MULTIMODAL PERCEPTION OF PRAISING AND BLAMING MANDARIN SPEECH BETWEEN THE INTERLOCUTORS WITH FRIENDLY OR HOSTILE RELATIONSHIPS

Shanpeng Li[1] and Wentao Gu[1,2]

[1]School of Chinese Language and Literature, Nanjing Normal University
[2]Nanjing Normal University and The George Washington University Joint Laboratory of Speech, Hearing and Rehabilitation Sciences
lspqdu@foxmail.com, wtgu@njnu.edu.cn

ABSTRACT

This work investigated the role of multimodality and relationship between interlocutors in perception of Mandarin praising and blaming attitudes. The perceptual experiment found that the differences of accuracy between praising and blaming attitude were larger in friendly relationship than that in hostile. Moreover, praising attitude was identified better in audio channel than that in visual channel, while blaming attitude was discriminated more in visual channel than that in audio channel. This suggests that the hostile relationship between interlocutors was conveyed mainly by facial expression, either in praising or blaming speech.
Keywords: multimodal speech perception, speaker relationship, attitude, praising, blaming.

1. INTRODUCTION

Although there have been a number of studies on the processing of emotions, little has been known about how the audio and visual channels contribute to the perception of social affects such as praising and blaming. Moreover, while most studies of emotions and social affects are focused on English, French, or Japanese, there are fewer studies on Mandarin. Therefore, this work is aimed to the perception of Mandarin attitudinal speech.

The perception of attitudinal speech is affected by many factors, such as sentence length, modality, and linguistic background of the listener. Lu et al. [6] found that sentence length could affect the identification rate of attitudes in Mandarin: the shorter sentence leads to the lower perceptual identification rate, except for the infant-directed speech. The effect of sentence length was also found in Japanese [10], though less regular. Gu et al. [4] found that the valence of attitudes also played roles in the perception of Mandarin attitudes. They reported that the positive attitude had more accuracy than the negative attitude only in the praising-blaming pairs. On the contrast, other attitude pairs, such as friendly-hostile, polite-rude, serious-joking, and confident-uncertain had a higher accuracy in negative than positive. Apart from these linguistic factors referred in previous studies, non-linguistic cues, like the role of relationship of interlocutor, has received less attention, where we just want to focus on.

Many studies of attitudes have been conducted (cf. [1, 2, 8]), including some cross-linguistic studies (cf. [11]), but most studies only dealt with the audio modality. Different modalities of stimuli can also change the result of perception. Specifically, audio and visual signals can transmit speech and facial expression about affects. When listeners identify attitudes, they can use different modalities to make a decision, like tone of voice, facial expression, and body gesture.

de Moraes et al. [3] investigated the role of multimodality in the perception of Brazilian Portuguese attitudes. Their results showed that for propositional attitudes, audio and visual modalities were equivalent, except for irony, while for social attitudes, audio played a less important role, receiving significantly lower scores than audio-visual and visual for all attitudes except seduction and politeness. Also, Hönemann et al. [5] found that, in German, audio played more important roles than audio-visual modality in declarative sentence and irritation.

The aim of this work is to study whether the relationships between interlocutors can influence the multimodal perception of praising and blaming attitudes in Mandarin speech. In the present study, we only examine one relationship between interlocutors, i.e., friendly and hostile. When speaker and listener are friendly, we called praising attitude as friendly praising while the blaming speech as friendly blaming. When speaker and listener are hostile, we called praising attitude as hostile praising while blame speech as hostile blaming. So totally, we discuss these four speech styles (2 relationships * 2 attitudes) to investigate the role of speaker relationship in the multimodal perception of praising and blaming Mandarin speech.

2. SPEECH MATERIALS

2.1. Speakers

Twelve Mandarin-speaking undergraduate students (6 M and 6F) from Nanjing Normal University participated in speech recording. Their mean age was 21.6 years (SD = 0.3). All speakers majored in broadcasting, and had some experiences in art performance, so they were expected to express very typical attitudes. All speakers received reasonable payments for their recording.

2.2. Materials

We designed twelve sentences, each of which had a conversational background and dialogues for each speech styles, so that speakers can immerse themselves in an affective situation and perform nearly natural and real affects. Here are some examples of target sentences:

“头一次见到像你这么刻苦的人！” (“I have never seen a person who works harder than you!”)

“你真是见多识广啊！” (“You really have wide knowledge and experience! ”)

2.3 Recording procedure

Prior to recording, all speakers were explained for these four speaking styles, and they had enough time to familiarize with the materials. During the recording, speakers sat in front of a display in a soundproof room, as showed in Figure 1. There was a cardioid microphone (Neumann U87Ai) placed 40 cm from their mouth for recording. The microphone was connected to an audio interface RME Fireface 800, which was connected to a computer outside the soundproof room.

There were also two cameras behind the display for video capture. One was an HD webcam (Logitech C310), collecting the interlocutor’s face to make a face-to-face online conversation. The other was a professional video recorder (HDR-PJCX510E), collecting the speaker’s facial expression.

Hand claps between each recording blocks recorded both by the camera and the microphone, allowed a post-processing that replacing the camera sound with the high-quality sound recorded by the Microphone, synchronize with the claps in Adobe Premiere 2.1.

All sound clips were recorded at 44.1 kHz, 16 bits, mono track, and all video clips were encoded with a 784 * 576 pixels’ resolution, MPEG-4 coded format and 50 frames per second in AVI files.

Figure 1: Recording setting.

3. PERCEPTION EXPERIMENT

3.1. Listeners

Sixteen Mandarin participants (8M and 8F) were recruited to the perceptual experiment. They were all graduate students at Nanjing Normal University, with mean age of 24.9 years (SD = .94). All of the listeners had normal hearing and no experiences on performance. They were reasonably paid for their participation.

3.2. Experimental procedure

Perceptual stimuli modalities had three conditions, namely audio-only, visual-only and audio-visual, presented by E-Prime 2.0. Each condition had 576 stimuli (12 speakers * 12 sentences * 4 speech styles). All stimuli were randomized within speakers for three conditions.

All participants were explained for the meaning of this four speaking styles, and took part in a practice test before the formal experiment to make sure that they had understood the perceptual task. In order to eliminate the disturbances among three modalities, all subjects listened audio-only and watched visual-only stimuli on the first day, then watched audio-visual stimuli almost at the same time on the second day.

After each stimulus was presented, the participants had five seconds to make a forced choice among five labels (“Friendly Praising”, “Hostile Praising”, “Friendly Blaming”, “Hostile Blaming”, and “I don’t know”). The next stimulus would be presented if subjects made a choice or five seconds passed.

3.3. Statistical analysis

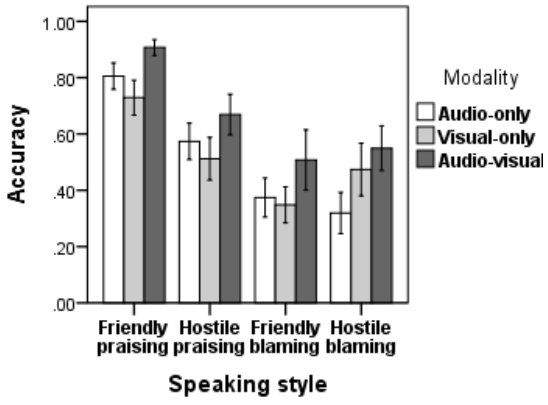
The rate of identification was analysed using the repeated-measured ANOVAs in SPSS 20.0. When there was a significant effect, a Bonferroni post hoc was further conducted for multiple comparison.

Before ANOVAs, we conducted a Friedman test for between-subject reliability, and it showed that there was no difference (Chi-square = 15.475, p = .162) among subjects. Then, we conducted an ANOVAs analysis with modalities, relationships (friendly, hostile) and attitudes (praising, blaming) as within-subject factors, gender as the between-subject factor to investigate how the speaker relationship affected multimodal perception of praising and blaming.

4. RESULTS

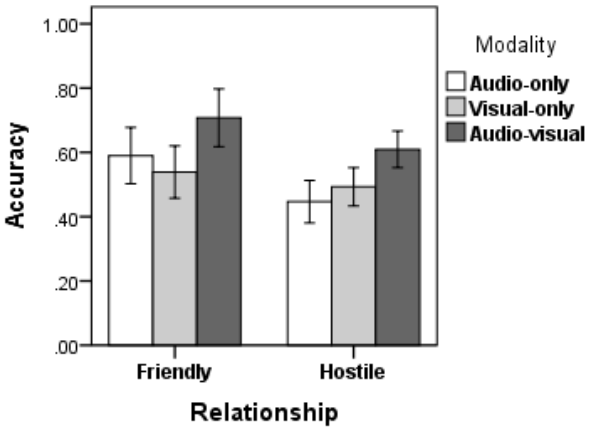
Confusion matrix for perceptual accuracy of four speaking styles by 16 participants was provided as Supplement data. According to the ANOVA analysis, Modality interacted with speech style significantly (F (6, 84) = 7.131, p < .001, η^2 = .337). Pairwise comparisons revealed that, as showed in Figure 2, friendly praising; hostile praising and hostile praising all had the had the lowest accuracy (72.9%, 51.2%, and 34.8%, respectively) in visual-only modality, while hostile blaming had the lowest accuracy (31.9%) in audio-only modality. Moreover, the accuracies in audio-visual were significant higher than that in audio only and visual only for friendly praising, hostile praising and friendly blaming, except for hostile blaming, where there was no significant difference between visual-only (47.4%) and audio-visual (54.9%).

Figure2: The mean and standard error of identification rates for speaking styles in three modalities.



Modality and relationship had a significant interaction effect (F (2, 28) = 7.825, p < .05, η^2= .359). As showed in Fig. 3, the identification rates of both friendly and hostile relationships in audio-only (59%, 44.7%) and visual-only (53.9%, 49.3%) were significantly lower than in audio-visual (70.7%, 60.9%), but the difference between audio-only and visual-only was not significant. Remarkably, the friendly relationship was identified better in audio-only than that in visual-only, while

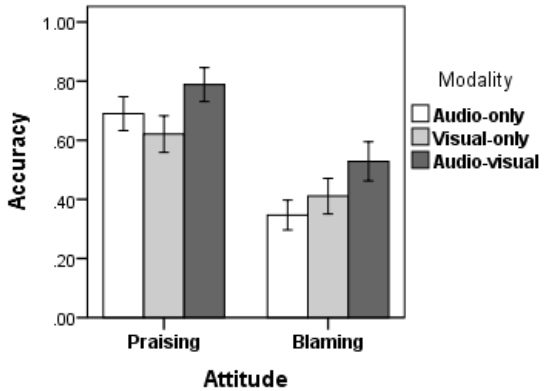
Figure 3: The mean and the standard error of identification rates for relationships in three modalities.



the hostile relationship had a higher identification rate in visual-only than that in audio-only.

Modality also interacted with attitude significantly (F (2, 28) = 9.742, p < .001, η^2 = .41). Post hoc analysis showed that the accuracy of praising in audio-only (69%), visual-only (62.1%) and audio-visual (78.8%) were significantly different between each other. Similarly, the accuracies of blaming in audio-only (34.7%), visual-only (41.1%) and audio-visual (52.9%) also showed significant difference between each other, even though the differences between audio-only and visual-only were marginally significant (p = .082). More importantly, as showed in Figure 4, the praising had significantly higher accuracy in audio-only than in visual-only, while the blaming had marginally higher accuracy in visual-only than that in audio-only.

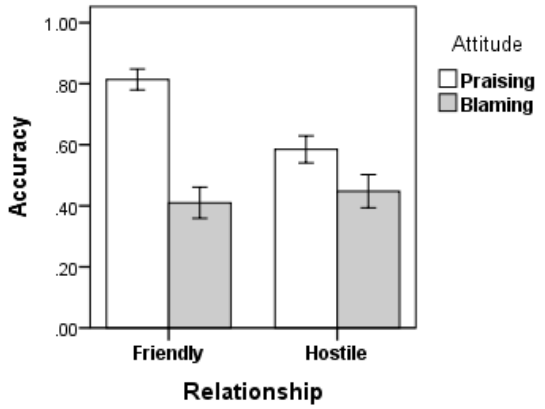
Figure 4: The mean and the standard error of identification rates for attitudes in three modalities.



There was also a significant interaction effect between relationship and attitude (F (1, 14) = 96.268, p < .001, η^2 = .873). As showed in Fig. 5, the difference in identification rates between praising and blaming was significantly larger in friendly (40.4%) relationship than that in hostile (13.7%).

Modality, relationship and attitudes had a significant three-way interaction effect ($F(2, 28) = 3.928$, $p < .05$, $\eta^2 = .219$) on accuracy. Pairwise comparisons revealed that when the relationship of interlocutors was friendly, as showed in Fig. 6, the

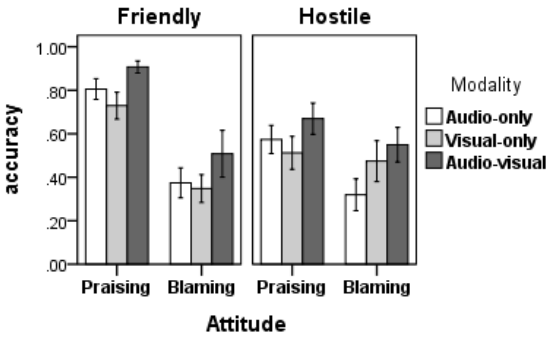
Figure 5: The mean and the standard error of identification rates for relationships and attitudes.



accuracy in audio-visual was significantly higher than that in audio-only and visual-only, both for praising and blaming. In addition, the blaming was identified better in audio only than that in visual only, while there was no difference for praising attitude.

Like friendly, when the speaker relationship was hostile, as showed in Figure 6, the accuracy in audio-visual was significantly higher than that in audio-only and visual-only, for both praising and blaming. Moreover, the blaming attitude was identified better in visual only than audio only, while there was on significant difference for praising.

Figure 6: The mean and the standard error of identification rates for attitudes in three modalities.



5. DISCUSSION AND CONCLUSION

The four speaking styles (i.e., friendly praising, hostile praising, friendly blaming, and hostile blaming) consisted of two relationships of interlocutors and two attitudes. The listeners decoded these four speaking styles with the help of different modalities usually. In our results, we found that the audio channel had advantages on recognition of friendly praising, hostile praising and friendly blaming while visual channel had advantages on recognition of hostile blaming. The reason may be that hostile blame in our stimulus, which was same as sarcasm, had inconsistent prosody and literal meaning, so that receivers need more contextual information (like facial expression) to verify.

The listeners identified different relationships and attitudes through different modalities. The positive relationship (friendly) and attitude (praising) were better identified in the audio channel than in the visual channel, while the negative relationship (hostile) and attitude (blaming) were better identified in the visual channel than in the audio channel. This may be because the speakers tend to express positive information by means of speech prosody and express negative information by facial expression.

We also found that the speaker relationship affected the recognition of attitudes. When the relationship was friendly, praising had a higher accuracy in audio-only than in visual-only. In contrast, when the relationship was hostile, praising had a higher accuracy in visual-only than in audio-only.

6. ACKNOWLEDGEMENT

This research was supported by the Major Program of the National Social Science Fund of China (13&ZD189) and the project for Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences (2017STD006).

7. REFERENCES

[1] Bänziger, T., & Scherer, K. R. 2005. The role of intonation in emotional expressions. *Speech Commun.* 46, 252–267.

[2] Campbell, N. 2005. Getting to the heart of the matter: Speech as the expression of affect rather than just text or language. *Lang Resources and Evaluation,* 39(1), 111–120.

[3] de Moraes J A, Rilliard A, de Oliveira Mota, B A, et al. 2010. Multimodal perception and production of attitudinal meaning in Brazilian Portuguese. *Proc. 5th Speech Prosody*, Chicago, IL, USA, Paper No. 340.

[4] Gu W, Zhang T, Fujisaki H. 2011. Prosodic analysis and perception of Mandarin utterances conveying attitudes. *Proc. 12th INTERSPEECH*, Florence, Italy, 1069-1072.

[5] Hönemann A, Mixdorff H, Rilliard A. 2014. Social Attitudes - Recordings and Evaluation of an audio-visual Corpus in German. *Proc. 7th ACUSTICUM*, Krakow.

[6] Lu, Y., Aubergé, V., & Rilliard, A. 2012. Do you hear my attitudes? Prosodic perception of social affects in

Mandarin. *Proc. 6th Speech Prosody*, ShangHai, China, 685-688.

[7] Menezes C, Erickson D, Han J. 2012. Cross-Linguistic Cross-modality Perception of English Sad and Happy Speech. *Proc. 6th Speech Prosody*, Shanghai, China, 649-652.

[8] Morlec, Y., Bailly, G., & Aubergé, V. 2001. Generating prosodic attitudes in French: Data, model and evaluation. *Speech Commun.* 33(4), 357–371.

[9] Rilliard A, Shochi T, Martin J C, et al. 2009. Multimodal indices to Japanese and French prosodically expressed social affects. *Lang and Speech*, 52 (2-3): 223-243.

[10] Shochi T, Auberg E V E R, Rilliard A. 2006. How prosodic attitudes can be false friends: Japanese vs. French social affects. *Proc. 3rd Speech Prosody,* Dresden, Germany, 693-696.

[11] Shochi, T., Aubergé, V., & Rilliard, A. 2007. Cross-listening of Japanese, English and French social affect: About universals, false friends and unknown attitudes. *Proc. 16th ICPhS*. 2097–2100.