Crosslinguistic Comparison on the Perception of Mandarin Attitudinal Speech

Wentao Gu¹, Ping Tang¹, Keikichi Hirose² and Véronique Aubergé³

¹ Nanjing Normal University, China
² The University of Tokyo, Japan
³ Centre National de la Recherche Scientifique, France

wtgu@njnu.edu.cn, tangp1990@gmail.com, hirose@gavo.t.u-tokyo.ac.jp, veronique.uberge@gmail.com

Abstract

This work investigated crosslinguistic perception of Mandarin utterances conveying six classes of attitudes, i.e., dominant/submissive, friendly/hostile, polite/rude, serious/joking, praising/blaming, and sincere/insincere. Five groups of subjects were tested: native Mandarin speakers, Japanese L2 learners of Mandarin, French L2 learners of Mandarin, native Japanese without Mandarin ability, and native French without Mandarin ability. A set of Mandarin attitudinal utterances elicited in role-play dialogues were used as stimuli. Perceptual experiments showed that native subjects performed the best in identifying attitudes, and L2 learners judged better than naïve foreigners. A complex interaction was found between attitude and the listener’s L1/L2 experience which was also closely related to culture. Also, the correlations between perceptual patterns and prosodic features were examined, and the results suggested that the prosodic cues for certain attitudes might be dependent on the listener’s language/culture background.

Index Terms: attitude, crosslinguistic perception, Mandarin, Japanese, French, L2 learner

1. Introduction

Among an increasing number of crosslinguistic studies on the production and perception of affective speech conveying emotions and attitudes, both universal and linguistic/culture-specific patterns have been reported.

For vocal emotions, similar inference rules were found across languages [1–4], and emotions in a spoken language could be identified by nonnatives fairly well [1–3]. Meanwhile, natives still perceived the best, and the perceptual accuracy decreased with language dissimilarity [1]. Similar tests were also conducted on attitudes or social affects. An investigation into the differences in cognitive distances for Chinese and French subjects showed that they shared common cognitive properties in 19 social affects except ‘question’ and ‘seduction’ [5]. It was also reported that Japanese and French differed substantially in perceiving polite and impolite Japanese speech [6]. While native speakers, L2 learners, and naïve foreigners showed different perceptual patterns on social affects, the difference between L2 learners at different levels was not significant [7, 8]. In comparison, social affects or attitudes show more language/culture dependent perceptual patterns than emotions since they are directly associated with social behaviors. Thus, the present work will be aimed at attitudes.

Some other studies examined acoustic correlates for social affective or attitudinal speech. Among others, Gu et al. [9, 10] found that speech rate was distinctive in each of five pairs of attitudes in Mandarin: friendly/hostile, polite/rude, serious/joking, praising/blaming, and confident/uncertain. Lu et al. [11] clustered 19 social affects in Mandarin into five sets, among which the differences were closely related to f0 and duration. Ofuka et al. [12] found that temporal features, including speech rate and duration of the final syllable, played important roles in perceiving politeness in Japanese. Chen et al. [13] argued that f0 features were more important than speech rate in Mandarin friendly speech. Cheang and Pell [14] found that both Cantonese and English used mean f0 as the primary acoustic parameter for conveying sarcasm, but in opposite manners – Cantonese raised mean f0 while English lowered it.

The aim of this study is to further examine how subjects with different L1 and L2 backgrounds perceive attitudes in Mandarin speech, and to find the relation between subjective perception and prosodic production. Taking Mandarin as the target, we selected Japanese and French as two contrastive linguistic backgrounds, which belong to the Oriental culture (like Mandarin) and the Occidental culture, respectively, and have shown different perceptual attributes for attitudes [15].

2. Speech data

Gu et al. [9, 10] differentiated attitudes from emotions in that attitude is an external expression imposed consciously by the speaker, while emotion as an internal state of the speaker is conveyed unconsciously in nature. Along this line of definition, they studied nine classes of attitudes commonly observed in daily communication [10]. Following this classification, the present work chose to investigate the following six classes of behavioral attitudes which are inherently associated with human interaction and hence are closely related to cultures:

- Class 1 (Dom/Sub): Dominant vs. Submissive
- Class 2 (Fri/Hos): Friendly vs. Hostile
- Class 3 (Pol/Rud): Polite vs. Rude
- Class 4 (Pra/Bla): Praising vs. Blaming
- Class 5 (Ser/Jok): Serious vs. Joking
- Class 6 (Sin/Ins): Sincere vs. Insincere

Within each class, we defined two extreme attitudinal labels (i.e., two poles). The former attitude in each class is regarded as ‘positive,’ while the latter is regarded as ‘negative,’ which basically coincides with our common evaluation.

For each attitude class, we designed 10 target sentences each composed of 6-12 syllables. In all classes except Classes 4 and 6, the sentences are literally neutral (i.e., not containing any words that is explicitly associated with a specific attitude) but at the same time can be expressed in opposite attitudes.
when embedded in different contexts. In Class 4 (Pra/Bla), some sentences are literally neutral, while most of them are literally praising but can pragmatically be blaming in a specific context, viz., ironic due to a mismatch between the lexical meaning and the conversational context. In Class 6 (Sin/Ins), most sentences literally use sincere words, but the oral expression can be sincere or insincere, depending on conversational situations – the insincere expression is a kind of lip-service that both the speaker and the listener understand.

Then, attitudinal speech data were collected by elicitation in a role-play. Following our conceptual distinction between attitude and emotion, attitudinal speech should in principle be elicited more naturally than emotional speech in a role-play, because attitude is expressed consciously in nature. For each target sentence in an attitude class, we designed two dialogues to elicit two opposite attitudes. Each dialogue consisted of 3 to 6 turns, within which the target sentence constitutes a turn by itself in the later part of the dialogue. Prompt texts were also given to elucidate the relationship between the two speakers.

The informants for speech collection were six native Mandarin speakers (3M+3F) at the age of 22-23. They were all undergraduate students in the major of broadcasting and hosting arts, all professional in pronunciation and skilled in vocal expression. The recording was done in a sound-proof studio. The dialogues were conducted by two informants naturally in a role-play manner. In addition, isolated target sentences were also recorded in a neutral reading style. Therefore, for each target sentence in an attitude class, the informants produced three versions (i.e., positive, negative, and neutral). Thus, there are 6 (attitude classes) * 3 (attitude levels) * 10 (sentences) * 6 (speakers) = 1080 target utterances.

3. Perceptual experiment

3.1. Subjects

We recruited 40 subjects with five different language backgrounds (4M+4F for each group) to participate in the perceptual experiment: native Mandarin subjects (CN), French L2 learners of Mandarin (F2), Japanese L2 learners of Mandarin (J2), naïve French without Mandarin ability (FR), and naïve Japanese without Mandarin ability (JP).

The CN subjects were graduate students in Nanjing Normal University at the age of 22-24. The F2 and J2 subjects were L2 Mandarin learners at the medium level, living in Nanjing at the age of 18-22; they had learned Mandarin for 1-2 years. The FR subjects were recruited at the GIPSA lab in Grenoble, France, aged 22-30 years old. The JP subjects were college students at the University of Tokyo, Japan, aged 19-25 years old. No subject had any reported impairment in listening or comprehension.

3.2. Procedure

A series of perceptual experiments were conducted in Nanjing (China), Tokyo (Japan), and Grenoble (France) separately. For each subject, the experiment consisted of six blocks, corresponding to six attitude classes, respectively. Each block tested 180 utterances and lasted for about half an hour. Different blocks could be run at different time.

The E-Prime software was used for stimulus presentation and response collection. The method of constant stimuli was adopted as test paradigm. For each attitude class, the 180 target utterances extracted from the dialogues were used as stimuli, which were randomized and then combined into a sound file with an inter-stimuli interval of 5 seconds. Then, the sound file was played back to the subject through headphones in a sound-proof studio. Within each 5s inter-stimuli interval, the subjects were requested to judge the attitude on a 5-point scale. Taking the “polite/rude” class as example, the scores were set as follows: 2 (obviously polite), 1 (somewhat polite), 0 (neutral), -1 (somewhat rude), and -2 (obviously rude). The same rule applied to other five classes. For each group of subjects, all instructions/questions presented in E-Prime were in their native languages. Before experiment, a training session was repeated until the subjects could give answers confidently.

3.3. Results

Figure 1 shows the rate of identification for attitudes in each class for all five groups of subjects. A score with correct polarity is regarded as correctly identified, e.g., “obviously polite” and “somewhat polite” are both correct judgments for polite speech. Consistently, in all attitude classes, CN gave the highest rate, followed by two L2 groups J2 and F2, while JP and FR gave the lowest identification. The two L2 groups differed mainly in “polite/rude” and “praising/blaming”, where J2 judged better than F2. The difference between JP and FR, however, depended highly on the attitude class.

![Figure 1: Rate of identification for positive/negative attitudes.](image1)

![Figure 2: Mean perceptual scores for positive and negative attitudes in six classes.](image2)

1335
The rate of identification still cannot give us a full picture. To have a better illustration, Fig. 2 shows the mean perceptual scores for both positive and negative attitudes in all six classes. For each bar, the parts above and below the abscissa indicate the scores for positive and negative attitudes, respectively. The absolute value indicates the perceived strength of the attitude. We observe almost the same pattern as shown in Fig. 1, i.e., the native group CN judged the best, followed by two L2 groups J2 and F2, while JP and FR judged the worst. The major exception exists in ‘sincere/insincere’, where L2 groups did not judge better than native foreigner groups, except that FR judged ‘insincere’ worst. Another minor exception exists in ‘praising’, where F2 showed an even lower score than JP.

For Dom/Sub, all groups gave consistently much higher absolute scores on the positive pole than on the negative pole, indicating that perception of ‘dominant’ is generally much easier than that of ‘submissive’. In contrast, for Fri/Hos and Pol/Rud, all groups gave consistently higher absolute scores on the negative pole than on the positive pole, indicating that perception of hostile or rude attitude is generally easier than that of friendly or polite attitude. The other three classes did not show any consistent preference between positive and negative attitudes, but the results were L1-dependent, e.g., two French groups judged ‘blaming’ better than did ‘praising’, whereas two Japanese groups behaved the opposite.

To get deeper statistical results, we conducted repeated-measures ANOVA on the mean absolute scores for two opposite attitudes in all six classes. Absolute scores were used such that the perceptual attributes on the two attitudinal poles can be compared directly. Here group (5 levels) was a between-subjects factor, while class (6 levels) and polarity (2 levels: positive vs. negative) were within-subjects factors.

The results showed that group had a significant main effect: F(4, 35) = 21.72, p < 0.001; class had a significant main effect: F(5, 175) = 36.13, p < 0.001; the group×class interaction was significant: F(20, 175) = 2.02, p < 0.01; the class×polarity interaction was significant: F(5, 175) = 22.05, p < 0.001; the group×polarity interaction was not significant: F(4, 35) = 1.49, p = 0.226; and the group×class×polarity interaction was not significant: F(20, 175) = 0.75, p = 0.772.

Bonferroni post-hoc test between groups showed that CN gave significantly higher absolute scores than all nonnative groups, while F2 and J2 gave significantly higher absolute scores than FR and JP. The difference between F2 and J2, and the difference between FR and JP, were not significant.

The group×class interaction was examined in more detail. For “praising/blaming”, CN gave significantly higher absolute score than all four nonnative groups. J2 gave significantly higher absolute score than FR, and no significant differences were found among other groups; further analysis on both polarities showed that the observed differences existed mainly in ‘praising’. For “sincere/insincere”, CN gave significantly higher absolute scores than four other groups, among which there were no significant differences; further analysis showed that the differences existed mainly in ‘sincere’, except that FR gave significantly lower absolute score than CN in ‘insincere’.

Although less important, the scores on neutral attitude were also analyzed by repeated-measures ANOVA, which showed that group had marginally significant main effect: F(4, 35) = 2.30, p = 0.078; class had a significant main effect: F(5, 175) = 17.96, p < 0.001; and the group×class interaction was also significant: F(20, 175) = 2.21, p < 0.01. Simple effect analysis showed that there were only two significant effects: J2 gave a lower positive score than F2 in Ser/Jok, and a lower positive score than CN and F2 in Sin/Ins. Therefore, there is little difference in perceiving neutral speech among groups, and the only two effects seem to be related to the following facts: (1) Without context, the sentences in Ser/Jok – though lexically neutral – can hardly be regarded joking, and neutral speech sounds calm which might be confused with ‘serious’, even for native listeners [9, 10]; (2) The sentences in Sin/Ins literally use sincere words which may lead neutral speech to sound ‘sincere’. With comparable Mandarin abilities, J2 and F2 were able to get lexical meanings equally well. However, J2’s lower scores to neutral speech than F2’s in these two attitudinal dimensions (even lower than CN’s in Sin/Ins) imply that Japanese are less misled by lexical meanings, suggesting that Japanese have higher expectation/attention to prosodic manipulation in the speech conveying ‘serious’ and ‘sincere’ due to higher sensitivity rooted in the Japanese culture.

4. Acoustic analysis

To further investigate the relationship between perceptual scores and acoustic characteristics, for each target utterance we calculated five f0 parameters including f0Min, f0Max, f0Range, f0Mean, f0Slope, and two timing parameters including SpRate (speaking rate, i.e., number of syllables per second) and FinalDur (duration of the utterance-final syllable). To eliminate inter-speaker f0 variations, we used the normalized metrics for f0Min, f0Max, and f0Mean. Defining f0MinRef and f0MaxRef as the averaged minimum and maximum f0s among all target utterances of the same speaker, respectively, we normalized the observed f0 parameters, e.g.: f0Min = f0Min_org − f0MinRef

Based on these f0 and timing parameters, two MANOVAs were conducted. The first MANOVA compared the parameters among three attitudinal levels in each attitude class, taking attitudinal polarity as an independent variable. The results are shown in Table 1, where asterisks indicate effects at the p < 0.01 level. The effect size η² was also calculated for each parameter. In each attitude class, the parameter giving the largest effect size is marked with double asterisks.

The second MANOVA analyzed the correlations between perceptual scores and prosodic parameters, taking the average score for each target utterance as the covariate. After covariance analysis, all significant correlations (p < 0.01) are shown in Table 1, either positive or negative. For each group, the parameter giving the highest absolute correlation with perceptual scores is marked with bold +/- symbols.

As shown, in Dom/Sub, Fri/Hos and Pol/Rud, timing parameters played more important roles in perception, whereas in Ser/Jok and Sin/Ins, f0 parameters, especially f0Mean and f0Slope, played more important roles. In particular, an L1 dependent pattern was found in Pra/Bla, i.e., CN and the two Japanese groups showed the highest correlation on f0Mean, while the two French groups showed the highest correlation on SpRate. This difference might explain why the French groups perceived less accurately than others in ‘praising/blaming’.

5. Discussion and conclusion

This study compared five groups of subjects with different language backgrounds in their subjective perception of six pairs of attitudes as well as the neutral expression in Mandarin.
Table 1. Effects of prosodic parameters and their correlations with perceptual scores.

<table>
<thead>
<tr>
<th></th>
<th>Dominant/Submissive</th>
<th>Friendly/Hostile</th>
<th>Polite/Rude</th>
<th>Praising/Blaming</th>
<th>Serious/Joking</th>
<th>Sincere/Insincere</th>
</tr>
</thead>
<tbody>
<tr>
<td>f0Min</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td>Effect</td>
</tr>
<tr>
<td>f0Max</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
<tr>
<td>f0Range</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
<tr>
<td>f0Mean</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
<tr>
<td>f0Slope</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
<tr>
<td>SpRate</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
<tr>
<td>FinalDur</td>
<td>CN</td>
<td>F2</td>
<td>J2</td>
<td>FR</td>
<td>JP</td>
<td></td>
</tr>
</tbody>
</table>

**Effects of prosodic parameters and their correlations with perceptual scores.**

- The results on neutral attitude were basically universal, except that J2 was less affected by lexical meanings in Ser/Jok and Sin/Ins. For positive/negative attitudes, the native group gave the highest identification, followed by L2 learners, while native foreigners without any Mandarin ability gave the lowest identification. All these results indicate the roles of language experience in identifying attitudinal speech; especially, L2 learners perform better after a period of language learning.

- In addition to this general tendency, some other language- and attitude-dependent perceptual patterns have been observed. Two special crosslinguistic differences deserve particular attention. First, the two French groups perceived ‘praising’ less accurately than the native and the two Japanese groups – even French L2 learners did not judge better than naive Japanese, suggesting that French are less used to the Chinese manner of ‘praising’ than Japanese. Second, for “sincere/insincere”, all four nonnative groups judged much worse than CN, suggesting that whether Chinese speakers express sincerity or not is particularly difficult for Japanese and French to differentiate; Especially, native French showed a perceptual bias to sincerity instead of insincerity in Mandarin speech.

- These two crosslinguistic perceptual patterns are closely related to cultural backgrounds. That is, the differences in perception are rooted in different strategies of production in their native languages/cultures. The first pattern implies that the manner of ‘praising’ may differ between the Oriental and Occidental cultures – Japanese judged better than French because Japanese and Chinese are both in the Oriental culture. The second pattern implies that the Chinese complex interpersonal culture, probably in close connection with the officialdom, is hard to be interpreted by foreigners, either Japanese or French. These two patterns elucidate that the perception of attitudes correlates not only with the ability on the target language but also with the L1 background.

- Moreover, the relatively low perceptual accuracy for L2 groups in these two classes may be explained from the design of target sentences. Unlike in other four attitude classes where the sentences are literally neutral, most sentences in Class 4 (Pra/Bla) and Class 6 (Sin/Ins) are literally associated with positive attitudes – hence the speech in negative attitudes are ironic. In this case, L2 learners’ perceptual judgment tended to be interfered by lexical meanings and hence made mistakes.

- The relation between subjective perception and objective prosodic parameters was also examined. In some attitude classes timing features gave the highest absolute correlation with perceptual results, while in other classes f0 features did. Especially, in Pra/Bla, the French groups correlated mainly on speech rate while other groups correlated mainly on mean f0; this might explain the perceptual differences between them.

- In summary, perception of attitudinal speech involves a complex interaction between attitude and L1/L2 experience which is also closely related to culture. Systematic correlations between perceptual attributes and prosodic parameters have been observed. For certain attitudes, the prosodic cues may be dependent on the listener’s language/culture background.

6. **Acknowledgements**

This work is supported jointly by the National Social Science Fund of China 10CYY009, 13&ZD189 (Major program), 13BYY009, and the key project funded by the Jiangsu Higher Institutions’ Key Research Base for Philosophy and Social Sciences (2010JDXM024).
7. References


