THE PENTA MODEL OF SPEECH MELODY: TRANSMITTING MULTIPLE COMMUNICATIVE FUNCTIONS IN PARALLEL

Yi Xu

Haskins Laboratories, New Haven, CT, USA
xu@haskins.yale.edu

ABSTRACT

Existing models of intonation typically define intonational components primarily in form and only secondarily in function. They also typically try to link observed F_0 contours directly to intonational meanings. The PENTA model of speech melody presented in this paper deviates from this tradition. First, it makes a clear separation between the meaning-bearing components of intonation, which are functionally defined, and the primitives of speech melody, which are defined purely in form (i.e., devoid of meaning) and readily implementable in articulation. Second, it specifies mechanisms for concurrent transmission of multiple intonational functions. Third, it specifies a continuous link between articulatory mechanisms of F_0 contour generation and the functional components of speech melody.

INTRODUCTION

An important goal in studying intonation is to identify its components and understand how they function in speech. Much of the research toward this goal is done by observing various aspects of the acoustic signals, including the fundamental frequency (F_0), amplitude, duration, voice quality, and spectral characteristics. Of these by far the most researched is F_0, which is the most direct correlate of speech melody. To identify tonal and intonational components from F_0, much effort has been devoted to figuring out how observed F_0 curves should be *divided* into individual intonational components. Various proposals have been made, as seen in a variety of intonation models. Despite extensive differences among them, however, most of the existing models of intonation share two critical assumptions. First, intonational components are defined primarily in form and only secondarily in function. Second, the form-defined intonational components are directly linked to meaning. These assumptions are behind proposed intonational components such as *nucleus*, *head* and *tail* in the British model (Cruttenden, 1997), *H* and *L tones* (manifesting as F_0 peaks and valleys) in the Autosegmental-Metrical (AM) model (Ladd, 1996; Pierrehumbert, 1980), *accent* and *phrase commands* in the command-response model (Fujisaki, 1988), and *complex F_0 shapes* that are either fully overt or stylized in the Bolinger model (Bolinger, 1989), the IPO model ('t Hart et al., 1990) or the tilt model (Taylor, 2000). Both of these assumptions, however, need to be reconsidered if we were to develop an effective comprehensive model of intonation.

First, intonational components should be defined in terms of function rather than form. As pointed out recently by Studdert-Kennedy (in press) about language evolution: “…what we need…is a model of language function. Form follows function, not function form”. This should be true for our understanding of various synchronic aspects of speech as well. That is, it is unlikely that communicative functions have evolved to serve preexisting linguistic forms. Rather, it is more likely that linguistic forms have evolved to serve various communicative functions, i.e., to convey different communicative meanings.

Second, the surface patterns of speech melody are not directly linked to communicative meanings. In searching for melodic components corresponding to meaning-conveying functions, we have noticed that observed F_0 contours rarely correspond directly to specific communicative meanings (Xu, 2004a). Rather, the functional components of intonation are detached from directly observable surface acoustic forms by at least three degrees of separation: *articulatory implementation*, *target assignment* and *parallel encoding*. The first separation is imposed by various physical properties of the articulatory system, which make it impossible for the surface forms to fully resemble the intended targets. The second separation is due to the fact that there are many language specific and often arbitrary rules that assign different articulatory targets to the same categorical unit. The third separation has to do with the fact that multiple layers of information are simultaneously transmitted, and they each leave a unique mark on the observed acoustic form (cf. Xu, 2004a for detailed discussion).

A comprehensive model of speech melody thus needs to satisfy at least three critical requirements. First, it has to make a clear separation between the meaning-bearing components of intonation, which are functionally defined, and the primitives of speech melody, which are defined purely in form (i.e., devoid of meaning) and readily implementable in articulation. Second, it has to specify mechanisms for concurrent transmission of *multiple* intonational functions. Third, it has to specify a continuous link between articulatory mechanisms of F_0 contour generation and the functional components of speech melody. In the following, I will show that the recently proposed Parallel Encoding and Target Approximation (PENTA) model attends to all these requirements (Xu, 2004b). As I will demonstrate, the PENTA model provides a framework through which a rich repertoire of communicative functions can be realized concurrently through F_0, with all the details of the F_0 contours still traceable to their proper sources via specific linking mechanisms.

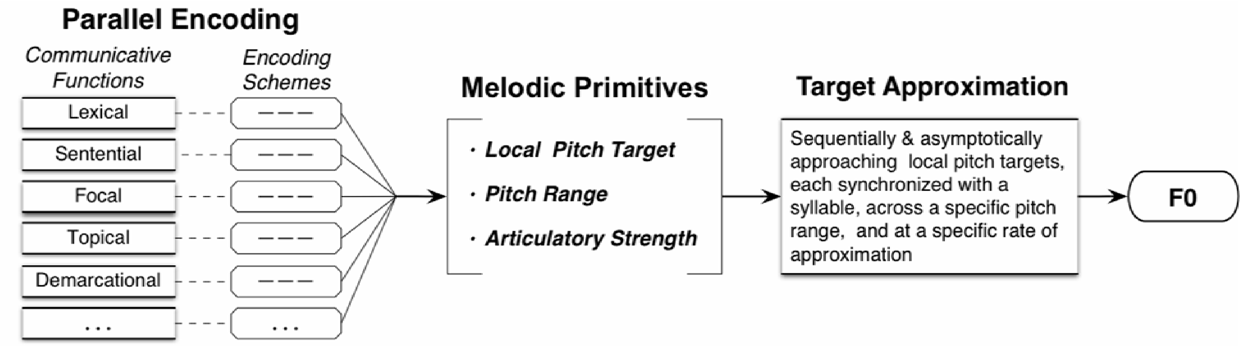

Figure 1. A sketch of the PENTA model. See text for explanations. The unnamed block at the bottom left indicates communicative functions yet to be identified.

THE PENTA MODEL

The PENTA model is based on two basic assumptions. First, speech melody is produced by an articulatory system whose physical and neurological properties impose various constraints on the way surface acoustic forms are generated. In particular, inertia limits the maximum speed of pitch change; and coordination of different articulators limits the degrees of freedom for the alignment of laryngeal movement with those of other articulators (Xu, 2004a; Xu & Sun, 2002;

Xu & Wang, 2001). Second, a multitude of communicative functions are concurrently conveyed through speech and perceptual parsing of these functions requires that each function be *uniquely* encoded. A diagram of the PENTA model is shown in Figure 1. The stacked boxes on the far left represent individual communicative functions. These functions control F_0 through distinctive *encoding schemes* (shown to their right) that specify the values of the melodic primitives, which include *local pitch target*, *pitch range* and *articulatory strength*. The values of the melodic primitives as stipulated by different encoding schemes can be specified both symbolically and numerically. Table 1 shows possible symbolic values of the melodic primitives.

Table 1. Possible symbolic values of the melodic primitives: *local target*, *pitch range*, and *articulatory strength*, which may be notationally distinguished from one another by [ ], %, | | and < >, respectively.

Local Target:	Regular target:	[high], [low], [rise], [fall], [mid]
	Boundary tone:	high%, low%, mid%
Pitch Range:	Height:	|high|, |low|, |mid|
	Span:	|wide|, |narrow|, |normal|
Articulatory Strength:		<strong>, <weak>, <normal>

As can be seen in Table 1, a local pitch target can be either static — when it is [high], [low] or [mid], or dynamic — when it is [rise] or [fall]. When a target is static, its relative pitch height is the only intended goal. When a target is dynamic, in contrast, the velocity of the pitch movement as well as the relative pitch height are the intended goals (Xu & Wang, 2001). Pitch range determines the frequency scope across which local pitch targets are implemented. It has two kinds of specifications: height and span. Height specifies the relative height of the pitch range, e.g., |high|, |low| or |mid|. Span specifies the width of the pitch range, e.g., |wide| or |narrow|. Articulatory strength specifies the speed at which a local pitch target is approached. When the strength is <strong>, the target is approached faster than when it is <weak>.

An example of encoding schemes for lexical tone and focus can be seen in Figure 2a, which shows an actual case of simultaneous transmission of lexical tones and focus in Mandarin, with decomposition of the functional components in the framework of the PENTA model. Displayed in the graphic part of Figure 2a are the mean F_0 curves of the Mandarin sentence "Māomǐ mō māomī" (tone sequence: HLHHH) said with and without initial focus (thick-solid/-thin curves), together with the average F_0 curve of an all-H sentence as reference (dotted line). The lexical tones are associated, via tonal rules specific to Mandarin, with the local targets [high] and [low], respectively, as shown in the *Tonal* tier below the F_0 plot. Evidence for these local targets has been demonstrated in Xu (1997, 1999) and Xu & Wang (2001). When the first (disyllabic) word of the sentence is focused, the encoding scheme of focus assigns a [wide] pitch range to the focused syllables, and a [narrow]+[low] pitch range to the post-focus words, as shown in the *Focal* tier below the F_0 plot. Evidence for such pitch range specifications has been reported for several languages, including English, Dutch, Shanghai and Mandarin (Cooper et al., 1985; Rump & Collier, 1996; Selkirk & Shen, 1990; Xu, 1999; Xu et al., 2004).

The symbolic representations of the local targets and pitch ranges also correspond to specific numerical values. The specific height and shape of the local pitch targets corresponding to the lexical tones are depicted by the short horizontal lines in Figure 2a, indicating that they can be represented numerically by simple linear functions. The pitch range adjustments by focus are indicated by the block arrows. The two unfilled block arrows on the left indicate a [wide] pitch range as compared to a [normal] range not explicitly depicted in the graph. The filled block arrow on the right indicates a [narrow]+[low] pitch range (though the [narrow] is not obvious because all the local targets in these post-focus words are static).

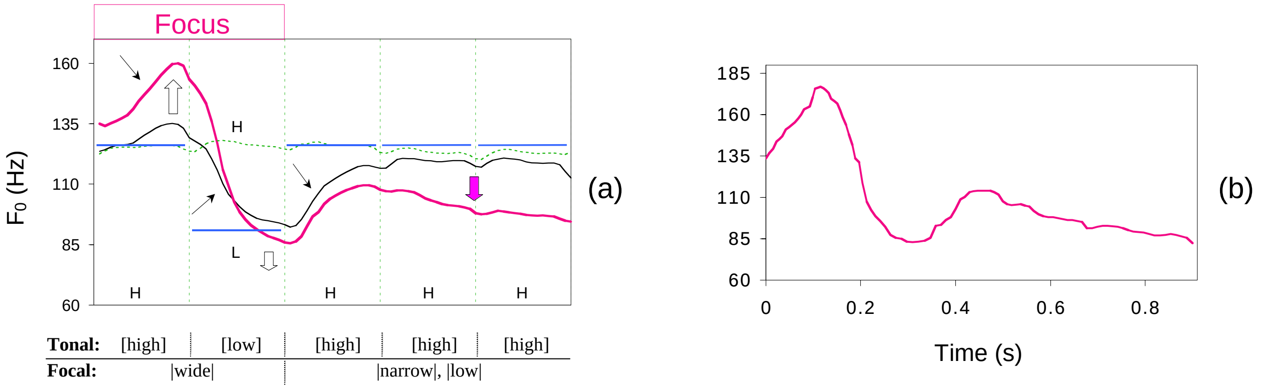

Figure 2. (a): Averages F_0 of 20 repetitions of the Mandarin sentences “Māomǐ mō māomī” [Cat-rice strokes Kitty], and “Māomī mō māomī” [Kitty strokes Kitty], by 4 male speakers (Data from Xu, 1999). Thick solid curve: focus on “Māomǐ”; thin solid curve: no focus; dotted curve: HHHHH with no focus. Vertical grids indicate nasal murmur onsets. Short horizontal lines represent hypothetical local pitch targets [high] and [low]. Line arrows point to F_0 transitions due to articulatory implementation. Unfilled block arrows indicate on-focus pitch range expansion. Filled block arrow indicates post-focus pitch range lowering and narrowing. X-axis: normalized time. (b): F_0 of a single repetition of the sentence “Māomǐ mō māomī” with focus on the first word “Māomǐ.”

In the PENTA mode, the melodic primitives are, *at the same time*, control parameters for the Target Approximation model that simulates articulatory implementation of the local targets (Xu, C. et al., 1999; Xu & Wang, 2001). A sketch of the Target Approximation model is shown in Figure 3. Through this model, the control parameters corresponding to the melodic primitives are turned into continuous F_0 contours through asymptotic approximation of local pitch targets, which are synchronized with their associated syllables. Under the constraint of maximum speed of pitch change (Xu & Sun, 2002), such asymptotic approximation often results in a long transition in the early part of a syllable, as illustrated in Figure 3. The asymptotic transitions are also apparent in syllables 1-3 in Figure 2a, as indicated by the line arrows. In addition, the target approximation process also produces the peaks in syllables 1 and 3 (the latter only when with initial focus), and the valley in syllable 3. Also seen in Figure 2a are the mechanical effects of downstep brought about by the L tone, which raises F_0 of the preceding H and lowers the F_0 of the following H (most obvious in the thin solid curve where the F_0 lowering after L cannot be attributed to an early focus) (cf. Xu, 1997, 1999 for detailed discussion). Thus through target assignment by the encoding schemes and target approximation by articulatory implementation, the functional components of intonation are eventually turned into continuous F_0 contours.

To contrast the functional decomposition of F_0 contours shown in Figure 2a, Figure 2b shows F_0 tracing of a single repetition of the Mandarin sentence “Māomǐ mō māomī” with focus again on the first word. The plot, however, is displayed without indication of the lexical tones, syllable boundaries or focus location. From such a plot, one can easily observe two peaks, a valley and an overall downtrend. These F_0 events, while clearly discernable, and having been the focus of many studies, do not seem able to directly tell us much about the functional components behind the detailed F_0 contours.

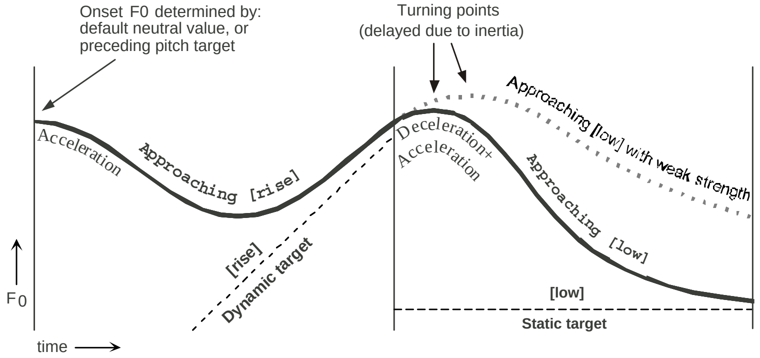

Figure 3. Illustration of the Target Approximation model. The vertical lines indicate syllable boundaries. The straight dashed lines represent *local pitch targets*. The solid curve depicts the F_0 contour resulting from *asymptotic approximation* of the pitch targets. (Adapted from Xu & Wang, 2001.) The dotted curve in syllable 2 simulates the effect of <weak> articulatory strength.

CONCLUSION

The PENTA model presented in this paper results from an attempt to understand speech melody through a function-oriented approach, which tries to define intonational components in terms of function rather than form. It results further from the recognition that the functional components of intonation do not directly correspond to surface acoustic forms, but are rather linked to them through distinctive encoding schemes specifying the values of the melodic primitives and an articulation process that implements these primitives. The PENTA model recognizes three melodic primitives — *local pitch target*, *pitch range* and *articulatory strength* — and treats them as both basic encoding elements for the communicative functions and control parameters for the articulatory system that generates F_0 contours. The PENTA model further assumes that the articulatory system generates F_0 by successively approaching syllable-synchronized local pitch targets, across specific pitch ranges, and with specific articulatory strengths. By so doing, a continuous link is specified between the articulatory mechanisms of F_0 contour generation and multiple functional components of speech melody that are transmitted in parallel.

Initial effort to quantify the Target Approximation part of the PENTA model was made in Xu, C. et al. (1999). Effort to quantify the entire model is currently underway.

ACKNOWLEDGEMENTS

This study is supported in part by NIH grant DC03902.

REFERENCES

Bolinger, D. (1989) *Intonation and Its Uses -- Melody in Grammar and Discourse*. Stanford, California: Stanford University Press.

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.

Cruttenden, A. (1997) *Intonation*. Cambridge: Cambridge University Press.

Fujisaki, H. (1988) A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology: Voice Production* (edited by O. Fujimura). New York: Raven Press, Ltd., 347-355.

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

Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.* Ph.D. dissertation. MIT, Cambridge, MA.

Rump, H. H. & Collier, R. (1996) Focus conditions and the prominence of pitch-accented syllables. *Language and Speech, 39*, 1-17.

Selkirk, E. & Shen, T. (1990) Prosodic domains in Shanghai Chinese. In *The Phonology-Syntax Connection* (edited by S. Inkelas and D. Zec). Chicago: University of Chicago Press, 313-37.

Studdert-Kennedy, M. (in press) How did language go discrete? In *Evolutionary Prerequisites of Language* (edited by M. Tallerman), Oxford: Oxford University Press.

't Hart, J., Collier, R. & Cohen, A. (1990) A perceptual Study of Intonation — An experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press.

Taylor, P. (2000) Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America, 107*, 1697-1714.

Xu, C. X., Xu, Y. & Luo, L.-S. (1999) A pitch target approximation model for F0 contours in Mandarin. In *Proceedings of The 14th International Congress of Phonetic Sciences*, San Francisco, 2359-2362.

Xu, Y. (1997) Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*, 61-83.

Xu, Y. (1999) Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics, 27*, 55-105.

Xu, Y. (2004a) Separation of functional components of tone and intonation from observed F0 patterns. *From Traditional Phonology to Modern Speech Processing: Festschrift for Professor Wu Zongji's 95th Birthday* (edited by G. Fant, H. Fujisaki, J. Cao and Y. Xu), Beijing: Foreign Language Teaching and Research Press**,** 483-505.

Xu, Y. (2004b) Transmitting Tone and Intonation Simultaneously — The Parallel Encoding and Target Approximation (PENTA) Model. *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing**,** 215-220.

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. & Wang, Q. E. (2001) Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33*, 319-337.

Xu, Y., Xu, C. X. and Sun, X. (2004) On the Temporal Domain of Focus. *Proceedings of International Conference on Speech Prosody 2004*, Nara, 81-84.