Perception on Synthesized Friendly Standard Chinese Speech

Aijun Li[+], Fangxin Chen[++], Haibo Wang[+], Tianqing Wang[+]
liaj@cass.org.cn; chfx@cn.ibm.com

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

Abstract*

This paper presents a perceptual experiment on the friendliness of the synthesized Standard Chinese. Based on the acoustic investigation of friendly speech, tonal pitch, phone duration and spectral energy distribution were adjusted in synthesis and the synthesized stimuli were subjected to perception test. It was found that: (1) Friendliness of synthesized speech could be achieved via adjusting the perceptually distinctive acoustic parameters; (2) Tonal pitch is the most prominent cue for better expression of friendliness. However, the optimal adjustment for friendliness is the combination of pitch, phone duration and spectral energy.

1. Introduction

The acoustic features of emotional speech have been widely investigated. Some new features are found useful for classifying some kind of emotional or expressive speech. For example, it was found in [2] that "spectral tilt serves as a good predictor of angry, anxious, bored and friendly." Here the spectral tilt refers to first harmonic subtracted from second harmonic, measured in dB, over a 30ms window centered over the middle of the vowel.

To improve the expressiveness of the existing TTS system for dialogue applicatons, we have been conducting an expressive speech research project investigating the acoustic aspects of affective states most relevent to the dialogue situation. Based on the perceptually classified friendly and neutral speech data, spectral, tonal and durational analyses were conducted at different phonetic levels. The study was reported in [1]. Following is a brief summary of that study on the acoustic analysis of friendly state speech for your easy reference.

For the friendly state speech, energy levels for different phonetic categories were increased to different degrees around the frequency range of 0-1k range. At the frequency range of 3-5 k, the energy levels were decreased significantly. There are slight energy level variations at the higher frequency ranges, but not that significant (see Figure 1).

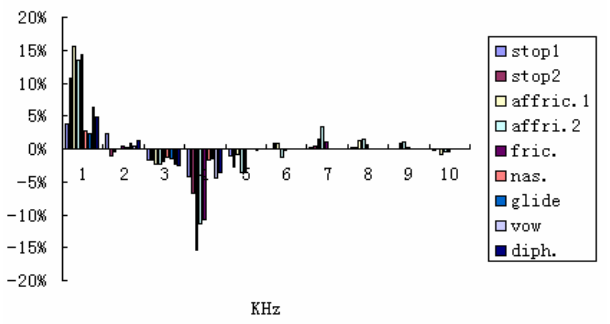

Figure 1: Averaged relative energy variation in different frequency bands for different phonetic categories of the *friendly* vs. *neutral* speech.

As for lexical tones of friendly speech, we found that (see Figure 2):

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

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. The pitch offset for Tone1 and Tone3 at IP final did not actually change much as compared with the neutral speech status (see Table 1). Awareness of this phenomenon is important in the modeling of lexical tones in the affective speech.

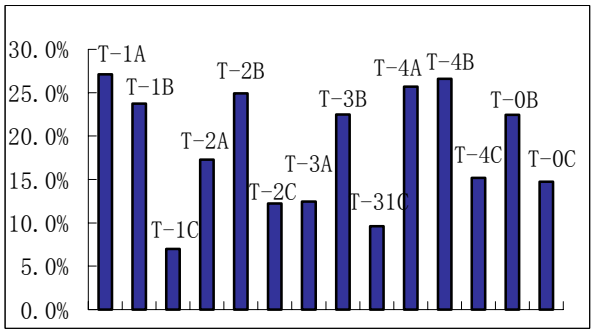

Figure 2: Averaged Pitch variations of different tones at IP initial (A), middle (B) and final (C) position for friendly vs. neutral speech.

With respect to phone duration, 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 degree of duration reduction was phone-dependent. Fricatives, vowels, diphthongs, triphthongs and nasal offset showed the biggest reduction in duration (see Figure 3).

Table 1: Pitch variations of different tone groups for friendly vs. neutral speech. F01, F02 and F03 represent the lexical tone onset, middle and offset respectively.

	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%

* 本文在国际语调和声调研讨会（TAL2004）上发表。

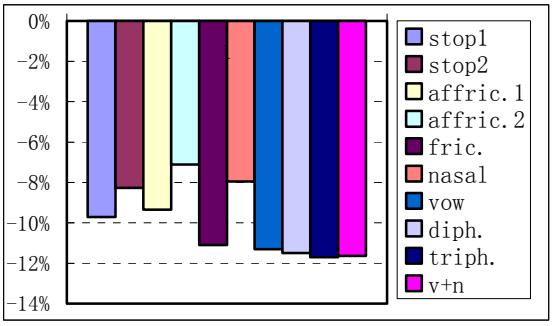

Figure 3: Duration reduction ratio of different phonetic categories in friendly vs. neutral speech.

How the above acoustic features contribute to friendly state perception? Which one is most prominent? How to adjust them in our existing TTS system to get more friendly speech? In this paper, we would like to answer these questions through our perceptual experiments on the synthesized utterances with nine different acoustic cue combinations.

2. Stimuli Preparation

Stimuli of friendly speech were synthesized by IBM's formant TTS system. First, the friendly state annotation mechanism was implemented in the text processing procedure. With the text annotated in friendly state, the corresponding acoustic model will be applied in the synthesis process. The user can insert the friendly state tag at any place of the input text and the ensuing text would be read in the friendly mode until the neutral state tag was encountered.

Table 2: Acoustic feature combinations used for synthesizing friendly speech (1-9)

P.	Features/method	Statement
1	D	Duration modified
2	DE	Duration and Energy modified (tilt=5)
3	E	Energy modified (tilt=5)
4	EP	Energy and Pitch modified (tilt =5)
5	EPD	Energy, Pitch and Duration modified
6	P	Pitch modified
7	PD	Duration and Pitch modified
8	EiP	Energy and Pitch modified (tilt=10)
9	EiPD	Energy, Pitch and Duration modified (tilt=10);
10	N	Default (Neutral)

The acoustic modeling of the friendly state was based on the lexical tone, phone duration and energy features of the friendly speech as illustrated in part 1. The lexical tone target register was adjusted according to the statistics of table 1. F01, F02 and F03 are the lexical tone onset, middle and offset for each syllable respectively. A, B, C refer to IP initial, middle and final position respectively. Tone 1,2,3,4 stands for 4 lexical tones. 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 (tone31), those followed by another Tone3 syllable (tone32), and those followed by other tone syllables (tone33). Durations were lengthened for different phone categories according to the statistics in Figure 3. For the energy change pattern at different frequency bands for the friendly state, as illustrated in Figure 1, the existing Formant synthesizer has no effective way to enhance or reduce energy at a particular frequency band. As an approximation, the tilt parameter in the formant synthesizer was used to suppress the higher frequency. Since the synthesizer was set at 11k sampling rate, increasing tilt value virtually reduced the energy levels at the higher frequency bands, which covers the frequency band around 2-5 kHz. Certainly this approximation of energy adjustment did not conform ideally what really happened for the friendly speech. The perceptual results concerning the energy adjustment should be interpreted accordingly.

To examine the degree of importance of the above three acoustic cues to the perception of friendly state, we synthesized the first dialogue in our data script into 10 passages with different acoustic cue combinations and subjected them to the perception tests as shown in table 4. The three acoustic elements are represented by D (duration), E(energy) and P(Pitch).

Passage 8 and 9 are the same as Passage 4 and 5 except that the tilt value was increased from 5 to 10. The purpose of varying tilt was to check if the degree of spectral tilt would affect the perception of friendliness.

Dialogue:

A. 你好,是花鸟公寓一号吗?
S1. 是啊,请问先生想购房吗?
A. 你们现在有新开盘的楼盘是吧?
S2. 有,咱们的一号楼是礼拜一刚开盘的,选择性还比较大.常说的户型都有.
A. 一号楼也是均价五千二吗?
S3. 对对对, 也是五千二一平米.
A. 交通方便吗?
S4. 交通应该是非常方便的. 它位于玉佛寺环岛西南侧。城铁和十七，四十，二零三路公交车站只要几分钟就走到了。
A. 小区有娱乐设施吗?
S5. 咱们是属于旅游休闲度假的高雅社区。当然少不了各种娱乐设施. 除了健身房，游泳池，温水游泳馆，和篮球场等，在社区的西北方还有一个高尔夫球场呢!
A. 行,那谢谢您啊!
S6. 没事,有兴趣的话您可以再打电话来.您过来看看也行啊.
A. 好嘞!再见.
S7. 再见!

---

3. Experimental Design

In each synthesized passage, utterences S1-S6 were selected for perception. For each sentence, we got 45 stimulus pairs from 10 passages (with different acoustic cue modifications), and total 6*45=270 tokens for all sentences. In the perceptual experiment, each token was presented to the subjects with a neutral utterence as a baseline.

Five native speakers (two male and three female graduate students), without hearing problems, age from 20-30, were recruited for the experiment. They were asked to rate the friendness of two utterances in each pair by comparing them with the corresponding 'neutral' utterance. No friendliness utterance marked as 1, weak friendliness as 2, and strong friendliness as 3.

A testing program was prepared for perception experiment. The tokens were played randomly and could be easily repeated by the subjects. After listening, they rated two sentences for each pairs by selecting the scores on the screen of the testing program. The perceptual results were recorded automatically into a text file for analysis.

4. Perceptual Results and Statistical Analysis

Figure 4 is the perceptual results for different acoustic cue combinations (methods hereafter) of five listeners. Table 3 shows the average score for each method.

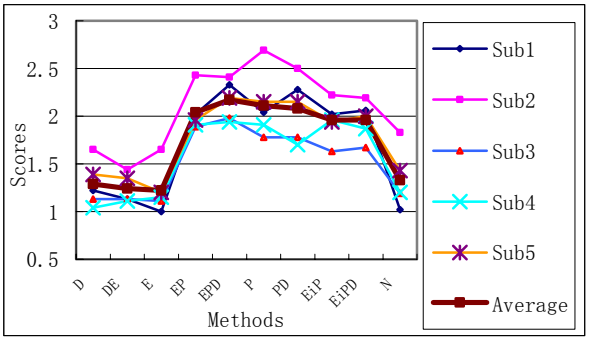

Figure 4: Perceived scores for different methods

ANOVA was conducted for testing the factors of methods, subjects and sentences that could affect the perceptual results. The statistic results indicated that all the factors including methods (df=9, F=167.6, P=0.0), subjects (df=4, F=102.3,P=0.0), and sentences (df=5, F=4.6,P=0.0) had significant effect on perceptual results.

Table3: The average score for 10 methods

Methods	Average score and dev. for all subjects	
D	1.29	0.49
DE	1.24	0.47
E	1.22	0.46
EP	2.04	0.57
EPD	2.17	0.61
P	2.11	0.62
PD	2.08	0.63
EiP	1.96	0.58
EiPD	1.96	0.59
N	1.33	0.57

Pairwise Comparisons showed that subject 3 and 4 had the same perceptual pattern while other three had slightly different patterns ($p<0.05$). POST HOC test for homogeneous subsets according to perceptual results is shown in table 4. Four groups are got ($P>0.05$), i.e. （D\DE\E\N）, (EP\EiP\EiPD), (EP\P\PD) and (EPD\P\PD). The last subset got the highest score.

Since the score of 2 or above stood for the friendly state, we can simply classify the perceptual results into two subsets (D\DE\E\N) and (EP\EiP\EiPD\EP\P\PD\EPD), with the first group less than 2 points and the second group more than 2 points. Interestingly we found that all the methods with pitch modification got higher scores than those without pitch modification. This implies that pitch contributes most to friendly expression. While modifying duration or spectral energy only, or in combination of both, could not produce friendly speech.

The higher scores occur at subset four as shown in table 4, indicating that the better way to synthesize friendly speech is to adjust pitch(P) or pith with duration(PD). The best is to change pitch, dration and spectrum simultaneously(FPD).

5. Summary and Conclusion

Based on our perceptual experiment, we come to the following tentative summary:

1. *Friendliness* can be achieved via tuning the right acoustic parameters in speech synthesis;
2. For Standard Chinese, pitch is the most prominent acoustic cue that contributes to the perception of *friendly* speech. Nevertheless, the optimal acoustic adjustment for *friendly* speech is the combination of pitch, speaking rate and spectral energy distribution.
3. Adjusting spectrum tilt solely does not affect much of friendliness perception. However, it played its rule for enhancing *friendliness* in combination with other acoustic cues. The same is with the speaking rate, ie. lengthening of phone duration;
4. Perceptual scores indicated that difference existed among the 5 listeners except subject 3 and 4 (pairwise comparison $P=0.102> 0.05$). But each listener had his or her own systematic patterns as shown in figure 4. Five perceptual curves have the same and agreement patterns but at different height, implying that the normalized pattern should be the same and perception results are reliable. However, further experiments should be done on dialogue level rather than sentences or single turns.
5. Figure 5 presents the perceptual results for 6 utterances in different feature combinations. Each utterance got its highest *friendly* score in different feature combination: S1 :P; S2 :EPD; S3 :EPD/PD;S4 : EPD/PD; S5 :P;S6 : EPD. This could be related to sentence acts such as interrogation or exclamation. Further research on this will be carried on.
6. Only the tonal register was adjusted for *friendly* speech in this study. But tonal contours could be different too in different emotional states: So the pitch modification for the friendly speech is far from satisfactory.

In conclusion, we achieved some preliminarily perceptual results onthe prominent acoustic cues in friendly speech synthesis. Further investigation at syntactic and prosodic structure level, as well as the glottal source, is necessary for a comprehensive understanding and synthesis of the friendly state speech.

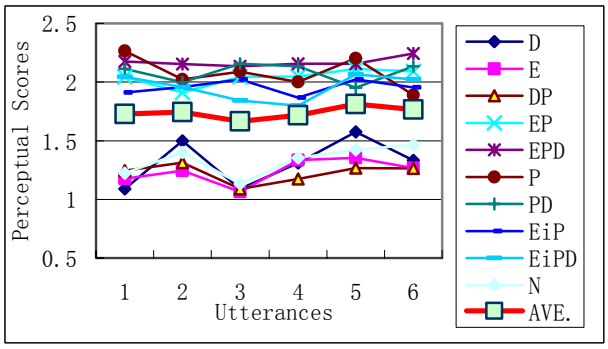

Fig 5: Perceptual results for 6 utterances by different feature combinations

Table 4: Means and 4 groups in homogeneous subsets

Methods		Subset			
		1	2	3	4
3.00	E	1.2222			
2.00	DE	1.2407			
1.00	D	1.2852			
6.00	N	1.3333			
10.00	EiPD		1.9556		
9.00	EiP		1.9556		
4.00	EP		2.0407	2.0407	
8.00	PD			2.0815	2.0815
7.00	P			2.1111	2.1111
5.00	EPD				2.1704
Sig.		.063	.138	.258	.116

6. Reference

[1] Fangxin Chen, Aijun Li, HaiBo Wang, Tianqing Wang ,2004. Qiang Fang, *Acoustic Analysis of Friendly Speech*, to appear in the proceedings of ICASSP2004.

[2] Jackson Liscombe, Jennifer Venditti, Julia Hirschberg, 2003. *Classifying Subject Ratings of Emotional Speech Using Acoustic Features*. Eurospeech 2003- Geneva, 725-728.