Audiovisual Perception of Expressions of Mandarin Chinese social affects by French L2 Learners

Yan Lu[1], Véronique Aubergé[2], Nicolas Audibert[3], Albert Rilliard[4]

[1] GIPSA Lab, CNRS, Stendhal University, Grenoble France
[2] LIG Lab, CNRS, Grenoble France
[3] Laboratoire de Phonétique et Phonologie, Univ. Sorbonne-Nouvelle/CNRS, Paris, France
[4] LIMSI-CNRS, Orsay, France
yan.lu@gipsa-lab.grenoble-inp.fr, Veronique.Auberge@imag.fr,
nicolas.audibert@univ-paris3.fr, albert.rilliard@limsi.fr

Abstract

This study focuses on confusions made by French L2 learners vs. native subjects in the perception of 11 audiovisual Mandarin Chinese social affects. Two groups of French L2 learners of Mandarin Chinese were selected according to their Chinese level : 9 beginners (A1) vs. 10 intermediate learners (A2). Subjects evaluated the 11 social affects in audio, visual and audiovisual condition. Comparison of confusions between learners of level A1 vs. A2 indicates few significant differences, mostly in audiovisual condition and without a clear gain for one group over the other. The comparison of French L2 learners pooled together vs. native speakers reference sheds light on major confusions to be targeted by specific methods and exercises. Cross-modality comparisons suggest a limited contribution of informations conveyed by acoustic prosody in the identification of audiovisual social affects by L2 learners.

Index Terms: social affects, attitudes, audio-visual perception, L2 prosody, Mandarin Chinese

1. Introduction

The face to face interaction functions of prosody are the main vector of the "socio-affective glue" that builds the communication channel [1] and are expressed following different cognitive processing levels [2]: from involuntary controlled expressions (emotions) to the voluntary control of the social affects of the speaker [3,4]: intentions, attitudes, social cues etc). During the face-to-face communication, people express their affects within the audio-visual speech prosody [5], and as social affects are constructed socially for and by the language, and prosodic realization of one specific social affect in a specific language may be ambiguous or unknown in the learner's language [6], the cross-cultural approach looks more convenient to spotlight the cultural specifications of the expression of social affects. Meanwhile, the social affects are generally learned in childhood within the language community in question and they can also be learned by the learners of a foreign language or a second language if they are different from the social affects in their mother language. But it is preliminarily necessary for foreign learners to recognize first the social affects expressed in the target language. Some studies have shown that the foreign language learners did not perceive the attitude expressed in the target language in the same way with the native people, and the attitudinal prosody needs to be taught in the L2 class [7, 8].

For these reasons, we intend in this study to examine in a cross-cultural context the perception of the audio-visual expressions of Chinese social affective prosody by both native subjects and French learners of Mandarin Chinese. The largest part of the analysis focuses on differences between native Chinese listeners and French learners for the perception of the same social affect, investigating the potential influence of the learners' language skill on their perception of these social affects. The relative contribution of the acoustic vs. visual modalities is also compared across groups of subjects. Finally, consequences for L2 teaching of differences in the perception of social affects between learners and native speakers will be discussed.

2. Method

2.1. Selection of social affects

A large audio-visual Chinese corpus (acted speech) of 19 social affects (each expressed on a set of utterances varying within length, syntax and tone location) was initially validated in an acoustic only perception experiment [9].On the basis of an acoustic cross-perception experiment by French naive listeners [10], 11 social affects were selected among the 19. These social affects were observed problematic for both native and foreign listeners during the previous perception experiments because of their "attractivity" (cumulated percentage of confusions from other attitudes to each one) and showed great difference in perception behavior between native subjects and foreign ones. Hence, they are supposed to be also problematic and difficult for French learners of Chinese in their face-to-face communication with native speakers. The 11 social affects selected are composed of the attitudes, intention or opinion of the speaker about what he says; the characteristics of the social relation implied in the interaction (e.g. "politeness") and the socio-cultural context of interaction (e.g. "infant-directed speech"). Table 1 presents the 11 Chinese social affects and their abbreviation.

Table 1. *Summary of the 11 social affects selected*

Social affects and abbreviation	
declaration (DECL)	obviousness (OBVI)
question (QUES)	neutral surprise (NEU-S)
irritation (IRRI)	politeness (POLI)
doubt (DOUB)	authority (AUTH)
contempt(CONT)	infant-directed speech (IDS)
disappointment (DISA)	

2.2. Audiovisual corpus

The audiovisual speech corpus is based on a 4-syllable long sentence, which is constructed to bear a literally neutral meaning but could be expressed with all social affects studied.

10.21437/SpeechProsody.2014-22

This sentence was performed with 11 social affects by one native Chinese female speaker, who speaks an unmarked standard Mandarin Chinese. The audio part of the corpus has been validated in [9], where all of 11 attitudes have been recognized over chance level. The sentence used in this experiment was better recognized for all social affects and considered the most representative of a prototypical expression of the targeted social affect. This sentence is "四天三夜" (si4 tian1 san1 ye4, "four days and three nights" in English). Thus, 11 short videos were used in the present experiment.

2.3. Subjects

A first group consisted of 30 Chinese native listeners (12 males and 18 females, mean age = 33.3) as the reference group of "optimal" performances of Chinese perception. Two groups of L2 learners are composed according to their results to a test of placement in Chinese language according to CECRL (the Common European Framework of Reference for Languages), taken at the beginning of the term. One group was composed of 9 French learners of Mandarin Chinese whose acquired Chinese level is A1: beginners, having taken less than 100 hours of teaching (1 male and 8 females, mean age = 20.7). Another group was composed of 10 French learners of level A2 – having taken less than 200 hours of teaching (3 males and 7 females, mean age = 22.3). All 19 French subjects study the Mandarin Chinese as foreign language in the LANSAD (Languages for the specialists of other disciplines) Department of University Stendhal-Grenoble 3.

Subjects in each group were divided randomly into two sub-groups of equal size with a different presentation order: audio only → video only → audiovisual in one sub-group, and video only →audio only →audiovisual condition in the other.

2.4. Perceptual evaluation protocol

Before the test, subjects were briefly presented the setting of the experiment and a description of each attitude with examples of situations in which it can happen. They took the test in a quiet room with closed headphone, using a graphical user interface developed with LiveCode® for stimuli presentation and answers collection. Each stimulus was presented once, in a different random order for each subject.

The Chinese audiovisual corpus was presented to subjects in three different conditions:

- Audio only (AU): in this condition, images were hidden. Subjects were instructed to listen to what the speaker said before answering.
- Video only (VI): in this condition, the sound was turned off. Subjects were instructed to carefully watch the facial and body movements of the speaker.
- Audiovisual (AV): in this condition, both modalities were presented simultaneously. Subjects were instructed to watch the facial and body movements of the speaker and at the same time to listen to her voice

For each stimulus presented, subjects were asked to judge which attitude was expressed by the speaker, by choosing among the eleven labels proposed.

3. Results

Subjects' answers were pooled into a confusion matrix for each group of listeners x presentation condition. In each presentation condition, confusion matrices cells values were compared against chance level, and between groups using chi-squared tests for comparison of proportions.

3.1. Native subjects performance

As expected from previous studies on this corpus with Chinese and completely naive French listeners, the native listeners performed better than French learners for all modalities of presentation of social affects. In audiovisual condition, they recognized significantly over chance level (9%) all attitudes except "contempt" (recognition rate: 33%). "Declaration" and "irritation" were significantly recognized over chance in all three conditions.

3.2. A1 vs. A2 learners

A first analysis was performed comparing confusions of L2 learners with level A1 vs. L2 learners with level A2, revealing few inter-group differences. The largest part of significant differences was found in audiovisual presentation condition. Compared to group A1, the group A2 confused significantly less question with obviousness and doubt with neutral surprise, outperforming both native subjects and A1 learners in their identification of the expression of doubt. Surprisingly, a better performance of group A1 was found in the identification of authority. However, a comparison with native subjects reveals that the confusion pattern observed in group A2 for expressions of authority is similar to native subjects confusions, authority being largely confused with obviousness.

In audio-only condition, the expression of doubt was significantly less confused with disappointment by learners of the group A2 compared to group A1. In visual-only condition, learners in group A2 confused significantly less contempt with irritation than group A1, but they confused significantly more doubt with irritation.

3.3. Native subjects vs. French L2 learners

Results reported supra do not picture a clear advantage in performances of the group of A2 learners vs. A1. In order to get more insight into major differences between native Mandarin Chinese subjects and French L2 learners, answers in groups A1 and A2 are pooled altogether in a 'L2 learners' group (19 subjects) in the following analysis.

Table 2, 3 and 4 summarize confusions by the 30 native Mandarin Chinese listeners and French L2 learners respectively in audio-only, visual-only and audiovisual presentation condition, all learners answers pooled. Each cell in the confusion matrices has two values, native subjects' performance (top) and L2 learners' performance (bottom). Values significantly different from chance level are flagged by stars on the right part. Stars on the left part of a cell indicate a significant difference between the group of native speakers and the group of French L2 learners.

For instance, the cell corresponding to neutral surprise (NEU-S, 3rd column, 3rd line) in the matrix diagonal in audio-only condition (Table 2) indicates that native Mandarin Chinese subjects identified neutral surprise at 60%, which is significantly higher than chance ($p<.01$), while L2 learners did not identify it better than chance (16%, $p>.05$). The left part of this cell also reports that the ratio of correct identification of audiovisual neutral surprise is significantly higher ($p<.01$) in the native subjects group than in the L2 learners group.

Table 2. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in audio-only condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

Audio	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	50% *	0%	0%	0%	0%	10%	33%	0%	7%	0%	0%
	32%	0%	0%	0%	0%	21%	21%	0%	26%	0%	0%
NEU-S	0%	** 60% **	10%	13%	0%	0%	10%	0%	0%	7%	0%
	0%	16%	21%	5%	0%	0%	32%	5%	0%	16%	5%
QUES	7%	13%	57% **	13%	3%	0%	* 0%	3%	0%	0%	3%
	0%	26%	32%	16%	0%	0%	16%	0%	0%	0%	11%
DOUB	3%	10%	23%	27%	0%	3%	10%	13%	0%	3%	7%
	11%	26%	16%	11%	0%	0%	0%	11%	0%	11%	16%
IDS	13%	0%	7%	7%	53% *	7%	7%	3%	0%	0%	3%
	16%	11%	0%	0%	68% **	0%	0%	0%	0%	0%	5%
POLI	43% *	0%	0%	0%	3%	** 30%	10%	0%	13%	0%	0%
	53% *	0%	0%	5%	16%	0%	5%	5%	0%	5%	11%
OBVI	27%	0%	3%	0%	0%	3%	40%	10%	13%	0%	3%
	32%	0%	5%	11%	0%	0%	16%	16%	21%	0%	0%
CONT	3%	17%	* 7%	30%	0%	0%	7%	27%	3%	3%	3%
	11%	5%	32%	16%	5%	0%	5%	11%	0%	11%	5%
AUTH	13%	0%	0%	0%	0%	0%	** 37%	3%	33%	10%	3%
	26%	0%	0%	0%	0%	5%	0%	11%	47% *	11%	0%
IRRI	10%	3%	3%	3%	0%	0%	13%	7%	13%	43% *	3%
	5%	11%	0%	5%	0%	0%	11%	5%	26%	32%	5%
DISA	* 37%	0%	0%	0%	3%	7%	10%	7%	7%	0%	30%
	5%	5%	0%	0%	5%	11%	5%	11%	0%	5%	53% *

Table 3. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in visual-only condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

Visual	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	50% *	0%	3%	3%	3%	10%	20%	0%	7%	0%	3%
	58% **	5%	0%	5%	0%	0%	5%	5%	11%	0%	11%
NEU-S	7%	23%	30%	13%	7%	10%	7%	0%	0%	0%	3%
	5%	11%	32%	5%	0%	0%	16%	0%	11%	5%	16%
QUES	33%	0%	10%	0%	10%	20%	17%	0%	10%	0%	0%
	53% *	5%	5%	0%	0%	21%	16%	0%	0%	0%	0%
DOUB	0%	3%	23%	50% *	0%	3%	0%	0%	0%	7%	13%
	0%	5%	11%	53% *	0%	0%	0%	0%	0%	21%	11%
IDS	10%	0%	7%	0%	27%	40%	13%	0%	3%	0%	0%
	32%	5%	5%	0%	21%	21%	11%	0%	5%	0%	0%
POLI	23%	7%	0%	0%	10%	50% *	10%	0%	0%	0%	0%
	21%	11%	0%	0%	5%	37%	21%	0%	0%	5%	0%
OBVI	* 3%	0%	3%	3%	0%	13%	37%	13%	13%	0%	13%
	26%	5%	0%	0%	0%	11%	47% *	0%	0%	0%	11%
CONT	10%	3%	7%	10%	0%	7%	10%	* 23%	0%	* 3%	27%
	11%	0%	0%	26%	5%	0%	0%	0%	0%	26%	32%
AUTH	20%	0%	3%	0%	3%	0%	20%	10%	33%	3%	7%
	11%	0%	0%	5%	0%	0%	32%	11%	37%	5%	0%
IRRI	0%	0%	3%	10%	0%	0%	0%	20%	0%	43% *	23%
	0%	0%	0%	11%	0%	0%	0%	26%	0%	37%	26%
DISA	7%	0%	7%	** 0%	0%	0%	0%	23%	0%	0%	63% **
	5%	11%	0%	21%	0%	0%	5%	11%	0%	0%	47% *

- For audio only modality: native subjects recognized significantly better neutral surprise and politeness than French learners; the native subjects confused more authority with obviousness and disappointment with declaration than French learners; on the other hand, French learners confused more contempt with question and question with obviousness than native subjects.
- For video only modality: native subjects recognized better contempt than French learners (A1: 0%, A2: 0%); French learners identified mistakenly more obviousness with declaration, contempt with irritation, and disappointment with doubt than native subjects.
- For audiovisual modality: native subjects identified better neutral surprise than French learners, who, on the contrary, recognized better infant-directed speech; native subjects confused more declaration with obviousness, doubt with question and contempt with doubt than French L2 learners. However for L2 learners, neutral surprise was more confused with irritation, question with obviousness and contempt with disappointment.

Table 4. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in audio-visual condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

AV	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	53% *	0%	0%	0%	0%	10%	* 33%	0%	3%	0%	0%
	79% **	0%	0%	0%	0%	11%	5%	0%	5%	0%	0%
NEU-S	7%	* 67% **	10%	17%	0%	0%	0%	0%	0%	** 0%	0%
	0%	32%	26%	0%	0%	0%	11%	5%	5%	21%	0%
QUES	10%	13%	47% *	23%	3%	3%	* 0%	0%	0%	0%	0%
	16%	32%	21%	11%	0%	5%	16%	0%	0%	0%	0%
DOUB	0%	10%	* 30%	47% *	0%	0%	0%	7%	0%	7%	0%
	0%	16%	5%	63% **	0%	0%	0%	5%	0%	5%	5%
IDS	10%	0%	0%	3%	* 57% **	17%	10%	3%	0%	0%	0%
	11%	0%	5%	0%	84% **	0%	0%	0%	0%	0%	0%
POLI	33%	0%	0%	0%	7%	50% *	10%	0%	0%	0%	0%
	16%	0%	0%	5%	11%	63% **	5%	0%	0%	0%	0%
OBVI	23%	0%	0%	0%	0%	3%	53% *	10%	10%	0%	0%
	21%	0%	0%	0%	0%	0%	63% **	5%	11%	0%	0%
CONT	3%	0%	13%	* 20%	0%	0%	7%	33%	0%	3%	* 20%
	5%	0%	11%	0%	0%	0%	5%	21%	0%	11%	47% *
AUTH	3%	0%	0%	0%	0%	0%	37%	0%	60% **	0%	0%
	0%	0%	0%	0%	0%	0%	16%	5%	74% **	5%	0%
IRRI	3%	0%	0%	0%	0%	0%	0%	27%	0%	63% **	7%
	0%	0%	0%	11%	0%	0%	0%	42%	5%	37%	5%
DISA	7%	0%	0%	0%	0%	0%	0%	7%	0%	0%	87% **
	0%	5%	0%	0%	5%	0%	0%	11%	0%	0%	79% **

Identification rates and confusions were also compared between presentation conditions using chi-square tests, for each group of subjects. For the sake of concision, those results are not reported extensively in this paper. Cross-condition comparisons indicate different multimodal strategies for the identification of social affects between groups. While native subjects tend to rely more on acoustic cues (with neutral surprise, question and infant-directed speech significantly better recognized in audio condition, and only disappointment better recognized in visual condition), this tendency is only partly reproduced in L2 learners performance (with question and infant-directed speech better recognized in audio condition, and doubt, politeness and obviousness better recognized in visual condition). Most differences between native subjects and L2 learners are found in the comparison of audiovisual vs. audio-only condition: while native subjects show a significant gain in audiovisual condition for authority and disappointment, L2 learners significantly benefit from the audiovisual information for declaration, doubt, politeness and obviousness.

4. Discussion and conclusion

This paper investigated the perceptual behavior of 19 French learners of Mandarin Chinese vs. 30 native listeners for 11 Mandarin Chinese social affects, presented in three different conditions: audio-only, video-only and audiovisual conditions. Meanwhile it also examined the correlation between the listeners' language skill and their perceptual behavior.

According to the results of analysis, the perception of all subject groups for the audiovisual modality shows the best scores for almost all attitudes [7]. Though differences were found between groups of learners with different level, the French learners in A2 level showed no clear advantage over the learners in A1 level. However, significant differences were found between native Chinese speakers and French L2 learners as a whole: the native listeners recognized better "neutral surprise" in both audio only and audiovisual modalities, and it was more confused with "question" in audiovisual modality by French learners; "politeness" was better recognized by native subjects in audio only modality; although "contempt" was relatively better recognized by native subjects in video only modality, in fact, the recognition rate of native subjects was not satisfactory (23%); "infant-directed speech" was unexpectedly better recognized by French learners than by native subjects. Limited differences in audio-visual condition between native Mandarin Chinese subjects and French L2 learners suggest that in a face-to-face communication context, L2 learners might compensate their relative inability to identify the acoustic prosodic correlates of social affects by relying more extensively on visual cues.

Meanwhile, French learners showed more difficulties in recognizing "politeness" in audio only condition, "question" and "contempt" in video only condition. "Politeness" was mostly confused with "declaration", that may be because of their similar prosodic characteristic. The facial expression of "question" was considered similar with that of "declaration", because both of them are basic communicative functions expressed by utterance modalities and are neutral in terms of affective sate. Acoustically, "question" was more confused with "neutral surprise" by French learners. These observations suggest that L2 teaching of Mandarin Chinese for French learners could benefit from integrating specific exercises on social affects, particularly concerning their acoustic realization with a focus on "neutral surprise" and "politeness".

5. Acknowledgements

This study was supported jointly by the French National IDEFI Innovalangues and the Major Program for the National Social Science Fund of China (13&ZD189).

6. References

[1] Aubergé V., Sasa Y., Robert T., Bonnefond N., Meillon B. (2013) "Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot". In proceedings of WASSS 2013, Grenoble, France.

[2] Aubergé, V., "A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP", Speech Prosody Proc., Aix-en-Provence, France, 151-155, 2002.

[3] Fónagy I. (1983). La vive voix. Essais de psycho-phonétique, Paris, Payot.

[4] Léon, P., "Précis de phonotylistique, parole et expressivité", Nathan, Paris, 1993.

[5] Barkhuysen, P., Krahmer, E. and Swerts, M., "Cross-modal perception of emotional speech", ICPhS Proc, Saarbruecken, Germany, 2133-2136, 2007.

[6] Shochi, T., Aubergé, V., and Rilliard, A., "How prosodic attitudes can be false friends: Japanese vs. French social affects", Speech Prosody Proc, Dresden, 692-696, 2006.

[7] De Meo, A. and Pettorino, M., "L'acquisizione della competenza prosodica in Italiano L2 da parte di studenti sinofoni", in E. Bonvino and S. Rastelli [Eds], La didattica dell'Italiano a studenti cinesi e il progetto Marco Polo, Pavia University Press, 67-78, 2011.

[8] Shochi, T., Gagnié, G., Rillard, A., Erickson, D. and Aubergé, V., "Learning effect of French prosodic social affects for Japanese learners of French language", Speech Prosody Proc, Chicago, IL, USA, paper 155, 2010

[9] Lu, Y., Aubergé, V. and Rilliard, A., "Do you hear my attitude? Prosodic perception of social affects in Mandarin", Speech Prosody 2012 Proc., 685-688, Shanghai, China, 2012.

[10] Lu, Y., Aubergé, V. and Rilliard, A., "Tonal Influences on the prosodic Cross-linguistic Perception of Mandarin Social Affects by French and Vietnamese listeners", the Third International Symposium of Tonal Aspect of Language Proc., Nanjing, China, 2012.