Perceptual Study for Emotional Speech of Mandarin Chinese

Ting Wang[1], *Hongwei Ding*[1,2], *Wentao Gu*[2]

[1]School of Foreign Languages, Tongji University, China
[2]Research Center for Language Information Technologies, Nanjing Normal University, China
{2011ting_wang, hongwei.ding}@tongji.edu.cn; wtgu@njnu.edu.cn

Abstract

This paper conducts a set of perceptual experiments together with an F_0 acoustic analysis on emotional speech of Mandarin Chinese. Sixty utterances spoken by two actors in five basic emotions (i.e. happiness, fear, anger, sadness, and boredom) as well as in the neutral style served as stimuli. The perceptual experiments showed the following results: (1) For Mandarin Chinese, the rates of identification for five basic emotions differ significantly, ranking in Sadness > Happiness > Anger > Fear > Boredom. (2) Perceptual confusion occurs mainly between fear and sadness, and between boredom and anger. There is a perceptual boundary within each pair. The similarity coefficient is 0.16 for fear and sadness, and 0.22 for boredom and anger. (3) The accuracy in the perception of emotions varies with the linguistic background of the listener; non-native subjects show a degraded perception. Moreover, acoustic analysis implies that F_0 feature is not the only cue for the perception of emotions in Mandarin Chinese.

Index Terms: Mandarin emotional speech, Perceptual boundary, Cross-linguistic perception

1. Introduction

Spoken language in daily-life communication carries not only linguistic information, but also nonlinguistic information such as the speaker's age, gender, social status, emotions, etc. Emotions conveyed in speech reflect the speaker's physiological and psychological states. In daily conversation, people may often understand what a speaker says but misunderstand the emotion in his/her expression, and vice versa. An in-depth perceptual study on emotional speech is necessary.

A number of studies have been conducted on emotional speech. For instance, Frank Enos and Julia Hirschberg offered an approach and for eliciting emotional speech [1]. Liscombe et al. conducted a study examining acoustic features of emotional speech and their use in automatic classification of emotional speech [2]. Douglas-Cowie et al. addressed four main issues in developing databases of emotional speech: scope, naturalness, context and descriptors [3]. Dang et al. conducted a comparative experiment on emotion perception among listeners from Japan, the United States and China. The target language was Japanese. The results from this study suggested that a wide range of human emotions might fall into a small set of basic emotions [4]. Abelin and Allwood conducted a cross-linguistic perceptual study using Swedish utterances. The results showed that emotions were interpreted with different degrees of success depending on the mother tongue of the listeners [5].

Among these previous studies, there are relatively few researches on the perception of emotional speech of Mandarin Chinese. Because Chinese is a tone language in which there is a complex interaction between tone and intonation, it is worthwhile to investigate the perception of emotional Mandarin speech by Chinese natives as well as by the subjects who are native in non-tone languages. In the current study, a set of perceptual experiments was conducted to study the following issues: (1) the rates of identification for different emotions; (2) the confusing patterns between emotions; (3) the perceptual boundaries between similar emotions; (4) cross-linguistic perceptual patterns. Meanwhile, the differences in F_0 features among five emotions were also analyzed.

The remaining parts of this paper are organized as follows. In Section 2, emotional speech material and listening subjects are described. Section 3 presents a set of perceptual experiments. In Section 4, an acoustic analysis on F_0 is given. Concluding remarks are given in Section 5.

2. Stimuli and Subjects

Five basic emotions were investigated in the current work: happiness, fear, anger, sadness and boredom. Neutral speech was also involved in the study for comparison.

2.1. Stimuli

Since linguistic information may have effect on the perception of emotions, we designed five sentences that are literally not associated with any particular emotion. All five sentences are composed of three disyllabic prosodic words - the first two words are fixed and only the third word varies in tones. The five sentences are listed below:

Table 1: *Target sentences*

Sentence 1	zhe4 shi4 ta1 de0 qi4 che1 (This is his car)		
Prosodic words	zhe4 shi4	ta1 de0	qi4 che1
Sentence 2	zhe4 shi4 ta1 de0 zhao4 pian4 (This is his picture)		
Prosodic words	zhe4 shi4	ta1 de0	zhao4 pian4
Sentence 3	zhe4 shi4 ta1 de0 xin1 fang2 (This is his new house)		
Prosodic words	zhe4 shi4	ta1 de0	xin1 fang2
Sentence 4	zhe4 shi4 ta1 de0 dian4 nao3 (This is his computer)		
Prosodic words	zhe4 shi4	ta1 de0	dian4 nao3
Sentence 5	zhe4 shi4 ta1 de0 xue2 xiao4 (This is his school)		
Prosodic words	zhe4 shi4	ta1 de0	xue2 xiao4

The speech material was collected by the method of acting. Two speakers (one female and one male) are employed to record the utterances. Both of them are professional actors skilled at emotional expression. The speech was recorded at a sampling rate of 16 kHz with a 16-bit precision. The two speakers uttered every sentence in each of the five emotions as well as in a neutral style, thus producing 60 utterances in total. Before recording each utterance, the speakers were shown a picture with prompt text eliciting the particular emotion. The

recording was monitored, and the speakers were asked to repeat an utterance until the recording quality was satisfying.

2.2. Subjects

Two groups of subjects were recruited for the perceptual tests on emotional speech. The first group (LG 1) consists of 20 Chinese natives (14 females and 6 males); none of them has any impair in hearing or emotion comprehension. The second group (LG 2) consists of 20 non-native subjects, half of whom are native in English and the other half are native in German; none of them can speak nor understand Mandarin Chinese.

3. Perceptual Experiments

Three experiments were conducted in this study.

3.1. Experiment 1: Perceptual pattern

The method of constant stimuli was adopted as the test paradigm. The 60 utterances (stimuli) were randomized and then presented to the subjects in LG 1 through headphones. The subjects were requested to answer the perceived emotion of each utterance, which could be selected from 'happiness', 'fear', 'anger', 'sadness', 'boredom', 'neutral', or 'uncertain' when failing to judge.

Figure 1 shows the rates of identification for happiness, fear, anger, sadness and boredom, which are ranked as follows:

Sadness > Happiness > Anger > Fear > Boredom

As a result, about 82.5% of happiness were identified correctly by native subjects in LG 1, and 64.5% for anger and 89.5% for sadness. The rates of identification are only 59.5% and 56%, for fear and boredom, respectively.

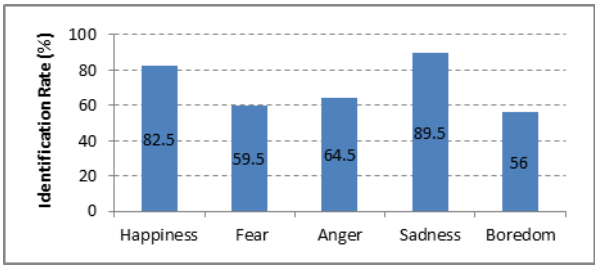

Figure 1: *Rates of identification for the five emotions.*

Sadness gives the highest rate of identification while boredom gives the lowest rate. This is consistent with the finding by Wang [6]. The rates of identification are significantly different among five basic emotions ($p < 0.05$).

Exp.1 also revealed the confusing pattern in perceving different emotions, as shown in Figure 2. In particular, fear is interpreted as sadness in a percentage of 12%, while sadness as fear in a percentage of 8.5%; anger is interpreted as boredom in a percentage of 24%, while boredom as anger in a percentage of 19.5%.

Thus, two pairs of emotions tend to be mixed up, i.e. fear vs. sadness, and boredom vs. anger. Exp.2 was conducted to further investigate the perceptual boundaries between fear and sadness, as well as between boredom and anger.

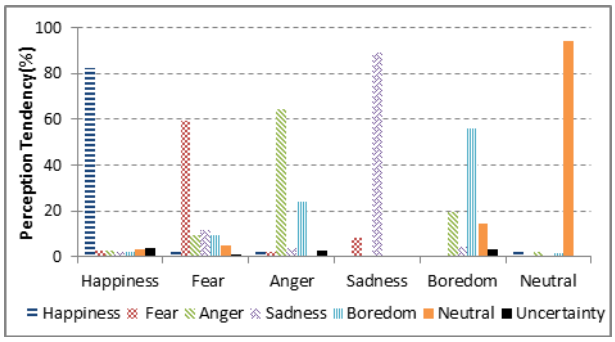

Figure 2: *Perceptual pattern by native Chinese subjects.*

3.2. Experiment 2: Perceptual boundaries

Assume a pair of sounds (A and B) with two different emotions. Normalize the duration of the two sounds first and then create a continuum of 11 sounds that changing gradually from A to B. The continuum is formed by scaling the amplitudes of the two sounds prior to the addition with weights that shift from A to B by using the toolkit of Praat. The continuum of 11 sounds was used as the stimulus.

The 11 sounds were randomized and then presented to the subjects in LG 1 for their judgment of emotions. The purpose of this experiment is to find the perceptual threshold between two different emotions.

Figure 3 shows the perceptual result for sadness-fear continuum. At the average F_0 of 21.8 St to 21.9 St, subjects' interpretations shift suddenly from sadness to fear. This is the perception boundary between sadness and fear. There exists a linear coefficient for the perception states between sadness and fear, which we mark as S (Similarity). The red trendline in Figure 3 illustrates that the coefficient S is 0.16.

For the continuum of boredom and anger in Figure 4, the perceptual boundary is between 22.81 St and 23.44 St. The red trendline shows that the coefficient S is 0.22.

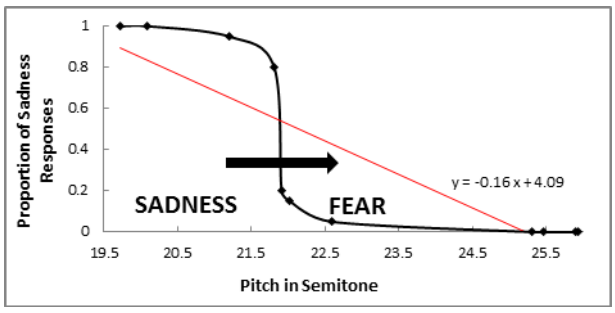

Figure 3: *Perception boundary between Sadness and Fear.*

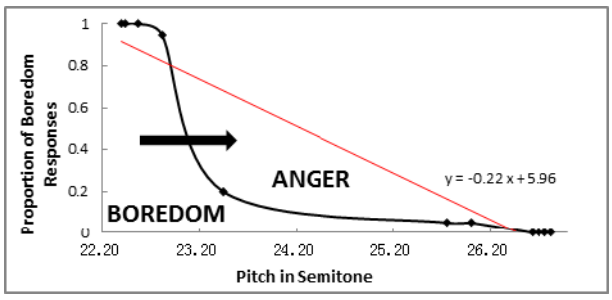

Figure 4: *Perception boundary between Boredom and Anger.*

3.3. Experiment 3: Cross-linguistic perception of emotions

In Exp.3, we conducted a comparative experiment on emotion perception across different linguistic backgrounds. The same perceptual test was done for LG 2, i.e. non-Chinese subjects whose native language is either English or German.

As can be observed from Figures 5 and 6, for non-native subjects, emotions in Mandarin speech are perceived with a lower accuracy. In Figure 5 for English subjects, sadness is perceived with the highest rate of 66.7%, while boredom with the lowest rate of 44.4%. The confusion occurs mainly between fear and sadness (bidirectional), and from boredom to anger (unidirectional). In Figure 6 for German subjects, sadness is perceived the best while boredom is perceived the worst; there are two major confusing pairs, i.e., fear vs. sadness, and boredom vs. anger. Figure 7 further shows the rates of identification for all three groups of subjects with different linguistic backgrounds.

A cross-linguistic comparison shows the following results:
(1) Generally, non-native subjects perceive vocal emotions less accurately than native subjects.
(2) Chinese, English, and German subjects share a similar ranking in the rates of identification of five emotions (highest for sadness and lowest for boredom), except for happiness. This indicates that subjects of different linguistic backgrounds share some common characteristics in the perception of vocal emotions, which is not surprising because emotion is nonlinguistic information.
(3) The most prominent degradation in the perception of emotions introduced by non-native subjects lies in the much larger confusion between fear and sadness. Further study is needed to find out whether this is a general case.

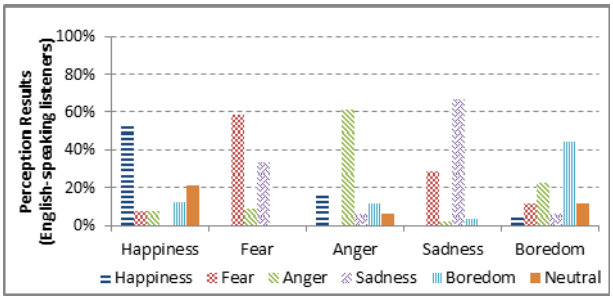

Figure 5: *Perceptual pattern by English subjects.*

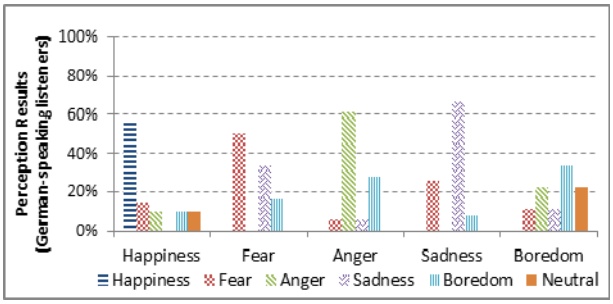

Figure 6: *Perceptual pattern by German subjects.*

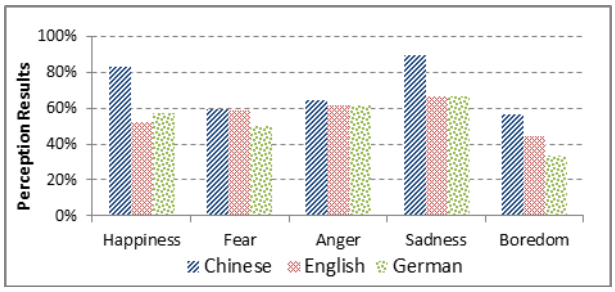

Figure 7: *Rate of identification of emotions by Chinese, English, and German subjects.*

4. Acoustic Analysis of F_0 features

The above experiments show the perceptual characteristics for vocal emotions. Here we further look into the acoustic features of emotional speech. Since F_0 is widely regarded as an important acoustic feature for emotional speech, we conducted an acoustic experiment to investigate the manifestation of F_0 in Mandarin emotional speech.

Among all 60 utterances, only 36 utterances with an identification rate over 80% in Exp. 1 were used for acoustic analysis. Average F_0 and F_0 range of the selected sentences were measured. The F_0 values were extracted using Praat and then manually corrected. All F_0 values are measured in semitone.

$$F_{st} = 12 * \log_2(f / f_r) \quad (1)$$

where F_{st} is the F_0 value in semitone, f is the F_0 value in Hz, and f_r is the reference value which is set at 55Hz for male and 80 Hz for female.

Table 2 summarizes the F_0 statistics for the speech in five basic emotions as well as in the neutral style. The reference value is the average F_0 of neutral speech. Happiness shows a very high average F_0, while sadness gives the lowest average F_0. The average F_0 is ranked as follows: happiness > fear > anger > boredom > sadness. This order is similar to the findings by Paeschke [7], Abelin [5] and Yuan [8]. The highest F_0 level occurs for fear, followed by happiness, sadness, anger and boredom. The order of pitch range is anger > happiness > boredom > sadness > fear, which is similar to the results by Yuan [8] and Gu & Lee [9].

According to the above acoustic analysis, happiness and fear are similar in having higher average F_0 and higher pitch level, while boredom and sadness are similar in having lower average F_0 and medium pitch range. Here comes to the conclusion that F_0 feature is not the only cue for the perception of similar emotions.

Table 2: *Statistics of F_0 features (St).*

Emotion	Average F_0	F_0 range	Mean Difference
Neutral	19.54	12.34	ReferenceValue=19.54
Happiness	26.40	14.35	6.86
Fear	25.80	7.87	6.26
Anger	23.60	15.55	4.06
Boredom	22.24	14.24	2.70
Sadness	21.74	10.09	2.20

5. Conclusions

In the current study, three perceptual experiments and an acoustic experiment were conducted for Mandarin speech in five basic emotions.

The first perceptual experiment on native Chinese subjects shows that the rates of identification for five basic emotions differ significantly and are ranked as follows: Sadness > Happiness > Anger > Fear > Boredom. The perceptual results with the five-emotion categories also imply a higher confusion between fear and sadness, as well as between boredom and anger.

The second experiment studies the perceptual boundary between fear and sadness, as well as between boredom and anger. For fear and sadness, at the average F_0 of 21.8 St to 21.9 St, subjects' interpretations change from sadness to fear. The similarity coefficient is 0.16. The similar pair of boredom and anger differs from each other at the average F_0 of 22.81 St to 23.44 St with a similarity coefficient 0.22.

The third perceptual experiment investigates cross-linguistic perception of emotional Mandarin speech. It shows that emotions are perceived with different degrees of success depending on the linguistic backgrounds of the listeners; native listeners no doubt give the best accuracy. For both native and non-native subjects, perception is the best for sadness and is the worst for boredom, and perceptual confusion occurs mainly between fear and sadness, and between boredom and anger.

A further acoustic analysis implies that F_0 feature is not the only cue for distinguishing emotions in Mandarin Chinese.

In further work, we need to conduct the perceptual experiments on a larger corpus of emotional speech, and more acoustic features need also to be investigated.

6. Acknowledgements

This research was supported jointly by the Innovation Program of Shanghai Municipal Education Commission (No. 12ZS030), the Jiangsu Social Science Fund (09YYB006), the Chinese National Social Science Fund (10CYY009), and the key project entitled 'Cross-linguistic Comparison of Speech Prosody, and Error Analysis and Automatic Assessment of Second Language Prosody' (2010JDXM024) from Jiangsu Higher Institutions' Key Research Base for Philosophy and Social Sciences.

7. References

[1] Enos, F., and Hirschberg, J., "A framework for eliciting emotional speech: Capitalizing on the actor's process", International Conference on Language Resources and Evaluation, 6–10, 2006.

[2] Liscombe, J., Venditti, J., and Hirschberg, J., "Classifying Subject Ratings of Emotional Speech Using Acoustic Features", EUROSPEECH, 725-728, 2003.

[3] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P., "Emotional speech: towards a new generation of databases", Speech Communication, 40: 33-60, 2003.

[4] Dang, J. and Li, A. et al., "Comparison of emotion perception among different cultures", Acoustical Science and Technology, 31(6): 394-402, 2010.

[5] Abelin, A. and Allwood, J., "Cross linguistic interpretation of emotional prosody", Proc. ISCA Workshop on Speech and Emotion, 110-113, 2000.

[6] Wang, H. and Li, A., "The construction of emotion corpus and perception experiments", Report of Phonetic Research, 2003.

[7] Paeschke, A., Kienast, A.M. and Sendlmeier, W.F., "F0-contours in emotional speech", Proceedings of the 14th International Congress of Phonetic Sciences, 1999.

[8] Yuan, J., Shen, L. and Chen, F., "The acoustic analysis of anger, fear, joy and sadness in Chinese", Proc. 7th International Conference on Spoken Language Processing, 2025-2028, 2002.

[9] Gu, W. and Lee, T., "Quantitative Analysis of F0 Contours of Emotional Speech of Mandarin", Proc. 6th ISCA Speech Synthesis Workshop, 228-233, 2007.