ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PROSODY AND TOPIC STRUCTURING IN SPOKEN DIALOGUE

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

Spoken Language Research Institute[1] & CREST, JST (Japan Science and Technology)[2]

yang@vowel.ucsb.edu or lyang@sprynet.com

ABSTRACT

Prosody is critical in conveying topic coherence and the salience of information in speech. In this study we propose that the overall coherence is brought about through pitch level structuring of phrases at both the local level of hierarchical phrase unit positioning and the global level of pitch baseline rise and fall as climax and resolution. Our results show that prosody has critical importance in conversation, and is crucial for topic segmentation, topic tracking, and information extraction.

Keywords: prosody, topic, coherence, discourse, emotion

1. INTRODUCTION

In spontaneous discourse, intonational patterns of phrases are a critical component in conveying the coherence of topic development and the salience of communicated information between participants. Unlike written or pre-planned discourse, where known information is presented in a connected, structured way to make a point or tell a story, in spoken discourse, unexpected events occur constantly because participants bring their own intentions and information states to the conversation, thereby bringing about continuous shifts of focus. The complex topic relationships at work, and the ongoing dynamic interactions between speakers contribute to a constant changing mix of more fragmented and more structured discourse, characteristic of spontaneous spoken discourse.

In this study, we propose that intonationally the overall coherence is brought about through the pitch level structuring of each phrase unit relative to other phrases. We show that the specific relationship of importance and development from one phrase to the next is accompanied by the specific change in general pitch level between phrases. In our view, this intonational positioning of units in the overall hierarchical structure is the key component for signalling topic relationship among phrases, and is critical for information extraction and interpretation.

2. SPEECH CORPUS

2.1 Speech Data

The corpus of this study consists of digitized spontaneous conversations and broadcast narratives in English and Mandarin Chinese, totaling about 6 hours of speech data. Our approach is to abstract from the surface shapes, discourse text, and context, to form a theory linking the surface shapes to these elements. 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. This technique allows us to detect and track important topic events and points of interest easily, and to form appropriate generalizations.

2.2 How Is Spontaneous Speech Different?

Why do we use natural discourse data instead of more controlled speech such as read or tasked-oriented data for our study? Research on spontaneous speech shows that read or planned speech differs from spontaneous conversation in many critical ways. One is the *degree of control*. Using controlled data can limit the focus to critical variables, but may also introduce artifacts and problems of generalization. Most utterances used in controlled settings are highly decontextualized as well as de-emotionalized, whereas real life situations always occur in a specific context and we respond to that particular situation accordingly. This context is an essential part of language interpretation and understanding. The second difference is *how varied* are the context and different types of situations which the data present. Read speech, task-oriented and elicited speech in general only represent a subset of the situations encountered in a normal conversation. Moreover, because of the artificial production process, controlled settings are likely to produce results which may not correspond to natural speech conditions, and consequently lead to poor system performance, as demonstrated in recent work on conversational speech synthesis and recognition [1,3]. Another difference is the *degree of interaction* among participants. In spontaneous conversation typically there is a high degree of interaction and involvement because the goals and progress of the conversation are determined mutually by the participants themselves, and the prosodic phenomena depend upon the complex nature of these elements. We believe that if the goal is to understand ordinary conversation and to achieve human-like quality in intelligent interactive systems, then it is important to investigate a level of complexity sufficient to model spontaneous speech.

3. INTER-PHRASE TOPIC STRUCTURE

3.1 Patterns of Topic Organization

Downstepping. Analysis of the pitch movements in our conversational data shows that topics are hierarchically organized by intonation. In our data, dialogue, episodes and topic initiation often begin with a high pitch level or expanded pitch range and a high amplitude, while endings or closings are marked by a low pitch level or narrowed pitch range and a low amplitude. The direction of pitch level movement is frequently downward i.e. from higher to lower between phrases within the topic or episode. Within topics, downstepping between phrases usually occurs when there is a natural elaboration of topic ideas which move towards a resolution. This process can be seen as progressive movement away from uncertainty, with each subsequent phrase closer to a final resolution. The degree of step lowering represents the degree of completeness and finality of the phrase relation in the topic hierarchy. This resolution can be of several types. Speakers often have an underlying goal, and each subsequent statement brings the speaker closer to that goal. The discourse resolution can also include a more explicit goal, as in the working out of a logical problem to a successful conclusion or in situations of mutually referencing shared knowledge. In all of these cases it can be seen that the progression from high pitch to low pitch correlated with elaboration of previous steps can be associated with a growing level of confidence and certainty.

Upstepping. Another important marker of hierarchical topic organization is upstepping, i.e. upward movement in pitch between phrases. We found that although topic development frequently follows a downstepping pattern, in spontaneous speech, new topics or subtopics are often introduced as more gradual and natural developments of previous topics, and topic initiation phrases can start at a low or more intermediate level and subsequent phrases develop in patterns of upstepping. This may occur because the unity of the narrative development takes precedence over the need to signal new information. Successive upstepping from such low topic initiation points often occurs as emotional elements unexpectedly enter the conversation or in situations of overcoming previous inadequacy or incompleteness. Our analysis indicates that upstepping and downstepping between phrases signal different cognitive activities, reflecting the different status of uncertainty and certainty, new information, and planning. In our data, upstepping between phrases often occurs in situations of cognitive uncertainty, as in self-reflection and doubt, in contrast to downstepping, which is typically associated with definiteness, finality and completion. In addition to occurring as natural progression to a new topic, upstepping frequently occurs in cognitively less planned situations, or when the speaker is getting progressively more involved with the topic.

Throughout conversation, upstepping and downstepping frequently alternate on a phrase-to-phrase basis, expressing changing uncertainty and certainty as new propositions emerge, are tested, and are resolved. Thus, the pitch level structuring of phrases not only signals the coherence relations among phrases but really represents the multi-tiered emotional and cognitive processes at work in spontaneous discourse.

Data Discussion. We illustrate what is happening at the inter-phrase level with respect to these topic structure patterns in Figures 1-2. What we have found in our data 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 [4].

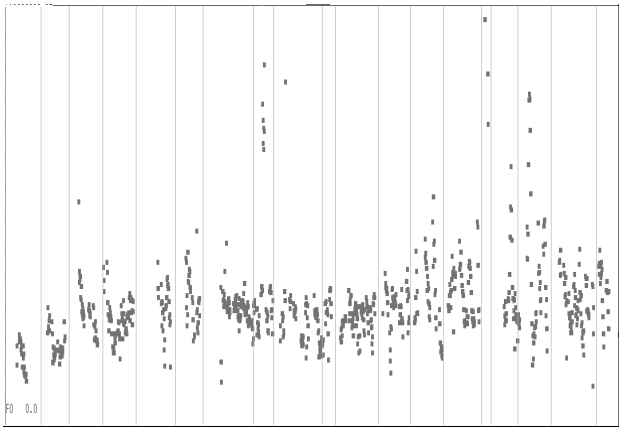


Figure 1: A mix of upstepping and downstepping phrasal movements correlate systematically with topic development in a section of English narrative data.

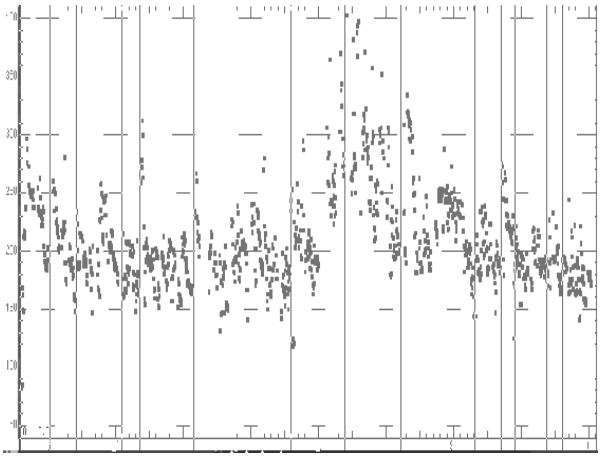
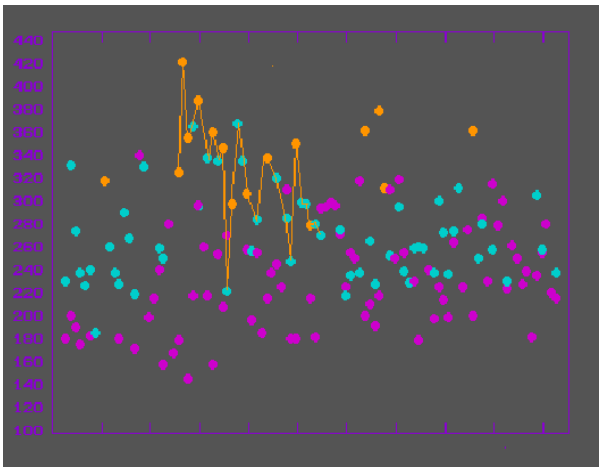
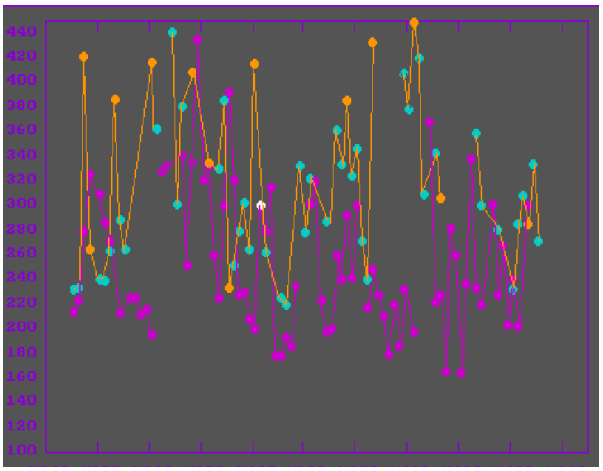


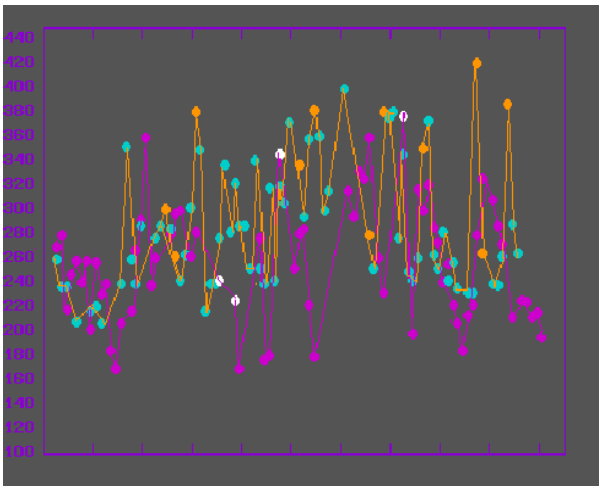
Figure 2: The striking contrast between the sequence of upstepping leading to a dramatic climax, and the following anti-climactic downsteps is evident when the whole narrative event is captured in one view.

The systematic nature of phrase-to-phrase intonation is evident in Figure 1, which shows the pitch movements of a 14-phrase subsection of a radio story. The overall rise-fall-rise-fall pattern is segmented naturally by the intonational direction, with each of the four sustained directional movements corresponding to development on a different aspect of the topic, with local peaks constituting narrative climaxes.

Figure 2 shows a very dramatic intonational hierarchy of upstepping followed by downstepping which correlates with the topic structure. Analysis of the data shows that in this section, each phrase functions to add new information to overcome the previous phrase until the speaker finally comes to the climax of the story. The speaker then gradually descends in pitch, elaborating further details on the downslope. The topic relations and pitch level patterns in these examples show that there is an underlying coherent organization of topic through intonation. Even when the topic is disorganized, the intonation signals systematically the coherence and unity of the discourse.







Figures 3-5: In top left, lower left, and top right order
A: magenta, white (interruptions) B: blue, yellow (interruptions)

4. EPISODIC TOPIC STRUCTURE

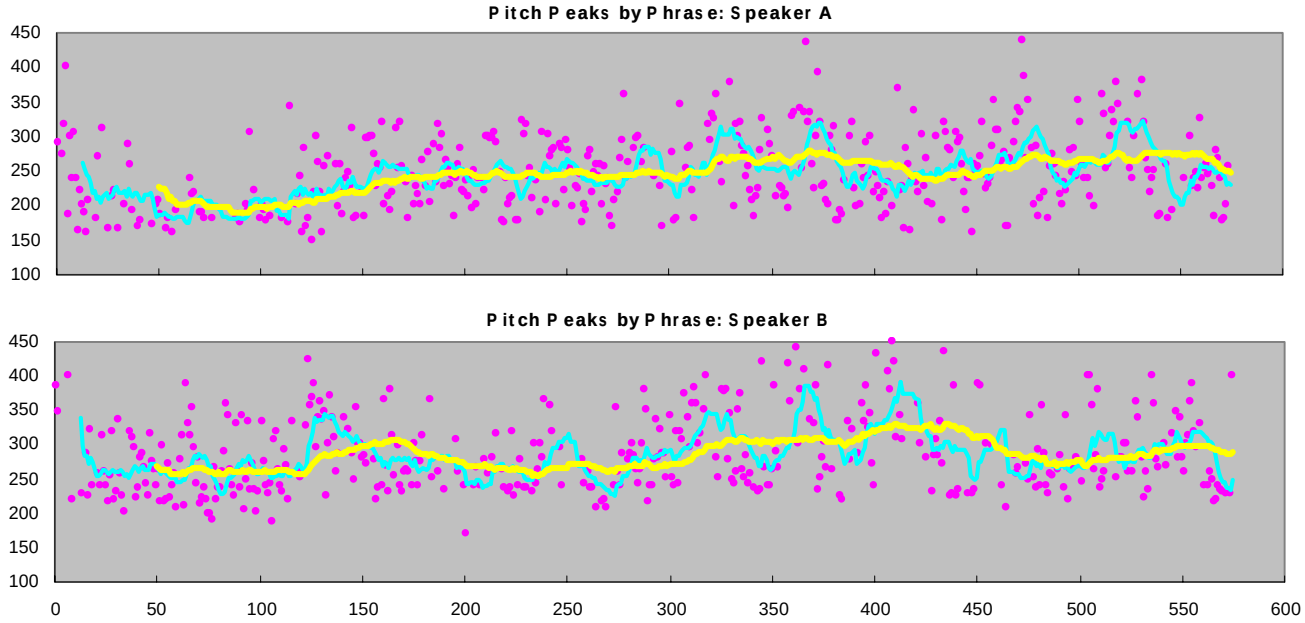
4.1 Climax and Resolution

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. Our data show that 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 following examples in Figures 3-5, in which the climax is reached when topic, discourse, and cognitive-emotional elements all come together.

Intensity and degree of uncertainty and certainty are significant determinants of topic direction and have a strong influence on the prosodic structure. In this part of the conversation (Figure 3), speaker A starts to talk about a conference she attended previously and speaker B mistook it to be the conference that she was interested in, so she initiates a series of short questions (in the form of interruptions) to confirm and clarify the information. The general pattern seen here is that speaker B encounters an initial high unsettled state of uncertainty and gradually progresses to a more settled and certain state, as the information becomes more complete, and this is clearly expressed in the overall downtrend in the pitch levels for these utterances. The internal structure, the ups and downs within this pattern, is also very revealing. At each utterance that expresses doubt and a need for clarification, there is a local rise in pitch, whereas those utterances which express acknowledgement and certainty are locally lower in pitch. The specific strength of signal needed varies systematically with the resolution of the differing interests and knowledge states of participants.

Taking a more extended view of our data shows that pitch movements also vary according to overall patterns of topic development and intensity of speaker involvement. Analysis of the discourse text shows that the rise-fall arc seen in Figure 4 also coincides with the development of a major subtopic that both speakers actively contribute to. This involvement is signaled by the large amount of dots at varying heights of both speakers. Both speakers' involvement reaches a peak of excitement roughly at the U320-U330 section, then gradually descends as speaker A gives more specific details in concluding the topic. The pitch levels of both speakers also converge and follow the same rise-fall pattern as interest in the topic increases and then is resolved. And this shows that there exists an overall systematic prosodic structure that integrates topic progression and speaker involvement through a process of climax and resolution.

What is happening in the conversation in Figure 5 is that one speaker (speaker B) begins to develop a topic that she is interested in but that had not been successfully communicated, and the interest level and the speaker's involvement are intensified as she attempts to overcome the mismatch, whereas speaker A's pitch movements are expressed in more uniform overall descending pattern. The descending part of the curve also coincides with the resolution of an issue that speaker B had been very uncertain about throughout that section of dialogue. This reinforces our view that at each level, prosody links speaker interactions, topic progression and cognitive state.



Figures 6a-b: Plots of 600 consecutive pitch peaks of both speakers. The upper plot represents speaker A, the lower plot 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, with 13-period and 50-period moving averages superimposed. Speaker B's pitch movements in the corresponding sections also reflect speaker involvement with the topic development.

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 with moving averages superimposed, shown in Figures 6a-b, we can see that phrases vary greatly in pitch height, but there are also arch-shaped patterns of a gradual rise of pitch level followed by a gradual fall. Such a pattern occurs from U275 to U500 (of B) over a sequence of about 225 utterances, which includes the sections discussed in examples 2-3.

We can see that there are about three extended rise-fall arches as speaker A develops different topics. These occur at approximately U100 to U275, U275 to U425, and U450 to U575, where our previous examples 1-3 also appear as prominent local structures in the short-term moving average. Note that even these arches are increasing in pitch height themselves. In Figure 6b, we can also see how speaker B's involvement with each topic is reflected largely synchronously in the corresponding sections. 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.

5. SUMMARY AND CONCLUSION

In this paper we have shown that topic structure in spontaneous conversation is manifested systematically at two interrelated higher levels. At the broadest level, our data exhibit long waves of pitch baseline rise and fall extending over large sequences of utterances, which are associated with general levels of psychological and interactive involvement with topic, and extended processes of climax and resolution. 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.

Our results show that prosody has critical importance in natural conversation, and that it is crucial that such systematic prosodic information present at global and local levels in spoken language be effectively incorporated in an integrated prosody and speech system to enhance performance in speech understanding projects such as topic detection and tracking, event segmentation, information extraction, emotion detection, and interactive spoken language systems.

6. ACKNOWLEDGEMENTS

We would like to thank Wallace Chafe of University of California at Santa Barbara, and Nick Campbell of ATR for their encouragement and support of this research. We would also like to thank Japan Science and Technology Agency for funding support.

7. REFERENCES

1. Campbell, W.N. 1996. "Synthesizing spontaneous speech, Sagisaka, Y., Campbell, W.N., and Norio, H., editors, *Computing prosody: computational models for processing spontaneous speech*, Springer-Verlag, 165-186.
2. Hirschberg, J. and Pierrehumbert, J. 1986. The Intonational Structuring of Discourse. 24th *ACL Proceedings*: 136-144.
3. Shriberg, E., Stolcke, A., Hakkani-Tur, D. & Tur, G. 2000. "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics", to appear in *Speech Communication* Vol. 32, Nos. 1-2.
4. Yang, L-C. 1995. *Intonational Structures of Mandarin Discourse*. Ph.D. Dissertation, Georgetown University.