Interactional and Pragmatics-Related Prosodic Patterns in Mandarin Dialog

Nigel G. Ward¹,², Yuanchao Li², Tianyu Zhao², Tatsuya Kawahara²

¹University of Texas at El Paso, ²Kyoto University

Abstract

The roles of prosody vary from language to language. In European languages prosody is largely involved in pragmatics, but this may be less true for other languages, especially tone languages. As a case study this paper examines Mandarin. Using telephone dialog data and a semi-automatic bottom-up analysis method based on Principal Components Analysis, we identify a dozen prosodic patterns in Mandarin which appear to have pragmatic and/or interactional significance. Examination of the overall fraction of prosodic variation explained by different factors also suggests that Mandarin uses prosody heavily for pragmatic functions in dialog.

Index Terms: conversation, spontaneous speech, tone, Chinese, superpositional modeling, unsupervised learning, prosodic constructions, dialog activities

1. Motivation

In European languages prosody is known to convey many pragmatic functions, and prosody-pragmatics mappings have been exploited in many applications, including turn-taking, user modeling, information retrieval, and speech recognition [1, 2, 3, 4, 5]. However tone languages may use prosody less for pragmatic purposes, and to the extent that this is the case, many techniques that use prosody will be less effective.

Whether this is true is an open question. On the one hand, it is often said that the heavy use of pitch for marking tone makes it less available for pragmatic functions, and indeed in tone languages such functions tend to be marked with particles. On the other hand, it is also said that in tone languages “the communicative use of sentence intonation seems to be as free as in nontone languages” [6], and for Mandarin specifically, the topic of this paper, prosodic correlates have been identified for several pragmatic purposes, including focusing, questioning, managing disfluency, marking topic boundaries, expressing emotion, and turn-taking [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Bottom-up work has also shown that there is much more to prosody in Mandarin than tone: in read speech “tone identities make up only 40-45% of output F0” [22].

This paper explores two questions:
1. What non-tonal functions does prosody have in Mandarin?
2. In Mandarin, what fraction of prosody is devoted to pragmatic functions, rather than tone?

We investigate using a method that supports the systematic examination of the major prosodic patterns of a language [23].

2. Data and Methods

As dialog is the realm of speech best suited for observing interactional and pragmatic functions, we chose to work with the CallFriend Mandarin corpus of telephone conversations [24]. We used 77 minutes of data: the first 15 minutes or so of files 0928, 4249, 4257, 4198, and 5975.

2.1. A Superpositional Model

We use a superpositional model, in which the observed prosody is assumed to be the result of summing the effects of simultaneously active underlying factors. Superpositional models have been used for many purposes, for many languages, as surveyed elsewhere [25, 23], including Mandarin [26, 27, 28, 29]. While superposition alone is clearly not adequate for modeling all aspects of prosody, a superpositional model can be a useful approximation.

Among the many possible ways to build a superpositional model, we chose Principal Components Analysis (PCA) because it has been useful in identifying pragmatics-related prosodic patterns in other languages [30]. Our strategy was to compute numerous low-level features at each point in the recordings, and then use PCA to discover the underlying patterns [25, 23]. Specifically, we computed features at 874,000 timepoints t, sampled evenly every 10 milliseconds throughout the dialogs, considering both tracks. At each timepoint these features characterized the activity in that vicinity, and all this data was fed to PCA to discover the underlying factors. Although PCA has some quirks, including a tendency to output factors with symmetric feature loadings, it has the advantage of simplicity.

This assumption-free strategy can be contrasted with that of [22] who started with a model of the “lower-level effects” of tone and then examined the residue when these were eliminated. We choose an assumption-free strategy because, despite many advances in the modeling of tone [31, 32], we are working with noisy and complex dialog data, where building a reasonable baseline model of the “lower-level effects” of tone would be an extreme challenge.

2.2. A Large Set of Time-Spread Features

To broadly characterize the pattern of activity in the vicinity of any point in time, t, we use features computed over windows that together span from about 3 seconds before t to 3 seconds after. The window sizes are roughly proportional to the distance from t, thus, for example, the most distant intensity window is 1.6 seconds long, from ~3.2 s to ~1.6 s, and the closest is 50 milliseconds long, from ~50 ms to 0 ms. There are windows for features of both speakers, enabling the discovery of joint patterns of prosodic behavior by both speakers, since the units of the two speakers are not generally aligned. We also do no time-warping or other stretch-
Table 1: Prosodic Feature Counts

<table>
<thead>
<tr>
<th>number of features</th>
<th>left track</th>
<th>right track</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>high pitch</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>low pitch</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>narrow pitch</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>wide pitch</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>intensity</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>creakiness</td>
<td>14</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>rate</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 1: Prosodic Feature Counts

Given the factors (dimensions) output by PCA, our next step was to distinguish those related to tone and those related to pragmatic functions. To do this we used a qualitative inductive method. At each timepoint in the data each factor has a value, which can be positive, negative, or zero. Most informative are the timepoints where a factor has an extreme value. For example, listening to timepoints where Dimension 4 had a high value, we were able to identify two patterns, one when the value on the dimension was positive and one when negative.

We also examined the loadings. For example, the features loading positively on Dimension 4 include the left-track speaker’s intensity over a short interval, as seen in the first line of Figure 1; this of course corresponds to the backchannel. The figure also shows a tendency for the right speaker’s pitch to go low about 300-400 milliseconds before the backchannel onset. (It is important to stress that the y-axes in these graphs are not based directly on the features; thus the line labeled “pitch height” is not directly the average or typical pitch height, but rather represents the difference between the loadings of the high-pitch and low-pitch features. Similarly the “pitch range” line shows the difference between the wide-pitch and narrow-pitch features. This is because simple pitch-height features are not robust enough, as noted earlier, and because of the use of PCA. Nevertheless human pitch is generally continuous, and in practice these lines always do indicate pitch contours that occur frequently in the data.)

Interpreting the patterns was time-consuming. For each we listened repeatedly to about a dozen examples of each, and considered the words said, their apparent pragmatic intention and effect, the larger context, and, of course, the tones. For some patterns we were able to find commonalities with few exceptions, but for others we were able to identify only tendencies or families of related functions. This is to be expected: in a superpositional model, the actual meaning is the sum of the meanings of all patterns simultaneously present, and at any specific timepoint the meaning of one pattern can obscure the contributions of another. Nevertheless, each of the descriptions below was valid for at least two thirds of the examples examined.
3. Prosodic Pattern and Interpretations

This section concisely describes some of the patterns observed. The complete loadings for each factor are at http://www.cs.utep.edu/nigel/mandarin/.

Factor 1 related to who was mostly talking.
1 positive talking The left-track person is silent while right-track person talks.
1 negative silent Conversely, the left-track person silent while the right-track person talks.

Factor 2 related to turn-taking, and was most strongly present, either positively or negatively, when one speaker stopped speaking and the other immediately started.
2 pos turn take Taking t = 0 as the middle of the turn transition, this pattern prototypically includes:
-1500 ms right-track person increases loudness, speaking rate, and creakiness
-800 ms right-track person begins to lower pitch
-400 ms right-track person’s loudness and pitch drop
+400 ms left-track person takes the turn, starting with moderately high pitch and creaky voice
+800 ms left-track person’s volume and pitch height abate, and the speaking rate increases
2 neg turn yield The converse pattern, where left-track person yields the turn and/or the right-track person takes the turn.

Factor 3 had, for both speakers, the same weightings for all features.
3p negative evaluation The speaker reports something that someone else has done or felt and assesses it as negative in some way. Within a region of relatively high pitch and fast speaking rate, the speaker shifts to use even higher pitch and increased volume and pitch range over about 1.5 seconds.
3n brief pause The speaker pauses for a second or so, while thinking how to express something.

Factor 4 reflected a joint behavior, as mentioned earlier.
4p new-information backchannel As seen in Figure 1, this pattern has several specific properties. Taking t = 0 as the center of the backchannel, and roughly noting the offset times seen, these include:
-800 ms: a region of increased volume and creakiness for a second or less, generally including the point of new information
-600 ms: a region of low pitch for a half second or less
-300ms: the end of the backchannel, starting moderately high and then falling in pitch, often some form of oh 哦
+300ms: the end of the backchannel
+500ms the original speaker’s resumed speech, starting with high volume and creakiness
4n new information and backchannel cue The converse pattern.

Factor 5 had loadings for both speakers the same.
5p no new information One or both speakers was speaking quickly, with lowish pitch, but providing no new information, often saying something that was already clear from the context. One speaker sometimes backchannelled, but, unlike those in Factor 4, these backchannels were quiet, short or fast, and low in pitch, for example ああ 哦，and were not aligned with the other speaker’s utterances.
5n topic progression One speaker introduced a new facet of the topic, for example progressing from talking about the other speaker’s interview to their own interview, or from talking about one person’s possible move to start a new job to the question of how her health condition could constrain the move. This pattern involved an overall slow speaking rate and a second or two of high pitch.

Factor 6 had opposite loadings for the two speakers.
6p high activity The left-track speaker continues speaking or starts or restarts a sentence, speaking quickly and with a short slightly higher pitch, while the right-track speaker is passive, at most producing a long backchannel or filler.
6n low activity The converse pattern, with the speaker roles reversed.

Factor 7 was prototypically a joint pattern, where one speaker bid for empathy and the other gave it.
7p empathy bid The right-track speaker bids for empathy, either showing exasperation or displeasure (signalled for example with a sigh like ai-ya 哎呀), or reporting being in a difficult situation, or inviting the other or asking a question and hoping for a favorable response. The pitch is high over a few seconds, and there is a sudden drop in loudness.
7n giving empathy The right-track speaker typically says something sympathetic or agreeable, with low pitch.

Factor 8 had loadings for both speakers the same.
8p engagement One or both speakers are engaged in a topic. Prosodically there is a second or less with moderate and falling volume, slight creakiness, and ending with a lower pitch. This sometimes involves modal particles such as うあ 哇，ba 吧，and a 啊.
8n topic exhaustion Both speakers are ready to end the current topic, and there is a short pause in the speech or a low-volume fast mumble.

Factor 9 had opposite loadings for the two speakers.
9p cue for response The left-side speaker asks the other a question, makes a suggestion, or otherwise invites or cues a short response. This involves a region of high creakiness and loudness, followed by a word with high pitch and increasing volume. The right-side speaker briefly answers, accepts, agrees or produces a backchannel to show understanding.
9n cued response The converse pattern.

Factor 10, with loadings for both speakers the same, comprises two apparently unrelated patterns.
10p contrast This involves a contrast, for example between some past situation and now, between one speaker’s location or situation and the other’s, or between an inference made by the listener and reality. This involves a sharp rise.
in pitch over about 800 milliseconds with a slight increase in creakiness, for example over the word fˇan zh `eng 反正 (anyway).

10n falling tone As seen in Figure 2, this involves a salient long pitch drop, and appears to involve emphasis of a word containing tone 4. There is also a slight decrease in volume and creakiness.

To mention briefly the apparent roles of two more dimensions: Factor 11 seems to involve detachment, where either the right or the left speaker is contributing only in a perfunctory way, without much warmth or interest. Factor 12 seems to involve some form of divergence, such as a difference of opinion or knowledge asymmetry between the speakers.

4. The Magnitude of Pragmatics-Related Prosody

Turning now to our second question, regarding the relative importance of the different functions of prosody, we first must acknowledge that any attempt to quantify this is rife with difficulties. Nevertheless, a very rough estimate can be obtained by considering the total amount of variation explained by factors with clear pragmatic functions, as summarized in Table 2. The sum of the variation in dimensions 2–12, excluding the half of dimension 10 that relates to tone, is 0.30. The total effective amount of variation, is not 1.0 but 0.77, excluding dimension 1, which relates mostly to the brute fact of which person is speaking. The former is 39% of the latter, and thus we estimate the pragmatic load of prosody as at least 39% of the total. For an even rougher estimate of the relative load of tone, we scanned the loadings of other dimensions and found that most factors below number 16 relate mostly to very localized pitch-height features; thus most of the remaining prosodic variation may be devoted to tone.

This is compatible with Tseng and Su’s finding that 55-60% of the variation is not explicable by individual tones alone, although the two results are not quite comparable, since we used dialog rather than read speech, we used unaligned features, and we included features beyond pitch height. While any such estimate must be regarded with skepticism, as it is dependent on the features chosen, these estimates do suggest that pragmatic functions are a major part of prosody’s role in Mandarin dialog.

5. Discussion and Summary

Interestingly, many of these pragmatic functions are also conveyed with prosodic patterns in other languages, and for at least two of the functions, empathy bids and backchannel cues, some of the prosodic components are the same also in English, Spanish, and Japanese. Investigation of possible universals is a prime topic for further research.

Another topic for future work would be to investigate these patterns using other methods. While our method, statistical pattern discovery plus subjective interpretation, is good for discovering new patterns and their possible significance, it is not adequate to establish either. Future work might reexamine the observed tendencies experimentally. Among other things, this might help pin down the exact timing and other properties of the components of these patterns, as our method gives only a blurred picture.

Nevertheless, our exploration suggests that Mandarin uses prosody for many pragmatic functions. Some details of the patterns found corroborate observations in the literature, for ex-

<table>
<thead>
<tr>
<th>var.</th>
<th>side</th>
<th>type</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23%</td>
<td>pos</td>
<td>neither</td>
</tr>
<tr>
<td>2</td>
<td>5%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>3</td>
<td>4%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>4</td>
<td>3%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>5</td>
<td>3%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>6</td>
<td>3%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>7</td>
<td>2%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>8</td>
<td>2%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>9</td>
<td>2%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td>10</td>
<td>2%</td>
<td>pos</td>
<td>prag</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tone</td>
</tr>
</tbody>
</table>

Table 2: Summary of Findings for the Top Ten Dimensions (Top 20 Patterns). Var. is the variance explained by each factor.

The sample regarding back-channeling, turn yielding, and turn starts [11, 9, 20]. Other patterns appear to convey functions not previously related to prosody in Mandarin, including negative evaluation, bids for empathy, topic progression, topic exhaustion, and contrast. We conclude that a significant fraction of the prosody of Mandarin dialog relates to pragmatic functions, and that developers interested in using prosody for various applications should not shy away from trying their methods for Mandarin.

6. Acknowledgments

This work was supported in part by a Fulbright Award and by DARPA under the Lorelei program. This work does not necessarily reflect the position of the Government, and no official endorsement should be inferred. We thank Wen Wang for comments.

7. References