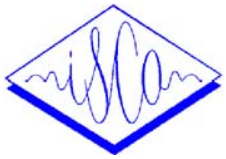


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# LEVELS OF PROSODIC REPRESENTATION IN SPOKEN DISCOURSE: AN EMPIRICAL APPROACH

*Richard Esposito[2] and Li-chiung Yang [1,2]*

University of California at Santa Barbara[1] & Spoken Language Research Institute[2]
U.S.A.
yang@vowel.ucsb.edu or lyang@sprynet.com

## ABSTRACT

In this paper we propose that prosodic structures in spontaneous discourse exhibit both linear and superpositional characteristics, and that these reflect the different scopes of multi-tiered emotional and cognitive processes. This multi-tiered structure encompasses the syllable and word level, inter-phrase movement, and extended pitch level baseline rise and fall. Analysis of the data suggests that integration of the 3 different prosodic levels within an overall prosodic model provides a critical link for the generation of natural-sounding interactive speech systems.

Keywords: prosody, discourse, multi-level, emotion- cognition



## 1. INTRODUCTION

### 1.1 Linear or Global?

New investigations into prosodic modeling have focused on the local and global aspects of intonation, and whether intonation is more adequately modeled as a linear or superpositional phenomenon [1, 2, 3, 4,5, 6, 7, 8]. This issue is usually defined in terms of the relation between word stress and sentence intonation. The linear approach tends to view the F0 patterns of an utterance as consisting of the sequence of word accent or pitch accent stings, whereas the superpositional approach views the same pattern as the word stress patterns superimposed on an overall independent sentence intonation pattern.

In the linear account of Pierrehumbert and Beckman, sentence intonation is composed of a linear sequence of characteristic pitch accents and boundary tones which signal different pragmatic functions. The advantage of the linear accounts is that they are simpler in that there is less need for look-ahead and only one layer of phonological rules to apply to produce the utterance. On the other hand, superpositionalists such as Grønnum, Fujisaki, Möbius, Campbell, and Bailly propose that a superpositional model allows the effects of several independent influences on intonation to be analyzed more clearly, and that there is adequate evidence that more global factors are at work.

The debate is mostly centered on sentence level intonation and its relation to local pitch accent patterns. In this paper we propose a multi-dimensional framework of intonation in spontaneous discourse based upon our research using natural discourse in Mandarin Chinese, and discuss what the intonational patterns may reveal about this issue.

## 2. METHODOLOGY AND SPEECH CORPUS

### 2.1 Speech Data

The corpus of this study consists of recordings of spontaneous conversation between native Chinese speakers in home settings. About 90 minutes of the speech data were digitized at a 22,500Hz sampling rate using the ESPS *waves+* program. For our analysis, we selected utterances from different points of the conversations, as well as continuous subsections of discourse. Data were annotated for discourse relations, topic structure, emotional-cognitive content and speaker turns.

In order to capture the different domains at which intonational patterns are manifested, data were analyzed both at the syllable and word levels as well as the inter-phrase level. The third level of our analysis focuses on how discourse flows over extended stretches of conversation. As a means of representing the intonational structures in discourse, we plotted the highest and lowest pitch points of roughly 600 continuous utterances, equivalent to a 20-minute dialogue segment, for each speaker to visualize the dynamic discourse flow patterns.

### 2.2 Why Spontaneous Discourse Data Is Important?

Why do we use natural discourse data instead of more controlled experimental speech for our study? One critical advantage of discourse data, in our view at least, is that we are able to expand our analytical scope and see distinct layers of different influencing factors. From our observation as well as supported by research on speech in general, components that are important or distinct at a lower level may not have the same status at a higher or more integrated level. The nature of the interacting variables is often changed depending upon the larger environment.

Another important reason for using discourse data is that by looking beyond word lists and sentences in isolated situations, we are better able to see how speech is organized and evolves through time. Discourse events are often connected in some way and relationships among events are established as the discourse progresses. Analysis based on the sentence level is unable to capture phenomena which depend upon the relationships in a large domain.

Moreover, in spontaneous discourse there are a lot of false starts, hesitations, repetitions, and other common speech

production phenomena, which may be significant to understand how language really works. Relationships among speakers are reflected in complex patterns of participant interactions and signaling of cognitive states. The development of the discourse itself undergoes rapid changes and frequent topic shifts, and rhythmic and emotional elements are very common. All of these things are reflected in the intonational patterns of spontaneous discourse. A clear understanding of how intonation functions in discourse not only contributes to the fundamental understanding of human communication but is also critical to the development of more robust and more natural human-computer interactive systems.

# 3. THREE LEVELS OF PROSODIC STRUCTURES

## 3.1 Level 1 : The Syllable/Word Level - the Within-the-Phrase Level

### *3.1.1 Focus, Cognitive and Emotional States*

One thing that becomes evident from looking at our data is that cognitive and emotional states are closely related to pitch variations of syllables and words within the phrase, and can have dramatic effects on tonal and intonational shape. Focus and cognitive and emotional states often affect both the pitch height and shape of syllables, words and the phrase itself in characteristic ways. The particular scope of the state is critical in determining the intonation contour of the phrase.

Our data indicate that the scope of cognitive-emotional states can vary, sometimes acting locally on single syllables and words, but they can also act globally on the entire phrase level. Figure 1 illustrates 3 different instances of the phrase *meiyou23* meaning "there isn't" having strikingly different contours under different emotional states. The contour in the beginning section of the example in speaker B's utterance *Meiyou meiyou bi zhe geng pianyi de a? Zhende a?* "There isn't there isn't anything cheaper than this? Really?" reflects the typical high-rise-steep-fall intonation pattern of surprise or incredulity [1, 5], whereas the flat intonation contour of the final example in speaker A's *Meiyou ta shuo* "There isn't, he said" reflects a mild matter-of-fact state.

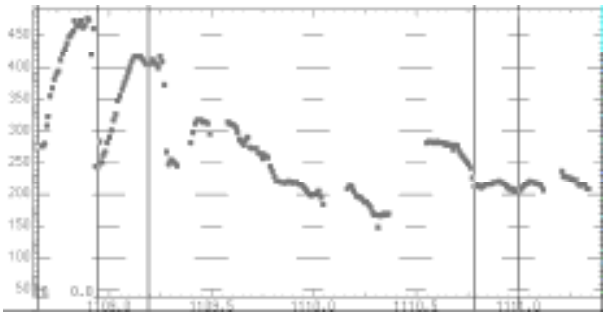
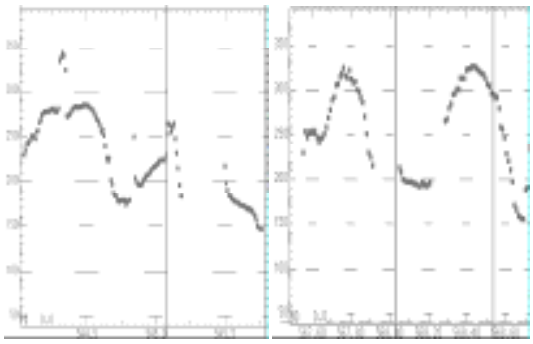


**Figure 1:** 3 instances of the phrase *meiyou* "there isn't".

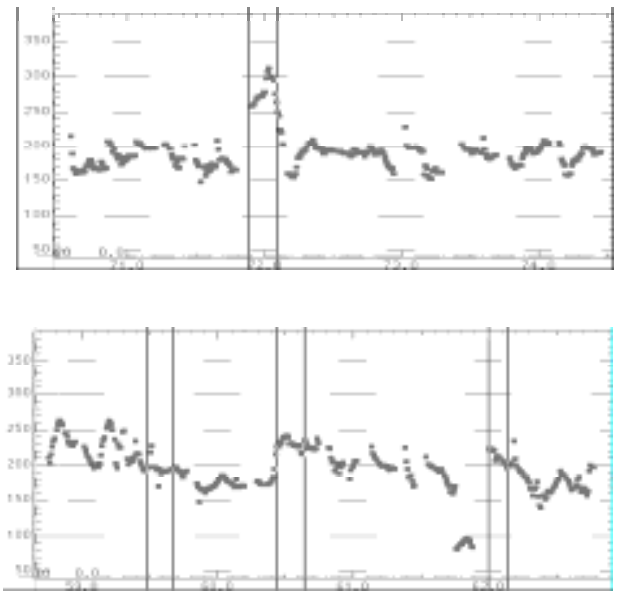
In these examples the scope of the emotion appears to be on the phrase as a whole. The rise-fall of the first 2 instances and the flatness of the last instance all occur over the entire phrase. Note that tones may change greatly to conform to the overriding intonational force as seen in this example.

The next set of pitch tracks in Figures 2a-b illustrates how scope can vary. The declining contour of *duo1* "many" in the first instance reflects the negative emphasis expressed by the speaker, whereas the arched shaped *duo* in the second example expresses exaggerated emphasis. Note the first example seems to have a more local scope, with the most prominent pitch changes on the syllables *hen duo*. In the second example, the scope is more global and strong enough throughout the phrase to cause a repetition of the contour.



**Figures 2a-b:** 2 instances of *hen duo* "very many". The utterance in 2a is *Ranhou neitian yinwei meiguoren hen duo* "Then because that day there were a lot of Americans" and the utterance in 2b is *You o! Hao duo o!* "There are! A whole lot!"

The examples seen in Figure 3a and Figure 3b illustrate nicely the contrast between a more integrated and a more isolated pitch pattern due to focus or specificity. *Tamen* "they" in the upper example is clearly separated in pitch level from the rest of the phrase.



**Figures 3a-b:** An example of an isolated vs. an integrated pitch pattern within the phrase. The vertical lines highlight the word *tamen* "they" in 2 close sections of a single narrative.

## 3.2 Level II: The Inter-Phrase Level

### *3.2.1 Topic Structure and Development*

The 2nd level of our analysis is concerned with what is happening at the inter-phrase level. What we have found here is that a structure of systematic hierarchical phrase level movements in discourse exists, and that these movements are pragmatically and cognitively meaningful. Specifically, topic structure and development are intonationally indicated by the phrase pitch height as well as direction of pitch step between phrases (see [9] for a detailed analysis).

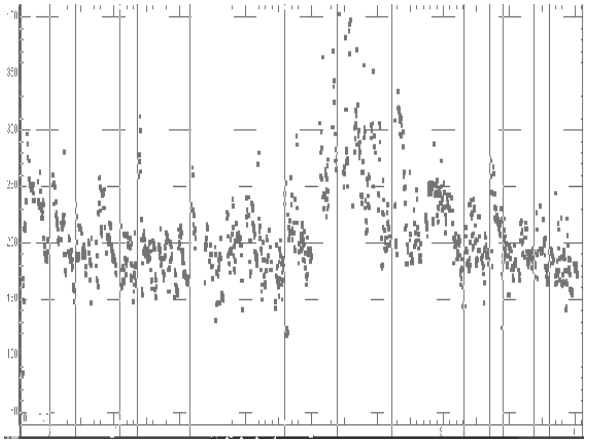


**Figure 4a:** The striking contrast between the sequence of upstepping leading to a dramatic climax, and the following anti-climactic downsteps is evident when the whole narrative event is captured in one view.

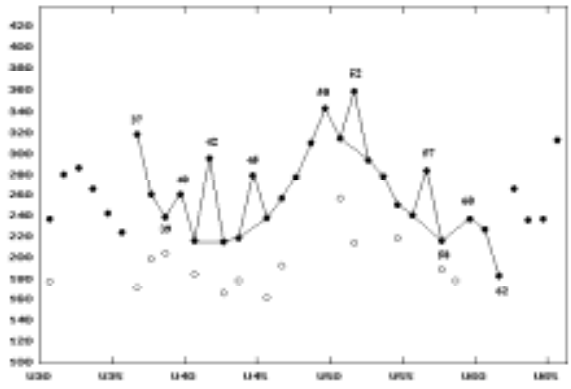


**Figure 4b:** The peak pitch plot mirroring the intonational structure of the compressed pitch track in Figure 4b.

In the systematic rising and falling structure seen in Figure 4b, there is a very orderly stepping hierarchy which correlates with the topic structure. Analysis of the data shows that in this section, each step functions to add new information to overcome the previous step until the speaker finally comes to the high point - the climax of the story. The speaker then gradually comes down in pitch, adding further details on the downslope. The 3 spurts of 60Hz seen in the peak pitch plot represent signals of a temporary break in the ongoing topic flow, at each spurt, the speaker interrupts the development to give a summarization of previous statements.

## 3.3 Level 3: Long waves

### *3.3.1 Extended Pitch Level Rise and Fall*

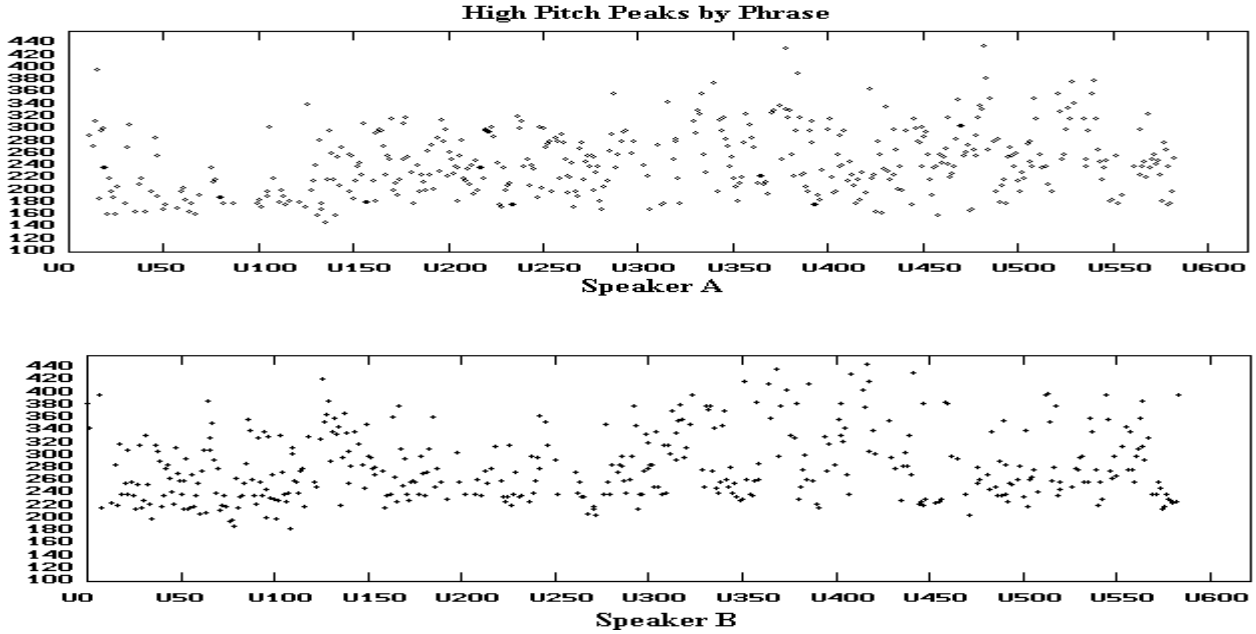
The coherent intonational organization of phrases which represents the phrase-to-phrase cognitive relationships of uncertainty and planning simultaneously manifests a process of climax and resolution. Climax and resolution patterns seem to be one of the most important recurring patterns of cognitive and emotional change which guide the development and intonational hierarchical structure of a conversation. A discourse climax is often mutually reached by participants in a conversation when the speaker successfully communicates and fully involves the hearer in the essential moral of the topic, at which point an intonational peak occurs. This pattern is illustrated in the previous example in Figures 4a-b, in which the climax is reached when topic, discourse, and cognitive-emotional elements all come together.

In our corpus, even larger scope climax and resolution structures at the global level of extended discourse sequences occur. If we look at the pitch peaks of phrases in our data over an extended section of discourse, as in the 600 phrases for each speaker portrayed in Figures 5a-b, we can see that phrases vary greatly in pitch height, but there also seem to be patterns of a gradual rise of pitch level followed by a gradual fall, that is, an arch shape. In the initial section of the conversation (U20 to U100), speaker B is the main speaker, with speaker A mostly providing feedback. At about U100, the conversation changes its course, and speaker A assumes the main speaker role, and we can see that there are about three extended rise-fall arches as speaker A develops different topics. These seem to occur at approximately U100 to U275, U275 to U425, and U450 to U575. Note that even these arches seem to be increasing in pitch height themselves. In the lower chart of Figure 5b, we can also see how speaker B's involvement with each topic is reflected in the corresponding sections.

These long waves pictured here occur over sequences of over 100 phrases each, and the rise-fall structure we showed in Figure 4b becomes difficult to detect at this larger scope. If we were to look at these long waves themselves in greater detail, we would also be able to identify patterns of rise-fall within the overall structures. The appearance of such climax and resolution patterns at several different levels of discourse suggests that the phenomenon of climax and resolution is a significant part of how topic organization, emotion and cognitive states, and long waves of episodic development are integrated intonationally.

# 4. A STRUCTURAL FRAMEWORK FOR INTONATION IN DISCOURSE

Based upon our data, we propose that it is useful to view intonation at 3 distinct but interrelated levels.




**Figures 5a-b:** Plot of 600 consecutive pitch peaks of both speakers. On the top is speaker A, on the bottom is speaker B. Three extended rise-fall arches can be seen in U100-U275, U275-U425, and U450-U575 in speaker A's pitch height movement. Speaker B's pitch movements in the corresponding sections also reflect speaker involvement with the topic development

1. At the broadest level, our data exhibit long waves of pitch level rise and fall extending over large sequences of utterances. In our analysis, these long waves constitute a broad-scope patterns of intonation, and are associated with general levels of psychological and interactive involvement with topic, and extended processes of climax and resolution.

2. Within these long waves, there is a second level of intonation which operates at the inter-phrase level. At this level, specific topic development occurs as utterances are intonationally positioned as phrase units relative to one another.

3. At the syllable, word and phrase level, a linear process of expressive, accentual and boundary marking progression occurs continuously. These elements in this process also interact with the specific accentual or tonal characteristics of the specific language.

Our data also suggest that intonational structures in spontaneous discourse exhibit both linear and superpositional characteristics, and these really reflect the different scopes of multi-tiered emotional and cognitive processes.

## 5. CONCLUSION

The fact that within phrase pitch contours ride along with these broader structures provides evidence that general pitch range and topic structure are superpositional. At the within phrase level, there seems to be a mix of both linear and superpositional elements. Expressive intonation with arbitrary scope can be seen as superimposed on the phrase. On the other hand, expressive intonation with a more localized scope, including focus, can be integrated in the realization of F0 in a linear fashion and this provides evidence for the linear view.

Our results provide support for the significance of the cognitive-discourse approach as proposed by Bailly and Campbell, as well as the linguistic intonational approaches of Beckman, if attention is focused on the varied scopes of different intonational components in the total process of expressive communication in spontaneous speech. A clear consideration of these different levels and their integration into an overall prosodic model is of crucial importance in the development of natural sounding spoken dialogue systems.

## 6. REFERENCES

[1] Bailly, G. 1998. The ICP Prosodic Model. http://www.icp.inpg.fr/cost258/prosody/ICP.

[2] Beckman, M. 1995. Local Shapes and Global Trends. In *ICPHS*:100-107.

[3] Campbell, W.N. 1997. "Synthesizing spontaneous speech, Sagisaka, Y., Campbell, W.N., and Norio, H., editors, *Computing prosody*, Springer-Verlag, 165-186.

[4] Fujisaki, H. 1995. A generative model for the prosody of connected speech in Japanese. *Annual Report of Engineering Research Institute*, 30:75-80, 1971.

[5] Hirschberg, J. and Pierrehumbert, J. 1986. The Intonational Structuring of Discourse. 24th *ACL Proceedings*: 136-144.

[6] Grønnum, N. 1995. Superpositional and Subordination in Intonation: A Non-linear Approach. *ICPHS*: 124-131.

[7] Ladd, D. R. 1995. "Linear" and "Overlay" Descriptions: An Autosegmental Metrical Middle Way. *ICPHS*: 116-123.

[8] Möbius, B. 1995. Components of a Quantitative Model of German Intonation. *ICPHS*: 108-115.

[9] Yang, L-C. 1995. *Intonational Structures of Mandarin Discourse*. Ph.D. Dissertation, Georgetown University.