# Perceptual Experiment and Acoustic Analysis of Chinese attitudes: A Preliminary Study

Ping Tang and Wentao Gu

Nanjing Normal University, China
tangp1990@gmail.com, wtgu@njnu.edu.cn

## ABSTRACT

The present work studied 13 types of Chinese attitudes and compared the confusion patterns between subjective and objective identification.

The listening experiment showed that the overall recognition rate for these attitudinal utterances was 46% by native subjects, while in discriminant analysis, the recognition rate was 25.4% based on five prosodic parameters (minimum f0, maximum f0, mean f0, f0 standard deviation, and speaking rate).

Cluster analysis of the subjective and objective confusion patterns showed some similarities between them. For example, "friendly" and "polite", "hostile" and "rude" were confused in both subjective and objective identification. However, some attitudes were prosodically similar, e.g., "neutral" and "sincere", but the subjects were able to distinguish them. On the contrary, neutral utterances were prosodically different from submissive utterances, while subjects still confused large amount of submissive utterances to "neutral".

**Keywords**: attitudes, prosody, Mandarin Chinese, cluster analysis, discriminant analysis

## 1. INTRODUCTION

The interests in affective speech have burgeoned over the past decades [1-4]. However, affective speech can actually be divided into emotional speech and attitudinal speech, while most previous studies paid attention on emotional speech rather than on attitudinal speech.

One feature of attitude is, unlike emotions which usually do not constitute bipolar pairs, attitudes can be bipolar, i.e., polite and rude. In previous works, Gu and Fujisaki [4] defined nine attitude pairs for Mandarin Chinese (i.e., friendly/hostile, polite/rude, serious/joking, dominant/submissive, sincere/ insincere, praising/sarcastic, willing/reluctant, confident/uncertain, concerned/indifferent), within each of which there are two opposite attitudes [4, 5]. As a follow-up study, the present study aims to investigate how native subjects perceive Mandarin utterances conveying these attitudes. Besides, since prosody plays an important role in expressing affective speech [6], the prosodic cues for these attitudinal utterances were also examined.

## 2. MATERIALS AND METHOD

### 2.1 Attitude categories

In the present study, we selected six attitude pairs:

Pair 1: Friendly vs. Hostile;
Pair 2: Polite vs. Rude;
Pair 3: Serious vs. Joking;
Pair 4: Praising vs. Sarcastic;
Pair 5: Dominant vs. Submissive;
Pair 6: Sincere vs. Insincere.

Within each attitude pair, we designed ten target sentences (6-10 syllable long), which were literally neutral (i.e., not containing any words that lexically imply a specific attitude or emotion) but at the same time can be expressed in opposite attitudes when embedded in different dialogues. For each target sentence, we designed two dialogues to elicit two opposite attitudes.

### 2.2 Attitudes elicitation

Two speakers (1 male and 1 female), who were both 24 years old graduate students at Nanjing Normal University, were recruited to produce the attitudinal utterances. Both of the speakers are native Chinese and they have the experience of public speaking, good at expressing themselves.

A role-play dialogue in a given scenario was designed for each utterance to elicit the target attitude effectively. Altogether, there were 260 utterances (13 types * 10 utterances * 2 speakers).

### 2.3 Perceptual and acoustic study

12 subjects (6 males and 6 females), who were all 22-25 years old graduate students at Nanjing Normal University, were recruited in the perceptual experiment.

There were 280 stimuli in total (20 stimuli for training and 260 for analysis) in the perceptual experiment. After one stimulus was presented, the

subject was asked to identify which attitude was just played by choosing one from 13 options.

Five acoustic parameters (i.e., f0Min, f0Max, f0Range, f0Mean and Speaking rate) of the whole utterance were extracted. To eliminate speaker difference, we chose a speaker's averaged minimum f0 among all utterances in neutral condition as the reference f0 (f0Ref) of that speaker. The f0 parameters were standardized as follows:

f0Min = (f0MinObserved − f0Ref) / f0Ref,

f0Max = (f0MaxObserved − f0Ref) / f0Ref,

f0Mean = (f0MeanObserved − f0Ref) / f0Ref,

f0RangeNorm = f0Max − f0Min.

## 3. RESULTS

### 3.1 Perceptual experiment

The overall recognition rate was 46.4%, six times higher than the chance level (7%); the recognition rates of each attitude are presented in the descending order in Fig. 1. Almost all utterances within each attitude were recognized above chance level (7.7% as indicated by the dash line).

**Figure 1**: Recognition rates for the 13 attitudes.

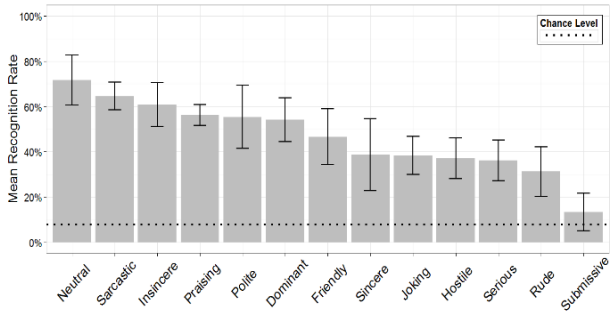


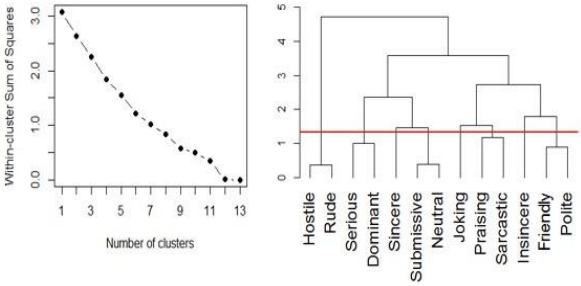

**Table 1**: Recognition patterns by native subjects for 13 attitudes.

| Attitude target | Percentages of responses | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fri | Pol | Pra | Jok | Sin | Ser | Hos | Rud | Bla | Ins | Com | Sug | Neu |
| Fri | **47** | 18 | 0 | 3 | 8 | 0 | 4 | 4 | 0 | 1 | 1 | 5 | 10 |
| Pol | 12 | **55** | 0 | 0 | 5 | 2 | 2 | 2 | 0 | 2 | 1 | 10 | 10 |
| Pra | 8 | 5 | **56** | 3 | 9 | 3 | 0 | 0 | 13 | 0 | 0 | 0 | 3 |
| Jok | 13 | 2 | 5 | **38** | 11 | 5 | 1 | 1 | 9 | 1 | 5 | 2 | 7 |
| Sin | 9 | 9 | 3 | 1 | **38** | 13 | 2 | 0 | 0 | 6 | 0 | 0 | 19 |
| Ser | 3 | 2 | 5 | 1 | 20 | **36** | 2 | 0 | 3 | 0 | 14 | 2 | 11 |
| Hos | 3 | 0 | 0 | 0 | 1 | 3 | **37** | **36** | 8 | 4 | 3 | 0 | 5 |
| Rud | 0 | 1 | 0 | 1 | 2 | 8 | **33** | **31** | 0 | 9 | 6 | 1 | 9 |
| Bla | 1 | 0 | 9 | 4 | 5 | 5 | 3 | 4 | **65** | 3 | 0 | 0 | 3 |
| Ins | 8 | 6 | 0 | 1 | 10 | 2 | 2 | 0 | 1 | **60** | 0 | 0 | 10 |
| Com | 0 | 0 | 0 | 0 | 0 | **23** | 5 | 3 | 0 | 0 | **54** | 0 | 15 |
| Sug | 4 | 15 | 0 | 0 | 8 | 6 | 0 | 0 | 0 | 2 | 8 | **13** | **44** |
| Neu | 2 | 8 | 2 | 0 | 3 | 4 | 2 | 0 | 3 | 2 | 2 | 1 | **72** |

A one-way ANOVA was conducted on the mean recognition rates as a function of attitude type. Result of the analysis indicated the effect of attitude type on recognition accuracy was significant when analysed by subjects, $Fs(12, 143)=11.535$, $p<0.001$, and by items, $Fi(1, 12)=110.15$, $p<0.001$. The recognition patterns for all attitudes are presented in Table 1, and the correctly recognized rates above three times of the chance level are highlighted with bold texts (attitudes are listed in the short form).

A hierarchy cluster analysis was conducted on the data in Table 1. The mean confusion rate between attitudes was referred as r, where 1-r was referred to as the perceived distance between attitudes. To optimize the number of clusters in the tree, the sum-of-squares explained by different numbers of clusters is presented in Figure 2, which suggests that the tree is proper to be divided into eight groups.

**Figure 2**: Hierarchy clustering of the perceived attitudes.




(1) "Hostile" and "rude": both of them were associated with negative and threatening signals, expressed an uncooperative intention in a drastic way. Therefore, they can be easily distinguished from other attitudes but hardly from one another.

(2) "Serious" and "dominant": "dominant" could be considered to show a higher social status, which were mostly been conveyed by a serious tone. In Lu's study, "authority" was largely confused with "irritation", where "serious" could be considered as a low-level irritation as well [7].

(3) "Sincere": although 19% sincere utterances were confused to "neutral", "sincere" has been isolated. Because two third of sincere utterances were confused to other attitudes, making it not clustered with any other single attitude.

(4) "Submissive" and "neutral": in order to avoid offending, speakers' submissive utterances were usually expressed with polite tones (i.e., 15% of submissive expression was confused for "polite"), or at least, neutral tones, while the later strategy seemed to be used more by Chinese speakers in the present study to convey "submissive" (44% of submissive expression was recognized as "neutral").

(5) “Joking”: no attitude was largely confused to “joking”, and it was not highly confused to any single attitude, making it isolated from others.

(6) “Praising” and “sarcastic”: “praising” indicated a positive evaluation, while “sarcastic” indicated a negative evaluation but was expressed by a semantic-praising and pragmatic-blaming style. Therefore, it was not easy to distinguish sarcastic from praising without any context.

(7) “Insincere”: “insincere” indicated that the speaker was unwilling or reluctant to do something or making a hypocritical promise, with a perfunctory tone, which was quite different from other utterances.

(8) “Friendly” and “polite”: both the two attitudes were expressed to show cooperative intention and were supposed to be characterized by breathy voice [7-8]. Besides, friendly and polite utterances were prosodic-unique, and could be well differentiated from other attitudes by prosodic cues (see 3.2.1).

### 3.2 Acoustic analysis

#### *3.2.1 Acoustic profile*

Averaged values of normalized f0 parameters and speaking rates are presented in Table 2.

To characterize the acoustic features of different attitudinal types, a one-factor MANOVA was conducted on attitudinal utterances. The “acoustic parameters” (5 parameters) served as dependent variables, and the “attitude type” (13 types) served as independent variable.

Results of the MANOVA were statistically significant according to Wilks’ Lambda (0.390), F (48, 1036) = 5.429, $p < 0.001$. Between-subjects tests showed that the effects of “attitude type” was significant on all five acoustic parameters with different effect sizes. To be specific, the influence of Attitude type was significant on SpRate: F(12, 248) = 9.896, $p<0.001$, $\eta^2 = 0.325$; f0Mean: F(12, 248) = 6.242, $p<0.001$, $\eta^2 = 0.233$; f0Max: F(12, 248) = 5.988, $p<0.001$, $\eta^2 = 0.225$; f0Min: F(12, 248) = 4.866, $p<0.001$, $\eta^2 = 0.191$; and f0Range: F(12, 248) = 4.332, $p<0.001$, $\eta^2 = 0.174$.

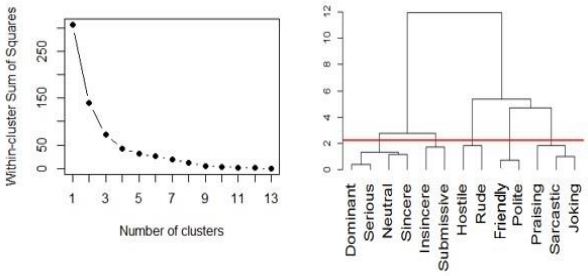
**Table 2**: Normalized fundamental frequency parameters and speaking rates (syllables/ second).

| f0Min | | f0Max | | f0Range | | f0Mean | | SpRate | |
|---|---|---|---|---|---|---|---|---|---|
| Rud | 0.25 | Hos | 2.42 | Hos | 2.31 | Hos | 1.13 | Rud | 6.81 |
| Fri | 0.21 | Rud | 2.29 | Iro | 2.15 | Rud | 1.10 | Hos | 6.65 |
| Pra | 0.19 | Iro | 2.24 | Rud | 2.04 | Pra | 0.97 | Pol | 6.29 |
| Pol | 0.17 | Jok | 2.19 | Jok | 2.02 | Iro | 0.96 | Fri | 6.25 |
| Jok | 0.17 | Pra | 2.06 | Pra | 1.87 | Jok | 0.95 | Pra | 5.86 |
| Ins | 0.11 | Pol | 1.85 | Sug | 1.76 | Pol | 0.87 | Sin | 5.86 |
| Hos | 0.11 | Sug | 1.83 | Sin | 1.73 | Fri | 0.84 | Com | 5.76 |
| Iro | 0.09 | Fri | 1.75 | Ser | 1.71 | Sug | 0.83 | Ser | 5.71 |
| Sug | 0.07 | Ser | 1.72 | Pol | 1.68 | Ins | 0.73 | Neu | 5.44 |
| Com | 0.02 | Com | 1.65 | Com | 1.63 | Neu | 0.70 | Iro | 5.39 |
| Ser | 0.01 | Sin | 1.64 | Neu | 1.58 | Ser | 0.69 | Sug | 5.29 |
| Neu | -0.04 | Ins | 1.56 | Fri | 1.55 | Com | 0.68 | Ins | 5.24 |
| Sin | -0.09 | Neu | 1.53 | Ins | 1.44 | Sin | 0.66 | Jok | 5.23 |

#### *3.2.2 Hierarchy cluster analysis based on acoustic parameters*

The hierarchy cluster analysis was conducted again on the acoustic parameters. The variance explained by different numbers of clusters suggested that the tree could be divided into five cluster, as cut at the red line in Figure 3.

**Figure 3**: Hierarchy clustering by 5 acoustic parameters.




(1) “Dominant”, “serious”, “neutral” and “sincere”: all these attitudes were related to low f0Min, f0Max and f0Mean, moderate f0Range and speaking rates.

(2) “Insincere” and “submissive”: both of them were associated with moderate f0Mean values and slow speaking rates.

(3) “Hostile” and “rude”: both attitudes showed the highest f0 values and fastest speaking rates, which made them easily be distinguished from others.

(4) “Friendly” and “polite”: utterances of friendly and polite were spoken with high f0Min and moderate f0Max, thus a low f0 variation, and fast speaking rates.

(5) “Praising”, “sarcastic” and “joking”: utterances of these attitudes were characterized by high f0Max, f0Mean and f0Range, which were lower only than utterances of hostile and rude.

#### *3.2.3 Discriminant analysis*

A discriminant analysis was conducted to estimate how well the five acoustic parameters can categorize the 13 Chinese attitudes. Pooled within-groups correlation test showed a high correlation between “f0Max” and “f0Range”. Besides, “f0Max” failed in the variables tolerance test ($p < 0.001$) and had been rejected in further analysis.

The discriminant analysis produced 3 significant canonical functions: the first function explained 73.2% of total variances and correlated positively with SpRate (r = 0.66); the second function accounted for 13.7% of variances and correlated positively with f0Min (r = 0.71) and f0Mean (r = 0.61), and correlated negatively with SpRate (r = -0.675); the third function explained 10.8% of variances and correlated positively with f0Range (r = 0.9).

In sum, the four acoustic parameters adopted by this model led to an accurate classification of 25.4% of the original tokens, and the classification rate varied across attitude types. In particular, 60% of rude speech, 50% of insincere utterances, 45% of hostile utterances and 40% of neutral utterances were correctly predicted by this model, while 25% of praising utterances, 20% of sincere, serious and sarcastic utterances were accurately categorized. In contrast, only 15% of friendly, polite and joking utterances were correctly classified. However, only 5% of the submissive utterances were classified and no dominant utterances were correctly classified.

## 4. DISCUSSION AND CONCLUSION

Obviously, the classification rates in discriminant analysis for all attitude types except "rude" and "hostile" were lower than the subject's recognition.

There are some similarities between the results of subjective and objective identification, i.e., "hostile" and "rude" had been isolated from other attitudes but confused with each other by native subjects and acoustic parameters, so did "friendly" and "polite", "serious" and "dominant".

However, "praising" and "sarcastic" could not be well distinguished by native subjects, while "joking", "sarcastic" and "praising" were not well differentiated by acoustic parameters. The distinction between "praising" and "sarcastic" could be cued not only by prosodic features but also by the match/mismatch between word expression and contextual situation. Therefore, when an utterance is dissociated from the context, attitudes of this kind may not be inferred reliably. In addition, the five acoustic parameters might not be adequate to characterize sarcastic from joking and praising, i.e., research indicated that the HNR, f0 standard deviation and overall reductions in mean f0 appeared reliable to distinguish sarcasm from sincere and humours [9].

Even though "submissive" and "insincere" shared similar f0 cues and speaking rates, subjects still isolated "insincere", because the way Mandarin speakers express insincere attitude is unique, i.e., subject of many "insincere" utterances was lengthened with a break between the subject and the predicate. In the present study, the average durational ratio of subjects to objects is 1.3 in insincere utterances, 1.0 in neutral utterances and 0.96 in submissive utterances.

On the contrary, the acoustic analysis showed that it is "sincere" rather than "submissive" that was more acoustically similar to "neutral", while subjects still confused a high proportion of submissive utterances for "neutral". For semantic meaning, "submissive" is more likely to be expressed in a polite tone (this is probably true to a certain extent, 15% of submissive utterances were confused for "polite"), while in the present study, although the corpus was well controlled, Mandarin speakers were still more likely to express submissive attitude in a neutral way.

In conclusion, native subjects could well categorize most attitude types defined in the present study, while prosody cues alone could categorize much fewer attitude types. Besides, the cluster analysis showed that native speakers tended not to use prosody information alone to perceive attitudes.

In the future, an additional experiment will be needed to test the cognitive distance between the terms of these attitudes because the confusions among attitudes may be due to the similarity in prosody and to the similarity on concept as well.

## 5. ACKNOWLEDGMENT

This work is supported jointly by the National Social Science Fund of China (10CYY009), the Major Program for the National Social Science Fund of China (13&ZD189), and the key project funded by the Jiangsu Higher Institutions' Key Research Base for Philosophy and Social Sciences (2010JDXM024).

## 6. REFERENCES

[1] Scherer, K. R. 1986. Vocal affect expression: a review and a model for future research. *Psychological Bulletin* 99, 143–165.

[2] Pell, M. D. 2001. Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America* 109, 1668–1680.

[3] Auberg é, V., Cathiard, M. 2003. Can we hear the prosody of smile? *Speech Communication* 40, 87–97.

[4] Gu, W., Fujisaki, H., 2013. Data Acquisition and Prosodic Analysis for Mandarin Attitudinal Speech. In: Peng, G., Shi, F. (eds.), *Eastward Flows the Great River: Festschrift in Honor of Professor William.* City University of Hong Kong Press, 483–500.

[5] Gu, W., Zhang, T., Fujisaki, H., 2011. Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes. *Proc. 12th INTERSPEECH*, Florence, 1069–1072.

[6] Hancil, S. 2009. *The Role of Prosody in Affective speech*. New York: Peter Lang.

[7] Lu, Y., Aubergé, V., Rilliard, A. 2012. Do you hear my attitude? Prosodic perception of social affects in Mandarin. *Proc. 6th Speech Prosody* Shanghai 685–688.

[8] Fangxin, C., Aijun, L. Haibo, W., Tianqing, W., Qiang, F., 2004. Acoustic analysis of friendly speech. *Proc. 29th ICASSP* Montreal, 569–572.

[9] Cheang, H. S., Pell, M. D. 2008. The sound of sarcasm. *Speech Communication* 50, 366–381.