INTERPRETING MEANING FROM CONTEXT: MODELING THE PROSODY OF DISCOURSE MARKERS IN SPEECH

Li-chiung Yang

Spoken Language Research Institute & Georgetown University

Abstract

In this paper we show how text, context and prosody of discourse markers reflect cognitive and discourse phenomena of uncertainty and certainty, intensity of emotional response, and interactive signals of knowledge state. We demonstrate how the subtle and finely differentiated meanings permeating spontaneous speech are communicated by prosodic variations and that it is the differences in shape that communicate the degree of uncertainty or certainty with respect to the speaker’s knowledge state, and cognitive status. We then propose a tripartite model of prosody and discourse markers to account for the contextual determination of discourse marker interpretation.

1. Introduction

1.1. Definition

In discourse, topic focus and the interest of participants are constantly changing, and many aspects of speech are involved in cooperatively coordinating the flow of topics and interests between participants. One class of words which are frequently used in this cooperative coordination are discourse markers. Discourse markers are words or phrases which are commonly used to signal the relationship between discourse units. They signal the relative status of discourse units by linking together phrases in a characteristic way. They include connectives and adverbials such as so, because, and, but, or, then, anyway, and now, feedback acknowledgments such as yes, yeah, unhum, and right, as well as particles such as ah, oh, well, hmmm, each marking the particular status of an utterance in the discourse.

1.2. Prosody and discourse markers

The importance of discourse markers arises because of their ability to provide immediate indicators of the salience of a communicative exchange in an effective and efficient way. Both prosody and discourse markers share the common function of bringing salience to communication, and this shared characteristic provides a strong motivation for exploring the link between prosodic form and discourse marker meaning. Prosody is critical in spoken language because it provides an additional dimension that communicates subtle and finely differentiated layers of meanings, and in some cases, prosody is the sole cue for disambiguating discourse structure and discourse interpretation.

2. Speech corpus

In this paper we investigate prosody and discourse markers based on a corpus of spontaneous conversations. We show how prosody contributes to the communication and interpretation of the multi-level meanings expressed in discourse markers, and suggest how the prosody of discourse markers fits within an overall system of discourse. Our approach differs from previous research in that we take an integrated approach of combining detailed qualitative discourse analysis with a quantitative corpus-based approach, utilizing both acoustic and discourse data. Our data consists of 6 hours of spontaneous conversations in Mandarin Chinese. The speech data were digitized and annotated for discourse relations, topic structure, discourse markers, and speaker turns. The acoustic measurements of f0, amplitude and duration were correlated with the specific characteristics of the discourse markers extracted from the corpus.

3. Multi-level meaning of discourse markers: how are they disambiguated?

Enumeration of discourse markers is very rich, and a great amount of progress has been accomplished in the exploring and in extensive description of the functions of discourse markers and discourse particles by researchers in the field (e.g., Schiffrin 1987, Chafe 1992, Fischer & Brandt-Pook 1998, among others). On the other hand, the prosodic functional correlates of discourse markers are not well-known or categorized, and discourse markers are underutilized in automatic speech processing, where their relationship signaling function shows great promise. In addition, the contribution of prosody to the differentiation of multi-functional discourse markers has not been estimated in terms of the relative contribution as compared with lexical meaning, syntax, and context. A key question in a prosodic approach to these problems is: What role does prosody play in distinguishing the lexical homophones, and how does prosody map different discourse markers to the same function? In addressing these problems, it is important to take a multi-disciplinary viewpoint, as valuable information has been obtained from many different approaches, and incorporate the best elements from each in an integrated optimal way.

3.1. Cognitive status, understanding, and degrees of agreement: the marker dui

In discourse, speakers often explicitly search for some indication of mutual agreement and understanding, and the hearer provides corresponding feedback or backchannels as expressions of understanding and interest. Both the speaker’s prompting and the hearer’s feedback are often done through the use of specific discourse markers such as um, unhum, uhuh, yeah, and right which are associated with specific pragmatic distinctions, marking new or old information, certainty and uncertainty, degree of comprehension, and emotional response. Such markers perform a specific signaling and monitoring function in the cooperative give-and-take of conversation, and
signal different nuances of participants’ intentions, as well as information on their reactions to the conversation flow.

One of the most important feedback markers in Mandarin discourse is the word *dui*, meaning ‘right’ or ‘yeah’. In conversation, *dui* frequently acts as a signifier of agreement and support, showing the speaker’s approval, judgment, and interest on the current topic. The expressive nuances that *dui* assumes are directly related to the specific interactive context, emotional intensity, and the simultaneous expression of other emotions that are consistent with *dui’s* underlying function. *Dui* is the most frequently occurring feedback word in our conversational corpus because it functions both as a signal of cooperation and as a marker of information receipt, two functions especially critical to a successful discourse.

Our data show that the speaker’s cognitive status and level of involvement are directly manifested in intonation. The degree of emotional intensity is commonly signaled by pitch range and pitch height. As shown in Figure 1, in the consecutive responses w-dui8, w-dui9, and w-dui10, the speaker is getting progressively more involved and her *duis* follow a corresponding progressively higher pitch pattern, from the gentle agreement of w-dui8 to the extreme level of emotional involvement and variation is also systematically indicated by the exaggerated emphasis of w-dui10. The intensity uniform stepwise increments in pitch level. Similarly, in Figure 2, both w-dui22 and w-dui23 are intense expressions but w-dui23 has a larger pitch range, exemplifying the speaker’s higher degree of emotional intensity. In contrast, w-dui24, an immediate follow-up confirmation of w-dui23, has a more gradual convex shape and a moderate pitch level and pitch range because of the speaker’s more normalized state.

Pitch shape characteristics such as *pitch slope* and *concavity* and *convexity* are very important features in distinguishing intonational meaning and cognitive states, with more moderate slopes associated with more tentativeness, and sharper falling slopes with greater definiteness. The concavity or convexity of slope is also critically related to the perceived degree of harshness or softness of the utterance, and these shape characteristics reflect the underlying expressive states that often arise from the discourse process itself. For example, w-dui13, w-dui14, and w-dui15 are shorter and have sharper slopes than w-dui1 and w-dui2. In this sequence, the speaker first started to express her opinion in w-dui13, but was interrupted, and in w-dui14 she restarts, so her pitch level is higher. In w-dui15, the speaker was just providing further confirmation after an explanation, so the *dui* here is low-pitched with a convex shape, corresponding to the more gentle agreement. W-dui17, w-dui18, and w-dui19 show a similar pattern of intensity variation, with a higher-pitched sharper slope *dui* perceived as more definite.

While most of the high pitched instances of *dui* in these two figures are intense and have a concave shape, the remaining *duis* in the mid and low pitch ranges exhibit mostly convex shapes. For example, w-dui2, w-dui3, w-dui5, and w-dui15 have a gradual pitch slope with a convex shape, expressing the speaker’s sympathetic understanding and confirmation. W-dui11, w-dui12 and w-dui31 are at the other extreme from harshness and definiteness and have the flattest slopes of all the instances. In these cases, the speaker reconfirms the hearer’s correct receipt of matter-of-fact information, and the neutral quality of these expressions is represented in the insignificant pitch range, low amplitude and pitch level, as well as the mild shape of these *duis*.

As markers such as *dui* function as critical interactive communicative elements, their specific pitch shapes can reflect both speaker-dependent characteristics and the specific role a speaker assumes in the local discourse or topic. The pitch shapes of *dui* for speaker K in Figure 3 exhibit an overall pattern that is clearly different than speaker W’s overall results, with a predominance of gradual convex shapes of moderately long duration; however, the similarity of speaker K’s *dui*’s to the convex pitch instances of speaker W suggest that the pattern exhibited may not be indicative of a consistent and distinct difference in pitch shape between speakers in similar contexts, but instead reflects different distributions of a range of emotions within the particular conversations examined, where similar individual emotions result in similar shapes. In this conversation, speaker K takes a more sympathetic, listener role in the topic, and this is expressed in the convexity and lengthening of pitch shapes for the majority of her *dui’s*. Such differences in distribution may certainly reflect speaker characteristics or presentation style, as mediated through a different distribution and range of emotional states. The consistency of forms across both speakers reflects such a process: convex shapes occur at medium to low frequencies for
both speakers, and consistently become more concave at higher pitch levels, indicating that, in this particular conversation, speaker K never attained the high emphatic emotion expressed by speaker W.

The examples presented here illustrate that discourse is a cooperative process, and that feedback markers play an essential role in the mutual influencing interactions of discourse participants, signaling cognitive, emotional and cooperative information flow. At each point in conversation, the hearer’s utterances provide important clues about their own evolving information states, their views of the topic under discussion, and the specific cognitive-emotional status, ensuring that the direction of the discourse flows in synchrony with both speaker’s states.

### 3.2. Dimensions of meaning: Surprise to dawning realization to acknowledgement

Intonation expresses fine gradations in meaning even when lexical information is largely absent, as in the case of the particle oh ‘o’ and its variant ah ‘a’, which communicate a range of uncertainty-based states, including doubt, surprise, acceptance, acknowledgement, and registering of information. Three basic patterns for oh are evident in the plot containing 28 instances of oh of one speaker in Figure 4. As shown, oh often expresses surprise in a rise-fall shape, with an arched and extended concave pattern communicating different intensities of dawning realization. It is the differences in shape, height, and duration that communicate the degree of uncertainty or certainty with respect to the speaker’s knowledge state, the intensity of emotion, and the effects of other co-occurring emotions.

Our data show that intense surprise causes a high rise in pitch. In contrast to the mild gradual arch shapes of dawning realization seen in s-oh13, s-oh15, and s-oh19, a sharper and narrower arch shape indicates the presence of surprise with co-occurring emotions, as in the high amazement of s-oh4 and the horror expressed in s-oh3. A lower pitch range often reflects acceptance and registering of information, with a lesser degree of surprise, as in s-oh11, s-oh13, and s-oh22, and a matter-of-fact acceptance of information that offers little challenge to the speaker’s knowledge state causes the pattern of nearly flat pitch slopes in s-oh10 and s-oh17. Emotions that are closely related to acceptance, such as sympathy and approval also tend to be expressed in a low pitch level.

S-oh1 and s-oh18 are at the other extreme of uncertainty, with rapid rises in pitch within a short time-frame exemplifying incomprehension, alertness, and a need for further information, in contrast to the completely realized acceptance of information accompanying more extended duration pitch shapes. The uncertainty in s-oh1 in particular stands out because of the convex steep rise to at a high pitch level, with nearly no subsequent fall, reinforcing the final incomprehension, while the moderate pitch and gradual rise of s-oh27 expresses the speaker’s doubt and heightened interest. By contrast, concavity of pitch shape is associated with greater comprehension, as in s-12, s-24 and s-28 which express surprise, interest and quick recognition upon encountering unexpected new information.

The functions of an interjection are closely related to the nature of discourse and individual speaking style. In contrast to the highly varying expressive obs of the previous speaker, the other speaker, speaker K’s obs in Figure 5 seem much more subdued and are much shorter on average as they function mostly as quick acknowledgement or quick recollection responses. Due to the particular nature of the discourse, there are many instances of reorientation, recalling, acknowledging, and sudden occurrence of ideas by the speaker, and this is why these particular shapes dominate, in contrast to the varied reactions to new information experienced by speaker S.

It is worth noting that even within all these short expressions of oh’s, there still exist finer variations in shape, height, range, direction, duration and intensity, and these variations are systematic and are related to interpolation, status of information, and emotional state. For example, there is the short falling type as in 3 and 16, slightly longer falling type as in 2, 6, 10 and 20, the curvy twist type as in 14, 17 and 22, the arch type as in 4, 9, and 11, and the more intense emotional type, characterized by wider pitch movement and longer duration, as seen in 1, 5, 7, and 18, for expressing the speaker’s disgust, sudden remembrance of an important or exciting event, and appreciative acknowledgment, respectively. How expressive a speaker is in a particular conversation also depends on the degree of topic relevance. This is exemplified in the striking extended oh of 8 expressing the speaker’s mock terror and protesting emotion on encountering an unexpected event.
4. A model of prosody and discourse markers

Based on consideration of the data from both analytical and experimental results of our research (Yang 1995, 2001), we propose a model of prosody and discourse markers to account for contextual determination of discourse marker meaning in spoken dialogue. Our model includes the following elements:

1) Identification of the elements of multi-functionality of discourse markers, that is, a classification of the major categories of functions performed by discourse markers. These include at least the following:

   a. Specific phrase relationships, e.g. continuation, specification, clarification, etc.
   b. Functions of the discourse marker when considered in its interactive role or as an attention marker
   c. Specific cognitive relationship to the speaker, such as different degrees of certainty or uncertainty.
   d. Specific emotional relationship of the phrase or phrases to the speaker, such as angry, happy, or worried, etc.

2) Identification of the factors that lead to a correct interpretation of the discourse marker. The model includes the following:

   a. Discourse context, referring to content of the conversation thus far.
   b. Position in phrase.
   c. Lexical meaning or conventional range of meaning usages for the specific discourse marker.
   d. Prosody, including pitch, amplitude, duration, and voice quality variations.
   e. Prosodic context, i.e., the knowledge of how the prosodic form in question fits into the prosodic context established so far.

3) Identification of the domain and extent of applicability of each of the factors in (2), that is, how much each factor in (2) contributes to a correct interpretation of each category of (1).

   Based upon analysis of our data, we postulate that discourse context, position in phrase, lexical meaning, and prosody all contribute to the correct interpretation of 1.a, the specific phrase relationship, and similarly for 1.b, but prosody plays a more prominent role in the case of 1.b. For both 1.c and 1.d, the correct interpretation of the specific cognitive or emotional relationship to the speaker, prosody plays the predominant role in our model, although lexical meaning and context will also help.

5. Conclusion and general implications

As elements located at the natural intersection of several different but complimentary relationship dimensions, discourse markers offer an intriguing and informative window from which to view discourse as a whole, where relationships among concepts, speakers, and referents are a predominant motivating force for discourse development. Relationships are expressed throughout discourse, whether through explicit linking of concepts to external referents, through differential weighting of words by emphasis, or by signaling of expressive or phrasal relationships. In our research on discourse, we consistently find that these relationships are expressed systematically by prosody. Because of their brevity and lexically limited character, discourse markers represent general prosodic principles in a crystalline and compact form, and as locus points for the signaling and interpretation of multi-level relationships, discourse markers provide an environment where the critical role and forcefulness of prosody in distinguishing and communicating meaning is fully achieved. Integrating such aspects of prosodic meaning into existing spoken language architectures therefore is an essential task to speech understanding projects such as conversational speech recognition and to the development of natural-sounding spoken language systems.

6. Acknowledgements

Portions of this work were done while I was a Visiting Researcher at ATR (Advanced Telecommunications Research Institute International). I would like to thank Japan Science and Technology Corporation for support of this research under the Expressive Speech Processing Project in the research area of “Information Processing for an Advanced Media Society”. In particular, I would like to thank Wallace Chafe for his support and encouragement of this research.

7. References