Do You Hear My Attitude?
Prosodic Perception of Social Affects in Mandarin

Yan Lu[1], Véronique Aubergé[1], Albert Rilliard[2]

[1] GIPSA Lab, CNRS, Stendhal University, Grenoble France; [2] LIMSI-CNRS, Orsay, France
{Yan.Lu, veronique.auberge}@gipsa-lab.grenoble-inp.fr, albert.rilliard@limsi.fr

Abstract

Social affects are, on the contrary to emotions, some voluntary controls expressed within the prosody, and contribute to build the meaning in the speech acts. The present work was set up to examine the perception of Chinese social affects by native subjects, but in the long term aims for the method of attitudinal prosody teaching. A speech corpus was designed with the variation of length, tone location and syntactic structures of utterances, and has been incorporated with 19 social affects. The perceptual experiment reveals that the social affects were globally well recognized; "declaration" and "disappointment" were the best recognized; "confidence" and "irony" were the less recognized. All social affects listed were clustered into seven groups, coherently in terms of cognitive/mental processing.

Index Terms: prosodic perception, attitudes, social affects discrimination, Mandarin Chinese

1. Introduction

The affects expressed in interactive speech imply two different levels of the speaker's cognitive processing [1]: the involuntarily controlled expressions of affects (so-called "emotions"), and the intentionally controlled expressions expressed through audio-visual prosody (so-called "social affects" or "attitudes"). On the contrary to emotions, social affects are strongly linked with a language, inside a given culture and are learned during infancy. They are an integral part of the language interaction building, i.e. communication. Different hypotheses have been set up about the organization of social affects [2, 3], and here we propose a close one:

- The attitude, intention or opinion of the speaker about what he says (he is involved in his speech [4]), – and the fact that he doesn't express (doesn't want to, doesn't have or isn't able to express) any attitude, is considered as an attitude in itself (like a simple declaration and question);
- The characteristics of the social relation (like social hierarchy/potency) implied in the interaction, e.g. politeness, authority;
- The socio-cultural context of interaction: typically for intimacy, infant-directed speech and seduction.

After [5, 6], many scholars have conducted studies of attitudinal prosody in different languages [7, 8, 9, 10, 11], and some of them had a special focus on the attitudinal prosody teaching [12], or cross-cultural comparison [13].

This paper is the first step towards the long-term aim at developing a didactic method to teach the French attitudinal prosody to Chinese learners. The present study is dedicated to investigating how prosodic social affects in Chinese can be perceived by native Chinese: how do they perceptively discriminate or confuse the prosodic expressions of 19 Chinese social affects (an explanation of the social affects is given). It will examine the distance between acoustic perception/concept understanding of these affects.

It is known [14, 13, 10] that the deep link between language and culture, as well as their values can vary both between two languages and cultures and inside a same language with the social role, education, age and gender. It will be observed if the gender identity, both innate and stereotyped can influence or structure the social affects.

2. The corpus

2.1. Phonotactic and linguistic cues

In order to compare the parameters implied in the variability of prosody, it was preferred to build and record a dedicated and controlled corpus instead of collecting spontaneous data.

The speech corpus was designed with consideration of length (in syllables), of tone location and of syntactic structure of the utterances, which were systematically varied in order to analyze further the variation of one parameter in the same context for the others. As the social affects could not be produced without reference to context, we described a specific context for each social affect in order to help the speaker to express them as naturally as possible. Utterances were constructed to bear a literally neutral meaning (i.e., not containing any words which imply a specific social affect nor emotion) but in the same time so they can be expressed with all the social affects studied. The corpus is built to propose variations on:

- the global length of utterances: from 1 to 9 syllables;
- the syntactic structure: from mono-word to complex structures;
- the location of the syntactic boundary;
- the value and location of the tone.

2.2. Selected social affects

This speech corpus is dedicated to performing 19 social affects after [6] and [1, 9 10, 13]. Before the perceptive test, every label of affects to be chosen by the listeners (see table 1) were precisely explained to the listeners by giving an example of a situation context, where the social affect happens.

Table 1. *Classification of social affects and their abbreviation*

	Social affects and abbreviation
(modalities)	declaration(DECL) question(QUES)
Attitudes	admiration(ADMI) confidence(CONF) irritation(IRRI) resignation(RESI) contempt(CONT) irony(IRON) doubt(DOUB) obviousness(OBVI) disappointment(DISA) neutral surprise(NEU-S) positive surprise(POS-S) negative surprise(NEG-S)
Social parameters	politeness(POLI) authority(AUTH)
Social context	seduction(SEDU) infant-directed speech(IDS) intimacy(INTI)

2.3. The corpus recording

The corpus was recorded by one native Mandarin female from Shaanxi province of China, speaking an unmarked standard mandarin Chinese. She is a language learning teacher, experimented in producing examples of "real speech". The recording was conducted in a sound-proof room and was saved both in video and audio file format. It has to be noted that another recording with a male Chinese speaker is under way in order to measure a possible effect of gender.

3. The perceptual experiment

To test the validity of the attitudinal speech corpus and to look into the perception and the recognition of attitudes, we designed this perceptual experiment. The listening subjects were composed of 30 native Mandarin Chinese, from different areas of China: 15 males and 15 females with an average age of 25.2 years. They're almost all postgraduate students or PhD students in Grenoble of France (except one male subject who works as computer programmer in an IT company in Grenoble), and none of them reported any listening and understanding disorder.

Table 2. *Example of sub-corpus*

Tones	Chinese	English	Structure
1	书	book	word
1-4	三万	thirty thousand	number
1-1-1-4	公交车站	bus stop	NG(4)+VG(0)
1-1-1-1-1-1-4	张医生帮她搬新书架	Doctor Zhang help her to move the new bookshelf	GN(3)+GV(3)+GN(3)

The complete corpus contains 152 utterances performed with 19 attitudes, that is too numerous to be globally evaluated in an acceptable perceptual task. Thus a sub-corpus was specifically selected, composed of sentence of length, tone location and syntactic complexity. 21 utterances were selected with four 1-syllable words, seven 2-syllable words, six 4-syllable sentences and four 9-syllable sentences. In the 4-syllable utterances, three syntactic structures were selected: nominal group; a 3-syllable nominal group as subject followed by a 1-syllable verb as predicate; a 1-syllable noun as subject followed by a 3-syllable verbal group as predicate. In 9-syllable utterances, two syntactic structures were selected: a 7-syllable nominal group as subject, a monosyllable verb as predicate followed by a monosyllable noun as object; the subject, the predicate and the object are all in three syllables. To vary the value of the tones, all tones are placed at every position of the 2-syllable words and only the second and the fourth tones were in the last syllable of the 4-syllable and 9-syllable utterances. Other syllables were always in the first tone.

Table 2 shows an example of the sub-corpus. The complete sub-corpus contains 399 stimuli, (i.e. 21 utterances produced with 19 different attitudes) in order to measure a potential tonal-syntactic effect and the length effect on the recognition of social affects.

All 399 target utterances (stimuli) were presented to the subjects through headphones in a quiet room and were introduced by a presentation of the experiment and a description of each social affect with examples of situations in which such a social affect can happen. The subjects listened only one time to each stimulus and have to choose the perceived attitude from the 19 proposed labels. The presentation order of the stimuli was randomized for each subject.

4. Analysis and results

4.1. Analysis of variance

An analysis of variance (completely randomized three-factorial design) was carried out on the data. The three fixed factors were the subjects' gender (G, 2 levels), the presented attitudes (A, 19 levels) and the sentences length (L, 4 levels). Each cell of this design contained at least 60 observations. The significance level was set at 0.01. Table 3 shows the results of the analysis of variance of each factor.

It is observed that the factor Attitude has a significant effect with the highest observed strength of effect (η^2), and the interaction between attitudes. The interaction between the factors Attitude & Length is also responsible of significant variation of results, and has a strong effect. Factors Gender and Length are significant at the 1% level, but does only explain a small part of the variance observed. The interactions Attitude & Gender, Gender & Length and Attitude & Gender & Length did not have a great influence on the perception of attitudes.

Table 3. *ANOVA's results – significant effects in bold.*

	Sum Sq	Df	F value	Pr(>F)	η^2
A	**253.43**	**18**	**86.2218**	**0.0000**	**0.693**
G	**1.97**	**1**	**12.0648**	**0.0005**	**0.005**
L	**16.19**	**3**	**33.0582**	**0.0000**	**0.044**
A*G	5.57	18	1.8960	0.0122	0.015
A*L	**80.30**	**54**	**9.1061**	**0.0000**	**0.220**
G*L	0.13	3	0.2556	0.8574	0.000
A*G*L	7.87	54	0.8929	0.6958	0.021

Through the mean recognition rate of 19 social affects and the mean recognition rate of social affects distinguished by stimulus's length and gender presented in figure 1, it is observed that for native Chinese listeners, almost all of the social affects were recognized above chance level, except *"*confidence*"* and they can be classified in the decreasing order (cf. Figure 1, top). The identification of social affects varies with the length of stimuli: according to the confusion matrix of attitudes by length, there is a clear separation between the stimuli of 1 syllable and the stimuli of 2, 4 and 9 syllables. The 1-syllable stimuli received lower recognition scores while the 4-syllable stimuli received the highest (the 9- and 2-syllable stimuli are just under the 4-syllable ones). But the graph of the mean recognition rate for social affect by length (figure 1, middle), shows that "infant-directed speech" and "irritation" were not in the global influence tendency of the utterance's length to the recognition of attitudes. For "infant-directed speech", the stimuli of 1 syllable and 2 syllables were better perceived than the stimuli of 4 and 9 syllables (who were mixed with "seduction"). It's possible to explain this with the fact that adults rarely address long and complex sentences to a young child. For "irritation", the 2-syllable stimuli were not well perceived, in comparison with other lengths, and they were specifically confused with "declaration" and "confidence. By listening carefully the stimuli of "irritation" in every length, we perceptually recognized a harsh and tensed voice used by the speaker. Therefore, the problem of misunderstanding is surely more related to global patterns than to the general voice quality. It

needs to be verified in an acoustic analysis, as well as the influence of the tonal variation.

The effect of gender on the recognition scores is not significant (figure 1, bottom – on the contrary of our hypothesis). This may investigate using a more specific and precise setting.

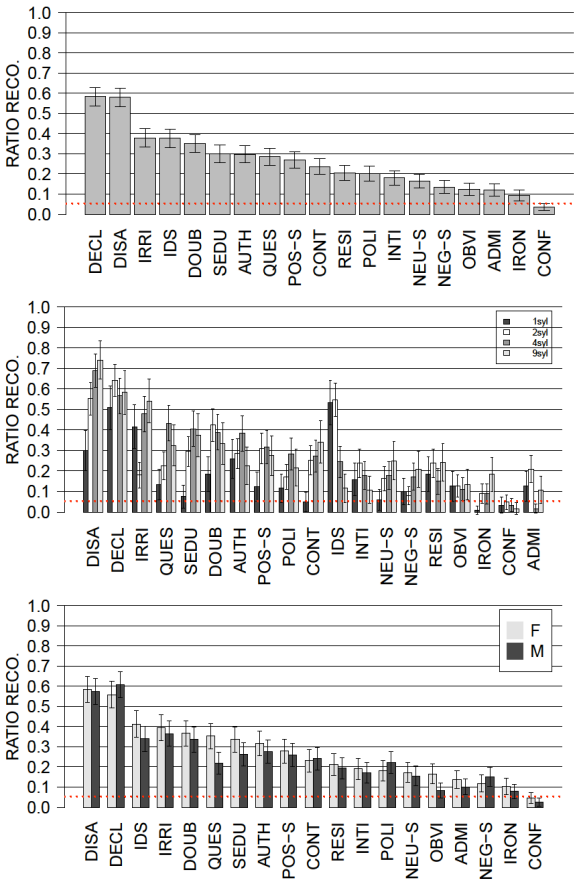

Figure 1: *The mean recognition for 19 social affects: rate per attitude (top), rate per stimuli's length (middle) and rate per gender (bottom)*

4.2. Clustering of attitudes

On the basis of the dispersion matrix obtained from the confusions made by subjects between attitudes, we can observe, on the basis of figure 2, the main confusion tendencies which are over twice the chance [3, 10] between the proposed attitudes:

Generally, the majority of attitudes were confused with "declaration", especially "confidence" (56.03%) and "politeness" (49.21%), but "declaration" itself was well recognized (58.4%) except some confusions with "obviousness". "Confidence" and "irony" were hardly recognized. "Neutral surprise", "negative surprise" and "positive surprise" were all confused with "doubt", which was confused with "question" and vice versa. "Contempt" and "irony" were confused one and the other and "contempt" was also confused with "declaration". "Infant-directed speech" was only significantly confused with "seduction". "Resignation" and "disappointment" were confused one another, and the former was significantly confused with the latter (43.56%).

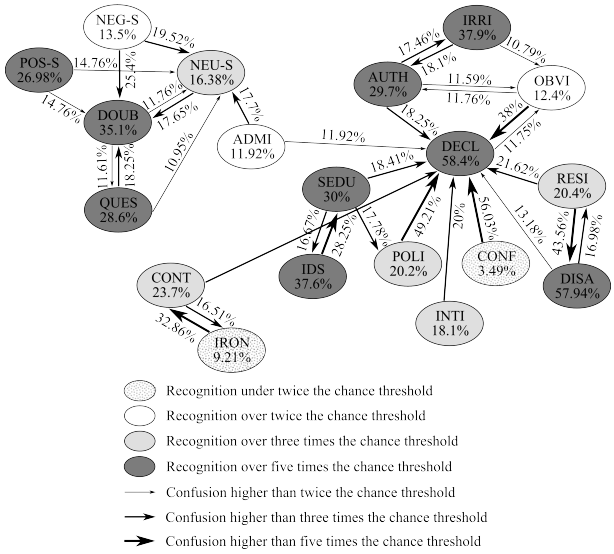

Figure 2: *The confusion table of social affects represented as a graph: percentages of identification of each social affect are in the nodes, the confusion rates are on arcs. Only the scores over twice the chance levels are noted.*

In order to measure the perceptive distances between each stimulus and identify the wider perceptual categories, a hierarchical cluster analysis was run on the dispersion matrix. Based on the dispersion matrix, perceived distances between presented attitudes were calculated by using the correlation between the rows (1-*r* is used as the distance, where *r* is the correlation between two rows). From this matrix of perceived distances, a hierarchical clustering algorithm was applied, in order to observe the main groups of attitudes. The confusions between attitudes are shown in figure 3. From this clustering, the native subjects perceptually grouped attitudes in seven generic groups [12]:

- "Admiration" and "positive surprise": these two attitudes are caused by a strong positive affect and relevant to high positive values in comparison with other positive attitudes.
- "Negative surprise", "neutral surprise", "question" and "doubt": these attitudes can be considered as linked with "unexpected, uncertain" mental states.
- A group of social context expressions composed of "infant-directed speech" and "seduction": both of them are specifically addressed to an interlocutor designed by the attitude itself: a young child for "infant-directed speech" and an affectively attractive adult for "seduction". These two attitudes use a breathy voice, as often noted, to imply "take care" or intimacy cues [2, 15].
- "Authority" and "irritation": "authority" could be considered as speech action to recall the speaker's social status that allows him to show his irritation, discontent or displeasure. Therefore, sometimes, they are blended each other.
- "Disappointment" and "resignation": both of them are related to mental states cognitively chained or close.
- A group of "impolite expressions" composed of "irony" and "contempt". The same grouping was showed in

Vietnamese [10]. They could be considered as two different steps of the same mental opinion.

- The other attitudes are grouped in what could be called "ways to declare": "declaration", "confidence", "politeness", "obviousness" except "intimacy" which isn't in the same case and has been supposed to be regrouped with the breathy voice attitudes.

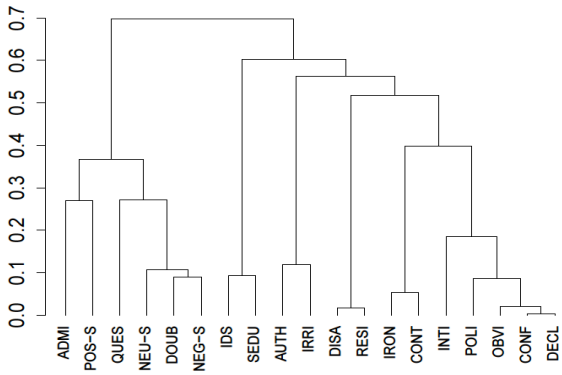

Figure 3: *Hierarchical clustering of the perceived social affects, based on R complete grouping criterion.*

5. Discussion and conclusion

This paper discriminates the social affects from emotion by considering that, on one hand, social affects can be voluntary controlled by the speaker and take part of the speech act cognitive processing, whereas emotions are involuntary produced "during" the speech acts. On the other hand, social affects are learned in a given culture. Therefore, we designed a Mandarin Chinese corpus of attitudinal speech for a perceptual experiment with 15 male and 15 female Chinese listeners.

The main aim was to validate the performances of this single speaker as a reference to study social affects: the attitudes have been recognized over the chance level, except "confidence". The attitudinal values and the length of utterances, as well as the interaction between them have significant effects on the recognition scores. The 1-syllable stimuli received the lowest scores while longer stimuli received clearly higher scores. No significant effect of gender is observed.

Observing the perceptual confusions made by subjects, it can be noted that the "declaration" attitude attracts most confusions, as it was already observed in Vietnamese [10] and English [9]. Moreover, it seems to be a cognitively complex task to recognize a perceived stimulus from one of the 19 attitudinal labels.

Thus, choosing "declaration" is a way to avoid false or uncertain answers without specifying any information about attitude. "Confidence" was poorly perceived by subjects and was mainly confused with "declaration", either because the performance of the speaker (the quality of the stimuli) was not good enough, or because it is a too complex task to get such a cue out of context from prosodic only parameters: visual cues may be of a great help to subject in their identification task, for some kinds of expressions [3, 9, 13].

According to the hierarchical clustering, seven groups of social affects were defined: it is to be verified by acoustic analysis, but by expert listening, we can make the hypothesis that inside a same cluster the affects are not acoustically closed but cognitively closed. The question is now to test this cluster as cognitive categories, inside or across languages, and to measure the prosodic distances within these cognitive distances, including the intonation of modalities (see declaration vs. question in [16].

A cross-cultural perception experiment of Mandarin Chinese attitudes by naive French is under way in order to compare the perceptual behaviour between native and foreign listeners, thus ascertaining the potential difficulties of French learners of Chinese during their language learning and in face-to-face communication with the Chinese people.

These two experiments will be repeated with a male Chinese speaker in order to test further the gender parameter. In parallel, the audio-visual perception processing will be explored, both with native Chinese and native French.

6. Acknowledgements

The corpus recording and cutting could not be conducted without the technical assistance of Christophe Savariaux and Lionel Granjon. This paper is partially supported by Chinese government within a PhD thesis grant.

7. References

[1] Aubergé, V., Gestalt, A., "Morphology of Prosody Directed by Functions", Speech Prosody 2002 Proc., 151-154, Aix en Provence, France, 2002.

[2] Wichmann, A., "Looking for attitudes in corpora", in L. E. Breivik and A. Hasselgren [Ed], Language and Computers, From the COLT's mouth ... and others', 247-261(15), 2002.

[3] de Moraes, J. A., Rilliard, A., Alberto, B. and Shochi, T., "Production and perception of attitudinal meaning in Brazilian Portuguese". Speech Prosody 2010 Proc., Chicago, USA, 2010.

[4] Daneš, F., "Involvement with language and in language", Journal of Pragmatics, 22, 251–264, 1994.

[5] Martins-Baltar, M., "De l'énoncé à l'énonciation: une approche des fonctions intonatives", Didier, Paris, 1977.

[6] Fónagy, Y., "La Vive Voix", Paris, Payot, 1991.

[7] Fujisaki, H. and Hirose, K., "Analysis and perception of intonation expressing paralinguistic information in spoken Japanese," ESCA Workshop on Prosody Proc., 254-257, Lund, Sweden, 1993.

[8] Mejvaldova, J., "Expressions prosodiques de certaines attitudes en thèque et en français: étude comparative", Université Paris 7 – Denis Diderot, Paris, 2000.

[9] Diaferia, M. L., "Les Attitudes de l'Anglais : Premiers Indices Prosodiques". Master thesis in Cognitive Science. National Polytechnique Institut of Grenoble, France, 2002.

[10] Mac, D. K., Aubergé, V., Rilliard, A. and Castelli, E., "How prosodic attitudes can be recognized and confused: Vietnamese multimodal social affects", SLTU, Penang, Malaysia, 2010.

[11] Gu W., Zhang T., and Fujisaki H., "Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes", Proceedings of Interspeech 2011, Firenze, Italy, 1069-1072, 2011.

[12] Shochi, T., Gagnié, G., Rilliard, A., Erickson, D. and Aubergé, V., "Learning effect of prosodic social affects for Japanese learners of French language", Speech Prosody 2010, paper 155, 2010.

[13] Shochi, T., Rilliard, A., Aubergé, V. and Erickson, D., "Intercultural Perception of English, French and Japanese Social Affective Prosody". in S. Hancil [Ed], The role of prosody in Affective Speech, 31-59, Linguistic Insights 97, Peter Lang AG, Bern, 2009.

[14] Cornaire, C., "La compréhension orale", 78-79, CLE International, Paris, 1998.

[15] Campbell, N., "Perception of Affect in Speech –towards an Automatic Processing of Paralinguistic Information in Pro". 8th International Conference on Spoken Language Processing (ICSLP), 881-884, Jeju, Korea, 2004.

[16] Jiahong, Y., "Perception of Mandarin intonation", Proceeding of International Symposium on Chinese Spoken Language Processing 2004, Hong Kong, China, 45-48, 2004.