THE EXPRESSION AND RECOGNITION OF EMOTIONS THROUGH PROSODY

Li-chiung Yang 1,2

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

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

ABSTRACT

Emotion is an integral component of human speech, and prosody is the principle conveyer of the speaker's state. In this study we show how specific emotional states are expressed in the prosody of spontaneous speech. The significance of prosodic meaning to communicating judgements, attitudes, and the cognitive state of the speaker makes it essential to emotion-intention tracking and to natural-sounding synthesis systems.

1. INTRODUCTION

Recent research has pointed to the centrality of emotion in all aspects of human activity. In spoken language, emotion is indispensable in the ongoing communication of ideas, feelings, and judgements of importance towards topic and towards participants, and guides the exchange of information through mutual coordination of states of uncertainty and knowledge. Prosody is of paramount importance in this process in providing a forceful and flexible additional dimension that distinguishes and communicates often complex and finely differentiated layers of meanings, and is fundamental to progress in increasing intelligence and responsiveness in interactive systems and natural-sounding speech synthesis.

2. SPEECH CORPUS

2.1 Data and Approach

In this study we investigate the expression and recognition of emotion through prosody by acoustically and perceptually analyzing natural interactive discourse data. Our approach differs from previous research in that we take an integrated approach of combining acoustic data from spontaneous conversation and experimental data from perceptual tests, with the goal of providing a more unified account of prosodic manifestations of emotions in natural discourse. The corpus consists of 6 hours of recorded spontaneous conversation in Mandarin Chinese as well as 20 segments extracted from the corpus for the perceptual experiment (results to be published separately due to space).

2.2 Why Use Spontaneous Speech Data?

We believe that it is crucial to study emotion using spontaneous speech because it is only in such speech that we will encounter the complex emotions occurring in real life. This complexity arises because of the high degree of involvement in what is being communicated and the goal-directed motivations of the participants. In spontaneous speech, the high degree of interactive involvement is expressed in a rapidly varying stream of complex emotional states, in contrast to more constrained expressions of emotions present in controlled speech. These emotions can be highly varied and span a much wider range than typically recognized. In addition, the emotions expressed in spontaneous speech can be very subtle and finely differentiated, and can occur to varying degrees of intensity as well as combined together because of the large number of contextual variables simultaneously at work. To achieve human-like quality in intelligent interactive systems, it is crucial to understand and model how humans act in normal communicative situations.

3. EMOTION AND PROSODY

Specific emotional and cognitive states such as disbelief, doubt, complaining, incomprehension, and puzzlement, can contribute greatly to intonation, and have systematic influences on the shapes of intonation. Our data show that in general, states such as continuation, expectation, hesitation, and uncertainty have a raising and lengthening effect, while states such as definiteness, finality, and negativity have a lowering effect on pitch. In addition, the degree of tentativeness or definiteness of an expression is often correlated with the steepness of pitch slope. Another important consideration is that emotions occur to different degrees of intensity, intensity being some measure of physiological change, and the intensity may determine the magnitude of intonational influence in an utterance. A given intonational tune may need to be specified as to intensity to adequately determine and interpret the resulting pitch realization.

3.1 Degree of Emotional Intensity

We plot different instances of the same expression dui “right” extracted from our discourse corpus as an illustration to show how variations in pitch shape and intensity of a common expression, a marker of agreement in this case, can give rise to different perceptions of cognitive-emotional state. The point that pitch range and pitch height variation can be correlated with the degree of emotional intensity is demonstrated in the series of dui's that occurred as consecutive responses, w-dui1, w-dui2, and w-dui3. In this sequence, the speaker (w) is getting progressively more involved and her dui's follow a corresponding progressively higher pitch pattern. In contrast to w-dui1, the pitch level of w-dui2 is much higher. The intense agreement is also signalled by a larger pitch range, corresponding to the greater emotional intensity. W-dui3 shows the same speaker’s further intensification to an extreme level of emotional involvement, and this strong emphatic and exaggerated effect is evident in the dramatic pitch range, forcefully signaling the degree of intensity involved. Pitch level often correlates with pitch range variation. As seen, in w-dui1, 2, and 3, the speaker’s pitch level is raised in uniform stepwise increments, reaching an unusually high level.
and pitch level, as well as the mild shape of this in the insignificant pitch range, short duration, low amplitude instances here. The soft quality of this expression is represented dui11, a quick and light response, is at the other extreme from s-dui9 and s-dui10 are similar in shape, slope, and effect. S- with a convex shape, giving an impression of gentle agreement. convex shape. For example, k-dui8 has a gradual pitch slope general falling pattern, the remaining w-dui3, w-dui4 are less extreme than the previous exaggerated dui, exemplifying that the degree of strength of this expression is less intense. Similarly, w-dui5, a follow-up confirmation to w-dui4, has a more moderate pitch level, pitch range and amplitude, because of the more normalized emotional level.

We have found that concavity and convexity of pitch slope are very important features in distinguishing intonational meaning and are related to the perceived harshness or softness of the utterance. While all of the instances of dui in Figure 1 exhibit a general falling pattern, the remaining duis exhibit a generally convex shape. For example, k-dui8 has a gradual pitch slope with a convex shape, giving an impression of gentle agreement. s-dui9 and s-dui10 are similar in shape, slope, and effect. S-dui11, a quick and light response, is at the other extreme from harshness and definiteness and has the flattest slope of all the instances here. The soft quality of this expression is represented in the insignificant pitch range, short duration, low amplitude and pitch level, as well as the mild shape of this dui.

3.2 Emphasis and Different States

Different emotions and different types of emphasis can lead to dramatically different prosodic manifestations of the same expression. This is shown in Figure 2, where different tokens of the phrase hen duo “very many” from different contexts were plotted. As seen, in w-duo1, both hen ‘very’ and duo ‘many' have a level shape, but they differ in pitch level and duration. The speaker’s prominent focus on duo is signalled by both the lengthened duration and by the sustained sound quality.

The prosodic realization of a dramatic exaggerated expression can differ from a more identification or informational type of emphasis. Instead of the level shape seen in w-duo1, the duo in w-duo2 is greatly modified and has a dramatic rise-fall arch shape, due to the exaggerated and persuading emphasis of the speaker. Comparison of w-duo2 and w-duo3 further shows the effect of different emotional intensities and the progression of focus on prosody. In w-duo3, the speaker’s emotion is more settled, both the pitch level and magnitude of pitch movement have correspondingly decreased, and the rising slope is smaller, symbolically representing the diminishing force. This suggests that the steepness of the pitch slope is very critical, and is correlated with the strength of emotion.

Pitch shapes can take a drastically different pitch direction in the presence of a strong intonational force as shown in k-duo4. In this utterance, the speaker is emphasizing hen duo with a negative emotion, expressing her disapproval of ‘too many’. In contrast to the level and arch shaped duos seen previously, the duo here has a distinct and complex falling shape. The perceptible pause of .19 sec between hen and duo also enhances considerably the expression of disapproval present. As compared to k-duo4, the hen in k-duo5 drops sharply, with a much higher pitch level and a large pitch range, signalling the greater focus. The light accusing tone here gives duo a rising shape. An even more striking rising duo is seen in s-duo6, where strong emphasis on duo and high involvement combine to give the sharp rising pitch slope within a very short time frame, exemplifying the speaker’s urgent and intense state.

The examples presented show that focus and emphasis can perturb the pitch relationship and play an important role in the overall intonational system; however, the specific realization or manifestation of the pitch relationship will depend upon the particular emotional relationship as well. It is the specific underlying expressive nature of a particular focus which ultimately determines the actual realized pitch shape.

3.3 Certainty and Uncertainty

Prosodic signals of definiteness and tentativeness are pervasive in discourse and critical to the development of a conversation. The degree of a speaker’s certainty or uncertainty on the content of the conversation is an important element in the participants' success in communicating their ideas. We illustrate in Figure 3 how remarkably the expression of definiteness and tentativeness can affect the prosodic manifestation, and show how rapidly emotional-cognitive states can change in discourse.

In this speech segment, the speaker is repeating the same phrase fanzheng ‘in any case', but the two instances of fanzheng have radically different intonational realizations: the first fanzheng has a rising-falling shape, whereas the second fanzheng has a
falling-rising contour instead, reflecting two different states, with the first one more definite and the second one more tentative. The resulting change in the degree of certainty causes the second *fanzheng* to move in the opposite direction, taking on a dramatic and lengthened rise, with an elevated pitch level and a greater pitch range. The third *fanzheng*, which occurred in the subsequent phrase, is at a much lower pitch level, has a narrow pitch range, and is also shorter in duration, indicating the speaker’s normalized emotional state and her desire to immediately return to the original topic.

3.4 Complaining

In contrast to tentativeness, uncertainty, and hesitation, which have a raising and lengthening effect on pitch, negative states such as disapproving, discontent and disgust commonly have a lowering effect on pitch. Negativity not only is associated with a low pitch level, but can also affect pitch shapes dramatically, exhibiting downward pitch inflections, as seen previously. Our data also show that expressions of negative emotions can differ greatly according to the particular emotion present. Inward-directed emotions such as resentment or unwillingness often have low pitch and narrow range, while a high-pitched negative expression can indicate the presence of an outward directed emotion such as protest.

Figure 4 shows an example of a complaining emotion. There is a clear intonational break which divides the utterance into two distinct patterns. The pitch level of the initial phrase *keshi meiyou* ‘but there wasn’t any’ is high and the pitch range of these words is also narrow, about 40Hz, and both of these contribute to a marked sustained whining quality which captures the protesting and plaintive emotion present. These words are said at a faster speed because of relative deemphasis. In contrast, the phrase *lengqi* ‘air conditioning is greatly expanded in pitch range and duration because of the greater emphasis. The steep pitch drop, the great lengthening and contemptuous air voice quality combine to give this utterance its strong blaming complaining quality. The contrast and progression from the initial plaintive and protesting sustained high pitch level to the negative and blaming emphasized syllables together convey the combination of elements inherent in the nature of complaining.

3.5 Surprise

Surprise is closely related to tentativeness, uncertainty and doubt, and is typically characterized by a rise-fall shape. The cognitive conflict between the pre-existing belief and the newly encountered awareness that occurs in surprise reflects a strong element of doubt and uncertainty. But surprise also contains an acceptance and belief that the new knowledge is true. Prosodically, the doubt and uncertainty is manifested in an initial rising pitch shape, and the acceptance gives the contour a declining pitch, reflecting the certainty of realization. The doubt and uncertainty together with the ultimate acceptance may give surprise its characteristic rise-fall pitch contour.

The expression of surprise and disbelief is often associated with a high pitch level, as shown in Figure 5. As seen, the speaker starts at a more moderately high pitch level, but immediately rises to a very dramatic high peak, reaching the upper extreme of her range, and then descends to the bottom of her pitch range. The extreme high pitch level gives the expression a very intense and forceful character, and the large pitch movement from very high to very low creates a steep descending slope, contributing greatly to the impression of disbelief and astonishment. There is also a stepwise lowering from the first *meiyou* ‘there isn’t’ to the second *meiyou* and to the rest of the phrase. This gradual lowering in pitch signals the normalization process of an initial strong emotion and represents the certainty that accompanies the acceptance and cognitive acknowledgement that are inherent to surprise.

The high degree of surprise is not only indicated by the unusually high pitch level, but is also present in the steep rising slope of the two *meiyou*, of a dramatic expanded pitch range. The extent of the rise is highly expressive and is associated with the high degree of the abrupt, urgent, unexpected doubt present. The significance of the rising slope is evident when we compare *meiyou* in this phrase with another *meiyou* in the following phrase by another speaker. In contrast to the striking rise of about 200Hz previously, *meiyou* here is very light and mild, and has a very slight rise and slight fall shape, all within a minimum pitch range of 20Hz only.
3.6 Surprise to Matter-of-Fact

A change in degree of emotional intensity often changes the nature of the emotion itself, such as from the emotion of surprise to matter-of-fact acknowledgement. Figure 6 plots several pitch manifestations of the expression zhende “really”, which is often used to express surprise. Both the high patterns of extreme surprise in k-zhende1 and s-zhende2 and the more moderate surprise in s-zhende3-4 exhibit the rise-fall slopes that exemplify the pattern of uncertainty and acceptance that is characteristic of surprise. By contrast, s-zhend5 and k-zhende6 are similar in having lower pitch level and flatter slopes, and both express light or matter-of-fact acceptance or acknowledgement, the result of an already completed normalization process subsequent to strong surprise in each case. The paired similarities seen in Figure 6 correlate with the different levels of emotion expressed, and this may suggest that differences in shape and level provide a systematic categorization of intonational meanings either locally or globally based.

3.7 Surprise to Dawning Realization

Intonation expresses fine gradations in meaning even when lexical information is largely absent, as in the case of the particle oh ‘oh’, which, like zhende, communicates 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 22 instances of oh of one speaker in Figure 7. Like zhende, 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, as in s-oh2, s-oh4, and s-oh5. 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.

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.

4. CONCLUSION

In this paper we have shown that prosody and emotion interact in systematic ways so that participants successfully communicate the many levels of finely differentiated meaning present in conversational speech. Emotion is an integral component of human speech, and prosody is the principle conveyer of the speaker's state and hence is significant in recovering information. The significance of prosodic meaning to communicating judgements, attitudes, and the cognitive state of the speaker thus makes it essential to speech understanding projects such as emotion and intention tracking and to the development of natural-sounding spoken language systems.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES