ABSTRACT

This paper discusses various sources of tonal variations in connected speech. It is argued that these sources are better understood when they are viewed as either voluntary or involuntary. Voluntary sources are those stemming from linguistic/paralinguistic demands, and involuntary sources from articulatory constraints. Linguistic/paralinguistic demands represent various communicative functions on the one hand, but are associated with articulation-specific pitch targets and pitch ranges on the other. These pitch targets and pitch ranges are what speakers actually intend to implement in their speech; but such implementation is constrained by the limitations of the articulators that actually produce the fundamental frequency of voice. Observed variations in F₀ contours in connected speech thus reflect different levels of linguistic/paralinguistic demands as well as their interaction with various articulatory constraints.

1. INTRODUCTION

An important goal of speech research is to find regularities in the seemingly highly variable acoustic manifestations of various linguistic functions in speech. For a tone language like Mandarin, for example, a major research objective is to understand how lexical tones are realized in utterances so that they can be reliably conveyed to the listener. This is by no means a trivial task, because it is well known that F₀ contours of tones vary extensively in connected speech (Chao 1968; Howie 1974; Kratochville 1985;
Wu 1990; Xu 1994). A seemingly straightforward approach to explaining the observed variability is to treat the variation as phonologization (or conventionalization) either explicitly or implicitly. Chao (1968), for example, states that $R$ ( = Rising tone = Tone 2; also $H$ = High Tone = Tone 1, $L$ = Low Tone = Tone 3, and $F$ = Falling Tone = Tone 4) often changes into $H$ when it is on the middle syllable of a trisyllable word in which the first syllable has $H$ or $R$ (e.g., in cōngyōubǐng [green onion pancake]. Wu (1990) regards polysyllabic tone sandhi patterns as the invariant units of intonation, thus in effect treating them as conventionalized pitch patterns. However, treating the conditional variation as phonologization or conventionalization assumes that the production of the variant form is deliberate on the part of the speaker. As found by Shih & Sproat (1992) and Xu (1994), despite substantial deviation from its citation form, the second $R$ in words like “cōngyōubǐng” does not really change into $H$. And, as shown by Xu (1997, 1999a; Xu & Wang, 2001), many seemingly stylized $F_0$ patterns in disyllabic words in Mandarin can be actually derived from the underlying form of individual tones by attributing the variability to articulatory transitions between adjacent tones. It therefore seems that these variant forms may not have been produced intentionally. Rather, they are probably produced due to certain articulatory pressure while the underlying phonetic targets in fact remain constant. Thus it appears that there is a difference between underlying targets that are actually intended by the speaker and articulatory constraints that shape the final realization of the targets in speech production. We may refer to the underlying targets as voluntary factors in speech production and the articulatory constraints as involuntary ones, to borrow the terms from ’t Hart, Collier, & Cohen (1990) (p. 39). As will be argued next, a distinction between the voluntary and involuntary factors in speech is not only possible but also necessary if we are to improve our understanding of invariance and variability in the $F_0$ contours of speech.

1.1 Dual Source of Phonetic Variations

Speech production is a voluntary act: the speaker voluntarily produces utterances in order to convey information to the listener. The information in an utterance is carried by various linguistic as well as
paralinguistic structures. To implement these structures, the speaker needs to produce the utterance in such a way that the underlying specifications of these structures are fulfilled. The act of fulfilling these linguistic/paralinguistic specifications when producing the utterance is thus deliberate and voluntary.

A voluntary act is not necessarily one that the speaker is consciously aware of. In fact, speakers most likely are not aware of the details of how various linguistic/paralinguistic information is conveyed. A speech act is voluntary only in the sense that it is *initiated* by the speaker.

Speech production is executed with the human vocal apparatus, which consists of the lungs, the larynx, and the vocal tract. Being a physical device, the vocal apparatus has many limitations. For example, none of the articulators can change its state instantaneously, and all articulators have limits as to how much and how fast they can be stretched, compressed, or displaced. These limits are not set by the speaker, nor are they specified in the information structure of the speech utterances. Rather, they are inherent to the articulators of the speaker. The effects of the articulatory limitations on speech production are therefore unintentional and *involuntary*.

It therefore seems that there are two distinct sources of phonetic variations — voluntary speech acts and involuntary articulatory constraints. They are orthogonal to each other in the sense that they are generated by quite different forces. Voluntary acts are brought forth by various linguistic/paralinguistic demands, while involuntary constraints by inherent properties of the articulators. Despite their mutual independence, however, the two sources interact in speech production, and they jointly determine the acoustic properties of speech utterances.

In the next section, this distinction will be applied to the analysis of tonal variations in Mandarin. Various factors that may contribute to the generation of $F_0$ contours will be recognized and classified into voluntary and involuntary ones.

1.2 Possible Sources of Tonal Variations in Mandarin

Various factors may contribute to the variations in the realization of lexical tones in Mandarin, and many of them have been considered by
previous studies (Howie 1974; Kratochville 1985; Shi & Zhang 1987; Wu 1990; Lin & Yan 1991; Xu 1994, 1997, 1998, 1999a). In the following, factors that may potentially affect F₀ contours in Mandarin are classified into two categories — voluntary and involuntary. The two lists are by no means complete, but they reflect our current knowledge about F₀ contour generation in Mandarin.

1.2.1 Voluntary Factors — Linguistic/Paralinguistic Demands

As mentioned above, voluntary sources of phonetic variations are generated by the need to convey communicative information. In Mandarin, the communicative information that can be expressed through F₀ patterns may be shown in the following list.

Lexical tone;
Prosodic structure;
Syntax;
Pragmatics;
Emotion.

Lexical tone. It has been generally agreed for a long time that tones in Mandarin are expressed mostly through F₀ contours (Liu 1924; Chao 1930, 1956, 1968; Bai 1934). This implies that much of the F₀ variation in a Mandarin utterance is attributable to lexical tones. This has been quantitatively shown to be the case recently (Xu 1999a).

Prosodic structure. Included here are the rhythmical and accentual structures of an utterance. They may have various impacts on the realization of tones. First, certain phonological tone-sandhi in Mandarin have been found to be governed by prosodic structure (Shih 1986). Second, certain phonetic tonal variations may also be related to prosodic structure of the utterance (Shih & Sproat 1992; Xu 1994).

Syntax. Although it has been commonly assumed that syntax has a direct role in intonation, few syntactic functions have been demonstrated definitively to play such a role (Shattuck-hufnagel & Turk 1996). Even the well-known question/statement dichotomy has been found not to be the ultimate determinant of final rise/fall in intonation (Bolinger 1989). Also as argued by Shih (1986), many conditional tonal variations are determined by
the prosodic rather than syntactic structure of an utterance. Nonetheless, some syntactic function may still be correlated with certain $F_0$ variations (Shattuck-hufnagel & Turk 1996). For Mandarin, Shih (1999) has found that verbs in general have lower average $F_0$ than other parts of speech in a sentence.

**Pragmatics.** The scope of pragmatics is at present far from being well defined (cf. Crystal 1997 for a detailed discussion). Nonetheless, in general pragmatics is concerned with discourse structure, conversational implicatures, suppositions, etc. (Crystal 1997). What is most relevant to $F_0$ analysis is the speaker’s choice of pitch patterns independent of lexical tone, prosody, syntax and possibly emotion. Such choices may be based on the speaker’s intention in information exchange as well as his/her evaluation of the hearer’s knowledge of the information. Based on such considerations, focus (also known as focal prominence, contrastive stress, emphatic stress, sentence-level stress, etc.) is probably controlled by pragmatics, though it has been in general treated as part of prosody (Shattuck-Hufnagel & Turk 1996), because the speaker’s choice of whether and where to put focus in an utterance is mostly based on the assessment of the information flow in the discourse (Bolinger 1989; van Heuven 1994). As will be shown, the influence of focus on $F_0$ contours in speech can be quite substantial. Similar to focus, topic initiation (also known as new topic, topic shift, etc.) is also likely to be controlled by pragmatics and its influence on $F_0$ has also been found to be quite extensive (Umeda 1982; Nakajima & Allen 1993).

**Emotion.** Of all the voluntary factors, emotion (also known as affect, attitude, etc.) is probably what we know the least. Although there have been a number of studies on the effect of emotion on $F_0$, (e.g., Grabe, Gussenhoven, Haan, Marsi, & Post 1997; Ladd, Silverman, Tolkmitt, Bergmann & Scherer 1985; O'Shaughnessy & Allen 1983; Protopapas & Lieberman 1997; Ross, Edmonson & Seibert 1986; Ross, Edmondson, Seibert & Chan 1992), findings about how emotion and attitude influence $F_0$ in speech are both vague and inconclusive. Nevertheless, there is little doubt that at least some of the emotional information is carried by $F_0$.

From the above brief discussion, we can see that although much is still unknown about the exact contribution of each of the voluntary factors to
the F₀ contours in Mandarin speech, the richness of the information they represent is self-evident. One would probably wonder how it is possible for F₀ contours, which is essentially one-dimensional, to carry so much information. To make things more complicated, all this is done with an articulatory system that has many inherent limitations.

1.2.2 Involuntary Factors — Articulatory Constraints

The articulatory limitations inherent to the F₀ production apparatus are due directly to specific properties of the individual muscles, bones, cartilages, and tissues involved in pitch production, such as their mass, stiffness, stretchability, and compressibility. However, as has been found in articulation studies (Abbs 1986 and many others), activities of individual muscles or articulators do not always correspond well with observed acoustic patterns. Instead, what corresponds better to the acoustic patterns is the combination of articulator movements (Abbs 1986). For the articulatory limits on F₀ production, therefore, we need to consider constraints that are due to the combined limitations of all involved anatomical structures, as shown in the following (probably still incomplete) list.

1. Overall pitch range of a speaker;
2. Vowel intrinsic pitch;
3. F₀ perturbation by consonants;
4. Maximum speed of pitch change;
5. Maximum speed of pitch direction shift;
6. Coordination of laryngeal and supralaryngeal movements;

*Overall pitch range of a speaker.* According to Fairbanks (1959), a speaker's conversational pitch range can span as much as two octaves. This means, for example, a male speaker with a minimum F₀ of 80 Hz may have a maximum F₀ of 320 Hz. Although few tone or intonation studies have specifically looked into it, the overall pitch range of speakers may conceivably be a potential limitation.

*Vowel intrinsic pitch.* As has been well established, other things being equal, different vowels are produced with different F₀ values (cf. Whalen & Levitt 1995 for a comprehensive summary). Such variations are mostly related to vowel height, with higher vowels having higher intrinsic pitch than
low vowels. Intrinsic pitch variations have also been found in Mandarin, i.e., the same tones are produced with different F₀ values under equal conditions (Shi & Zhang 1987). It has also been found, however, that intrinsic pitch differences tend to become smaller in connected speech (Ladd & Silverman 1984).

\textit{F₀ perturbation by consonants.} Consonants may affect the F₀ contours of both preceding and following vowels. It is well known that voiceless consonants may raise the F₀ of the following vowel, and certain voiced consonant may lower the F₀ of the following vowel, but both effects are temporally quite local (Lehiste & Peterson 1961; Howie 1974; Hombert 1978; Rose 1988). On the other hand, however, consonantal effect on the F₀ of preceding vowels is not well understood, although there have been report that there is no effect at all (van Santen & Hirschberg 1994). But informal observation has revealed certain very local lowering effects of stop consonants on the F₀ of the preceding vowel in Mandarin (Xu 1996). Shih (1999) has also noticed a similar effect.

\textit{Maximum speed of pitch change.} As reported by Ohala & Ewan (1973) and Sundberg (1979), when asked to change pitch by 6 semitones in the shortest amount of time possible, speakers (who are not professional singers) need at least 80-90 ms (females being a bit faster) to complete 75% (i.e., the fastest central portion) of the change when raising pitch and 70-75 ms when lowering pitch.

\textit{Maximum speed of pitch direction shift.} Not only is there a limit on how fast speakers can change pitch in one direction, i.e., raising or lowering (as reported by Ohala & Ewan 1973 and Sundberg 1979), but also there should be a limit on how fast they can shift pitch movement direction, i.e., to change from raising F₀ to lowering it, or the reverse. This is because, when produced by a physical system like the larynx, a pitch turn requires both deceleration and acceleration of the pitch movement. So, the time needed for producing a rise-fall contour should be more than the minimum pitch-raising time plus the minimum pitch-lowering as reported by Ohala & Ewan (1973) and Sundberg (1979). This has been found to be indeed the case in a recent study in which both English and Mandarin speakers were asked to shift pitch up and down as quickly as possible (Xu & Sun, 2000). Regardless of the
magnitude of the pitch shift, the time the speakers used to complete 100% of the pitch change was found to be roughly twice as long as the time they used to complete the middle 75% of the pitch change.

*Coordination of laryngeal and superlaryngeal movements.* Kelso (1984) asked human subjects to perform a simple task of wagging two fingers (one in each hand) together. At low speed, they could start the movement cycles of the two fingers either simultaneously, i.e., with 0° phase shift, or with one finger starting earlier than the other by half a cycle, i.e., with a 180° phase shift. At a high speed, however, they could move the two fingers together only with 0° phase shift. Schmidt, Carello & Turvey (1990) further found that the same happened when two people were asked to oscillate their legs while watching each other’s movement. Based on such findings, these authors suggest that (a) there is a deep-rooted biological tendency to coordinate one’s movement with the environment whenever pertinent, regardless of whether the environment is within the same person or between persons, (b) the 0° phase angle is the most stable phase relation between two coordinated movements, and (c) at high speed, the only way to temporally coordinate two movements is to lock their phase angle at 0°, i.e., making them oscillate in full synchrony. Similar to concomitant finger wagging and leg swinging, the articulatory movements for producing laryngeal events — F₀ and super-laryngeal events — consonants and vowels, may also be highly coordinated. In fact, the coordination should be much better than that found in those limb movement experiments, because speech production is much more highly skilled than the arbitrary movements performed in those studies. This coordination, however, may actually constitute an important constraint on the alignment of tone and syllable.

1.3 Intonation — What Is Left?

It has been a long tradition in English intonation research to assume that the basic intonation patterns are defined in terms of forms. These forms are specified in systems consisting of nucleus, head, tail, pre-head, and so on (Palmer 1922; O'Connor & Arnold 1961; Crystal 1969; Cruttenden 1997). These components are all defined in terms of pitch height and pitch movement rather than in their communicative functions. In recent years,
there have been attempts to adopt similar systems for describing Mandarin intonation (e.g. Shen 1994). As will be shown subsequently, however, a substantial portion of the intonation attributed to these components can be equally accounted for by the voluntary communicative functions listed in 1.2.1 and their interaction with the involuntary factors in 1.2.2. To the extent the pitch variations described by the conventional systems can be explained in terms of specific communicative functions, there may be little left that still needs to be accounted for by any purely form-oriented intonation framework.

Given our current limited knowledge, it is impossible to examine the effects of all the factors shown in the two lists given above. Nevertheless, recent progress in tone and intonation studies has made it possible for us to see a much clearer picture than before of how some of the voluntary and involuntary factors interact with one another in producing F₀ contours in Mandarin. In the following discussion, a summary of the new findings and their implications on our understanding of F₀ in general will be presented.

![Figure 1. Mean F₀ contours of four Mandarin tones in the monosyllable /ma/ produced in isolation by 8 male speakers. (Adopted from Xu, 1997)](image)

2. REALIZING ADJACENT TONES IN CONNECTED SPEECH

Figure 1 displays mean F₀ curves of the four Mandarin tones produced in isolation in the syllable /ma/ (Xu 1997), which are similar to
those found in other studies of Mandarin tones (e.g., Lin 1965; Howie 1976). These citation forms are not necessarily equivalent to the underlying forms of the tones. Nonetheless, since they were produced without any influence from adjacent tones, deviation from them when the tones are produced in context should tell us about the influences of the tonal context. These influences may be phonological in nature, or may be due to certain articulatory constraints.

Figure 2. Effects of preceding tone on the F0 contour of the following tone in Mandarin. In each panel, the tone of the second syllable is held constant, while the tone of the first syllable varies among H, R, L and F. The vertical lines represent the syllable boundaries (at the onsets of initial nasals). (adopted from Xu 1997)

2.1 How Tones Are Produced In Sequence

In order to minimize the effect of consonants on F0, Xu (1997) used the syllable sequence /mama/ as the tone carrier to examine how two tones are produced next to each other in Mandarin. Figures 2 and 3 display the mean F0 contours averaged across eight male speakers obtained in the study. In Figure 2, the tone of the second syllable is kept constant while the tone of the first syllable varies among the four lexical tones. Compared to their F0 contours in isolation, various changes have occurred in the four tones. The
first kind of changes are those that occur regardless of what the adjacent tone is. For example, the final F0 in F in the first syllable always drops only to the mid range rather than to the bottom as in isolation. The L in the first syllable (when not followed by another L) no longer has the final rise as seen in isolation. Apparently, these two changes are not directly related to the specific properties of the adjacent tone. We can refer to them as *positional variations*.

![Figure 3](image)  
*Figure 3. Effects of following tone on the F0 contour of the preceding tone in Mandarin. In each panel, the tone of the first syllable is held constant, while the tone of the second syllable varies among H, R, L and F. The vertical lines represent the syllable boundaries (at the onsets of initial nasals). (adopted from Xu 1997)*

The second kind of change is apparently categorical. When both syllables carry L, the F0 contour in the first syllable becomes almost identical to that in a RL sequence. This is the well-known tone sandhi phenomenon in Mandarin (Chao 1968; Wang & Li 1967).

The third kind of changes occur in the second syllable but are directly related to the F0 contour of the preceding tone. We can refer to them as *carryover variations*. These changes can be seen clearly in Figure 2. Specifically, the initial portion of the F0 contour of the same tone varies substantially with the tone of the first syllable. These variations are gradient
in that the onset $F_0$ of the second syllable is virtually determined by the ending $F_0$ of the first syllable, and that after the onset the $F_0$ gradually approximates a contour that somewhat resembles the citation form of the tone. Thus $H$ gradually becomes high-level, $R$ gradually becomes mid-rising, $L$ gradually becomes low (missing the final rise as explained above), and $F$ gradually becomes high-falling.

The fourth kind of changes occur in the first syllable but are directly related to the tone of the second syllable. We can refer to them as anticipatory variations. These changes can be seen in Figure 3. Unlike carryover variations, however, anticipatory variations are much smaller in magnitude. Furthermore, these variations are dissimilatory: the lower the $F_0$ minimum in syllable 2, the higher the $F_0$ maximum in syllable 1.

Finally, in RR and RL sequences, the $F_0$ peaks occur not inside syllable 1 which carries the first R. Rather, they occur soon after the boundary between the two syllables. This phenomenon has been referred to as peak delay (Xu 1998, 1999a, 1999b, 2001).

2.2 Possible Mechanisms Of Tonal Variations In Disyllabic Sequences

The five kinds of $F_0$ variations described above apparently have very different sources. The tone sandhi phenomenon is likely due to a phonological rule that changes the articulatory target of the tone so that it is produced in virtually the same way as $R$ (but see Xu 1993 for a small residue difference). Such interpretation is supported by well designed systematic perception tests (Wang & Li 1967). This means that $L$ in the first syllable in this case is deliberately produced with a tonal target not substantially different from that of $R$. The positional variations may also be due to a deliberate change in the tonal target itself. That is, the speaker voluntarily drops the final rise in $L$ and the final low dip in $F$.

2.2.1 Carryover Variations

In contrast, the carryover variations are not likely due to a voluntary change of the tonal target. Both the gradient variations in the initial $F_0$ of the second syllable and the gradual approximation of the proper pitch contour indicate that speakers are probably trying to realize a constant target but
have to somehow overcome certain articulatory constraints. One of such constraints is the \textit{maximum speed of pitch change}. The mean duration of syllable 1 in Xu (1997) is 176 ms and that of syllable 2 190 ms. In Figure 2, when H is preceded by L, its F\textsubscript{0} starts around 95 Hz and its highest F\textsubscript{0} later in the syllable reaches about 125 Hz. That is an interval of about 5 semitones. Due to the \textit{maximum speed of pitch change}, raising F\textsubscript{0} by 5 semitones should take at least 85 ms by male speakers according to Sundberg (1979). That is almost half of the syllable duration. If the initial acceleration and final deceleration in a pitch change are taken into consideration, it should take almost twice as long to complete the pitch change (Xu & Sun, 2000). It thus seems that the initial portion of the F\textsubscript{0} contour in syllable 2 is mostly an inevitable transition between the ending F\textsubscript{0} of syllable 1 and the underlying pitch target of syllable 2.

The situation is even more complicated when syllable 2 carries R or F. Presumably, both of them require pitch movements as part of their tonal specifications. In a LF sequence (Figure 2, lower-right panel), for example, the initial portion of the F\textsubscript{0} contour in the second syllable is a rise of about 7 semitones (92-135 Hz) and the final portion is a fall of about 4 semitones (135-106 Hz). According to Sundberg (1979), the former should take at least 90 ms and the latter 75 ms. Taking into consideration that there also have to be deceleration and acceleration near the turning point between the two movements (Xu & Sun, 2000), it seems that the speakers may have almost reached their limit when producing the pitch contour in the LF sequence.

The \textit{carryover variations}, therefore, seem to be a natural result of realizing a tone under the limitation of \textit{maximum speed of pitch change} and \textit{maximum speed of pitch direction shift}, assuming that the implementation of a tone has to start from the ending F\textsubscript{0} of the preceding tone.

\textbf{2.2.2 Alignment Of Tone And Syllable}

The above discussion seems to suggest that tones and syllables are tightly aligned with one another. Indeed, as can be seen in Figure 2, a tone may have very different starting F\textsubscript{0} when preceded by different tones, but it always follows a course that continually approaches its underlying target. The approximation is fast at first but slows down over time. And the
underlying target seems to be best approximated by the end of the syllable. Furthermore, as can be seen in Figure 3, when preceded by the same tone, the starting points of different tones at the syllable boundary are all in the vicinity of the ending point of the preceding tone, and they then diverge from that vicinity to approach their own specific targets.

The phenomenon of peak delay mentioned earlier may, at first glance, suggest that tone-syllable alignment is not always so consistent. But this phenomenon in fact further highlights the consistency of the tone-syllable alignment. To understand this point, we need to take into consideration the constraint imposed by the maximum speed of pitch direction shift. Based on the aforementioned observation that a tone is being implemented continually throughout its carrying syllable, and that its underlying target is best approximated by the end of the syllable, R should have the steepest rising slope near the end of the syllable, which is indeed the case as can be seen in Figures 2 and 3. When the following tone is R or L, pitch should start to go down immediately after the syllable boundary. However, since it takes time for the larynx to complete the pitch direction shift, the highest F0 occurs somewhat after rather than right at the syllable boundary. In other words, just as in other bi-tonal sequences, in a RL or RR sequence, the syllable boundary is where the implementation of the first tone ends and that of the second tone starts; and the slight peak delay is just a natural consequence of making a sharp F0 turn at the syllable boundary under the constraints of the larynx. Support for this interpretation was reported in later studies (Xu 1998, 1999a, 1999b, 2001).

But why, then, are tones aligned so tightly with syllables? Part of the reason might have to do with coordination of laryngeal and superlaryngeal movements — the last articulatory constraint listed in Section 1.2.2. Because speech production is a highly skilled activity, the implementation of tone and syllable is probably highly coordinated and a stable phase relationship between the two regularly maintained. At the rate of usually over five syllables per second, speech production involves fairly rapid articulatory movements. It is possible that at such a rate only the 0° phase angle can be stably maintained between tone and syllable. Hence, the rigid tone-syllable alignment found in recent studies (Xu 1997, 1998, 1999a, 2001) may be at
least partially due to the coordination between laryngeal and supralaryngeal articulatory movements.

2.2.3 Anticipatory Variations

The anticipatory effect has been found in a number of languages (Gandour, Potsuk & Dechongkit 1994 for Thai, Hyman 1993 for Enginni, Mankon, and Kirimi, Laniran 1992 for Yoruba, Laniran & Gerfen 1997 for Igbo; and Xu 1993 for Mandarin), and has been given different names, such as anticipatory raising, regressive H-raising, and anticipatory dissimilation. The underlying mechanism for the anticipatory effect, however, is still unclear. It is unlikely to be due to any categorical changes in the underlying target, because its magnitude is very small (mean difference = 3.14 Hz in Xu 1997). There have been various accounts for the underlying mechanism of anticipatory dissimilation, but none has been satisfactory (Gandour, Potsuk, Dechongkit, & Ponglorpisit 1992; Xu 1997; Xu & Wang 1997). In particular, it is not clear whether the effect is due to speakers’ voluntary change of pitch targets, or merely due to certain articulatory constraints. Without such information, a clear understanding of this effect is not possible.

2.3. Summary Of Adjacent-Tone Realization In Connected Speech

As demonstrated so far, most of the pitch contour variations due to adjacent tones are now readily explainable in terms of the interplay between voluntary speech acts — to produce the lexical tones with distinct underlying pitch targets, and articulatory constraints — maximum speed of pitch change, maximum speed of pitch direction shift, and coordination of laryngeal and supralaryngeal movements. The only exception is anticipatory variations, whose underlying mechanism still remains a mystery.

3. IMPLEMENTING TONE AND FOCUS CONCURRENTLY

Whether it should belong to prosody or pragmatics, there has been a general consensus that focus is conveyed mainly through variations in F₀ (Bruce, & Touati 1992; Cooper, Eady, & Mueller 1985; Eady, & Cooper 1986; Jin 1996; Liberman, & Pierehumbert 1984; Pierehumbert 1980; Rump, & Collier 1996; Xu 1999a). This may potentially be a problem for tone
languages like Mandarin, because tones are also conveyed mostly through F0. How, then, can lexical tones and focus in Mandarin be realized concurrently in an utterance? As found in Jin (1996) and Xu (1999a), focus and tones are realized by varying different aspects of F0 contours in Mandarin. Figure 4 illustrates the major effects of both focus and tone when other factors are kept constant (Xu, 1999a).

Figure 4. F0 variations due to tone and focus in Mandarin: selected mean F0 curves of two sentences produced by four male speakers. The thin line is from a sentence with a tone sequence of HHHHH (Māomī mō māomī), and the thick line HLHLH. (Māomī mō mādāō). (adopted from Xu 1999a)

To quote Xu (1999a, pp. 94-95), “[i]n general, tone identities are implemented as local F0 contours, while focus patterns are implemented as pitch range variations imposed on different regions of an utterance. The pitch range of tonal contours directly under focus is substantially expanded; the pitch range after focus is severely suppressed (lowered and compressed); and the pitch range before focus does not deviate much from the neutral-focus condition. Thus, there seem to be three distinct focus-related pitch ranges: expanded in non-final focused words, suppressed (lowered and compressed) in post-focus words, and neutral in all other words.” Furthermore, “the effect of focus is much more than just fine adjustment of the local tone contours. Rather, the adjustments are fairly substantial. In the post-focus region, the pitch range is sometimes suppressed so severely that different tone contours are hardly distinct from one other.”
Figure 4 also demonstrates that the role of articulatory constraints should always be taken into consideration if we are to fully understand the concomitant realization of tone and focus. For example, when the pitch ranges of the adjacent identical tones are different due to focus, apparent F₀ transitions will occur between these tones. Thus in Figure 4, there is a sharp rise between the second and third syllables in the HHHHH sequences in the lower right panel, and a sharp fall both between the second and third tones in the HHHHH sequence in the lower left panel and between the third and fourth tones in the HHHHH sequence in the lower right panel. Similar to the transitions due to tonal context as discussed earlier, these transitions also occur mainly in the early portion of the syllable, indicating that the tone-syllable alignment is still maintained. Apparently, the maximum speed of pitch change and coordination of laryngeal and superlaryngeal movements both play a role here, just as they do in the implementation of lexical tones.

4. POSSIBLE MECHANISMS OF DOWNSTEP AND DECLINATION

It is nowadays almost impossible to talk about intonation without mentioning downstep and declination, because both have been reported for so many languages and are virtually considered universal. Downstep refers to the phenomenon that in a HLH sequence, the second H is lower in F₀ than the first, which has been found in both tone languages (Stewart 1965; Meeussen 1970; Hyman 1973; Hyman & Schuh 1974; Clements & Ford 1979; Shih 1988; Manfredi 1993) and non-tone languages (e.g. Pierrehumbert 1980; Poser 1984; Pierrehumbert & Beckman 1988; Prieto, Shih, & Nibert 1996). Declination refers to the tendency for F₀ to gradually decline over the course of an utterance (Cohen & ’t Hart 1965; Cohen, Collier, & ’t Hart 1982), which has been reported also for both tone languages (Shih 1997) and non-tone languages (Pike 1945; Cohen & ’t Hart 1965; Maeda 1976; Cohen, Collier, & ’t Hart 1982; Cooper & Sorensen 1981; Ladd 1984). Following the line of thought developed in the present paper, one would naturally wonder, are downstep and declination voluntary or involuntary functions? Or, are they natural outcomes of certain voluntary or involuntary functions and/or their interactions?
4.1 Downstep

In Figure 4, the downstep phenomenon can be clearly seen. In the upper left panel, for example, the second H in the HLHLH sequence has much lower F0 than the first. As discussed earlier, however, we know that L exerts two different influences on the surrounding H: raising the F0 of the preceding H (anticipatory variation) and lowering the F0 of the following H (carryover variation). The combination of the two effects therefore makes the first H much higher than the second. Having realized that there are actually two separate mechanisms involved, we should no longer consider downstep as a holistic effect, though we may still use the name as a cover term. 3

4.2 Declination

Figure 4 also demonstrates that when occurring repeatedly, downstep may generate an overall down trend in F0 over the course of an entire utterance. Since declination is also known as an overall F0 downtrend, there might be certain similarities or possibly some overlap between the two phenomena. Indeed, Pierrehumbert (1980) and Liberman and Pierrehumbert (1984) report that much of the time-dependent lowering in English can be accounted for by a downstep model with an accent-by-accent decay of a constant ratio. Prieto et al. (1996) further suggests that declination is probably equivalent to a series of downsteps.

Yet a third factor that may contribute to the overall downtrend in an utterance is focus. As can be seen in the two lower panels in Figure 4, whenever there is a non-final focus in an utterance, there is always a substantial F0 difference between the initial and final H in addition to those due to downstep. Such difference is due to post-focus pitch-range suppression as well as (when focus is on the first word) on-focus pitch-range expansion.

Finally, it has been proposed that an exceedingly high F0 peak at the onset of the first sentence of a paragraph is used as a beginning signal for new topics, or is produced to draw listeners’ attention (Umeda 1982). More systematic discourse analyses have indeed found high F0 values related to topic initiation (Nakajima & Allen 1993). After the high initial F0 due to a
new topic, $F_0$ usually falls fast at first, but slows down gradually (Umeda 1982). Such a pattern is similar to the exponential decay reported for declination in several languages (Pierrehumbert 1980; Liberman and Pierrehumbert 1984; Gelfer, Harris, Collier, and Baer 1985; Prieto et al. 1996; Shih 1997). It is possible that when initiating a new topic the speaker substantially raises the pitch range of the very first word (or the first word that can be stressed) in a sentence, and then let the pitch range drop freely (i.e., without active lowering) to a relatively neutral level. As a consequence, the overall $F_0$ level would appear to decay exponentially: dropping fast at first, slowing down gradually, and finally virtually leveling off. It is also conceivable that, in the absence of any proper discourse context, as in many experimental settings, speakers read each sentence as if introducing a new topic, thus giving it a high initial $F_0$.

To summarize section 4, there has been strong evidence that downstep is probably due to the combined effects of anticipatory variation and carryover variation, and that declination may be due to the combined effects of downstep, focus, and topic initiation, whichever is applicable.

5. OTHER INTONATION VARIANTS?

The $F_0$-affecting factors discussed so far have mostly been studied extensively and their effects relatively well documented. There are also many other factors that are suspected of affecting $F_0$ but whose effects and functions are not yet well established. Among them are certain prosodic functions, pragmatic functions, and emotional functions. For prosodic functions, although there have been many (mostly phonological) studies on the prosodic structures of speech, their phonetic manifestations, especially in terms of $F_0$, are not yet well studied. Nonetheless, Shih & Sproat (1992) find that the R-variation described by Chao (1968) can be mostly attributed to the prosodic weakness of the syllable. It seems that the prosodic structure of an utterance may specify how vigorously each tone is being implemented, as well as how much time is assigned to each tone-carrying syllable, thus further determining how fully a tonal target is approximated. For understanding the pragmatic functions, there have been some promising attempts for English (Pierrehumbert & Hirschberg 1990; Hirschberg & Ward
1992), but not yet for Mandarin. Even for English, many more melodic patterns suggested by non-experimental studies (e.g., O’Connor 1961; Crystal 1969; Bolinger 1986, 1989) are probably also related to pragmatics, but have not been studied experimentally. As for emotional functions, though there have been many attempts (e.g., Uldall 1960; Ladd, Silverman, Tolkmitt, Bergmann & Scherer 1985; O'Shaughnessy & Allen 1983; Protopapas & Lieberman 1997; Ross, Edmonson & Seibert 1986; Ross, Edmondson, Seibert & Chan 1992), little has been firmly established so far.

6. SUMMARY AND FURTHER DISCUSSION

The foregoing discussion demonstrates how some of the observed F0 contour patterns in connected speech, Mandarin in particular, can be attributed to the combined effects of factors that are, as shown in Section 1.2., either voluntary or involuntary. The following summary briefly outlines how these factors may help shape the F0 contours in speech.

(1) Lexical tone specifies the underlying pitch target of each syllable.

(2) Under the articulatory constraint of coordination of laryngeal and supralaryngeal articulation, adjacent tones are produced in synchrony with their host syllables, and are implemented in such a way that the underlying target of each tone is continuously approximated within the duration of the host syllable. Due mostly to maximum speed of pitch change, the best approximation of a tone is achieved in the later portion of the syllable, and the target is often not fully reached. Due to maximum speed of pitch direction shift, the F0 peak associated with a tone often occurs in the next syllable.

(3) The prosodic structure of an utterance probably specifies how vigorously each tone is being implemented. It probably also specifies how much time is assigned to each syllable, thus further determining how fully a tonal target is approximated.

(4) Focus (when applicable) specifies the pitch ranges of the pre-focus, on-focus, and post-focus words in an utterance.

(5) New topic (when applicable) raises the pitch of the first accented word in an utterance.

(6) Similar to focus and new topic, other pragmatic functions probably
further specify more global F₀ patterns of an utterance.

(7) Emotion probably further modifies certain aspects (mostly likely, global settings) of pitch variations in speech. But exactly how such modifications are done awaits future research.

Having demonstrated how different aspects of F₀ variation can be attributed to factors that either have identifiable communicative functions (i.e., voluntary), or are merely articulatory (i.e., involuntary), we can now revisit the issue of phonologization mentioned in INTRODUCTION. Of particular relevance is the question of when we should invoke the phonologization interpretation when looking at a particular pattern of phonetic variation. As defined by Crystal (1997), phonologization is “a term used in historical phonology for a process whereby sounds which were formerly allophones develop a contrastive status (become phonemic) through the loss of their conditioning environments.” This definition, however, seems quite conservative considering how the term has actually been used in the literature. Frequently, not only are true phonemic neutralizations referred to as phonologization, so are conditional allophonic variations. But even if one adopts a looser definition, the following assumption seems inevitable:

When a phonetic variation is phonologized (or conventionalized), its production becomes deliberate.

In other words, when referring to a particular phonetic variant as a phonologized allophone, we imply that the variant is produced by the speaker with an articulatory target that is deliberately different from its canonical form. Though it is yet difficult to probe into the brains directly, the actual articulatory target the speaker has in mind may be assessed indirectly. For example, as discussed earlier, the R-variation in Mandarin described by Chao (1968) can be largely attributed to the combined effects of tonal context and prosodic weakening. In other words, what is deliberately changed is not the underlying tonal target, but the prosodic weight assigned to the R-carrying syllable, which allows articulatory constraints to bring greater distortion to the tone than in prosodically stronger positions. Given that the surface variation of the tone can be largely explained by factors other than the change of the tonal target itself (Xu 1997, 1999a), and given that listeners can successfully discover the original tones despite severe
superficial distortions (Xu 1994), it would be parsimonious to assume that a phonologization has not taken place. It is possible, therefore, that many other reported cases of phonologization may also turn out to be not truly phonological when reevaluated under the same principle.

Another closely related question is about the nature of the underlying target. An extensive discussion of this notion was done in Xu & Wang (2001). Defined briefly, a pitch target is the smallest pitch unit that is articulation-specific. According to Xu and Wang, while lexical tones are phonemes, the pitch targets associated with them are not. A phoneme is a symbolic unit which operates at an abstract level, and they are not necessarily articulatorily specific. The L in Mandarin probably has different articulation-specific targets when produced in isolation and in context. The final rise seen in isolation is likely to be articulatorily targeted. In context, in contrast, it is unlikely that the final rise is still part of the articulatory target of L.

It may be further speculated that articulatory targets are related to the proper phonemes through arbitrary association, and the arbitrariness of this association probably gives rise to certain idiosyncracies of various languages/dialects. For each phoneme, a speech community just choose through convention certain articulatory target(s) along an articulatorily possible continuum. Thus, language-/dialect-specific features are arbitrary and hence voluntary in the sense that speakers in the same community deliberately use the same specific articulatory target(s) to implement a phoneme. Viewed in this light, it is just natural that the exact articulatory value of a phoneme or even a phone transcribed using the same phonetic symbol differs among languages and/or dialects, as has been pointed out by Ladefoged (1980).

To conclude, it has been argued in this paper that there are many different sources of tonal variations in connected speech. These sources are better understood when they are viewed as either voluntary or involuntary. Voluntary sources are those stemming from linguistic/paralinguistic demands, and involuntary sources from articulatory constraints. Linguistic/paralinguistic demands represent various communicative functions on the one hand, but are associated with articulation-specific pitch targets and pitch ranges on the other. These pitch targets and pitch ranges are
what speakers actually attempt to implement in their speech; but the implementation is constrained by the limitations of the articulators that actually produce the fundamental frequency of voice. Observed variations in $F_0$ contours of in connected speech thus reflect different levels of linguistic/paralinguistic demands as well as their interaction with various articulatory constraints.

NOTES
1. There has been much new development in intonational phonology in recent years as represented by Pierrehumbert (1980) and Pierrehumbert & Beckman (1988). Nevertheless, these new systems are also primarily form-oriented.
2. Not reported in Xu (1997), but recomputed for the present paper.
3. Dowstep is sometimes used to refer to deliberate pitch lowering associated with certain syntactic or morphological functions, as in certain African tone languages. In those cases, pitch lowering is part of the linguistic information and hence should be considered voluntary.
4. Even when only phonemic neutralization is assumed to be true phonologization, in many cases, establishing a case of phonologization takes no more than simply writing out a rule in the form of $A \rightarrow B / C \_\_\_\_$, or a verbal description to the same effect.

REFERENCES


______. 1989. *Intonation and Its Uses -- Melody in Grammar and


LIN, Maocan 林茂灿. 1965. Yingao xianshiqi yu Putonghua shengdiao

LIU, Fu 刘复. 1924. Sisheng Shiyan Lu 四声实验室. Shanghai: Qunyi Shushe 群益书社.


XU, Yi, and WANG, Q. E. 1997. Components of intonation: what are
linguistic, what are mechanical/physiological? Presented at International Conference on Voice Physiology and Biomechanics, May 29-June 1, Evanston, Illinois.


连续 语流中 声 调 变 异 的 来源
许毅
西北大学

本文讨论影响连续 语流 中 声 调 变 异 的 各 种 因 素，并 试 图 论 证，要理
解 这些 因 素 最 好 把 它 们 分 为 主 观 和 客 观 因 素。主 观 因 素 来 自 于 语 言 功 能，
客观 因 素 来 自 于 发 音 器 官 的 局 限。语 言 功 能 一 方 面 对 应 于 各 种 交 际 需 求，
一方面又跟具体的音高目标和调域相对应。这些音高目标和调域是说话
人力 图 实 现 的 直 接 目 标，但 是 他 们 的 努 力 总 是 受 到 发 音 器 官 的 种 种 局 限。
因 此，我 们 在 连 续 语 流 里 观 察 到 的 基 频 曲 线 反 映 的 是 不 同 层 次 的 语 言 功
能 跟 各 种 发 音 局 限 相 互 作 用 的 结 果。

* This study was supported in part by NIH grant DC03902.