Proceedings of Speech Prosody 2006, Dresden, Germany

Speech prosody as articulated communicative functions

Yi Xu

Department of Phonetics and Linguistics, University College London, London
Haskins Laboratories, New Haven
yi@phon.ucl.ac.uk

Abstract

Speech prosody, just like the segmental aspect of speech, conveys communicative meanings by encoding functional contrasts. The contrasts are realized through articulation, a biomechanical process with specific constraints. Prosodic phonology or any other theory of prosody therefore cannot be autonomous from either communicative functions or biophysical mechanisms. Successful modeling of speech prosody can be achieved only if communicative functions and biophysical mechanisms are treated as the *core* rather than the *margins* of prosody.

1. Introduction

It is often assumed in prosody research, explicitly or implicitly, that the prosodic units commonly investigated are equivalent or at least analogous to phonemes in the segmental aspect of speech. Thus pitch accents, prominence, phrase tone, boundary tone, etc., are viewed as the phonological or functional units that need to be understood in experimental and theoretical investigations, and synthesized and recognized in technological applications. A fundamental property of the phoneme is the ability to distinguish meanings. Segmental phonemes, for example, can differentiate words. Thus "ferry" and "very" in English are distinguished by [f] versus [v], and these consonants are therefore viewed as distinct phonemes in the language. In contrast, /r/ in the two words is known to substantially vary in pronunciation across speakers, but the variations are not considered to be phonemic because they do not change the identity of the words.

When it comes to prosody, however, the criteria for determining whether an observed difference is phonemic often become obscured. This is best illustrated by the widespread AM model of intonation, in the words of Pierrehumbert [12:59]:

> In the literature, one can distinguish two approaches towards the problem of establishing which intonation patterns are linguistically distinct and which count as variants of the same pattern. One approach attacks the problem by attempting to deduce a system of phonological representation for intonation from observed features of F0 contours. After constructing such a system, the next step is to compare the usage of F0 patterns which are phonologically distinct. The contrasting approach is to begin by identifying intonation patterns which seem to convey the same or different nuances. The second step is to construct a phonology which gives the same underlying representation to contours with the same meaning, and different representations to contours with different meanings. … The work presented here takes the first approach, in fact, it stops at the first step in the first approach.

Thus whether the proposed phonological units actually distinguish meanings is viewed as largely irrelevant when the AM theory was first established. Although there have been later attempts, e.g. [13], to take "the next step," communicative functions are never treated as part of the "grammar" of intonational phonology, c.f. [14]. Similarly, many other prominent approaches also treat form as primary and function as secondary, as pointed out by [8].

Furthermore, also treated as marginal in these approaches are the articulatory mechanisms. The widespread assumption is that prosodic units directly correspond to observed forms, be it F_0 turning point, F_0 height and slope, or impressionistic perceptual prominence.

In this paper I argue for a major shift from conventional approaches, based on evidence that has been steadily accumulating over the years: Communicative functions and biophysical mechanisms of speech should be treated as the *core* rather than the *margins* of prosody.

2. Communicative functions as essence of prosody

"We speak to be heard in order to be understood" [6:13]. Being understood, however, has been interpreted in different ways. A widespread view is that understanding speech is done not by directly accessing the meaningful components, but by first processing an abstract structure — the phonology. This phonology has its own internal grammar that is largely independent of meanings [12]. As we have just seen, the application of this view has lead to a major deviation from the basic notion of phoneme. What we need to recognize is that although the meaningful contrasts in speech are realized through form, the link between function and form is never broken, as otherwise successful communication would not have been possible. In fact, I would like to argue that communicative meanings are not just *important* for prosody. They are the *essence* of prosody. In other words, what we need is no less than a major conceptual shift: The grammar of prosody should be about how communicative meanings are linked to acoustic forms rather than about how a formally defined phonological structure manifests itself based on a set of function-independent rules.

But communicative functions, as is well known, can be quite elusive. How, then, can we determine what is functional and what is not? The key, I believe, is to first recognize that we do not need to wait for an all-encompassing theory of prosodic meaning before starting to identify individual meaning-contrasting functions. After all, the identification of segmental phonemes as well as lexical tones is done well before any serious investigation of the impact of prosody on the phonetic realization of segments and tones has been carried out, and we now know that such impact is not trivial [2,3]. Thus it should be just as possible to identify prosodic functions one at a time. In fact, this has been done over the last few decades as reviewed in [20]. From these efforts, some

lessons can be drawn which can be summarized into a few general principles:

1. *Specificity*. Prosodic components should be defined in terms of communicative functions that are as specific as possible about what they contrast and about their temporal domains of operation.
2. *Mutual-exclusivity*. Each function should have a unique "encoding scheme" which has at least one predominant characteristic not overlapped by other functions. This means that once an observed pattern has been attributed to a particular function, it should not be reattributed to another function, unless there is clear evidence that they can both remain operative despite the overlap.
3. *Elicitability*. For a function to be verifiable, there needs to be at least one way of reliably eliciting it under experimental conditions. An unelicitable function is an unproven function.
4. *Audibility*. A functional contrast in a language must have reached certain perceptual threshold, otherwise it would not have been operational. While the precise values would depend on the nature of the function, there are some reasonable thresholds. For example, the identification rates for focus and question/statement both can be well over 80% [10, 23].

In light of the above principles, we can see that many of the prosodic components proposed and investigated over the years are predominantly formal, such as *pitch accent, prominence, boundary tone, phrase tone, rhythm*, etc. These components all have relatively explicit formal definitions, but rather vague functional definitions. Pitch accents, for example, are by definition related to both focus and lexical stress [12]; thus the specificity and mutual-exclusivity principles are both violated. Also, the audibility principle is violated because the agreement on the types of pitch accents is rather low even among highly trained ToBI labelers [18].

3. Encoding communicative functions through articulation

Because there are no other ways for speakers to manipulate the acoustic output of speech but to operate the articulatory system, communicative functions, including those related to prosody, have to be encoded through articulation. But how can an indirect articulatory process reliably encode communicative functions? Lexical tone, which most would agree is functional, seems to provide a key. As can be seen in Fig. 1, the F_0 contours of a tone vary extensively with the preceding tone. But by the end of the syllable, they all have converged to a quasi-linear line, as indicated by the arrow. The uniformity of the linear line despite the contextual variability suggests that it is directly linked to the underlying form of the tone, while the consistency of the converging movements reveals how the articulatory encoding is actually done: through approximation of the underlying target with a unidirectional movement [21]. There is also evidence that such target approximation is synchronized with the syllable, even in a non-tonal language like English [22]. These findings have motivated the Target Approximation (TA) model, which simulates the basic articulatory process of pitch production as *syllable-synchronized sequential target approximation* [20,21].

Based on the TA model, the presence and implementation of local pitch targets are *obligatory*, and prosodic functions can be encoded by specifying various aspects of the target approximation process, including the height and slope of the target, the speed at which the target is approached, and the pitch range and the time allotted to the target. Evidence for the use of all these aspects has been reported [20,21].

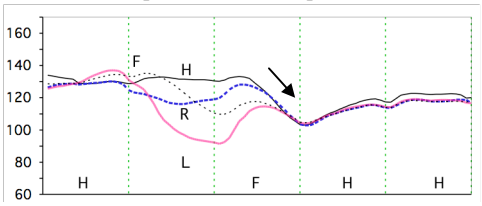

Fig. 1: *Mean F_0 contours of Mandarin F tone following four different tones. Adapted from [19].*

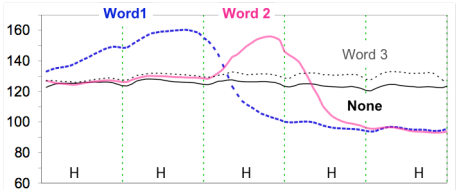

Fig. 2: *Mean F_0 contours of a Mandarin all-H sentence spoken in four focus conditions. Adapted from [19].*

The specificity principle stated earlier requires that each communicative function have a specific temporal domain of operation, as determined by the nature of the function. For example, in a tone language like Mandarin, the temporal domain of a lexical tone is a syllable, as each monosyllabic morpheme is independently assigned a tone. In contrast, the temporal domain of focus includes not only the focused item, but also all the items that are "out of focus" [23]. In Fig. 2 we can see that when a non-final word in a sentence is focused, not only is its own pitch range expanded, but also the pitch range of all the post-focus words is suppressed. For perception, the post-focus pitch range suppression is just as important as the on-focus pitch range expansion [11,16,23].

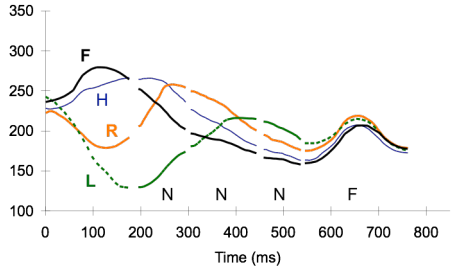

Fig. 3: *Mean F_0 contours of a sequence of Mandarin neutral tones following four full tones. Adapted from [1].*

Conversely, seemingly holistic acoustic patterns may have arisen from units with separate temporal domains. In Fig. 3, although the three Mandarin neutral tones seem to form a continuous F_0 contour, they each actually have a separate temporal domain, as each is linked to a different morpheme [1].

Clarifying the temporal domain of operation may also help elucidate the notion of planning. That is, what happens *within* the temporal domain of a function is due to *execution* rather than to planning. In question intonation, for example, F_0 increases nonlinearly toward the end of the sentence [10],

as shown in Fig. 4. The smaller rise before the larger final rise is not due to planning, but due to *early progression* of the function itself. Also, in Fig. 3 F_0 of the first and second neutral tone drops gradually not in anticipation of the turning point several syllables ahead, but due to slow movements toward the targets of the current syllables [1].

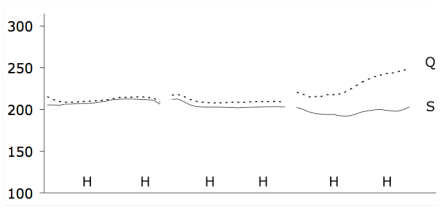

Fig. 4: *Mean F_0 contours of a Mandarin all-H sentence spoken as statement or question. Adapted from [10].*

4. Articulatory and functional models of prosody

The importance of articulation and communicative function has already been recognized over the years, as reflected in models that emphasize either the articulatory or functional aspects of prosody.

4.1. Articulatory models

Articulatory models are those that try to simulate the articulatory process of speech prosody. In these models, parameters do not control surface F_0 directly, but aspects of hypothesized articulatory mechanisms that affect F_0.

The command-response model [4] simulates F_0 contours as responses of a critically damped second order system to two types of muscular commands, accent commands and phrase commands, which are idealized step functions and impulse functions, respectively. The system responds to these commands by generating F_0 that rises or falls exponentially in the direction of the commands and then returns to the baseline. The model assumes no internal restrictions on the timing of the commands.

The Stem-ML model [7] describes F_0 contours as resulting from realizing underlying tonal templates with different amounts of muscle forces under the physical constraint of smoothness. The smoothness constraint makes the connection between adjacent templates seamless, and the varying muscle forces determine the degree to which the shape of each template is preserved in the surface F_0 under the influence of adjacent as well as distant tones.

4.2. Functional models

Functional models are those that try to simulate speech prosody as realizations of specific communicative functions. In these models, parameters typically correspond to proposed communicative functions rather than to units in an autonomous phonological structure.

The Thorsen model [17] describes Danish intonation as consisting of local tonal accents superimposed on a linear global intonation curve whose slope conveys whether the sentence is a statement or question.

The Kiel model [8,9] characterizes German, English and Dutch intonation as communicative functions realized as various terminal intonation patterns and different types of emphasis realized through either F_0 or intensity.

The IF (Intonation Functions) model [5] provides a notational system for representing the functions of intonation in terms of degrees of prominence and types of prosodic boundaries.

5. PENTA: An integrated model

If communicative functions and articulatory mechanisms are equally indispensable for speech, effective prosodic modeling should simulate both the articulatory process of pitch production and the process of encoding communicative meanings. This understanding has motivated the Parallel Encoding and Target Approximation (PENTA) model [20], a diagram of which is shown in Fig. 5.

The stacked boxes on the far left represent individual communicative functions which constitute the primary input to the model. They are parallel to each other with no hierarchical organizations, since the meanings they represent are independent of each other.

The communicative functions are manifested through distinctive *encoding schemes* (second stack of boxes from left), which are either universal or language specific. Being abstract and formal, these encoding schemes may appear to resemble the formal units in intonation phonology and other conventional approaches. They differ from the latter, however, in being always linked to specific functions, and are thus neither self-defining nor hierarchically organized.

The encoding schemes then specify the values of the melodic primitives (middle block): *pitch target*, *pitch range*, *articulatory strength* and *duration*, which are, at the same time, control parameters of the TA model that simulates the articulatory process as *syllable-synchronized sequential target approximation* [21].

To implement the PENTA model, qTA, a quantitative version of the TA model, has been developed [15]. It is a feedback controlled overdamped second order system driven

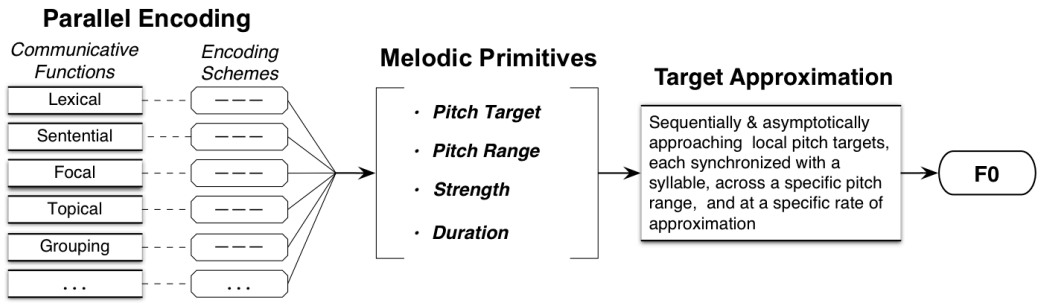

Fig. 5: *A schematic sketch of the PENTA model [20].*

by underlying pitch targets. Fig. 6 shows the display of an interactive Java implementation of the model. As can be seen, the surface F_0 contours continually approximate successive underlying pitch targets (short straight lines) which shift abruptly at each syllable boundary. The control of the surface F_0 is not direct, but through the manipulation of three parameters: target height, target slope, and natural frequency of the system which determines the speed of each target approximation.

Due to the nature of the second-order system, the final state of each syllable, in terms of F_0 height, velocity, and acceleration, is transferred to the next syllable as its initial state. Such transfer not only produces the surface continuity, but also generates phenomena such as peak delay and carryover assimilation, which can be substantial in the case of neutral tone (see the similarity of the three N tones in Fig. 6 to those in Fig. 3) and unstressed syllables, for which the articulatory strength is likely weak [1, 22].

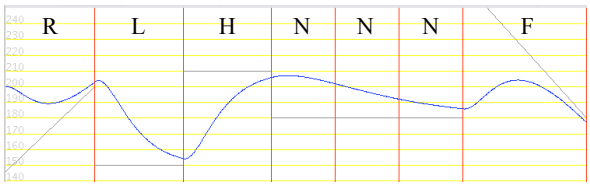

Fig. 6: *Display of an interactive Java implementation of the qTA model (http://www.phon.ucl.ac.uk/home/yi/qTA/). The neutral tone syllables (N) are assigned weak strength.*

The qTA model thus provides a realistic simulation of the "Target Approximation" module in Fig. 5, with which various encoding schemes can be implemented. Tests have been carried out on a dataset containing systematic variations in tone and focus [19]. The results in terms of root mean square error and correlation coefficient were comparable to those of other models even when the trained parameters were applied to novel sentences and novel speakers in the dataset [15].

6. Conclusion

Although the importance of both communicative meanings and biophysical properties of the articulatory system has been recognized by some models, in general the dominant theories of speech prosody have focused primarily on the observed prosodic forms as if they are largely autonomous from function and articulation. In this paper I have argued that it is time for a major conceptual shift, i.e., communicative functions and biophysical mechanisms of speech should be treated as the *core* rather than the *margins* of prosody. Following this view, prosodic components are defined and organized by communicative functions that are parallel to each other. These functions can be transmitted only by manipulating control parameters of an articulatory process: *syllable-synchronized sequential target approximation*. The recently proposed PENTA model is an initial step toward this conceptual shift. The ease with which tone and focus are simulated in our recent quantitative testing [15] has demonstrated an advantage of articulatory-functional approaches over the conventional formal approaches to speech prosody.

7. References

[1] Chen, Y.; Xu, Y., in press. Production of weak elements in speech -- Evidence from f0 patterns of neutral tone in standard Chinese. To appear in *Phonetica*.

[2] Cho, T., 2004. Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *J. Phonetics* 32, 141–176.

[3] Fougeron, C.; Keating, P. A., 1997. Articulatory strengthening at edges of prosodic domains. *J. Acoust. Soc. Am.* 101, 3728-3740.

[4] Fujisaki, H., this session. The roles of physiology, physics and mathematics in modeling prosodic features of speech.

[5] Hirst, D. J. 2005. Form and function in the representation of speech prosody. *Speech Communication* 46, 334-347.

[6] Jakobson, R.; Fant, G.; Halle, M., 1963. *Preliminaries to Speech Analysis*. Cambridge: MA: MIT Press.

[7] Kochanski, G.; Shih, C., 2003. Prosody modeling with soft templates. *Speech Commun.* 39, 311–352.

[8] Kohler, K. J., 2004. Prosody Revisited — FUNCTION, TIME , and the LISTENER in Intonational Phonology. *Proc. Speech Prosody 2004*, Nara, Japan, 171-174.

[9] Kohler, K. J., this session. What is emphasis and how is it coded?

[10] Liu, F.; Xu, Y., 2005. Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation. *Phonetica* 62, 70-87.

[11] Mixdorff, H., 2004. Quantitative tone and intonation modeling across languages. *Proc. International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, 137-142.

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

[13] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*. P. R. Cohen, J. Morgan and M. E. Pollack. (eds.) Cambridge, Massachusetts: MIT Press**,** 271-311.

[14] Pierrehumbert, J., 2000. Tonal elements and their alignment. In *Prosody: Theory and Experiment — Studies Presented to Gösta Bruce*. M. Horne. (eds.) London: Kluwer Academic Publishers**,** 11-36.

[15] Prom-on, S.; Xu, Y.; Bundit T., this volume. Functional-oriented articulatory modeling of tones and intonations.

[16] Rump, H. H.; Collier, R., 1996. Focus conditions and the prominence of pitch-accented syllables. *Lang. Speech* 39, 1-17.

[17] Thorsen, N. G., 1980. A study of the perception of sentence intonation — Evidence from Danish. *J. Acoust. Soc. Am.* 67, 1014-1030.

[18] Wightman, C. W., 2002. ToBI or not ToBI. *Proc. Speech Prosody 2002*, Aix-en-Provence, France. pp. 25-29.

[19] Xu, Y., 1999. Effects of tone and focus on the formation and alignment of F0 contours. *J. Phonetics* 27, 55-105.

[20] Xu, Y., 2005. Speech Melody as Articulatorily Implemented Communicative Functions. *Speech Commun.* 46, 220-251.

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

[22] Xu, Y.; Xu, C. X., 2005. Phonetic realization of focus in English declarative intonation. *J. Phonetics* 33, 159-197.

[23] Xu, Y.; Xu, C. X.; Sun, X., 2004. On the Temporal Domain of Focus. *Proc. Speech Prosody 2004*, Nara, Japan, 81-84.