /j/ the point where F3 had a minimum, and for /r/ the point where F3 moved up. The inconsistency in their methodology highlights the difficulty in segmenting the approximants based on spectral patterns. In the other published studies on duration measurement, the methods of determining the temporal interval of approximants are only vaguely mentioned or even not reported at all (Campbell & Isard, 1991; Crystal, 1982; Crystal & House, 1988, 1990; van Santen, 1992). Turk, Nakai, and Sugahara (2006) explicitly advised against using approximants in experiments designed to study durational patterns.

But the difficulty in segmenting approximants actually highlights an even more general problem. That is, if we do not know for certain the temporal interval of approximants based on their spectral patterns, how can we be so sure about the temporal interval of other, more "obvious" segments, as implied in our coarticulation assumption? Of course, it could be argued that, since coarticulation is ubiquitous, determining the exact temporal interval of any segment is pointless. The problem with this argument is that it uses coarticulation as justification for the lack of understanding of accurate segmentation when coarticulation itself calls for explanation. In fact, so far, coarticulation is used as no more than a cover term for a large amount of commonly observed phenomena whose descriptions are based largely on the conventional definition for the temporal interval of segments.

[1]The lateral [l], also known as an approximant, is not considered in the study because, when serving as an initial consonant, it involves landmarks similar to those of nasals due to partial closure of the oral cavity, and hence is supposedly less problematic in terms of acoustic segmentation. The findings of the present study are nevertheless applicable to the lateral.

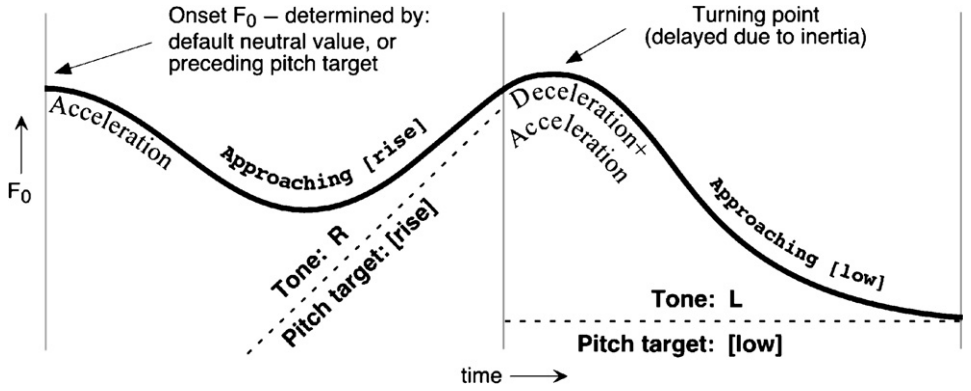


Fig. 1. A schematic sketch of the target approximation model (TA). The dashed lines represent underlying pitch targets, which is dynamic in syllable 1 but static in syllable 2. The thick curve represents the F_0 contours that result from articulatory implementation of the pitch targets. The vertical lines represent syllable boundaries at which the underlying targets abruptly shift from one to another.

The definition of the temporal interval of a phonetic unit does not have to depend on abrupt spectral shifts, however. In tone research, it has been found that a lexical tone is articulatorily realized by continuously approaching its underlying pitch target (Xu, 1997, 1998, 1999, 2001). This finding has led to the Target Approximation model (TA) (Xu & Wang, 2001), a sketch of which is shown in Fig. 1. In TA, it is assumed that each tone has an ideal pitch pattern (dashed lines in Fig. 1) largely resembling its conventional description, e.g., a dynamic rise for the rising tone[2] and a static low for the low tone, and that the production of a tone is to continually approach the target throughout the temporal interval designated for its implementation. Thus in TA the laryngeal effort to approach the ideal pitch pattern of a tone constitutes the entire articulation of the tone. In other words, the start of the articulatory effort to approach the tonal target *is* the onset of the tone, and the end of such effort *is* the offset of the tone. In this model, although the preceding tone always has carryover influence on the following tone (both in terms of displacement and velocity (Chen & Xu, 2006; Xu, 2001), the articulations of the two adjacent tones are not overlapped, because by definition the articulatory effort to approach the first tone has ended by the time F_0 starts to move toward the target of the second tone. In Fig. 1 this is illustrated by the *instantaneous* shift of targets at the syllable boundary.

For consonants and vowels, a similar understanding is represented by the task dynamic model (TD) (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). TD assumes that the primitives of consonants and vowels are articulatory gestures, which are defined as articulatory movements toward static vocal tract configurations. However, unlike tones, for which most of the articulatory execution is directly reflected in the F_0 contours, the goal-reaching articulatory movements for segments frequently result in abrupt spectral shifts rather than uninterrupted smooth movements. Because of this, it has been suggested that the articulatory period of a segment starts earlier than its acoustic period (Bell-Berti & Harris, 1981; Bell-Berti & Krakow, 1991), as illustrated in Fig. 2 of Bell-Berti and Harris (1981). Such a suggestion, however, seems to imply that the initial movement toward a segmental target does not leave any acoustic trace. But it is well known that it does. For example, in a V_1–C–V_2 sequence, formants start to move toward the C or even V_2 well before the stop closure (Öhman, 1966). These movements are commonly known as the V–C formant transitions (Fant, 1973; Öhman, 1966; Stevens, 1998). Thus, the start of a gestural movement *is* acoustically manifested. The complication arises at the end of goal-reaching movement for segments like stops and nasals. This is because the tightest vocal tract constriction does not occur at the onset of the oral closure, but rather sometime during the closure (Löfqvist & Gracco, 1999; Westbury & Hashi, 1997). Thus, for these sounds the end of the gestural movement is typically hidden behind a quasi-steady-state acoustic pattern.

[2]It is assumed in TA, based on empirical data (e.g., Xu, 1998, 1999, 2001), that some underlying targets are intrinsically dynamic. But the dynamic characteristics of a target are not equivalent to the dynamic trajectories in the surface form. The former is specified by the communicative function that uses the target to contrast with other targets, whereas the latter is the consequence of articulatorily implementing a target, which is either static or dynamic, in a given context. See Xu and Wang (2001) for detailed discussions.

Approximants like [j] and [w] present a rather different case. When they are intervocalic, no complete closure of the vocal tract is involved, and the formant movements are largely smooth and continuous. This means that the goal reaching movements in intervocalic approximants are in effect *acoustically transparent*, just like in the case of lexical tones.[3] Applying the definition for temporal intervals according to TA, the start of the formant movement toward its ideal pattern should be interpreted as the onset of the approximant, and the end of this movement the offset of the approximant. Following this account, the onsets of the formant transitions in an approximant and a stop or nasal are largely similar. The offset of the formant transition in an approximant, however, should be aligned somewhat earlier than the offset of a stop or nasal closure. This is because the offset of the closure is actually the moment when the opening articulatory movement toward the following vowel, after going on for a short while, has just resulted in the parting of the major articulators (the lips in the case of [m] and [b], and the tongue tip and the alveolar ridge in the case of [n] and [d]) according to previous articulatory studies (e.g., Löfqvist & Gracco, 1999; Westbury & Hashi, 1997). This hypothesis, of course, requires empirical support. What is particularly needed is an independent time reference with which the relative timings of consonants with different degrees of acoustic transparency can be compared to each other.

Interestingly, a potential source of time reference has recently emerged from research on tone and intonation. It has been found in an increasing number of languages that certain F_0 events related to lexical tone or focus are consistently aligned with acoustic landmarks of some segmental sounds. For example, as shown in Fig. 2a, in the Mandarin syllable sequence "ná mǎ" (where "ˊ" and "ˇ" denote the rising and low tones), an F_0 peak usually occurs inside the nasal murmur of the initial [m] of the second syllable (Xu, 1999). In Chichewa, F_0 peaks occurs consistently right after the offset of the syllable carrying the high tone if the syllable is pre-penult (Kim, 1999). In Dutch, the onset of an "accent-lending" F_0 rise is always aligned with the syllable onset (Caspers & van Heuven, 1993; Ladd, Mennen, & Schepman, 2000). In Greek, an F_0 maximum "is very precisely aligned just after the beginning of the first postaccentual vowel" (Arvaniti, Ladd, & Mennen, 1998, p. 23). In an English pre-nuclear accent, the F_0 peak occurs around 40 ms after the offset of the stressed syllable at normal speech rate (Ladd, Faulkner, Faulkner, & Schepman, 1999). Xu and Xu (2005) have found that in an English statement, an F_0 valley always occurs very close to the onset of a stressed syllable whether or not the syllable is focused (i.e., emphasized), as indicated by the right arrow in Fig. 2b.[4] Similar alignment patterns have been found in other studies (e.g., D'Imperio, 2001, for Neapolitan Italian, Frota, 2002, for European Portuguese; Grabe, 1998, and Atterer & Ladd, 2004, for German; Grabe, Post, Nolan, & Farrar, 2000 for British English; Prieto, van Santen, & Hirschberg, 1995, for Mexican Spanish; Silverman & Pierrehumbert 1990 for American English; Xu, 1998, 2001 for Mandarin Chinese). Also Xu and Xu (2005) have found that in English the F_0 peak location of a stressed syllable under focus is consistently aligned either before or after the syllable offset, as long as factors such as vowel length, position of stressed syllable in word and position of word in sentence remain constant. In particular, the F_0 peak always occurs before the offset of a focused stressed syllable when the vowel is long and the stress is word final, as indicated by the left arrow in Fig. 2b. Somewhat similar patterns have been reported for Dutch (Schepman, Lickley, & Ladd, 2006).

While the exact nature of such F_0-segment alignment is beyond the scope of the present paper (cf. Xu, 2005; Xu & Liu, in press for detailed discussion), what the recent findings tell us is not only how tonal events are aligned to segmental events, but also how segmental events are aligned to tonal events. Because of this, F_0 alignment may be used as a time reference for determining segmental alignment in cases where ambiguity is severe. To be able to do so, however, one must also be aware of the variability in the exact F_0 alignment across languages, dialects or even within the same dialect. In Beijing Mandarin, for example, the exact F_0 alignment in each syllable depends on the lexical tone, tonal context and focus location (Xu, 1998, 1999, 2001). In English, the exact F_0 alignment depends on lexical stress, phonological length of the vowel, location of stressed

[3]When an approximant like [ɹ] is clustered with another consonant as in "try", part of the [ɹ] articulation is hidden by the stop closure, and so it would no longer be acoustically transparent.

[4]Here *focus* is often referred to as the *nuclear accent* in the conventional *intonational phonology* (Ladd, 1996). But as argued in Xu (2005), whereas focus is primarily defined in function, nuclear accent is defined primarily in form. As demonstrated by Xu and Xu (2005), nuclear accent, as a formally defined unit, confuses between rather different functions such as focus and lexical stress.

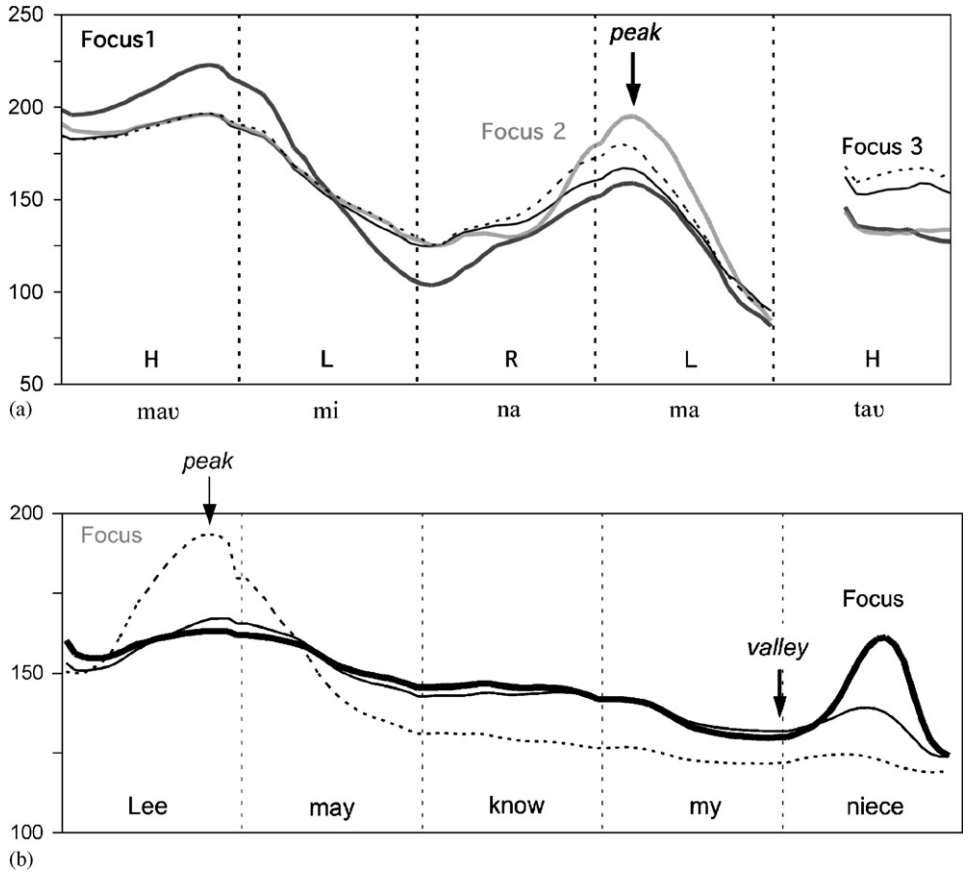


Fig. 2. (a) Mean F_0 contours (each averaged over 48 tokens by 8 speakers) of the Mandarin sentence "māomǐ ná mǎdāo" [Cat-rice holds the sable]. The four curves differ from each other in terms of focus locations, as indicated by the labels. The vertical dashed lines indicate boundaries of consonant and vowel segments. H, L and R represent the high, low and rising tones, respectively. Adapted from Xu (1999). (b) Mean F_0 contours (each averaged over 49 tokens by seven speakers) of the English sentence "Lee may know my niece" spoken with no focus (thin curve), initial focus (dashed curve) and final focus (thick curve). Adapted from Xu and Xu (2005).

syllable in a word, location of a word in a sentence and focus condition of the sentence (Xu & Xu, 2005). There are also reported differences in the exact F_0 alignment across languages and dialects (Atterer & Ladd, 2004). In addition, although there has been evidence that the *underlying* tone-syllable alignment remains constant regardless of intrinsic characteristic of the segment such as intrinsic duration, voicing and aspiration, the measured F_0 alignment may have systematic variations due to aerodynamic effects and the maximum speed of pitch change (Xu, 2001; Xu & Sun, 2002; Xu & Wallace, 2004; Xu & Xu, 2003, 2005). Despite the variability, nevertheless, the consistency of F_0 alignment remains high as long as the afore-mentioned factors are kept constant.

The present study is therefore an attempt to use F_0 alignment as a time reference to explore the temporal interval of segments. Specifically, we will compare the F_0 alignment patterns between initial nasals and approximants in two languages, Beijing Mandarin and General American English, with the goal to test the validity of defining the temporal interval of a segment as *the time period during which the target of the segment is being approached*, where the target is the ideal form of the segment in terms of articulatory state and/or acoustic correlates. Following the discussion made earlier, two main hypotheses will be tested. (A) The onset of formant movements toward the consonantal target (i.e., the start of the V-to-C transition) occurs at roughly the same time in an approximant as in a nasal. (B) The offset of formant movements toward the approximant target (i.e., the end of the V-to-C transition) occurs later than the onset of a nasal murmur but earlier than the offset of the murmur.

2. Experiment 1

This experiment set out to test in Mandarin the two main hypotheses of the study outlined in the Introduction. This was to be done by comparing the spectral alignment of approximants [j] and [w] with that of initial nasals using F_0 turning points as the time reference. The basic strategy was to make the comparisons as direct as possible by finding pairs of disyllabic sequences in which the turning points in F_0 and formant movements were the least ambiguous and maximally free from the extraneous factors such as variations in lexical tone, tonal contexts, focus location, consonant voicing and speaker dialect. We constructed pairs of disyllabic sequences in which (a) F_0 would make a sharp turn near the initial consonant of the second syllable due to specific lexical tones, (b) F1, F2, and/or F3 would make two sharp turns about this consonant, so that the formant movements toward the consonant would be clearly separated from those toward the flanking vowels, and (c) other than the initial approximants and nasals being compared, the tonal and segmental compositions were identical.

2.1. Material

Eight word pairs were used as testing material, as shown in Table 1. Some of them are real words, others either meaningful phrases or nonsense sequences. They are nevertheless all easily pronounceable by native speakers.

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 an approximant. The approximant and the nasal in each pair share similar places of articulation: [m]/[w], and [n]/[j].
2. The tone of the first syllable has a different ending pitch from the beginning pitch of the following tone, e.g., R (rising) tone is followed by L (low) tone, and F (falling) tone is followed by H (high) tone. This is to guarantee that F_0 makes a sharp turn near the initial consonant of the second syllable (cf. Xu, 1998, 1999, 2001).

Table 1
Chinese word pairs used in Experiment 1

Pair	Chinese character	Literal English translation	Pinyin	IPA w/o tone	Tone sequence
1	白麻	White hemp	bái má	paɪ ma	RR
	白娃	White kid	bái wá	paɪ wa	RR
2	白马	White horse	bái mǎ	paɪ ma	RL
	白瓦	White roof-tile	bái wǎ	paɪ wa	RL
3	薄牛	Thin ox	báo niú	paʊ niou	RR
	薄油	Thin oil	báo yóu	paʊ jiou	RR
4	薄纽	Thin button	báo niǔ	paʊ niou	RL
	薄友	Cold friend	báo yǒu	paʊ jiou	RL
5	败骂	Scold in frustration	bài mà	paɪ ma	FF
	败袜	Decayed socks	bài wà	paɪ wa	FF
6	拜妈	Greet mother	bài mā	paɪ ma	FH
	拜蛙	Greet frog	bài wā	paɪ wa	FH
7	抱拗	Persist stubborn	bào niù	paʊ niou	FF
	抱幼	Hold baby	bào yòu	paʊ jiou	FF
8	抱妞	Hold girl	bào niū	paʊ niou	FH
	报忧	Report troubles	bào yōu	paʊ jiou	FH

Most are nonsense words, although the morphemic meanings of the characters are provided. The tone marks "– ˊ ˇ ˋ" denote the H (high), R (rising), L (low) and F (falling) tones, respectively.

3. The rhyme of the first syllable consists of a diphthong whose F2 movement is in the opposite direction of the transitional movement toward the locus of the following consonant. This is to guarantee a sharp formant turn near the end of the first syllable.[5]
4. The rhyme of the second syllable is a vowel or diphthong whose F2 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. Subjects

Two male and two female native speakers of Mandarin served as subjects. Their age ranged from 25 to 46. Three of them (two females and a male) were born and raised in Beijing. The fourth one was raised speaking standard Mandarin. None of them reported any history of speech or hearing 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. The microphone was a head-worn type (Countryman Isomax hypercardiod) and was placed approximately 1 in away from the left side of the subject's lips.

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 within the carrier sentence "yào xiě __ zhège cí" [must write the word ___]. Subjects were instructed to say the target sentences at a normal rate. The sentences were presented in random order, and a different order was used for each subject.

Twelve repetitions of each word were recorded, half with carriers and the other 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 44.1 KHz, and were later down-sampled to 22.05 KHz.

2.4. Measurements

F_0 and formant analyses were done using a procedure that uses a custom-written script for Praat (Boersma, 2001) and a custom-written C program. First, the Praat script was run to display the spectrogram of each utterance (with Praat's default spectrogram settings) together with a TextGrid for manually adding event labels. The event labels are shown in Fig. 3 and their meanings are explained in the following:

W_{start}—word onset (starting at stop release) (hand labeled),

F_{turn1}—formant turn 1: this is the turning point of F2 (maximum or minimum F2) near the end of syllable 1,

N_{start}—nasal murmur onset in the nasal group (hand labeled),

$F0_{turn}$—F_0 turning point (maximum or minimum F_0) in the vicinity of the possible syllable boundary,

N_{end}—nasal murmur offset in the nasal group (hand labeled),

F_{turn2}—formant turn 2: this is the F2 (or F3 if F2 was too weak) maximum or minimum in the vicinity of the possible syllable boundary in the approximant group (hand labeled but algorithmically finalized),

W_{end}—word offset (hand labeled).

As indicated above, W_{start}, N_{start}, N_{end}, and W_{end} were manually placed, using both spectrogram and waveform as references. The rest of the event labels were placed algorithmically with the following procedures.

[5]It has been pointed out to us that many acoustic studies have shown that the F2-locus (especially in the context of labials) is not really a point but can span a fairly wide range (e.g., Fant, 1973; Kewley-Port, 1982; Lehiste & Peterson, 1961), and that the F2-target of an [aɪ] diphthong need not be that different from the so-called F2-locus of a following bilabial, given that the 2nd target of a diphthong is so often undershot (e.g., Gay, 1968). While these concerns are certainly valid, what is needed for the design to work is only *sufficient* directional change in the F2 movement due to the difference between the desired ending value of the diphthong and the desired locus of the consonant. The fact that measurable formant movements were produced by our subjects demonstrated the feasibility of the design.

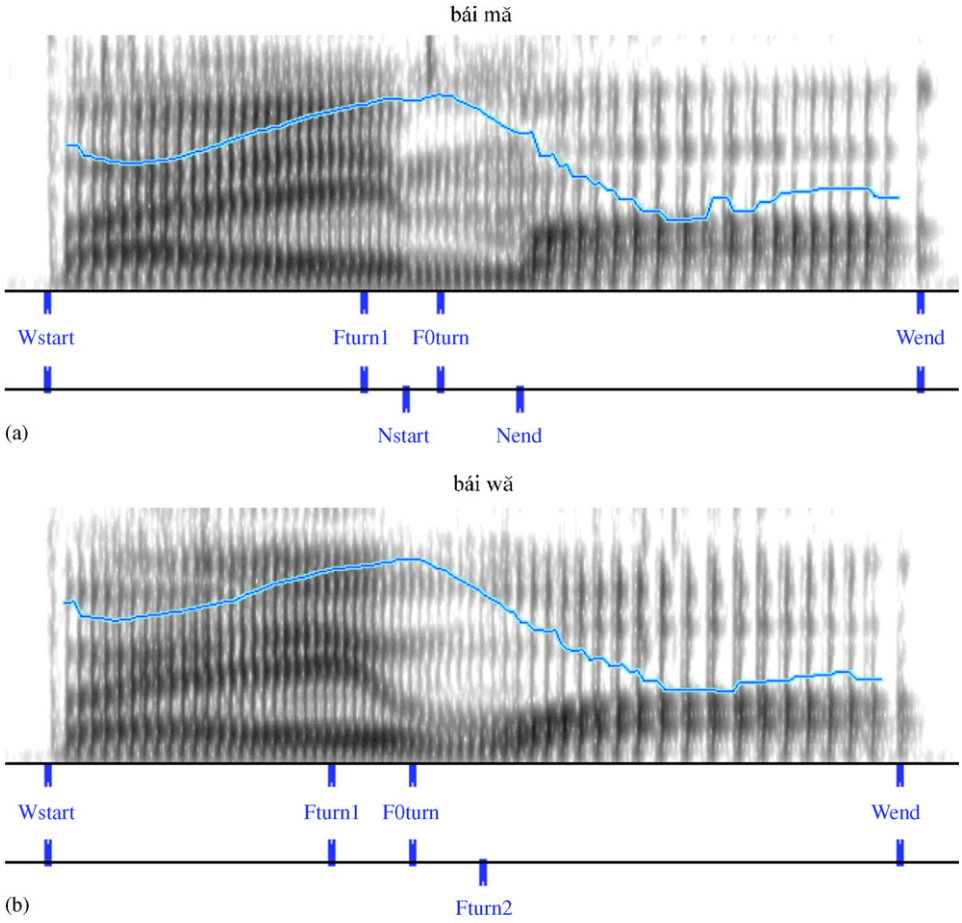


Fig. 3. Illustration of event label placement in Experiment 1. (a) Spectrogram, pitch track (thin curve, generated by Praat) and event labels of bái mă (white horse; tones: R L). (b) Spectrogram, pitch track and event labels of bái wă (white roof-tile; tones: R L).

To locate $F0_{turn}$, first, the afore-mentioned Praat script generated markers of individual vocal cycles using the "To PointProcess (periodic)" command in Praat and displayed the markers together with the waveform. The markers were then manually rectified for missed cycles and double markings. The script then converted the marked vocal-cycles to raw F_0 curves. Next, the C program smoothed the raw F_0 curves with a trimming algorithm (Xu, 1999) and then located the F_0 turning point between W_{start} and W_{end}.

To locate F_{turn1} and F_{turn2}, first LPC formant tracks were generated by the Praat script, using the "To Formant (burg)" command in Praat. In most cases, the default parameters for the command were used (max. number of formants = 5, maximum formant = 5500 Hz, window length = 0.025 s, time step = 25% of window length, pre-emphasis from 50 Hz). In a few cases where there were apparent tracking errors, "maximum formant (Hz)" was adjusted. The C program then located the two formant turning points using appropriate search ranges. F_{turn1} was located between W_{start} and N_{start} (or temporally hand-labeled F_{turn2}); and F_{turn2} was located between F_{turn1} and W_{end}.

The C program finally computed the following values.

F_{turn1}-to-$F0_{turn}$—time lapse from F_{turn1} to $F0_{turn}$ (= $F0_{turn}-F_{turn1}$),
$F0_{turn}$-to-N_{start}—time lapse from $F0_{turn}$ to N_{start} (= $N_{start}-F0_{turn}$),
$F0_{turn}$-to-N_{end}—time lapse from $F0_{turn}$ to N_{end} (= $N_{end}-F0_{turn}$),
$F0_{turn}$-to-F_{turn2}—time lapse from $F0_{turn}$ to F_{turn2} (= $F_{turn2}-F0_{turn}$).

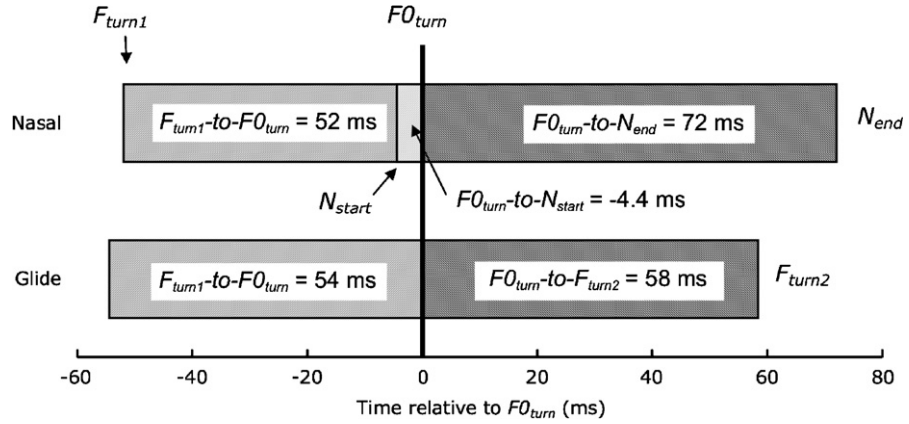


Fig. 4. Mean values of F_{turn1}-to-$F0_{turn}$, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2}, averaged across all four subjects. The F_0 turning point ($F0_{turn}$) is plotted at time 0, which serves as the reference point for all other values.

2.5. Analyses and results

Fig. 4 is a summary plot of the mean values of the measurements, including F_{turn1}-to-$F0_{turn}$, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2}. In the figure, the F_0 turning point ($F0_{turn}$) is plotted at time 0 and the other measurements plotted relative to it. Plotted this way, the time relation among the measurements provides information for determining the points in an approximant that are analogous to the onset and offset of a nasal murmur.

First, the values of F_{turn1}-to-$F0_{turn}$ were similar in the nasal group and the approximant group (nasal: 52.0 ms (SE = 4.47), approximant: 54.0 ms (SE = 3.44)). A repeated-measures ANOVA with tone (R/F) and consonant (nasal/approximant) as independent variables did not show a significant effect of consonant on F_{turn1}-to-$F0_{turn}$. The effect of tone on F_{turn1}-to-$F0_{turn}$, however, was significant ($F(1,3) = 38.8$, $p < 0.01$), with a greater value for R tone than for F tone (61 vs. 45 ms). This agrees with what was found previously (Xu, 1999). The similarity of F_{turn1}-to-$F0_{turn}$ between the nasal group and the approximant group means that, in both cases, the formant transition toward the initial consonant of the following syllable starts at about the same time relative to the F_0 turning point.

Second, as shown in the upper bar in Fig. 4, the mean value of $F0_{turn}$-to-N_{start} is negative, indicating that, on average, the F_0 turning point occurred after the onset of the nasal murmur. This agrees with previous reports on the alignment of the dynamic tones in Mandarin (Xu, 1998, 1999, 2001). The value of $F0_{turn}$-to-N_{start} 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-measures ANOVA with tone (R/F) and consonant (nasal/approximant) as independent variables showed that this difference was marginally significant ($F(1,3) = 18.6$, $p < 0.05$), indicating that the turning point occurred earlier in F tone than in R tone. This again confirms that pitch falls are faster than pitch rises.

Third, if a point in the approximants equivalent to the nasal murmur onset were to be located in Fig. 4, it would be on average 4.4 ms before the F_0 turning point. As we can see in Fig. 4 as well as in Fig. 3, this point would not be at any obvious segmental landmark. In particular, it would be about 62 ms (58 + 4 ms) ahead of F_{turn2}, where the formants have the most extreme values for the approximant.

Finally, the landmark point F_{turn2} appears to be close to, but somewhat earlier than, N_{end}, i.e., the offset of the nasal murmur in the second syllable. A repeated-measures ANOVA with tone and consonant as independent variables shows that $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2} are significantly different ($F(1,3) = 19.1$, $p < 0.05$).

2.6. Discussion

Using F_0 alignment as reference, and with direct comparisons between initial nasals and approximants, Experiment 1 revealed that first, the timings of F_{turn1} in the approximants and the nasals are virtually identical

(about 2 ms apart); second, the point in an initial approximant analogous to the onset of a nasal murmur was much earlier (about 62 ms on average) than F_{turn2}, the major landmark of the approximant; and third, F_{turn2} in an approximant was close to but slightly earlier than the offset of the nasal murmur. These results are consistent with the two main hypotheses outlined in the Introduction: (A) The onset of formant movements toward the consonantal target occurs at roughly the same time in an approximant as in a nasal. (B) The offset of formant movements toward the approximant target occurs later than the onset of a nasal murmur but earlier than the offset of the murmur.

3. Experiment 2

Experiment 2 was designed to examine whether similar alignment patterns could be found in English. Though having no lexical tones, English does use F_0 extensively to express various communicative meanings (Bolinger, 1986, 1989; Fry, 1958; Goldsmith, 1981; Ladd, 1996; Pierrehumbert, 1980; Xu & Xu, 2005). Furthermore, it has been found that certain F_0 turning points are quite consistently aligned to initial consonants in English (Ladd et al., 1999; Pierrehumbert & Steele, 1989; Xu & Xu, 2005). In particular, focus has been found to be associated with rather consistent F_0 patterns (Bolinger, 1958; Cooper, Eady, & Mueller, 1985; Xu & Xu, 2005). Experiment 2 thus uses focus-related F_0 alignment to test the two main hypotheses of the study in English by comparing the spectral alignment of approximants [j], [w] and [ɹ] with that of initial nasals using F_0 turning points as time references.

3.1. Material

Fifteen phrases were used as testing material. They were divided into six comparison sets, each consisting of a nasal phrase and an approximant phrase (which contained the glide [j] or [w] or, in the first three sets only, retroflex [ɹ]), 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 consists of a diphthong whose F2 movement is in the opposite direction of the transitional movement toward 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 consists of 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 was paired with two alternate leading questions to elicit focus (emphasis) on either the first or the second word. 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. The complete phrase list is shown in Table 2.

What's that?	My WHEEL.
Whose wheel?	MY wheel.

Table 2
English testing material used in Experiment 2

Early focus		Late focus	
Whose meal?	MY meal	What's that?	My MEAL
Whose wheel?	MY wheel	What's that?	My WHEEL
Whose reel?	MY reel	What's that?	My REEL
Whose mail?	MY mail	What's that?	My MAIL
Whose whale?	MY whale	What's that?	My WHALE
Whose rail?	MY rail	What's that?	My RAIL
Whose mike?	MY mike	What's that?	My MIKE
Whose wife?	MY wife	Who's that?	My WIFE
Whose rice?	MY rice	What's that?	My RICE
Who knew it?	YOU knew it	I what?	You KNEW it
Who uses it?	YOU use it	What should I do?	You USE it
Old name or new name?	NEW name	New name or New title?	New NAME
Old Yale or New Yale?	NEW Yale	New Yale or New Harvard?	New YALE
Old novel or new novel?	NEW novel	New novel or new movie?	New NOVEL
Old Yahoo or New Yahoo?	NEW Yahoo	New Yahoo or new Google?	New YAHOO

New novel or new movie?	New NOVEL.
Old novel or New novel?	NEW novel.

Based on acoustic data reported by Xu and Xu (2005), in a focused monosyllabic word with an initial nasal, an F_0 valley would occur around the onset of the nasal murmur, and an F_0 peak would occur around the middle of the vowel. In the AM theory of intonational phonology, these F_0 valleys and peaks have been given the phonological status of L and H* or LH* (Ladd, 1996; Pierrehumbert, 1980). Xu and Xu (2005) have found, however, that F_0 peaks and valleys do not necessarily correspond to unitary underlying units. Because the discussion of the exact nature of these peaks and valleys is beyond the scope of the present study (cf. Xu, 2005 and Xu & Xu, 2005 for extended discussion), we will refer to them simply as F_0 turning points.

3.2. Subjects

Three female and two male speakers of American English served as subjects. All of them grew up either in the Midwest or California and their dialects belong to the so-called General American variety, and none of them had noticeable regional accents. They were graduate or undergraduate students at Northwestern University. Their age ranged from 22 to 28. None of them reported any history of speech or hearing disorders.[6]

3.3. Recording

The recording location and environment were the same as in Experiment 1. 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 list was repeated nine times and each with a different random order. A different randomization was used for each subject.

[6]None of the subjects pronounced "whale" and "wheel" with a voiceless fricative.

3.4. Measurements, analyses and results

The measurements and how they were taken were the same as in Experiment 1 and so are not repeated here. Fig. 5 shows examples of the event labels in the nasal (a, d), glide (b, e) and retroflex (c, f) groups, with both early focus (a–c) and late focus (d–e).

Because the F_0 patterns are very different in the phrases with early focus and in those with late focus, the two focus conditions were analyzed separately. Fig. 6a is a summary plot of the mean values of $F0_{turn}$-to-F_{turn1}, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2} of all the phrases with early focus. In the figure, $F0_{turn}$ is plotted at time 0 and the other measurements plotted relative to it. The mean values of $F0_{turn}$-to-F_{turn1} are similar in the three consonant phrases, with the values in the glides 4 ms shorter, and those in the retroflex 9 ms longer than those of the nasals. A one-factor repeated-measures ANOVA did not show a significant consonant effect on $F0_{turn}$-to-F_{turn1}. Thus the hypothesis that the formant transition toward the initial consonant of the following syllable started at about the same time after the F_0 peak is not rejected.

In the nasal phrases, the mean value of $F0_{turn}$-to-N_{start} is about 62 ms. So, on average, a point in the approximants equivalent to the onset of the nasal murmur should have occurred about 62 ms after the F_0 peak. Looking at Fig. 6a, we can see that this inferred location is well ahead of the landmark point F_{turn2}: 50 ms earlier in the glides (112–62 ms), and 58 ms earlier in the retroflex (120–62 ms).

Next we look at the possibility, as we did in Experiment 1, that F_{turn2}, the second turning point of F2, is earlier than N_{end}, i.e., the offset of nasal murmur. On average, $F0_{turn}$-to-F_{turn2} in both the retroflex and glides is shorter than $F0_{turn}$-to-N_{end}. A one-factor repeated measures ANOVA showed a significant difference between $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2}: $F(2,8) = 8.808$, $p < 0.01$. A Student–Newman–Keuls post hoc test showed that the difference between the nasal and glide was significant (at $\alpha = 0.01$), while the difference between the nasal and the retroflex was not.

Fig. 6b is a summary plot of the mean values of F_{turn1}-to-$F0_{turn}$, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2} of all phrases with late focus. The F_0 turning point was plotted at time 0 and the other events were plotted relative to it. The valley is known to occur around the boundary between the two words (Xu & Xu, 2005). One subject (subject 5), however, did not produce any F_0 valley in most of his late-focus utterances. The plot and the subsequent analysis thus do not include data from this subject.

As can be seen in the top bar, the mean value of $F0_{turn}$-to-N_{start} is −10 ms. This negative value indicates that, on average, the F_0 valley occurred after the onset of the nasal murmur in the nasal phrase. This agrees with the findings of Xu and Xu (2005) for stressed syllables under focus. It is a bit later than what is reported by Ladd et al. (1999) for stressed syllables not under focus. This value of $F0_{turn}$-to-N_{start} suggests that the point in the approximants analogous to the nasal murmur onset should also be about 10 ms before the F_0 turning point. Looking at Fig. 6b, we can see that this inferred location is well ahead of the landmark point F_{turn2}: 51 ms earlier in the glides (41 + 10 ms), and 53 ms earlier in the retroflex (43 + 10 ms).

A potential problem in using F_0 valley as an alignment reference is that the relative location of $F0_{turn}$ varies depending on the initial consonants in word 2, as can be seen in the left half of the plot in Fig. 6b. A one-factor repeated-measures ANOVA showed the effect of consonant to be significant, $F(2,6) = 11.839$, $p < 0.01$. A Student–Newman–Keuls test showed that the difference between the nasal and glide and that between the nasal and the retroflex are both significant (at $\alpha = 0.01$ and 0.05, respectively). To understand these differences, which were not observed in Experiment 1 for Mandarin or in Fig. 6a for the early-focus phrases in English, we note that an aerodynamic factor may have played a role in the location of the F_0 valley in this case. The production of [w] and [j] involves a narrow constriction either at the lips or between the tongue blade and the hard palate. If the constriction is sufficiently narrow, i.e., when the cross-sectional area at the constriction is equal to or smaller than the cross-sectional area at the glottis, transglottal pressure is likely reduced (Stevens, 1998, pp. 35–37). Other things being equal, a reduced transglottal pressure would lead to a lower F_0 (Ladefoged, 1967; Ohala, 1978; Titze, 1989). Although the lowering was not always robust, when occurring during an interval in which F_0 is generally low, it was often enough to pull the F_0 valley toward the point where the oral constriction is the narrowest, as can be seen in Fig. 5e and f. The problem is further exacerbated by the well-known effect of

articulatory strengthening at the onset of a prosodic domain (Beckman & Edwards, 1994; de Jong, 1995; Edwards, Beckman, & Fletcher, 1991; Fougeron & Keating, 1997; Harrington, Fletcher, & Roberts, 1995; Pierrehumbert & Talkin, 1992), by which the constriction in an approximant would be tightened even further

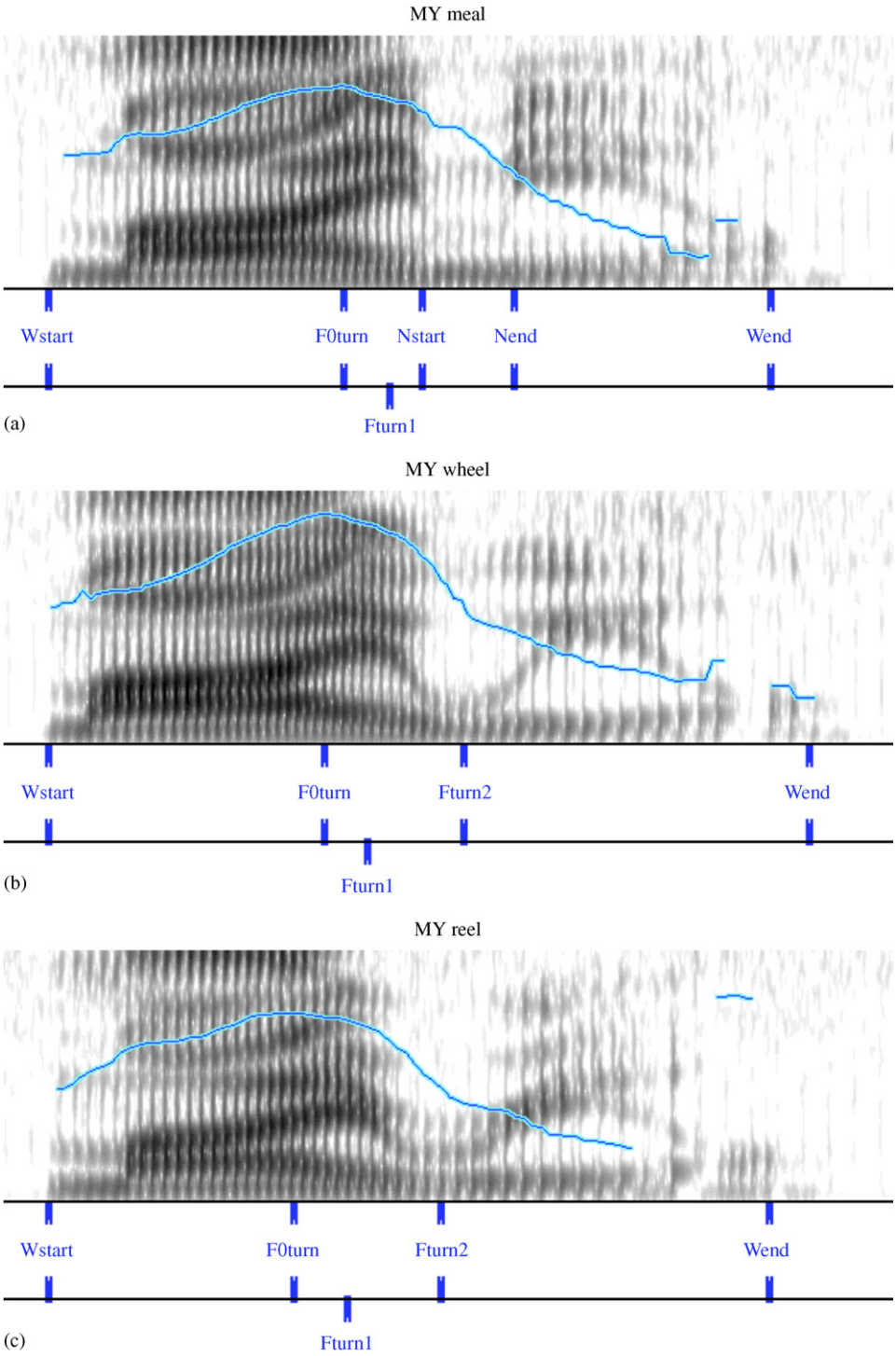
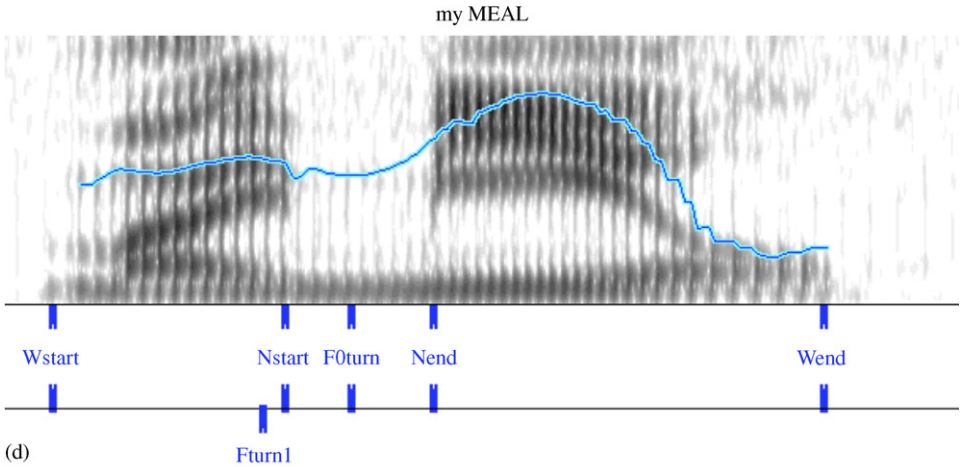
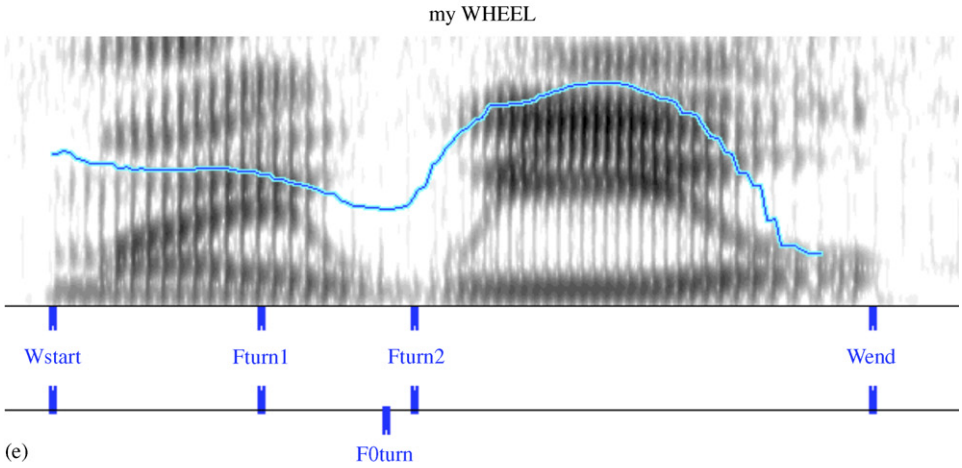


Fig. 5. Illustration of event label placement in "MY meal" (a), "MY wheel" (b), "MY reel" (c), "my MEAL" (d), "my WHEEL" (e) and "my REEL" (f). The pitch tracks (thin curves) were generated by Praat.





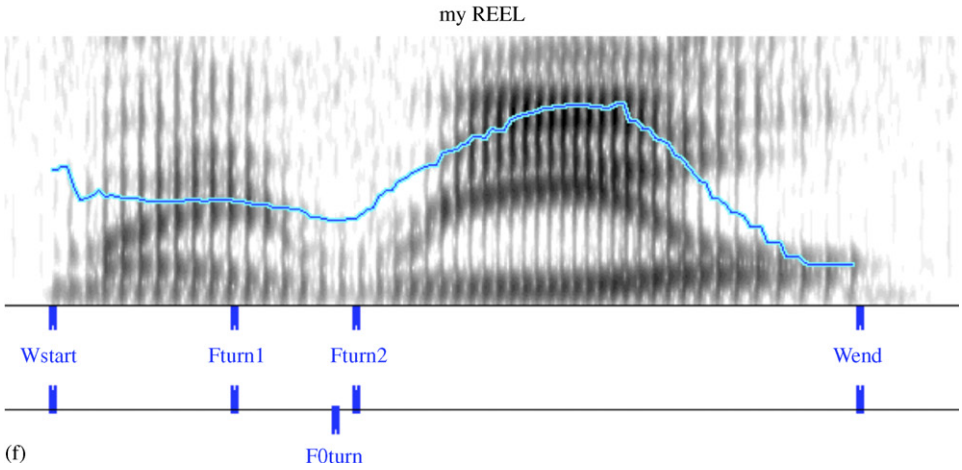


Fig. 5. (*Continued*)

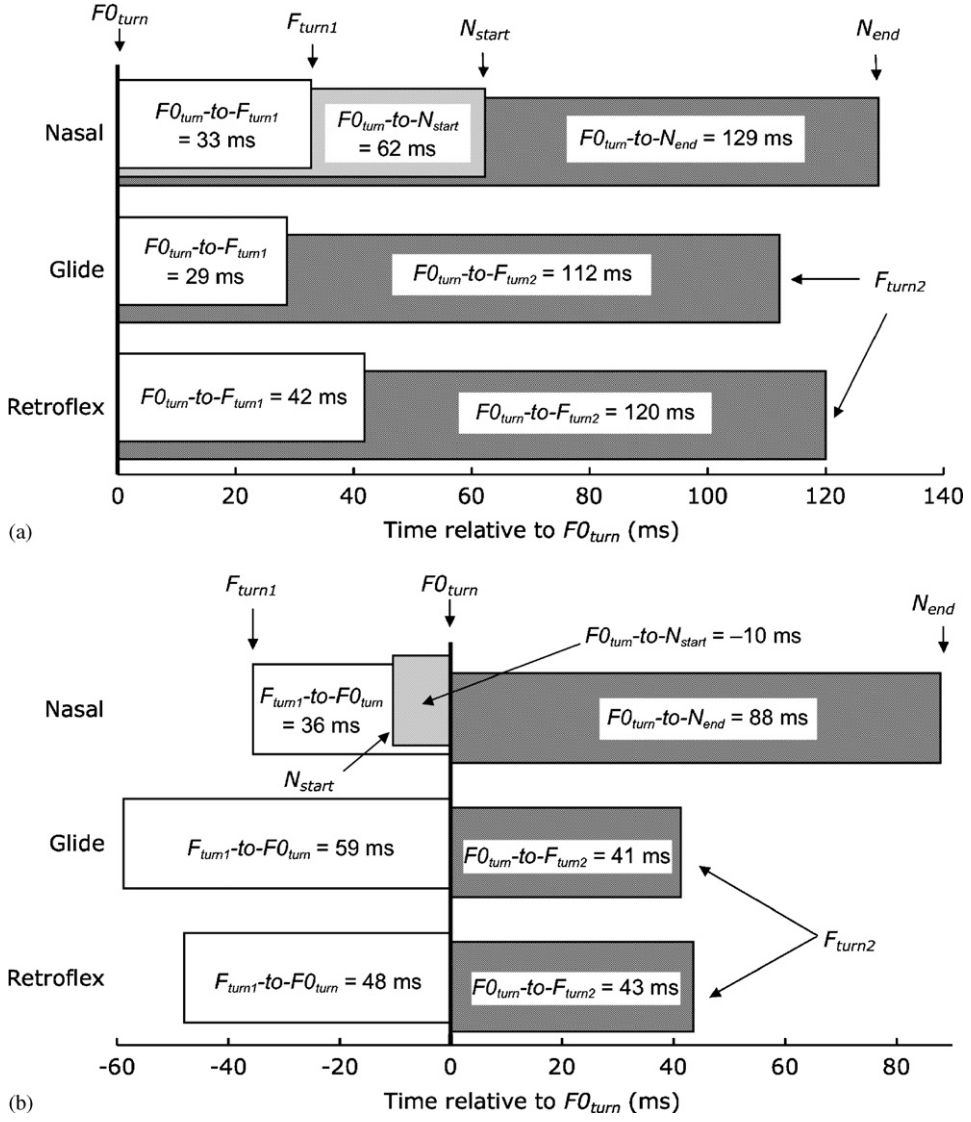


Fig. 6. (a) Mean values of $F0_{turn}$-to-F_{turn1}, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2} of all phrases with early focus, averaged across five subjects. (b) Mean values of F_{turn1}-to-$F0_{turn}$, $F0_{turn}$-to-N_{start}, $F0_{turn}$-to-N_{end} and $F0_{turn}$-to-F_{turn2} of all phrases with focus on the second word. In both graphs, the F_0 peak ($F0_{turn}$) is plotted at time 0, which serves as the reference point for all other measurements.

when the syllable is under focus. Therefore, an aerodynamic mechanism was likely competing with other mechanisms (e.g., the synchronization mechanisms proposed by Xu & Wang, 2001) in determining where the F_0 valley occurred.[7]

This conjuncture can be tested by examining the relationship between F_{turn1}-to-$F0_{turn}$ and F1, because the narrow constriction that supposedly pulls the F_0 valley toward the formant turning point should also lower F1. Thus F_{turn1}-to-$F0_{turn}$ would be partially predictable by F1: the lower the F1 minimum, the greater the value of F_{turn1}-to-$F0_{turn}$, i.e., the greater the likelihood that the F_0 valley coincides with formant extremes.

[7]The aerodynamic effect described here was not critical in Experiment 1 because there [j] and [w] were initial consonants of the second syllables in disyllabic words/phrases which are known to involve less degrees of vocal tract constrictions than those of the first syllables (Xu, 1986). The effect is also irrelevant to the case of early focus in the current experiment because there the F_0 turning point in question occurred in the vowel of word 1.

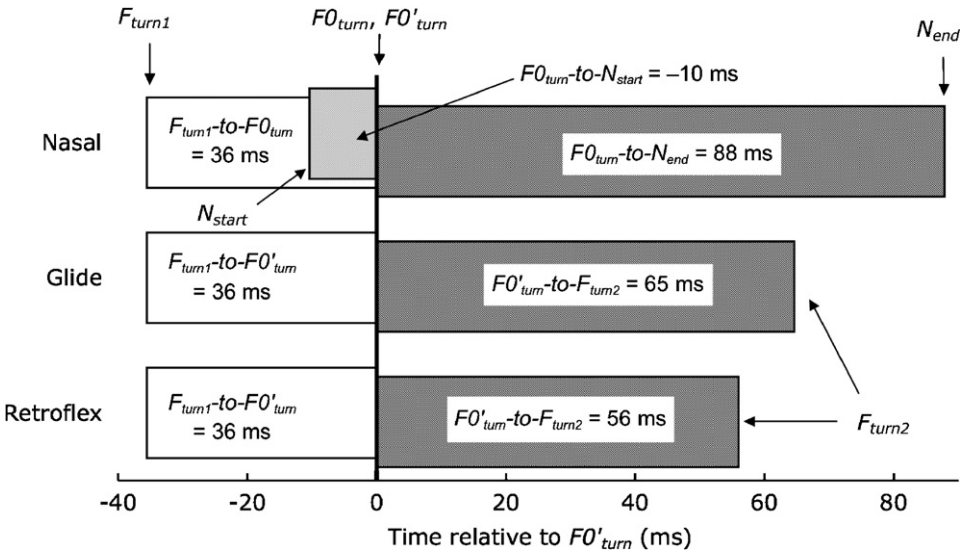


Fig. 7. Mean values of similar measurements as in Fig. 6(b), except that all the three groups are now aligned to F_{turn1} rather than $F0_{turn}$. Here $F0'_{turn}$ represents the points in the glide and retroflex that directly match $F0_{turn}$ in the nasals. In the graphs, however, the F_0 valleys ($F0_{turn}$, $F0'_{turn}$) are again plotted at time 0 for ease of comparison with Fig. 6.

A regression analysis was performed with F1 minimum (obtained from the same LPC formant tracks from which we obtained F_{turn1} and F_{turn2}, using a C program) as predictor and F_{turn1}-*to*-$F0_{turn}$ as the dependent variable. The regression result was highly significant ($F(1,287) = 37.083$, $p < 0.0001$). The R^2 was 0.115, indicating that 11.5% of the variation in F_{turn1}-*to*-$F0_{turn}$ can be predicted by F1. Note that the effect of a narrow oral constriction is only to pull the F_0 valley toward the point of formant extremes. Thus F1 is not and should not have been the sole predictor of F_{turn1}-*to*-$F0_{turn}$.

The pulling effect of the reduced transglottal pressure on the F_0 alignment actually creates a bias in favor of the second main hypothesis of the study, namely, F_{turn2} is later than N_{start} but *earlier* than N_{end}, as can be seen in the right half of Fig. 6b. One way to reduce this bias is to realign the approximants and the nasals according to F_{turn1} rather than $F0_{turn}$, as shown in Fig. 7. We may refer to the points in the approximant corresponding to $F0_{turn}$ in the nasals as $F0'_{turn}$. As can be seen, after the shift, the values of $F0'_{turn}$-*to*-F_{turn2} in the approximants (65 and 56 ms) are still shorter than that of $F0_{turn}$-*to*-N_{end} in the nasals (88 ms). A one-factor repeated-measures ANOVA was conducted and a significant effect of consonant was found ($F(2,6) = 27.36$, $p = 0.001$). A Student-Newman-Keuls post hoc test showed that $F0_{turn}$-*to*-N_{end} is significantly longer than $F0'_{turn}$-*to*-F_{turn2} in both glide and retroflex (at $\alpha = 0.01$).

3.5. Discussion

Using F_0 turning points associated with focus in English as time reference, Experiment 2 yielded two main results similar to those of Experiment 1. First, the timings of F_{turn1} in the approximants and the nasals are similar, although the similarity was somewhat affected by an aerodynamic effect in the cases of late focus. Second, the locations of formant extrema in initial approximants were later than the nasal murmur onset of initial nasals (50–58 ms in early focus, and 51–53 ms, or 66–75 ms with adjustment for aerodynamic effect, in late focus), but earlier than the nasal murmur offset (9–17 ms in early focus, and 45–47 ms, or 66–75 ms with adjustment for aerodynamic effect, in late focus). These results, similar to those of Experiment 1, are largely consistent with the two main hypotheses of the present study as outlined in the Introduction, namely, (A) the onset of formant movements toward the consonantal target occurs at roughly the same time in an approximant as in a nasal; and (B) the offset of formant movements toward the approximant target occurs later than the onset of a nasal murmur but earlier than the offset of the murmur.

4. General discussion

The issue of the temporal interval of segments is of critical importance for understanding the basic mechanisms of continuous speech. In particular, the understanding of coarticulation is contingent on knowing when a segment starts and when it ends, as without this knowledge the discussion of degrees of overlap between segments is extremely hard if not impossible. This is clearly evident in the long-drawn debate over the nature of coarticulation, as reviewed in detail by Farnetani and Recasens (1999) and Kühnert and Nolan (1999). The difficulty in defining the temporal interval of segments, though rarely recognized, has been the lack of independent time reference, which has limited the identification of the segmental boundaries to self-referencing: judging the whereabouts as well as the overlap of segments based on acoustic patterns or articulatory movements that are intrinsic to the segments themselves. It is therefore highly desirable to find a time reference that is relatively independent of the acoustic or articulatory landmarks of segments but is nevertheless based on events that are temporally closely related to segmental events.

There has been recently accumulating evidence that certain F_0 and segmental landmarks are consistently aligned with each other temporally. In particular, in both English and Mandarin, certain F_0 events, such as peaks and valleys associated with lexical tones in Mandarin (Xu, 1999) and focus in English (Xu & Xu, 2005), have been found to be consistently aligned with spectral landmarks such as the onset and offset of the nasal murmur. In the present study, we took advantages of these findings in an effort to explore the temporal interval of segments. We used the F_0-segment alignment in two experiments to test two hypotheses stemming from TA originally developed for lexical tones (Xu & Wang, 2001): (A) the onsets of formant movements toward consonant places of articulation are temporally equivalent in initial approximants and initial nasals, and (B) the offsets of formant movements toward the approximant place of articulation is later than the nasal murmur onset but earlier than the nasal murmur offset. The results of the two experiments have largely confirmed these hypotheses. This confirmation provides support, as will be explained next, for a new but more explicit definition of the temporal interval of a segment, namely, *the temporal interval of a segment is the time period during which the target of the segment is being approached*, where the target is the ideal form of the segment in terms of articulatory state and/or acoustic correlates.

Based on this definition, the temporal interval of approximants such as [j], [w] and [ɹ], which are conventionally considered to be difficult to segment in intervocalic positions, is in fact acoustically transparent. For example, in the spectrogram of "my wheel" in Fig. 8a, the formant movements can be divided into four intervals with the divisions as indicated by the arrows. During the first interval, the formants move toward a pattern that is appropriate for [ɪ], i.e., the final element of the diphthong [aɪ].[8] During the second interval, the formants move toward a pattern appropriate for [w]. During the third interval, the formants move toward a pattern appropriate for [i], although this movement seems to have reached an asymptote. And, during the fourth interval, the formants move toward a pattern appropriate for [ɬ]. Note that, conventionally, the interval between the first and second arrows would be viewed as a region of articulatory overlap between [aɪ] and [w], and the interval after the last arrow would be viewed as a region of overlap between [i] and [ɬ].

According to the new definition of the temporal interval of segment, which differs from the conventional views discussed in the Introduction, the four intervals are four contiguous but discrete time periods during which the ideal targets of [aɪ], [w], [i] and [ɬ] are approached one after another, *without overlap*.[9] In other

[8]In fact, the underlying target being approached here is likely to be intrinsically dynamic, i.e., one that is similar in nature to dynamic tonal targets such as [rise] and [fall] proposed in Xu and Wang (2001). The dynamic nature of the underlying target of [aɪ] can be seen if we compare the different renditions of "my" in Fig. 5. When under focus (a–c), the duration of the diphthong [aɪ] is much longer than when it is not under focus (e–f). However, with longer duration, it is the relatively steady-state portion of the diphthong that is lengthened, whereas the most dynamic portion is kept near the end of the syllable. This delay in the dynamic portion of a trajectory resembles what has been found in R tone in Mandarin (Xu, 1998). That is, as syllable duration gets longer due to reduced speech rate, the onset of the rise in R tone is delayed, as if to guarantee that the most dynamic portion of the F_0 contour occurs by the end of the syllable. This is despite the fact that even at slow speech rate, F_0 starts its initial drop in R tone in a H–R sequence immediately after the syllable boundary, indicating that the synchrony between tone and syllable is maintained regardless of speech rate.

[9]As will be discussed two paragraphs later, theoretically [w] should in fact be totally overlapped with [i]. However, the overlap is prevented in this case because the two adjacent targets require conflicting movements of the same articulators (both tongue blade and

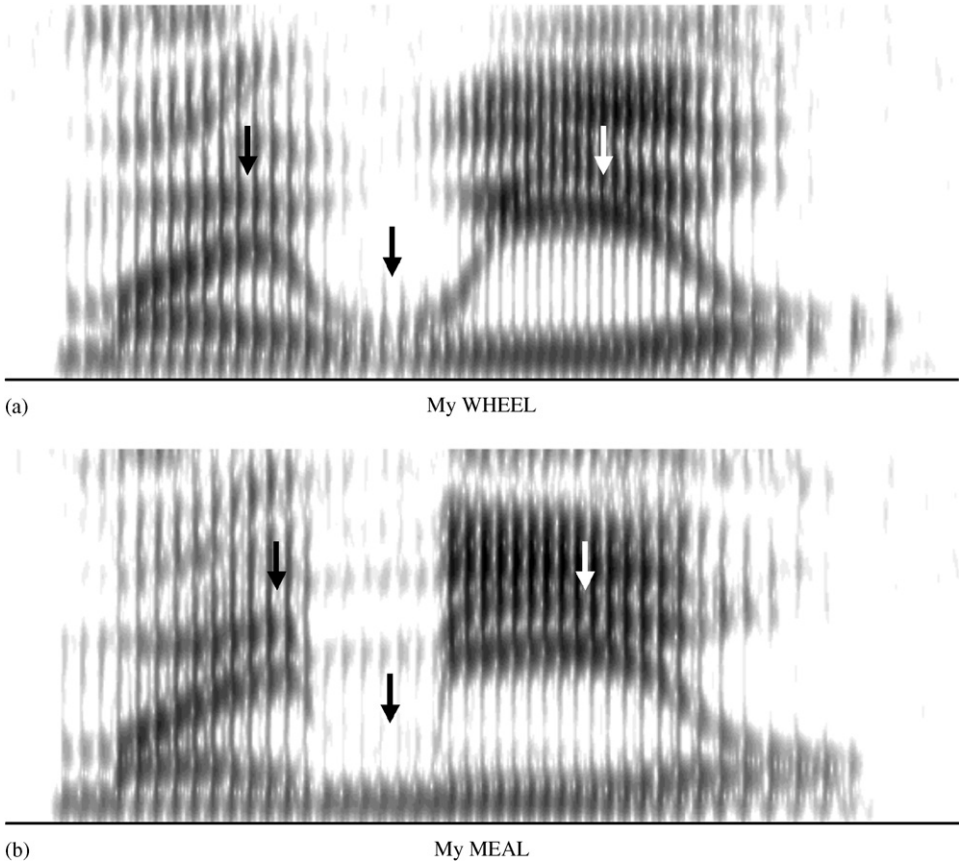


Fig. 8. Spectrograms of "my WHEEL" (a) and "my MEAL" (b). The arrows mark the points where presumably the articulatory movement toward one target terminates and that toward the next starts.

words, whenever an articulator or a group of articulators working in synergy are involved in making conflicting movements for successive sounds, be it consonants, vowels or tones, there does not have to be temporal overlap of articulatory executions by the same articulator. The involved articulator simply finishes one task before taking on the next. In fact, we are not the first to suggest this view. Bell-Berti (1993) has argued that each segment (be it consonant or vowel) has a specified velum position and the velum movement toward any particular position does not start until the movement toward the previous segment is terminated. This understanding is also relevant for the nasal sequence in Fig. 8b. There the point indicated by the first arrow is where the oral cavity starts to change its shape toward one that is appropriate for the bilabial nasal [m]. But the movement toward the air-tight labial closure is not completed at the nasal murmur onset. Rather, it is still at a high velocity at that moment, as found by Löfqvist and Gracco (1999) for [b] and [p]. Löfqvist (2002) has argued that the articulatory targets for the stops lie beyond the positions of contact between the opposing articulators so as to guarantee an air-tight seal. Thus, just as the articulatory movement toward the tightest constriction of a stop is not achieved until well after the onset of the stop closure, the tightest constriction of a nasal is also not achieved until well after the onset of the nasal murmur.

Furthermore, the nasal murmur offset is not really the end of the movement toward the articulatory goal of a nasal. Rather, that movement is terminated sometime *before* the offset of the nasal murmur, i.e., at the moment when the oral closure becomes the tightest, as found by Westbury and Hashi (1997) for nasals and Löfqvist and Gracco (1999) for stops, and further confirmed by the present results. Thus the end of the nasal

(*footnote continued*)
tongue dorsum). As found by Wood (1996), if any part of the consonantal movement directly conflicts with that of the following vowel, the conflicting parts of the movements are sequenced rather than blended into a single compromised movement.

murmur is actually the moment when the articulatory movement toward the following vowel, after going on for a short while, has just resulted in the parting of the lips (in the case of [m]) or the release of the full contact between the tongue tip and the alveolar ridge (in the case of [n]). In other words, just as the onset of the nasal murmur should no longer be considered as the onset of the nasal consonant, the offset of the nasal murmur should no longer be considered as the offset of the nasal consonant.

Perhaps the most important implication of this new understanding is that the amount of articulatory overlap a theory of speech production needs to assume could be significantly reduced. First, since the articulatory movement *away from* a target is no longer considered as part of the gesture for realizing the segment, none of the "carryover coarticulation" needs to be understood as due to an intended articulatory overlap. Second, the V-to-C formant transitions in a V#C (# = syllable boundary) sequence no longer need to be understood as due to an overlap between the vowel and the consonant, because they are part of the temporal interval of the consonant rather than part of the vowel. The movement toward the vowel target has ended when the formants, hence the underlying articulation that generates them, start to move toward the consonant. This understanding thus contrasts with other models in which many linguistically meaningful motor events are assumed to be bidirectional, i.e., consisting of both onset and release, or movements both to and from the target (Browman & Goldstein, 1986, 1989, 1992; Fujisaki, Wang, Ohno, & Gu, 2005; Goldstein & Fowler, 2003; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989; Steriade, 1993; van Santen & Möbius, 2000).

Furthermore, based on the classic findings of coarticulation (Kozhevnikov & Chistovich, 1965; Menzerath & de Lacerda, 1933, as cited by Kühnert & Nolan, 1999; Öhman, 1966), the V-to-C transition in a V_1#CV_2 sequence is not only toward the underlying target of C, but also toward that of V_2. So, the articulation of the initial consonant and the following vowel is concurrent till the moment when the tightest consonant closure is reached. After that, only the vowel articulation continues, which then terminates when the formants change directions again, e.g., at the third arrow in Fig. 8a and b, and turn toward the following coda consonant or the initial consonant as well as the first vowel of the next syllable. Again, the temporal interval of the vowel should have ended by the time of the third directional change in formant movements (third arrow).

A further implication of the new definition of the temporal interval of segment is that when listeners hear the second vowel in a V_1#CV_2 sequence "during" V_1, as found in many perception studies (e.g., Fowler, 1984; Martin & Bunnell, 1982), what they have heard is actually the spectral movement toward both C and V_2. That is, when the formants are moving in the direction of V2 well before the C closure, the temporal interval is already that of V2 rather than V1. Hence they are not hearing acoustic cues from anticipatory coarticulation of V_2 with V_1, but from the articulation of V_2 itself. The only caveat about this interpretation is that the perceptual studies of anticipatory coarticulation typically measure the temporal scope of perceptual anticipation in proportional time rather than real time (e.g., Fowler, 1984; Magen, 1997; Martin & Bunnell, 1982). It thus remains an open question as to whether listeners can actually hear the "anticipatory coarticulation" much earlier than the start of the V-to-C transition.

Finally, there has been some evidence that the movement away from a segment may not always provide highly useful perceptual information about the segment. For example, van Son and Pols (1999) showed that, first, the identification rate of a vowel is fairly low when listeners heard 50 ms of acoustic signal around the point where the formants are closest to the vowel target, second, inclusion of part of the acoustic signal *after* the formant turning point did not improve the vowel identification, while including the part of the C-to-V transition did improve it. This seems to suggest that the newly defined temporal interval of a segment is also where highly relevant perceptual information about the segment is located.[10]

The only contextual variations that cannot be fully explained away is the kind of long-distance vowel-to-vowel coarticulation through intervening [b] (which presumably has no tongue shape specification) as reported by Magen (1997) and Whalen (1990). But calculation by Fowler and Saltzman (1993) has shown that the actual distance of anticipation in Magen's (1997) data is not very different from that reported by Bell-Berti

[10]This may not be the case with consonants, for which transitions into the following vowels are known to provide highly relevant information (e.g., Cooper, Delattre, Liberman, Borst, & Gerstman, 1952 and many later studies). This issue therefore needs to be further investigated in future research.

and Harris (1981) and Bell-Berti and Krakow (1991). Still, this is an issue that needs to be explored in future research with more accurate timing control than has been applied so far.

5. Conclusion

We have argued in the present paper that there is a need to reconsider the conventional, though largely implicit, definition of the temporal interval of segments, which refers mainly to the acoustic *consequences* of articulation, especially those involving abrupt spectral shifts, i.e., landmarks. As an alternative, we have proposed that the temporal interval of a segment be defined as *the time period during which the target of the segment is being approached*, where the target is the ideal form of the segment in terms of articulatory state and/or acoustic correlates. The validity of the alternative definition was supported by the results of two experiments that tested two hypotheses derived from the definition: (A) the onsets of formant movements toward consonant places of articulation are temporally equivalent in initial approximants and initial nasals, and (B) the offsets of formant movements toward the approximant place of articulation is later than the nasal murmur onset but earlier than the nasal murmur offset.

The new and explicit definition of the temporal interval of segments is important for the understanding of coarticulation. If defined as the overlap of the temporal interval of neighboring segments, coarticulation occurs definitively only between an initial consonant and the following vowel in a CV syllable. Other than that, there is no carryover coarticulation of any kind, as the articulatory movement *away from* the target of a segment is by definition outside of the temporal interval of the segment, and there is no anticipatory coarticulation of either C or V_2 with V_1 in a V_1#CV_2 sequence, as the formant transitions toward C and V_2, by definition, occur after rather than during V_1. Note that these new understandings may be further extended to issues relating to the general temporal organization of speech at the syllable level, as has been explored by Xu and Liu (in press). Discussion of those issues, however, is beyond the scope of the present paper.

Finally, the findings of the present study demonstrate that F_0 events, being relatively independent of the segmental events, yet likely governed by similar organizational principles that govern segments, can be widely used in future research as convenient time references for understanding the temporal relations among different articulatory and acoustic events. A caveat in this regard, however, is that whether and where an F_0 turning point occurs in a syllable is jointly determined by the pitch target associated with the lexical tone or lexical stress and the pitch targets in the surrounding syllables (Xu, 1998, 1999, 2001; Xu & Xu, 2005). Therefore, the consistent F_0-segment alignment utilized in the present study happens strictly at the level of the syllable. As such it does not directly index any higher level temporal process such as rhythm. Nevertheless, the consistency in the alignment patterns may be used to assure the reliability of temporal measurements for higher level processes, provided that all the local factors known to attribute to the alignment are properly controlled.

Acknowledgements

This work is supported in part by NIH Grant DC03902. Part of the results of Experiment 1 was presented at the Seventh International Conference on Spoken Language Processing, Denver, and part of the results of Experiment 2 was presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain. We owe our thanks to Ken de Jong and two anonymous reviewers for insightful comments that have helped sharpen our view. We also thank Volker Dellwo for helpful comments on an earlier version of the manuscript.

References

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics*, *36*, 3–25.

Atterer, M., & Ladd, D. R. (2004). On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. *Journal of Phonetics*, *32*, 177–197.

Beckman, M. E., & Edwards, J. R. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7–33). Cambridge: Cambridge University Press.

Bell-Berti, F. (1993). Understanding velic motor control: Studies of segmental context. In M. K. Huffman, & R. A. Krakow (Eds.), *Nasals, nasalization, and the velum* (pp. 63–85). San Diego: Academic Press.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica*, *38*, 9–20.

Bell-Berti, F., & Krakow, R. A. (1991). Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America*, *90*, 112–123.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9(10)), 341–345.

Bolinger, D. (1958). A theory of pitch accent in English. *Word*, *14*, 109–149.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Palo Alto: Stanford University Press.

Bolinger, D. (1989). *Intonation and its uses—Melody in grammar and discourse*. Stanford, California: Stanford University Press.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155–180.

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*, 37–47.

Caspers, J., & van Heuven, V. J. (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica*, *50*, 161–171.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech—Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica*, *63*, 47–75.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*, 597–606.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *Journal of the Acoustical Society of America*, *77*, 2142–2156.

Crystal, D. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, *72*, 705–716.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, *83*, 1574–1585.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, *88*, 101–112.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*, 491–504.

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. *Speech Communication*, *33*, 339–356.

Edwards, J. R., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, *89*, 369–382.

Fant, G. (1973). *Speech sounds and features*. Cambridge: MIT Press.

Farnetani, E., & Recasens, D. (1999). Coarticulation models in recent speech production theories. In W. J. Hardcastle, & N. Hewlett (Eds.), *Coarticulation: Theory, data and techniques* (pp. 31–65). Cambridge: Cambridge University Press.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*, 113–133.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, *36*, 359–368.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1730–1741.

Fowler, C. A., & Saltzman, E. L. (1993). Coordination and coarticulation in speech production. *Language and Speech*, *36*, 171–195.

Frota, S. (2002). Tonal association and target alignment in European Portuguese nuclear falls. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology VII* (pp. 387–418). Berlin: Mouton de Gruyter.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

Fujisaki, H., Wang, C., Ohno, S., & Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech Communication*, *47*, 59–70.

Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, *44*, 1570–1573.

Goldsmith, J. A. (1981). English as a tone language. In D. Goyvaerts (Ed.), *Phonology in the 1980's* (pp. 287–308). Ghent: Story-Scientia.

Goldstein, L. M., & Fowler, C. (2003). Articulatory phonology: A phonology for public language use. In A. S. Meyer, & N. O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*. Berlin: Mouton de Gruyter.

Grabe, E. (1998). Pitch accent realization in English and German. *Journal of Phonetics*, *26*, 129–143.

Grabe, E., Post, B., Nolan, F., & Farrar, K. (2000). Pitch accent realization in four varieties of British English. *Journal of Phonetics*, *28*, 161–185.

Hardcastle, W. J., & Hewlett, N. (Eds.). (1999). *Coarticulation: Theory, data and techniques*. Cambridge: Cambridge University Press.

Harrington, J., Fletcher, J., & Roberts, C. (1995). Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics*, *23*, 305–322.

Joos, M. (1948). Acoustic phonetics. *Journal of the Acoustical Society of America*, *24*(Suppl.), 1–136.

Kent, R., & Minifie, F. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, *5*, 115–133.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant/vowel syllables. *Journal of the Acoustical Society of America*, *72*, 379–389.

Kim, S.-A. (1999). Positional effect on tonal alternation in Chichewa: Phonological rule vs. phonetic timing. In *Proceedings of annual meeting of Chicago linguistic society*, vol. 34, Chicago, pp. 245–257.

Kozhevnikov, V. A., & Chistovich, L. A. (1965). *Speech: Articulation and perception*. Washington, DC: Joint Publications Research Service.

Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. In W. J. Hardcastle, & N. Hewlett (Eds.), *Coarticulation: Theory, data and techniques* (pp. 7–30). Cambridge: Cambridge University Press.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *Journal of the Acoustical Society of America*, *106*, 1543–1554.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America*, *107*, 2685–2696.

Ladefoged, P. (1967). *Three areas of experimental phonetics*. Oxford: Oxford University Press.

Lehiste, I., & Peterson, G. (1961). Transitions, glides and diphthongs. *Journal of the Acoustical Society of America*, *33*, 268–277.

Löfqvist, A. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America*, *111*, 2811–2827.

Löfqvist, A., & Gracco, L. (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America*, *105*, 1864–1876.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *25*, 187–205.

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel–stop consonant–vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 473–488.

Menzerath, P., & de Lacerda, A. (1933). *Koartikulation, Seuerung und Lautabgrenzung*. Berlin, Bonn: Fred. Dummlers.

Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 5–39). New York: Academic Press.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *JASA*, *41*, 310–320.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*, 693–703.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD dissertation. Cambridge, MA: MIT [published in 1987 by Indiana University Linguistics Club, Bloomington].

Pierrehumbert, J., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, *46*, 181–196.

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. J. Docherty, & R. Ladd (Eds.), *Papers in laboratory phonology II: Gestures, segments, prosody* (pp. 90–117). Cambridge: Cambridge University Press.

Prieto, P., van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, *23*, 429–451.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review*, *94*, 84–106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333–382.

Schepman, A., Lickley, R., & Ladd, D. R. (2006). Effects of vowel length and "right context" on the alignment of Dutch nuclear accents. *Journal of Phonetics*, *34*, 1–28.

Silverman, K. E. A., & Pierrehumbert, J. B. (1990). The timing of prenuclear high accents in English. In J. Kingston, & M. E. Beckman (Eds.), *Papers in laboratory phonology 1—Between the grammar and physics of speech* (pp. 72–106). Cambridge: Cambridge University Press.

Steriade, D. (1993). Closure, release and nasal contours. In M. Huffman, & R. Krakow (Eds.), *Nasals, nasalization and the velum: Phonetics and phonology* (pp. 401–470). New York: Academic Press.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: The MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *Journal of the Acoustical Society of America*, *85*, 901–906.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 1–28). Berlin, New York: De Gruyter.

van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*, 513–546.

van Santen, J.P.H., & Möbius, B. (2000). A quantitative model of f0 generation and alignment. In: A. Botinis (Ed.), *Intonation: Analysis, modelling and technology* (pp. 269–288). Kluwer Academic Publishers.

van Son, R. J. J. H., & Pols, L. C. W. (1999). Perisegmental speech improves consonant and vowel identification. *Speech Communication*, *29*, 1–22.

Westbury, J., & Hashi, M. (1997). Lip–pellet positions during vowels and labial consonants. *Journal of Phonetics*, *25*, 405–419.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3–35.

Wood, S. A. J. (1996). Assimilation or coarticulation? Evidence from the temporal co-ordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics*, *24*, 139–164.

Xu, C. X., & Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association*, *33*, 165–181.

Xu, Y. (1986). Acoustic–phonetic characteristics of junctures in Mandarin Chinese. *Zhongguo Yuwen [Journal of Chinese Linguistics]*, 353–360 (in Chinese).
Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61–83.
Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, *55*, 179–203.
Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*, 55–105.
Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica*, *58*, 26–52.
Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*, 220–251.
Xu, Y., & Liu, F. Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics*, in press, to appear.
Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, *111*, 1399–1413.
Xu, Y., & Wallace, A. (2004). Multiple effects of consonant manner of articulation and intonation type on F0 in English. *Journal of the Acoustical Society of America*, *115*(Pt. 2), 2397.
Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*, 319–337.
Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159–197.