Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes

Wentao Gu[1]*, Ting Zhang*[1]*, and Hiroya Fujisaki*[2]

[1]Institute of Linguistic Science and Technology, Nanjing Normal University, Nanjing, China
[2]The University of Tokyo, Tokyo, Japan

wtgu@njnu.edu.cn, tingtina_zhang@126.com, fujisaki@alum.mit.edu

Abstract

After differentiating attitudes from emotions, the present work investigates prosodic manifestations and perceptual attributes of Mandarin utterances conveying various attitudes. A speech corpus was designed to incorporate five classes of attitudes: friendly/hostile, polite/rude, serious/joking, praising/blaming, and confident/uncertain. Perceptual experiment reveals two different patterns between intended and perceived attitudes. Statistical analysis of prosodic features shows that speech rate is distinctive in all five classes, while utterance-level F_0 height and F_0 range are distinctive only for some classes. Moreover, F_0 features in the words carrying sentential stress are more distinctive than utterance-level settings. The relation between perception and acoustics is also examined.

Index Terms: attitude, Mandarin, F_0, duration, perception

1. Introduction

Expressiveness of speech is an important topic, not only in the study of phonetics and social psychology, but also in the technology of high-quality speech synthesis. It entails both paralinguistic and non-linguistic information added onto the linguistic information contained in the corresponding text.

In the last two decades, there have been a great number of studies on expressive speech conveying various emotions, attitudes, and intentions. While psychologists have tried to differentiate the meanings of these terms precisely, only very few linguists have paid attention to differentiate 'emotional' and 'attitudinal' speech. Among others, Couper-Kuhlen defined emotion as a speaker state and attitude as a kind of behavior [1], while Wichmann complemented the definition of attitude by incorporating 'propositional attitude' which is a function of opinion, belief or knowledge [2].

Fujisaki classified the information conveyed by speech into linguistic, paralinguistic, and non-linguistic information [3]. Paralinguistic information modifies or supplements the linguistic information of the text and is consciously controlled by the speaker, whereas non-linguistic information concerning speaker's physical or psychological states is not related to the linguistic information of the text and is not consciously controlled by the speaker. Along this line of definition, attitude is paralinguistic, whereas emotion is non-linguistic.

Although emotion and attitude interact in a complex way, in the present study we make a distinction between them in the following way. Emotion is an interior state of the speaker. It is unconsciously conveyed in nature, though it can also be acted. Attitude, on the other hand, is an exterior expression made consciously by the speaker, either associated with intentions to act (named 'behavioral attitude') or related to opinions/beliefs (named 'propositional attitude'); the former is dependent of interaction while the latter is not. Attitude does not necessarily reflect emotion, though it may partly be aroused from emotion.

There have been a number of studies investigating acoustic correlates and perceptual cues for expressive speech, though in many of them emotions and attitudes are mixed or confounded. In fact, only very few studies dealt exclusively with attitudinal speech as defined here. Among others, Fujisaki and Hirose compared contours of confident, interrogative, exhortative, hesitative, as well as neutral speech of Japanese, and examined the accuracy of perceptual recognition [4]. Polite speech has received fairly more studies, e.g., Ofuka et al.'s acoustic comparison on polite and casual speech of Japanese question sentences showed consistent differences in speech rate and utterance-final F_0 movement [5]. Also, Li and Wang's acoustic and perceptual study on friendly speech of Mandarin revealed that pitch is the primary cue for the expression of friendliness, while duration is hardly related [6].

Wichmann argued that the majority of attitudes can only be explained using pragmatic analysis/inference and do not have direct acoustic correlates, because the same prosodic feature can be attitudinally neutral, positive or negative depending on a complex interaction between prosody, text and context [2]. Despite agreeing to this theoretical point of view, we still deem it practically necessary to look into the prosodic correlates of attitudinal speech for the purpose of expressive speech synthesis, in which we think the demand for conveying attitudinal meanings is generally more realistic than for conveying emotions. Although there is no direct prosodic mapping of attitudes without reference to the context, a statistically appropriate and perceptually recognizable prosody would be helpful for the expressiveness of synthesized speech, even if slightly exaggerated.

It is known that both segmental and suprasegmental (prosodic) features play roles in conveying emotions and attitudes. The latter, however, is generally regarded to be primary. Prosodic features include intonation (F_0), duration, intensity, and voice quality. The present study shall be dedicated to investigating the intonational and durational characteristics of attitudinal speech of Mandarin Chinese. Meanwhile, the perceptual outputs of the intended attitudinal meanings will also be examined.

2. Speech Corpus Collection

The studies on attitudinal speech used to compare speech of a certain attitude label with neutral speech, e.g. [5, 6]. However, unlike emotions which usually do not constitute bipolar pairs (e.g., there is not a basic emotion 'unhappy' or 'unsurprised'), attitudes can usually be bipolar. Hence, a class of attitudes can be defined with two opposite poles or a few labels between the poles. Such contrastive study may give clearer results.

Instead of giving a systematic classification of attitudes, the present study shall only investigate five classes of attitudes which are commonly encountered in daily conversation:

Class 1: Friendly vs. Hostile;
Class 2: Polite vs. Rude;
Class 3: Serious vs. Joking;
Class 4: Praising vs. Blaming;
Class 5: Confident vs. Uncertain.

The first four classes of attitudes are towards the listener (hence 'behavioral attitudes'), whereas the fifth class is

towards the content of speech (hence 'propositional attitudes'), indicating whether the speaker is confident or not of what he/she is saying. Within each class, we defined two opposite labels (i.e. two poles), though in reality these attitudes can be expressed in a continuum of degrees. On the basis of general evaluation, the former attitude in each class can be regarded as 'positive,' and the latter as 'negative.'

We term the above five classes of expressions as attitudes rather than emotions, because they are apparently not interior feelings – they cannot be embedded in the frame "he/she is feeling …" Instead, they are either associated with behaviors (classes 1 to 4), or related to beliefs or opinions (class 5). They are all consciously expressed by the speaker. For example, being friendly does not necessarily imply whether the speaker is feeling happy or not; being polite does not imply any emotions of the speaker, for he/she can behave politely in any emotional state; and apparently, confident or uncertain is not related to emotions but dependent on knowledge.

There are generally three methods of collecting expressive speech: simulated speech, elicited speech in a role-play, and spontaneous speech, as are listed in ascending order of naturalness yet descending order of control. For a tradeoff between naturalness and control, the present study uses elicited speech, for which the texts are controlled and hence comparison can be made on the basis of the same linguistic information. Meanwhile, the naturalness of elicited speech is guaranteed in a role-play for which the scenario is given. Most importantly, attitudinal speech should in principle be elicited more naturally than emotional speech, because attitude is exterior and consciously expressed in nature – it is less acted.

We designed the scenarios in two different ways. For behavioral attitudes which require interaction, dialogues were designed. For propositional attitudes which are independent of interaction, only monologues were designed.

For each of the four classes of behavioral attitudes, we designed 15 short sentences composed of 6 to 12 syllables. In classes 1 to 3, the sentences are literally neutral (i.e., not containing any words that imply a specific attitude or emotion) but at the same time can be expressed in two opposite attitudes when embedded in two different dialogues. In class 4, some sentences are literally neutral, while the others are literally praising but can pragmatically be blaming in a specific context (viz., 'ironic' or 'sarcastic' due to the mismatch between word expression and contextual situation).

Then, for each sentence in the four classes, we designed two dialogues to elicit two opposite attitudes, respectively. Each dialogue consists of 2 to 5 utterances. The prompt texts were also given to elucidate the relationship between the two speakers.

For class 5, we designed 15 short declarative sentences each composed of 6 to 10 syllables. The sentences are literally neutral, without implying whether the speaker is certain or not of the content. For each sentence, two versions of prompt text were given to inform the speaker whether to be certain or not. There are no dialogues for this class.

The subjects are 16 native Mandarin speakers, 8 males and 8 females around the age of 20. They are undergraduate students in the major of broadcasting and hosting arts, all professional in pronunciation and skilled in vocal expression.

For each target sentence in the four classes of behavioral attitudes, two dialogues, together with the isolated sentence in neutral reading, were recorded. The target sentences in both dialogues were uttered by the same subject. In dialogue recording, the content of speech was allowed to be altered slightly from the text, *except for* the target sentences. Thus, in each of the five classes of attitudes, each target sentence was uttered in three versions: two opposite attitudes together with neutral reading.

The recording was conducted in a sound-proof room after the subjects had got familiar with the scenarios. To keep the attitudinal expression consistent, the speech sharing the same attitude was recorded in one session.

Only the target utterances in the corpus are examined. We have analyzed the speech data of six subjects, on which our research results will be based. The F_0 values were extracted using Praat, and were then smoothed and interpolated (for voiceless intervals) to obtain continuous F_0 contours. Syllable segmentation was done manually by visual inspection of waveform and spectrogram. Ignoring durational differences, the time-normalized F_0 contour was obtained by extracting an equally spaced 10-point sequence of F_0 values in each syllable.

3. Perceptual Experiment

Attitudes may be perceived differently from the intended ones. To test the validity of the attitudinal speech corpus and to investigate the relationship between expression and perception of attitudes, we conducted perceptual experiment, for which the listening subjects are six native speakers of Mandarin, 3 males and 3 females. They are all graduate students around the age of 20, without any impairs in hearing and comprehension.

The method of constant stimuli was adopted as the test paradigm. All 1,350 target utterances (stimuli) were combined randomly into 90 sound files, each composed of 15 stimuli with an inter-stimuli interval of 10 seconds. These sound files were presented to the listening subjects through headphones in a sound-proof room. Within each 10-second inter-stimuli interval, the listeners were requested to give two answers:

(1) perceived attitude (chosen from the attitude labels in the given class of attitudes, including 'neutral;' or 'unsure' when failing to judge);
(2) the word carrying sentential stress, if any.

Before the experiment, a training session was repeated until the listening subjects could give the answers confidently.

Table 1 shows the perceptual results for all five classes of attitudes. In the table, '+', '–', and '0' represent two opposite attitude labels and neutral speech, respectively. In a given class, for each intended attitude label (including neutral), the percentages of perceived attitude labels are calculated over all six listening subjects. The values on the diagonal (as shaded) indicate the rates of coincidence between the intended and perceived attitudes.

Table l. *Percentages of perceived attitude labels (%)*

Attitude class	Intended	Perceived			
		+	0	–	unsure
Friendly/Hostile	friendly (+)	85.2	10.7	3.1	0.9
	neutral (0)	10.2	83.3	5.6	0.9
	hostile (–)	3.0	4.3	92.6	0.2
Polite/Rude	polite (+)	88.6	10.8	0.5	0
	neutral (0)	15.8	81.1	2.7	0.3
	rude (–)	2.5	4.2	92.5	0.8
Serious/Joking	serious (+)	87.6	9.0	2.8	0.6
	neutral (0)	22.8	76.3	0.7	0.2
	joking (–)	4.3	0.4	95.4	0
Praising/Blaming	praising (+)	81.0	4.2	14.1	0.8
	neutral (0)	8.7	86.7	4.4	0.2
	blaming (–)	16.7	4.6	78.0	0.8
Confident/Uncertain	confident (+)	85.0	14.4	0.6	0
	neutral (0)	22.0	77.6	0.2	0.2
	uncertain (–)	1.5	1.3	97.2	0

It is observed that the perceptual patterns are quite different between class 4 and other classes:

(1) In all classes except 4, the rates of coincidence are consistently in the order of "negative > positive > neutral." This indicates that the attitudes conveyed in speech are easily perceived. The negative attitudes are perceived even more accurately, indicating that the expression of negative attitudes here are more prominent.

(2) In class 4, the rates of coincidence are in the order of "neutral > positive > negative," exactly reverse to the order in other classes. In particular, the rate of coincidence for 'blaming' is much lower than in other classes.

(3) In all classes except 4, perceptual confusion occurs mainly between positive attitudes and neutral. In particular, the rate of identifying neutral reading as 'serious' or 'confident' is noticeably high. This may be because that neutral reading gives a placid style which resembles the speaking style in serious attitude – in fact neutral reading can be regarded as somewhat serious in nature. Also, neutral reading of declarative sentences is assertive in nature, hence resulting in the similarity between neutral and confident.

(4) In class 4, perceptual confusion is mainly between positive and negative ones. It is to be noted that many sentences in this class are literally praising but verbally can be blaming (i.e., ironic or sarcastic) when spoken in a specific context. The distinction between praising and blaming may be cued not only by prosody but also (even mainly) by the mismatch between word expression and contextual situation. Thus, when an utterance is disassociated from the context, attitudes of this kind may not be inferred reliably. This explanation is a support of Wichmann's argument [2].

Table 2 lists the numbers of perceived sentential stress in all target utterances (henceforth we list '+' and '–' adjacent to each other in the tables, for the contrast between two opposite attitude labels is the main point of interest). A word is deemed to carry sentential stress when more than half of the listening subjects agreed in judgment. As shown, more sentential stresses are perceived in attitudinal speech than in neutral one. Also, attitudes can transfer the position of sentential stress. For example, hostile speech and blaming speech tend to add stress at utterance-medial positions, while the speech expressing uncertainty tends to add an utterance-final stress.

4. Prosodic Analysis

It is well known that emotions tend to affect the utterance-level prosodic settings [7, 8]. For example, a wide F_0 range is usually associated with active emotions, while a narrow F_0 range is associated with passive emotions. A similar study can be applied to attitudinal speech.

Table 2. *Numbers of perceived sentential stress*

Attitude class	Attitude label	non-final	final	Total
Friendly/Hostile	friendly (+)	40	44	84
	hostile (–)	53	28	81
	neutral (0)	4	27	31
Polite/Rude	polite (+)	30	18	48
	rude (–)	37	21	58
	neutral (0)	23	9	32
Serious/Joking	serious (+)	38	37	75
	joking (–)	37	45	82
	neutral (0)	24	30	54
Praising/Blaming	praising (+)	23	52	75
	blaming (–)	44	38	82
	neutral (0)	18	35	53
Confident/Uncertain	confident (+)	20	55	75
	uncertain (–)	15	67	82
	neutral (0)	15	36	51

Three utterance-level prosodic features were calculated: mean syllabic duration (implying speech rate), mean F_0 height, and F_0 range over an utterance. They were then averaged over all utterances from all six subjects (i.e., over 90 samples) for each attitude label. Table 3 lists the average measurements for all attitude labels in all five classes of attitudes. The height and range of F_0 were both calculated and averaged in the logarithmic domain, though the F_0 height presented in the table has been converted back to Hz for the sake of clarity.

Multiple paired *t*-tests with Bonferroni adjustment were conducted between each pair of attitude labels within each class. The prosodic features for the two utterances sharing the same text sentence and the same subject but differing in attitude label are paired in the *t*-tests. In Table 3, the values darkly shaded indicate *no* significant difference between the two opposite attitudes, while the values lightly shaded indicate that neutral state is at least *not* significantly different from one end of attitude in the same class. The differences between all others are statistically significant ($p < 0.05$), and in fact most of them are extremely significant ($p < 0.001$).

The results of statistical analysis of prosodic features here are only slightly correlated with the results of perceptual experiment. For instance, the relatively high rate of perceptual confusion between serious and neutral may partly be ascribed to insignificant differences in F_0 height (both are 172) and in F_0 range (1.09 vs. 1.17). The relationship between acoustic and perceptual results, however, is not obvious, suggesting that there may be other acoustic cues for attitudinal speech.

The quantitative results in Table 3 are summarized below:

(1) Friendly/Hostile: friendly speech is slower and of narrower F_0 range than hostile one; both are faster and of higher F_0 than neutral speech. The comparison on friendly and neutral speech coincides with [6].

(2) Polite/Rude: polite speech is slower than rude one; both are faster and of higher F_0 than neutral speech.

(3) Serious/Joking: serious speech is faster and of lower F_0 and narrower F_0 range than joking one.

(4) Praising/Blaming: praising speech is faster and of higher F_0 than blaming one; both have higher F_0 than neutral speech.

(5) Confident/Uncertain: confident speech is faster and of lower F_0 than uncertain one; both have higher F_0 than neutral speech.

Table 3. *Mean values of prosodic features*

Attitude class	Attitude label	Syl.dur (sec.)	F_0 height (Hz)	F_0 range (octave)
Friendly/Hostile	friendly (+)	0.166	184	1.09
	hostile (–)	0.153	180	1.29
	neutral (0)	0.173	165	1.19
Polite/Rude	polite (+)	0.165	181	1.19
	rude (–)	0.154	180	1.19
	neutral (0)	0.178	169	1.25
Serious/Joking	serious (+)	0.167	172	1.09
	joking (–)	0.178	195	1.38
	neutral (0)	0.179	172	1.17
Praising/Blaming	praising (+)	0.176	195	1.27
	blaming (–)	0.190	182	1.34
	neutral (0)	0.180	171	1.26
Confident/Uncertain	confident (+)	0.180	190	1.39
	uncertain (–)	0.199	234	1.30
	neutral (0)	0.191	176	1.21

As a result, speech rate is distinctive in all the five classes of attitudes, F_0 height is distinctive only for serious/joking, for praising/blaming, and for confident/uncertain, while F_0 range is distinctive only for friendly/hostile and for serious/joking.

In the above, we only compared utterance-level prosodic settings, which however are not the only prosodic correlates for the expressiveness of speech. Prosodic details in minor units may sometimes play more important role in conveying emotions or attitudes. In particular, prosodic contrasts in stressed syllables are more prominent than general. Previous studies have shown that stressed words carry more identifiable acoustic features for emotions than unstressed words [7].

Hence, in each class of attitudes, we calculated local F_0 features for the words carrying sentential stress consistently in all three attitude labels. Table 4 lists the average measurements for the minimum F_0, maximum F_0, and mean F_0 values in stressed words among six subjects.

By comparison, the differences in F_0 are more conspicuous in stressed syllables than in general. In all five classes of attitudes, F_0 of stressed word in attitudinal speech is consistently higher than in neutral speech. In the classes of friendly/hostile and praising/blaming, F_0 of stressed word is higher in positive attitude, while in other three classes F_0 of stressed word is higher in negative attitude. Moreover, it is observed that the contrasts in maximum F_0 are more conspicuous, while the differences in minimum F_0 are relatively small. This indicates that the topline F_0 is more characteristic of attitudinal expressions.

Table 4. *Average F_0 features for the words carrying sentential stress consistently in all three attitude labels*

Attitude class	Attitude label	min F_0 (Hz)	max F_0 (Hz)	mean F_0 (Hz)
Friendly/Hostile	friendly (+)	157	260	214
	hostile (–)	153	243	199
	neutral (0)	139	228	184
Polite/Rude	polite (+)	167	254	223
	rude (–)	177	298	233
	neutral (0)	157	229	197
Serious/Joking	serious (+)	152	246	196
	joking (–)	157	294	219
	neutral (0)	134	215	172
Praising/Blaming	praising (+)	141	307	219
	blaming (–)	146	274	198
	neutral (0)	129	234	181
Confident/Uncertain	confident (+)	141	239	191
	uncertain (–)	161	329	241
	neutral (0)	128	219	170

5. Conclusions

We differentiated attitude from emotion by defining attitude as an exterior expression imposed consciously by the speaker. Along this line of definition, attitudinal speech can be naturally elicited in a context, being less acted or contrived than for emotional speech. Thus, a corpus of attitudinal speech of Mandarin was designed, incorporating four classes of behavioral attitudes and one class of propositional attitudes.

Perceptual experiment shows two patterns between the intended and perceived attitudes. Except for praising/blaming where irony is involved, the rates of perceptual recognition are consistently in the order of "negative > positive > neutral," where positive and neutral are apt to be confused. Statistical analysis of utterance-level prosodic features reveals that speech rate is distinctive in all five classes of attitudes, F_0 height is distinctive only for serious/joking, for praising/blaming, and for confident/uncertain, while F_0 range is distinctive only for friendly/hostile and for serious/joking. A further comparison on the words carrying sentential stress reveals that F_0 features in stressed words are more distinctive than utterance-level settings in all five classes. The relation between acoustics and perception has also been discussed.

Obviously, much remains to be explored. More classes of attitudes are to be defined and studied. Prosodic analysis can be done in more detail and more reliably when more speech data from more subjects are available. In addition to surface prosodic features, a model-based analysis will also be applied to continuous F_0 contours, as already done for emotional speech of Mandarin [8]. Besides, voice quality, as another important acoustic correlate of attitudinal speech, will be examined in future work.

6. Acknowledgements

This work is supported jointly by the Chinese National Social Science Fund (10CYY009), Jiangsu Social Science Fund (09YYB006), and "211" leading academic discipline project "Linguistic Science & Technology Innovation and Platform Construction" from Nanjing Normal University.

7. References

[1] Couper-Kuhlen, E. *English Prosody*, London: Edward Arnold, 1986.

[2] Wichmann, A., "The attitudinal effects of prosody and how they relate to emotion," *Proc. ISCA Workshop on Speech and Emotion*, pp. 143-148, Newcastle, North Ireland, 2000.

[3] Fujisaki, H., "Prosody, models, and spontaneous speech," In: Sagisaka, Y., Campbell, N., Higuchi, N. (eds.), *Computing Prosody*, 27-42, New York: Springer-Verlag, 1996.

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

[5] Ofuka, E., McKeown, J. D., Waterman, M. G. and Roach, P. J., "Prosodic cues for rated politeness in Japanese speech," *Speech Communication* 32(3): 199-217, 2000.

[6] Li, A. and Wang, H., "Friendly speech analysis and perception in Standard Chinese," *Proc. ICSLP*, pp. 897-900, Jeju, Korea, 2004.

[7] Zhang, S., Ching, P. C. and Kong, F., "Acoustic analysis of emotional speech in Mandarin Chinese," *Proc. ISCSLP*, pp. 57-66, Singapore, 2006.

[8] Gu, W. and Lee, T., "Quantitative analysis of F_0 contours of emotional speech of Mandarin," *Proc. 6th ISCA Speech Synthesis Workshop*, pp. 228-233, Bonn, Germany, 2007.