Acoustic Analysis of Friendly Speech

Fangxin Chen[+], *Aijun Li*[++], *Haibo Wang*[++], *Tianqing Wang*[++], *Qiang Fang*[++]
liaj@cass.org.cn

[+] IBM China Research Laboratory
[++] Phonetic Lab. of the Linguistics Institute, Chinese Academy of Social Science

Abstract*

To provide acoustic information of affective speech for computer modelling, this study conducted an investigation on the friendly Mandarin Chinese speech. Based on the perceptually classified friendly and neutral speech data, spectral, tonal and durational analyses were conducted at different phonetic levels. The distinctive acoustic features of friendly speech were sifted out and followed by tentative discussions.

1. Introduction

Natural speech carries not only linguistic, but also para-linguistic information, such as the speaker's social, physiological and affective status. Thanks to the significantly improved Text-to-Speech (TTS) intelligibility, current TTS engines can reasonably well convey the literal meaning of an input text. However, the expression of para-lingusitic information of a text, especially with respect to the speaker's affective states, still remains a big challenge to the existing TTS technology.

With the awareness of the intricacy of the term *affect* in its connotation with sociology, psychology and linguistics [1], this study treated the issue only from the viewpoint of speech technology. *Affect* was considered as consisting of two subcategories: *emotion* and *attitude*. *Emotion* refers to the speaker's 'internal feeling', such as *joy*, *angry*, *sad* and *surprised* etc. While *attitude* to the speaker's external behavior', such as *friendly*, *appreciative* and *apologetic*, etc.

There were reports on acoustic investigation of the basic *emotion* states (such as happiness, anger, sadness, neutral etc) and the initial modeling of those states in the TTS systems [2][3], . The acoustic study of the more complicated affective states, especially with the *attitudinal* aspect of speech, however, were little touched, mostly due to their fuzziness in both perception and acoustics manifestation, as well as their strong dependency on speech context. Nevertheless, the expressiveness involved in real dialogue situation is complicated and the ability to model affective speech beyond the basic emotion states could be crucial for the success of a TTS dialogue system. To improve the expressiveness of the existing TTS system for dialogue applications, we have been conducting an expressive speech research project investigating the acoustic aspects of affective states most relevant to the dialogue situation. This paper reports the initial results on the *friendly* speech, Other affective states are still under investigation, and will be reported when the results are ready.

2. Experimental Design

As mentioned, the investigation of the *friendly* speech is one part of our expressive speech research project. Therefore, the expeimental design was under the project's global scheme. Initially eight commonly encounted *emotional* and *attitudinal* states in dialogue situations were identified. They were: *friendly, excited, appreciative, apologetic; confused, surprised, complaining* and the default state of *neutral.* Based on the above eight states, fifteen short dialogue passages were designed in a hypothetical dialogue application context to stimulate the desired affective speech, with each passage focused on one or two target affective states. The data script designing also took the segmental and tonal balance into consideration.

Two male drama school students were initially recruited for the dialogue recording. The dialogue passages were distributed to the speakers for rehearsal before recording. During recording, the speakers were instructed to carry the dialogue naturally in the defined affective modes rather than simply recite the text. Since the speaker could deviate from the text for more natural conversation, the recorded affective speech was transcribed back to text again. With that text, the speakers repeated the recording. That measure guaranteed the recorded dialogues for both the neutral and affective speech are exactly the same. All the recordings were done together with the laryngegraph signals in a soundproof room, sampled at 22k. The recordings were then cut into individual sentences. A perceptual test was performed to classify those sentences into the eight defined affective categories. There was an additional category when the listener thought none of the listed eight categories was suitable. In that case, the listener needed to fill out any other affective states he/she thought to be. As a measure to reduce the semantic effect of the sentence on listener's decision of the affective states, each sentence read with intended affective state was paired with its corresponding neutral state sentence. Listeners were first asked to listen to the randomized sentence pairs and answer the question: whether the two sentences in the pair were basically the same with respect to either the affective state, or degree of affect. If there was no significant difference, either in affective category, or in degree of affect, this pair of sentences was considered as containing no perceptually distinctive acoustic information for a particular affective state beyond the literal meaning, and subjected no further acoustic investigation.

Otherwise, the listener needed to further classify the affective states of the two sentences in the pair. One should always be in the neutral state, and the other any of the other affective states. Each pair of sentence was listened by six linguistically experienced listeners. Sentences with four or more listeners identified as the same affective state grouped together. From a total 1123 neutral vs. expressive sentence pairs, 185 sentence pairs were identified as perceptually distinctive *friendly* vs. *neutral* speech.

We assumed that in the perception test, the filtering measure applied in the first step had significantly reduced the possible semantic effects of the sentences, and the listener's judgment in the second step on the affective category of the sentences was mostly based on the acoustic, rather than the semantic or/and context information. That was the justification for the ensuing acoustic analyses of the friendly speech based on the data sets classified that way.

3. Acoustic Feature Analysis

The perceptually classified *friendly* vs. *neutral* speech data set was first undergone annotation. The boundaries at phone, syllable, prosodic word, prosodic phrase and intonation phrase levels, the pause levels, the syllable stress levels and the lexical tone types of the syllables and the phonetic transcription were all

* 本文在ICASSP2004上发表。

manually labeled. The pitch values were extracted automatically and edited with manual checking to ensure accuracy.

3.1 Spectral Features

Spectral analysis was conducted at phone category levels. The spectral analyses in this study mainly focused on the relative energy level variation at different frequency ranges due to the change of speech mode, rather than the overall energy level difference between different speech modes though we were aware that there existed overall energy level difference between the *friendly* and *neutral* speech.

The long-term average FFT energy value was computed with 10 bands covering 10k frequency range for each phone of both friendly and neutral speech data. Each band was 1k Hz. All the phones were grouped into eight phonetic categories: *unaspirated stops* (stop1), *aspirated stops* (stop2), *fricatives (fric), unaspirated affricates (affri.1), aspirated affricates (affri.2), nasals, vowels (vow) and diphthongs (diph)*. For each phonetic category, the averaged energy difference scores at each band were computed based on the following formula:

$$aE_i = (\sum_{k=1}^{n}(A_{ki}/A_{kT} - B_{ki}/B_{kT}))/n \qquad (1)$$

where aE_i is the averaged *ith* band energy variation ratio of a particular phonetic group; *n* is the total number of phones in that phonetic category; A_{ki} is the *ith* band energy value for the *kth* friendly state phone; A_{kT} is the total energy value of the *kth* friendly state phone; B_{ki} and B_{kT} are the corresponding energy values for the *kth* neutral state phone.

Table 1: Energy variation ratios at different frequency bands for different phonetic categories of the *friendly* vs. *neutral* speech.

	1kHz	2kHz	3kHz	4kHz	5kHz	6kHz	7kHz	8kHz	9kHz	10kHz
stop1	0.037	0.024	-0.016	-0.043	-0.010	0.001	0.003	0.001	0.000	-0.002
stop2	0.108	-0.010	-0.016	-0.068	-0.028	0.008	0.004	0.001	0.000	0.000
affric.1	0.155	-0.004	-0.024	-0.154	-0.009	0.008	0.015	0.012	0.008	-0.008
affri.2	0.134	0.001	-0.024	-0.114	-0.036	-0.013	0.033	0.014	0.009	-0.004
fric.	0.144	0.003	-0.019	-0.106	-0.035	-0.002	0.010	0.007	0.003	-0.004
nas.	0.027	0.002	-0.012	-0.016	-0.001	0.000	0.000	0.000	0.000	0.000
glide	0.024	0.008	-0.015	-0.015	-0.001	0.000	0.000	0.000	0.000	0.000
vow	0.063	0.004	-0.022	-0.044	-0.001	0.000	0.000	0.000	0.000	0.000
diph.	0.048	0.012	-0.025	-0.035	0.000	0.000	0.000	0.000	0.000	0.000

Table 1 lists the computed energy variation ratios at different frequency bands for different phonetic categories as referenced to the total energy of the phone.

3. 2. Lexical tone Features

MC is a lexical tone language. Its intonation could be seen as the concatenation of a sequence of *context-dependent* lexical tones. The *context* here means the lexical, syntactic and semantic position of the current tone carrier syllable is in. To examine the intonation for friendly MC speech, then, the task became the comparison of lexical tones at certain *context* with their corresponding tones in the neutral speech state. There are four MC lexical tones in citation form: *high level (tone1), rising (tone2), falling-rising (tone3) high falling (tone4).* Taking the tone *sandhi* phenomenon into consideration, we further divided Tone3 into three subcategories: those Tone3 syllables located at the end of a prosodic phrase ($tone3_1$), those followed by another *Tone3* syllable ($tone3_2$), and those followed by other tone syllables ($tone3_3$). The reason for the sub-division of *Tone3* was that *Tone3* has different pitch representations when it is followed by different tones, or it is at a prosodic boundary. *Neutral* tone has no fixed pitch pattern in citation form, but it has relatively stable pitch patterns in continued speech. Therefore, neutral tone was also considered as an independent tone category in this study. Consequently, there were seven tone categories divided. Taking the syntactic position effect into consideration, the seven tone categories were further classified into twenty-one tone groups based on their respective syllable position in an intonation phrase (IP). IP initial was the syllable at the very beginning of an IP, IP final was the syllable at the very end of an IP. All the other syllable positions were treated as IP middle. We were aware that other syntactic and semantic factors could also potentially affect the pitch patterns of a syllable, such as the word stress, nuclear accent etc, but they were not considered in this initial investigation to avoid too much complication.

All the syllables in the *friendly* and *neutral* speech data were classified into the twenty-one tone groups, and the pitch was extracted for each syllable with ten F0 values. Based on the ten F0 values, we got the mean, pitch onset (*F01*), middle (*F02*) and offset (*F03*) for each syllable. *F01* was the average of the first three F0 values of the syllable, F02 of the next four and *F03* of the last three values. The *friendly vs. neutral* difference F0 scores for each pair of syllables were calculated according to the following formula.

$$aF_0 = (\sum_{k=1}^{n}(F_{0Ak}/F_{0Bk} - 1))/n \qquad (2)$$

where aF_0 is the averaged pitch variation ratio for a particular tone in a particular tonal and syntactic positions; n is the syllable number for a particular tone in a particular tonal and syntactic positions; F_{0Ak} and F_{0Bk} are the pitch values of the friendly and neutral speech for a particular tone in a particular tonal and syntactic positions respectively.

The averaged pitch variation results are listed in Table 2, where A, B, C refer to IP initial, middle and final position respectively. For example, Tone1A refers to Tone1 at IP initial, and so forth. All the tone group results were subject to T-tests for the validity of the variation. The value marked with* in Table 2 indicate that the variation in that category was not statistically significant from the T-test result.

Table 2: Pitch variations of different tone groups for *friendly* vs. *neutral* speech.

	F01	F02	F03	F0 Mean
Tone1A	26%	27%	28%	27%
Tone1B	23%	24%	24%	24%
Tone1C	10%	8%	*3%	7%
Tone2A	16%	18%	17%	17%
Tone2B	25%	25%	25%	25%
Tone2C	14%	11%	12%	12%
Tone31C	17%	9%	*2%	10%
Tone32A	15%	12%	10%	12%
Tone32B	25%	25%	23%	24%
Tone33A	13%	13%	13%	13%
Tone33B	22%	21%	19%	21%
Tone4A	21%	25%	31%	26%
Tone4B	25%	28%	28%	27%
Tone4C	19%	18%	13%	15%
Tone0B	20%	23%	25%	22%
Tone0C	13%	15%	16%	15%

3. 3. Temporal features

Temporal comparison of the *friendly* vs. *neutral* speech was based on phonetic categories, which include *Unaspirated Stops* (Stop1), *Aspirated Stops* (Stop2), *Unaspirated Affricates* (AF1),

Aspirated Affricates (AF2), *Fricative* (Fric.), *Nasal Onset* (Nas), vowel with *nasal Offset* (V+N), *Vowel* (V1), *Diphthong* (V2) and triphthong(V3). The phone category duration variation scores are listed in the following table.

Table 3: Duration reduction ratio of different phonetic categories in *friendly* vs. *neutral* speech.

	%
Stop1	-9.7
Stop2	-8.3
AF1	-9.4
AF2	-7.1
Fric.	-11.1
Nas	-8
V+N	-11.6
V1	-11.3
V2	-11.5
V3	-11.7

3.4. Articulatory features:

Another interesting phenomenon observed in friendly speech data is the occasional adding of breathy glottal sound at the starting of friendly conversation as illustrated in Figure 1.

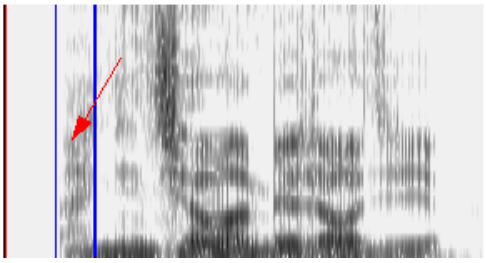

Figure 1: A case of breathy glottal sound at the beginning of a friendly speech.

The spectrum between the blue bars pointed by the arrow in the above figure is the part of that additional breathy glottal sound, which is about 90 ms in duration. That kind of glottal sound was found in both speakers' friendly speech. The phonetic realization of glottal sound were different speaker wise, but they mostly happened at the starting of a dialogue, rather than in the middle of a conversation.

4. Discussion

For the friendly speech, energy level was increased by different degrees around the frequency range of 0-1.5k range for different phones. At the frequency range of 2.5-5 k, the energy level was decreased. There are slight energy level variations at the higher frequency ranges, but not that significant.

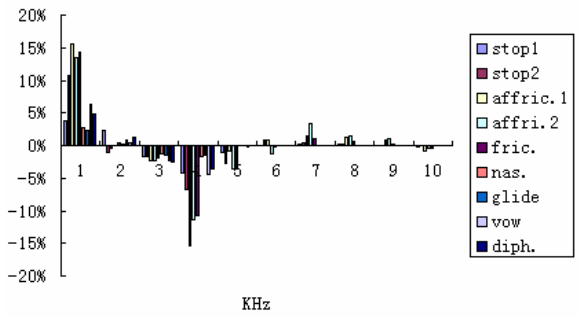

Figure 2: Averaged relative energy variation in different frequency bands for different phonetic categories of the *friendly* vs. *neutral* speech.

As for lexical tones of *friendly* speech, we can observe from Figure 3 that:

- The averaged *pitch mean was moved up for all the lexical tones;*
- The averaged *pitch mean* at either *IP initial* or *middle* was raised more as compared with that at the *IP final* for all the lexical tones;

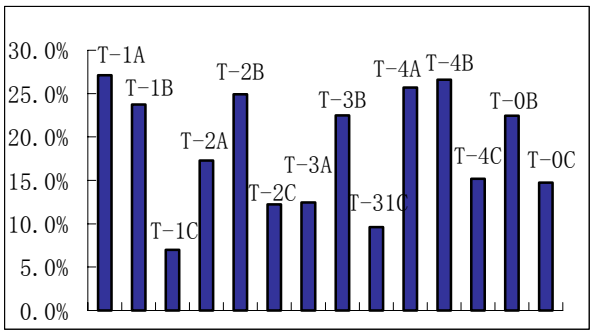

Figure 3. Averaged Pitch variations of different tones at IP initial (A), middle (B) and final (C) position for *friendly* vs. *neutral* speech.

The averaged *pitch mean* moving up for all the tones, however, does not mean that it was a simple moving up of the pitch Baseline for all the tones in the friendly speech mode. As shown in Table 2, the pitch offset for Tone1 and Tone3 at IP final did not actually change much as compared with the neutral speech state. Awareness of this phenomenon is important in our future modeling of lexical tones in the affective speech.

With regard to phone duration, as indicted in Figure 4, the averaged durations of all the phone categories were significantly shortened for the *friendly* speech as compared with their counterparts in the neutral speech. The duration shortening was phone-dependent. *Fricatives*, *vowels*, *diphthongs*, *triphthongs* showed more reduction in duration as compared with other phone categories.

The last , but not the least important acoustic phenomenon is the glottal sound occationally added at the beginning of a friendly conversation. One interpretaton is that a speaker uses that kind of sound to signal friendness or intimacy to the listener(s). This phenomenon could be cross-languages and worth further investigation.

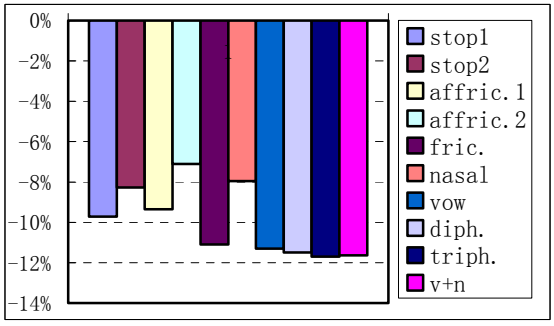

Figure 4. Duration reduction ratio of different phonetic categories in *friendly* vs. *neutral* speech.

5. Conclusion

The distinctive acoustic features of the friendly speech from this investigation were based on the perceptually classified affective speech dataset of two male speakers. Currently we are working on collecting and analyzing two female speaker data for comparison. We also plan to model the above acoustic features of the friendly speech in our parameter-based TTS system. With the perception test on the synthesized speech, we can validate the results came up with the present study.

Acknowledgement: Thanks to M. Pichney and E. Edie of IBM Watson Lab. for sharing with us their expressive speech research methodology .

6. Reference

[1] Wichmann, A., 2002, The attitudinal effects of prosody, and how they relate to emotion, ISCA Workshop on Speech and Emotion.

[2] Schroder, M., 2001, Emotional speech synthesis : a review, Eurospeech 2001, Scandinavia.

[3] Yuan, J., Shen, L., Chen, F., 2002, The acoustic realization of anger, fear, joy and sadness in Chinese, Speech Prosody 2002, France.