Prosodic Profiles of Social Affects in Mandarin Chinese

Yan Lu1, Véronique Aubergé2, Albert Rilliard3

1 GIPSA Lab, CNRS, Stendhal University, Grenoble France; 2 LIG Lab, CNRS, Grenoble France 3 LIMSI-CNRS, Orsay, France

Yan.lu@gipsa-lab.grenoble-inp.fr, Veronique.Auberge@imag.fr, albert.rilliard@limsi.fr

Abstract

This work examines the production side of social affects in Mandarin Chinese, with the aim of extracting the more prominent patterns of acoustical variations. Results are then compared to previous perception data obtained on the same expressions. The \( F_0 \) intensity and duration characteristics of 76 utterances conveying 19 prosodic attitudes are statistically examined in this study. All attitudes are regrouped into 5 clusters according to their prosodic features. The result of the statistical analysis shows that the prominent differentiation between clusters is mostly related to \( F_0 \) and duration parameters; some similarities are noted between the clustering of attitudes from acoustic features and from perceptual confusions obtained in previous experiments; inside each cluster, some attitudes show typical characteristics in \( F_0 \) and duration.

Index Terms: prosodic attitudes, social affects, acoustic parameters, Mandarin Chinese

1. Introduction

Since the voice is considered as a carrier of affective signals in human speech, vocal cues, and especially prosodic cues, shall have an important role in the expression of affective nuances, which ensures some interaction functions like situation cues, mental states and processing, intentions, attitudes and emotions. [1] distinguished the socio-affective expressions (or expressions of attitude), which can be voluntarily controlled, from the expressions of emotion, which cannot be. Meanwhile, both emotional expressions and socio-affective expressions are often conveyed by the prosodic variations, which influence significantly the interpersonal interaction and social communication [2].

Many empirical studies demonstrated that people decode the acoustic signal conveying emotional expressions and attitudes with only voice samples (e.g. [3, 4, 5, 6]). On the other hand, many scholars have been engaged in finding out how affective signals are encoded in voice, with special focus on the acoustic measures of emotion encoding (e.g. [7, 8, 9, 10, 11]). The acoustic variables which have been widely measured in the literature are mostly fundamental frequency, energy (or intensity), duration (or speech rate), harmonics, stress, intonation, timbre, etc. Fundamental frequency, intensity and duration are the most classical acoustic parameters used as correlates of prosody [10], and it was commonly accepted that the fundamental frequency play an important role to signal affect, intention, or emotion [12].

Although [13] has claimed that the affective prosody universally exists in every language, there is also no denying that the expression of affective prosody varies from one language to another, and many studies have been conducted for the specific purpose of investigating the prosodic characteristics of affective speech specifically in Chinese (e.g. [14, 9, 11, 15, 16]).

Following the example of these previous studies, we will intend in the present work to identify the main prosodic profile of 19 social affects expressed in Mandarin Chinese by statistically separating them into different clusters. The characteristics of these expressions on \( F_0 \), intensity and duration parameters, both at the sentence and the syllable levels will be took into consideration. Meanwhile, this work also aims at finding explanations for the perceptual confusions observed between these social affects during previous perception experiment [6], because acoustic proximity is thought to be possibly explicative of certain confusions because they remain important cues for encoding and decoding studies [17].

2. Method

2.1. Corpus of Chinese social affects

Based on research on attitudes in Chinese and other languages [18, 19, 20, 3, 4], 19 Chinese daily encountered attitudes have been selected for our study. The speech corpus of this work contains four sentences performed with these 19 attitudes by one native Chinese female speaker speaking an unmarked standard Mandarin Chinese. The corpus has been perceptually validated in [6], where almost all attitudes have been recognized over the chance level (except “confidence”). The sentences analyzed in this paper received the best average recognition score across all attitudes in each length (monosyllable, disyllable, 4-syllable and 9-syllable). These sentences can be considered as the most representative of a prototypical expression of the targeted attitude. Table 1 lists the sentences composing the corpus and Table 2 presents the 19 Chinese attitudes which will be analyzed in the present work.

Table 1. Sentences composing the corpus.

<table>
<thead>
<tr>
<th>Chinese</th>
<th>Pinyin</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>姑妈休假</td>
<td>gu1 ma1 kua1</td>
<td>aunt will go on holiday.</td>
</tr>
<tr>
<td>张医生来了</td>
<td>Zhang1 y1 sheng1 lai2</td>
<td>Doctor Zhang will come.</td>
</tr>
<tr>
<td>王医生他三姑妈休假</td>
<td>Wang2 y1 sheng1 tal san1 gu1 ma1 xiu1 jia4</td>
<td>Doctor Wang’s third aunt will go on holiday.</td>
</tr>
<tr>
<td>学生</td>
<td>xue2 sheng1</td>
<td>student</td>
</tr>
<tr>
<td>教我来</td>
<td>jiao1 wo1 lai2</td>
<td>teach me</td>
</tr>
<tr>
<td>老师</td>
<td>lao1 shi1</td>
<td>teacher</td>
</tr>
<tr>
<td>医生</td>
<td>y1 sheng1</td>
<td>doctor</td>
</tr>
<tr>
<td>学校</td>
<td>xue2 giao2</td>
<td>school</td>
</tr>
<tr>
<td>教育</td>
<td>jiao1 yu4</td>
<td>education</td>
</tr>
</tbody>
</table>

10.21437/SpeechProsody.2014-13
2.2. Measurements of prosodic features

According to some studies on emotional expressiveness in Chinese and other tonal languages, the global intonation form of sentence is often linked to its expressive function [16; 21] and the variation of the initial and final movements of $F_0$ contour is more significant in characterizing the $F_0$ contour of different attitude than the movement in middle [22]. Therefore, we will observe in this work both global prosodic characteristics of social affects in sentence level and the specific cues at the end and beginning of sentence.

The majority of the values of prosodic parameters were automatically extracted from the 76 stimuli using the PRAAT software. Each sentence was previously hand-labeled with specific cues at the beginning and end of sentence. The SPSS software was used to carry out the statistical analysis. Table 2 shows the 10 prosodic parameters calculated on the sentence level and the variation of the initial and final movement were also measured, as well as the mean intensity of sentence. $F_0$ values were extracted by cross-correlation method and are expressed in semi-tone (with reference to 1 Hz). The intensity values are expressed in decibel (dB). The decimal logarithm of duration (in milliseconds) is used here [14]. Two other parameters were calculated manually in Excel:

- $F_0$ range (in semitones): the difference between the maximum and the minimum value of $F_0$ [16].
- $F_0$ slope (semitones/s): is here defined as the direction and rate of $F_0$ change. It is the difference of the mean $F_0$ of the last syllable to the mean $F_0$ of the first syllable divided by sentence duration [14].

The means for selected acoustic parameters of vocal utterances conveying 19 attitudes were calculated before being statistically analyzed. Table 3 shows the 10 prosodic variables extracted from the audio samples.

3. Results

3.1. Hierarchical clustering of attitudes based on principal component analysis

Combining a principal component analysis and an agglomerative hierarchical cluster analysis, this method has the advantage of clustering the individuals with less noise [23]. Therefore, a principle component analysis (hereafter referred to as PCA) was performed as a preprocessing step before the cluster analysis. After having measured the sampling adequacy with KMO & Bartlett’s test (KMO = 0.614; Bartlett’s Sig. < 0.001), the PCA was carried out on the dataset with 19 individuals (attitudes) and 10 prosodic variables. The measures were standardized during the procedure. The SPSS software was used to carry out the analysis.

The result of the PCA is presented in Table 4: two principal components were extracted, which explained cumulatively almost 82% of the variance. Table 5 shows the matrix of components after rotation (the “varimax rotation” method was used here). The main results indicate that the first principle component is more linked to $F_0$ parameters, intensity register and the duration of the last syllable, while the second one is linked to the duration parameters and the slope of $F_0$. The factor scores of the 19 attitudes on the two first components were then used as a new variable on which was performed a hierarchical cluster analysis based on an agglomerative procedure. The Ward’s minimum variance criterion was used to calculate the distance between clusters, while the squared Euclidean distance of observations was
defined as their distance. To determine the optimal number of clusters from the hierarchical tree, we referred to the “elbow criterion” based on the variance explained by each clusters [24]. Figure 1 shows the dendrogram of the hierarchical clustering and the graph of the between-inertia in function of the number of clusters on the top right corner. Observing the inertia graph, we thought it would be reasonable to consider 5 clusters.

![Dendrogram using Ward Linkage](image)

Figure 1. Dendrogram resulting from a hierarchical clustering of 19 social affects. The red line marks the clusters resulting in the observation of inertia reduction presented in the graph on the top right corner.

The last step of this analysis consists in visualizing the 5 clusters’ position on the two principal components, with the aim of highlighting how these attitudes disperse on these two dimensions, and of looking at how the acoustic dimensions allow separating or grouping them together.

As showed in Figure 2, the majority of attitudes are projected along the axis in the plot, while certain ones are in the extremity of the axis: “positive surprise”, “negative surprise” and “neutral surprise” are marked by their highest values on dimension 1, while “disappointment”, “resignation” and “confidence” the lowest values; “infant-directed-speech” shows the highest values in dimension 2, at the opposite of “intimacy”. The coordinates of some attitudes almost overlap on the graph, and it is the case of “positive surprise” and “negative surprise”, “disappointment” and “resignation”, “seduction” and “politeness”, “authority” and “declaration”. This proximity of distance between two attitudes implies the similarity of their prosodic features. That may be an important cue to explain some confusion patterns observed during the perceptual experiment.

As regards the 5 clusters obtained in hierarchical cluster, they are well separated on the two principal dimensions (cf. Figure 2):

- Cluster 1 groups “admiration”, “infant-directed speech”, “contempt” and “doubt”. These attitudes show some high values in both F0 and duration parameters.
- Cluster 2 groups “authority”, “declaration”, “resignation” and “disappointment”. On the contrary to the first group, they have low values in both F0 and duration parameters.
- Cluster 3 groups “obviousness”, “seduction”, “politeness” and “confidence”. These attitudes show some similarities with the members of cluster 2 in F0 phenomenon, but differ from them with their higher values in the dimension of duration.
- Cluster 4 groups “irony” and “intimacy”, which are quite separated from other attitudes and characterized by their extremely low values in duration parameters.
- Cluster 5 groups “positive surprise”, “negative surprise”, “neutral surprise”, “question” and “irritation”. They look more similar to the first group in F0 parameters, but quite different in duration dimension.

![Figure 2](image)

Figure 2. Representation of the 19 attitudes clustered in 5 groups on the first two principal components of the PCA. Numbers under labels refer to the cluster they belong to.

### 3.2. Differentiation of attitudes clusters

The analysis of the hierarchical clustering gives a global impression of the dispersion of attitudes and of their clustering according to their prosodic features. But it is also necessary to observe the prosodic characteristics of each cluster — in other words, how do the clusters differentiate prosodically one from another. Consequently, another analysis was done by comparing the means of each cluster across all variables. The results are detailed in Figure 3:

1. For the F0 measures (F0 range, F0 register of sentence, mean F0 of the first and the last syllable, F0 variation and F0 slope), cluster 1 and 5 have higher values, while cluster 2 and 3 have the lower ones. Cluster 5 displays the highest values on F0 range, F0 register of sentence, mean F0 of the first and the last syllable; cluster 1 has the highest value in F0 variation; cluster 2 has the lowest values on almost all F0 variables except F0 slope. Cluster 4 has the highest negative value of F0 slope. All clusters show slight differences in F0 register of sentence. On the other hand, it can be found that the difference between cluster 1 and 5 is linked to the duration of the first syllable and of the sentence, with higher values for cluster 1 than for cluster 5. Cluster 3 and 2 also present some differences in duration measures, with higher values for cluster 3 than for cluster 2, and they are especially different in duration of the last syllable.
(2). For duration measures (duration of the first syllable and the last syllable, sentence duration), cluster 1 shows the highest values, while cluster 2 and 4 shows the lowest ones. Cluster 3 and 5 are in between and the latter has a very high value of duration of the last syllable just next below cluster 1. It is worth noting that cluster 4 and 2 does not show remarkable difference in duration measures neither in F0 measures. Regarding the differences between cluster 3 and 5, we found that their differences mostly concern F0 measures: the values of cluster 5 are apparently higher than that of cluster 3.

(3). Concerning intensity measure (intensity register of sentence), no clear patterns of differences between clusters were found.

4. Discussion and Conclusions

In the present work, we investigated how the attitudes could be clustered according to their prosodic features by observing the acoustic parameters of F0, intensity and duration. The corpus of attitudinal speech contained 4 sentences of different length conveying everyone 19 Chinese attitudes and it had been perceptually validated during a previous experiment. Although the present study is still preliminary and the data involved is not large, some interesting and valuable results have been obtained.

First of all, the result of the hierarchical clustering ran on principal components gives a separation of attitudes into two main groups (cf. Figure 1) and this separation is basically based on the characteristics of fundamental frequency of attitudes (cf. Figure 2). One group is composed of attitudes which have high pitch level (e.g. “positive surprise” and “admiration”) and large pitch span (e.g. “question” and “doubt”); the other is composed of attitudes whose pitch level and pitch span are lower and narrower, (e.g. “declaration” and “politeness”). Such a higher and wider pitch span for surprise and admiration may be related to hypothesis of the “effort code” postulated by Gussenhoven [25]. A similar separation between the studied attitudes has been found in the perception experiment where the main distinction was observed between “assertive” attitudes like “politeness” and “obviousness”, as well as the confusion between “disappointment” and “resignation”. Of course, some perceptual similarities remain unexplained with only these acoustic cues: for example, the confusions between “infant-directed speech” and “seduction”, “authority” and “irritation”. In these cases, one analysis of voice quality seems necessary, because voice quality, as the fourth dimension of prosody [26], is an important aspect of the affective expression.

The F0 slope did not exhibit important difference across clusters, except for cluster 4 (“irony” and “intimacy”) which is distinguished by its highest value. All of the clusters are homogeneous in intensity register. Hence, we can summarise the salient prosodic features of each attitude cluster essentially in function of their F0 and duration profile: the cluster 1 (“admiration”, “infant-directed speech”, “contempt” and “doubt”) shows a very long duration; on the contrary, the cluster 4 (“irony” and “intimacy”) shows the lowest duration values and the highest for F0 slope. The cluster 5 (“positive surprise”, “negative surprise”, “neutral surprise”, “question” and “irritation”) is typically marked by large F0 range and high F0 level, while the cluster 2 (“authority”, “declaration”, “resignation” and “disappointment”) by lower F0 values. Cluster 3 (“obviousness”, “seduction”, “politeness” and “confidence”) is characterized by low F0 values and in particular the lowest F0 slope.

Some differences inside clusters also deserve our attention: “infant-directed speech” shows the longest duration and “intimacy” the shortest duration; positive, negative and neutral surprises are marked by their high F0 values while “resignation” and “disappointment” show the lowest F0 values of all attitudes; “confidence” differs from the other attitudes of cluster 3 by lower F0 values and a longer duration.

Another acoustic analysis about voice quality of the same audio samples is under way in order to investigate the potential influence of voice quality to the perception of the attitudes in question.

5. Acknowledgements

This work is partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

![Figure 3. Mean values of the five clusters across all prosodic parameters. Numbers represent the clusters; bars the mean values for a given parameter.](image-url)
6. References


