Friendly Speech Analysis and Perception in Standard Chinese

Aijun Li, Haibo Wang
liaj@cass.org.cn

Phonetics Laboratory, Institute of Linguistics, Chinese Academy of Social Sciences

Abstract*

Previous analyses on friendly speech were made using dialogues without differentiating their linguistic functions. This paper reports an analysis on declarative and interrogative sentences respectively. Pitch and duration of prosodic words were statistically analyzed and compared concerning factors of their positions and stresses. Based on the acoustic investigation of friendly speech, tonal pitch and prosodic duration distributions 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 important cue for better expression of friendliness; (3) Only adjusting duration is no use for expressive friendly speech; (4) Interrogative sentences got higher perceptual results than declarative sentences; (5) A high boundary tone for interrogative sentence was usually used by speakers to express friendly speech.

1. Introduction

Many investigations have been carried out on emotional or expressive speech from aspects such as voice quality and prosody[4-7]. To improve the expressiveness of the TTS system for dialogue applications, we conducted an expressive speech research project investigating the acoustic aspects of affective states most relevant to the dialogue situation [1]. Based on the perceptually classified friendly and neutral speech data, spectral, tonal and durational analyses were conducted at different phonetic levels. A perceptual experiment was made for synthesized dialogues with different acoustic parameter combinations [2]. Based on the perceptual experiment, we've got the following results:

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 plays its rule for enhancing friendliness in combination with other acoustic cues. The same is with the speaking rate, i.e. lengthening of phone duration.
4. Perceptual scores indicate that each listener has his or her own systematic. Five perceptual curves have the same patterns in agreement, but at different height, implying that the normalized pattern should be the same and the perception results are reliable.
5. Fig. 1 presents the perceptual results for 6 utterances in different feature combinations. Each utterance gets 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 speech acts such as interrogation or exclamation.
6. Only the tonal register is adjusted for friendly speech in this study. But tonal contours could be different too in different expressive and emotional states.

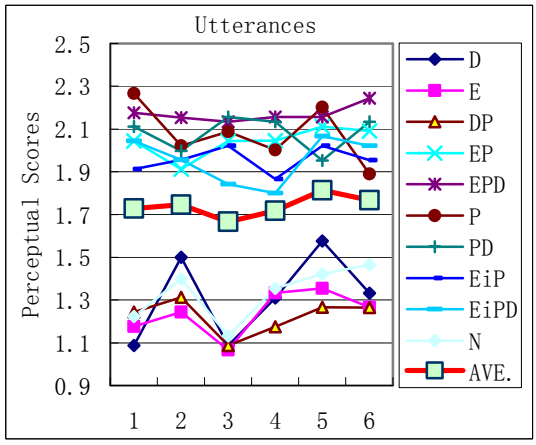

Fig. 1: Perceptual results for 6 utterances with different feature combinations. [D: duration, P: pitch, E: energy (title=5), Ei (title=10)]

As we have known, people use many interrogative sentences for communication and in expressing their attitude. In the IBM-CASS dialogue corpus [1], there are 88 interrogative sentences in 272 turns. In the previous study, all the utterances were analyzed without considering intonation or stress. In the present study, firstly, we'd like to take a further look at the acoustic features for declarative and interrogative sentences of friendly speech. Then, some utterances are synthesized by modifying different acoustic features for neutral speech to conduct a perceptual experiment. The purpose is to find out which parameter(s) is (are) the more important for synthesizing friendly speech for declarative and interrogative sentences.

2. Acoustic Analysis

2.1 Corpus and annotation

The material used in this paper comes from the IBM-CASS expressive speech corpus collected by the Phonetic Laboratory of the Linguistics Institute, Chinese Academy of Social Sciences [1, 2]. The sampling rate is 22 KHz, and the data precision is 16 bits. Two channel signals were recorded through Microphone & EGG Plata-graphic device. Utterances in this experiment were selected through perceptual experiment as described in [1]. Seven naïve subjects having no hearing problems took part in the experiment. Firstly, the sentence counterparts of both neutral and friendly utterances having above 71% agreement were chosen. Then phonetic annotation for both segmental and supra-segmental layer was

* This research endeavor was funded by the National Science Foundation, Project No.60275015. The previous studies [1] and [2] were funded by IBM China Research Center. [本文在国际口语处理会议ICSLP2004上发表。]

made by C-ToBI [3]. And all the utterances were divided into declarative and interrogative sentences as shown in Table 1.

Table 1: Distribution of utterance selected from CASS-IBM expressive corpus.

Sentence type	Speaker1/ male	Speaker2/ male	Speaker3/ female	Speaker4/ female	Total
Declarative	21*2	16*2	18*2	22*2	292
Interrogative	9*2	7*2	21*2	32*2	

2.2 F0 normalization

F0 is normalized to eliminate inter-speakers effect for pitch analysis. F0 values were first extracted using Praat in semitone with a reference frequency of 100Hz. (http://www.fon.hum.uva.nl/praat/). Then each speaker's maximum and minimum pitch value were calculated after all the error pitch points were revised. For each speaker the F0 was normalized to 100 according to the following formula:

$F0i,nor = 100*(F0i-F0imin) / (F0imax-F0imin)$

where F0imin and F0imax are speaker i's minimum and maximum F0 values in all his/her utterances. F0i is the speaker's pitch value in semitone and F0i,nor is speaker i's normalized F0 value.

2.3 Acoustic analysis on prosodic word

For each prosodic word, its top and bottom values of the normalized F0, the stress information and duration were extracted according to the annotation files. Table 2 shows the extracted features for the analysis. Here we only take care of the sentence stress (major prosodic phrase stress) without considering the minor phrase or word stress.

The average value of F0max and F0min concerning different context features are depicted in Fig. 2 for neutral speech and figure 3 for friendly speech. We can see from these two figures that:

- In neutral speech of declarative sentences, prosodic words present different pitch range and register patterns in different stress and positions. The prosodic word with sentence stress has higher pitch register and broader pitch range than those without. The F0 declination can be obviously observed. (Fig. 2, left part)
- In neutral speech of interrogative sentences, prosodic words have higher pitch register for initial and final words with sentence stress than those without. Because most of the interrogative sentences have question marks at the end, their intonations present declination patterns too. (Fig. 2 middle part)
- In neutral speech, one-word interrogative sentences have higher pitch register than one-word declarative sentences. (Fig. 2, right part)
- In neutral speech, interrogative sentences have higher pitch register than those of the declarative sentences. The systemic difference occurs from the beginning of the utterances.
- In friendly speech of interrogative sentences, prosodic words have higher pitch register for the first and the last words bearing sentence stress than those without. Because most of the interrogative sentences chosen have question marks at the sentence final, their intonations present declination patterns too. (Fig. 3, left part)
- In friendly speech of declarative sentences, prosodic words have higher pitch register for the first and the last words bearing sentence stress than those without. Their intonations present declination patterns too. (Fig. 3, middle part)
- In friendly speech, one-word interrogative sentences have higher pitch register than one-word declarative sentences. (Fig. 3, right part)
- An interesting phenomenon is that the boundary tone of interrogative sentence has become H% as compared to its neutral counterpart.
- By comparing Fig. 2 with Fig. 3, friendly speech has higher pitch register for both declarative and interrogative sentences.

Table 2: Features extracted according to the annotation files.

Features	Meaning	
Expressive states (N/F)	N: neutral speech	F: friendly speech
Sentence type (I/S)	I: interrogative	S: declarative
Stress (S3/S0)	S3: major prosodic phrase (sentence) stress	S0: others
Position (P1-P3)(D)	P1-P3: Prosodic word at the initial, the middle and the final position	D: one-word sentence
F0max (wi)	Maximum value of the normalized F0 of the i^{th} word	
F0min (wi)	Minimum value of the normalized F0 of the i^{th} word	
Duration: Di	Di: the duration of word wi	

To get parameters for synthesizing friendly speech, we calculated the differences between friendly speech and the neutral speech by the following formulae:

$\Delta F0max_{i,j} = F0max,friendly(w_i, s_j) - F0max,neutral(w_i, s_j)$
$\Delta F0min_{i,j} = F0min,friendly(w_i, s_j) - F0min,neutral(w_i, s_j)$
$\Delta D_{i,j} = (Dfriendly(w_i, s_j) - Dneutral(w_i, s_j)) / Dneutral(w_i, s_j)$

Where w_i is the i^{th} word of sentence s_j, $\Delta F0max$ is difference of the two corresponding maximum pitch values between friendly utterance and the neutral utterance, ΔD is the durational reduction rate for each word counterparts of friendly and neutral speech.

We've got the statistical parameters from the above formulae concerning positions, sentence types and stresses in table 3. There are four stress combinations for each prosodic word counterpart: sentence stresses for both neutral and friendly utterances: S3-S3; no sentence stresses for both: S0-S0; sentence stress for friendly speech but no for the neutral one: S3-S0; sentence stress for neutral speech but no for the friendly one: S0-S3.

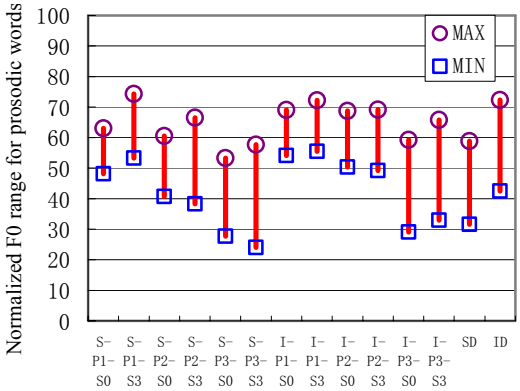

Fig. 2: F0 range of prosodic words in declarative and interrogative sentences for neutral speech

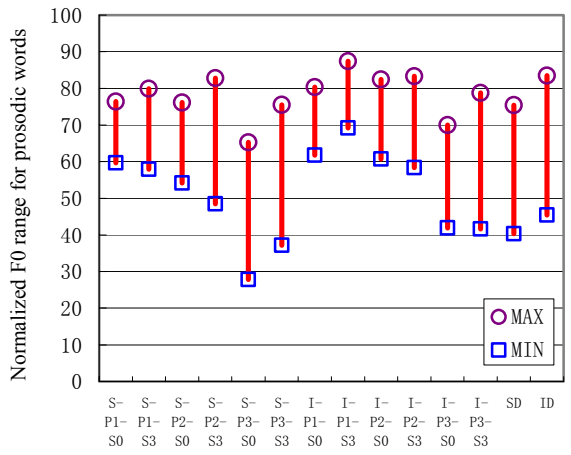

Fig. 3: F0 range of prosodic words in declarative and interrogative sentences for friendly speech.

We can see from Table 3 that top F0 values of prosodic words have bigger difference than the bottom values between friendly and neutral sentences. The biggest difference is over 25% of the whole pitch range for top values whereas less than 20% for bottom values. The top F0 values have bigger difference for declaratives than those for interrogative sentences while the bottom F0 values have smaller difference. Obviously, friendly speech is faster than neutral speech.

Table 3: The statistical parameters for pitch and duration of prosodic words.

Sentence type	position	Stress Combination	Average ΔF0max	Average ΔF0min	Average ΔD
S	D	S3-S3	16. 4979	8. 6932	. 8969
	P1	S0-S0	12. 1461	10. 1829	. 8884
		S3-S3	17. 7153	5. 0286	. 9774
		S3-S0	11. 0289	9. 6482	. 8901
		S0-S3	21. 2994	19. 1140	. 8009
	P3	S0-S0	11. 9180	1. 5148	. 9080
		S3-S3	17. 3435	10. 1161	. 9093
		S3-S0	20. 7789	10. 1117	. 9146
		S0-S3	2. 4561	−. 0237	. 9000
	P2	S0-S0	15. 1450	13. 6209	. 8568
		S3-S3	22. 2854	13. 1711	. 9403
		S3-S0	26. 1980	8. 3158	. 9051
		S0-S3	6. 2690	4. 2710	. 8430
I	D	S3-S3	11. 1607	2. 9307	. 9859
	P1	S0-S0	11. 6361	7. 4265	. 9963
		S3-S3	16. 6419	15. 7343	. 8807
		S3-S0	9. 9259	5. 5210	1. 3987
		S0-S3	9. 8314	17. 4452	. 9268
	P3	S0-S0	10. 2865	11. 7164	. 9443
		S3-S3	12. 9736	11. 4709	. 9542
		S3-S0	20. 7887	2. 9956	. 9745
		S0-S3	3. 7298	1. 4272	1. 0605
	P2	S0-S0	13. 3843	9. 9559	. 9223
		S3-S3	15. 3940	12. 1857	. 9309
		S3-S0	14. 2413	9. 9429	1. 1293
		S0-S3	13. 9816	8. 6644	. 8497

3. Perceptual Experiment on Synthesized Speech

3.1 Utterance and parameters for synthesis

We selected 8 declarative and 5 interrogative sentences with different stress positions for 7 male and 6 female utterances. Table 4 shows an example for parameters for neutral speech and the corresponding friendly speech which were calculated in Table 3. PW stands for the prosodic word. H and L are top and bottom F0values of prosodic word measured using Praat for neutral utterance, H' and L' are top and bottom F0 values calculated according to Table 3 based on H and L. ΔD is the duration reduction rate of each prosodic word got from Table 3 according to stress positions for friendly speech.

Table 4: Acoustic parameters for neutral speech and the corresponding friendly speech

Utterance	F0 (ST)	PW1	PW2	PW3
1MB4 常说的户型**都**有 (male voice with sentence stress at the final position)		常说的	户型	**都**有
	H	3	1.18	2.76
	L	-2.19	-3.16	-4.41
	H'	5.87	4.76	6.86
	L'	0.2165	0.06	-2.02
	ΔD	0.8884	0.8568	0.9093

3.2 Synthesized stimuli

In order to find which acoustic feature or feature combinations are the most important in synthesizing friendly speech, the stimuli of friendly utterances were synthesized by adjusting parameters of the corresponding neutral voice through Psola synthesizer in Praat. Three acoustic combinations adopted are (1) D: modifying duration parameters for each word as shown in table 4; (2) P: modifying pitch values of prosodic word according to table 4; (3) PD: changing P and D parameters simultaneously.

Fig. 4 shows a synthesized interrogative sentence with the modified pitch (stylized pitch is for friendly speech, the gray dotted line is the original neutral speech) and duration. Note that the boundary tone of the interrogative sentence is set to H% while its counterparts is L%.

Finally 3* 13=39 synthesized stimuli were got and together with the original neutral and friendly counterparts, we've got 39+26=64 stimuli for perception.

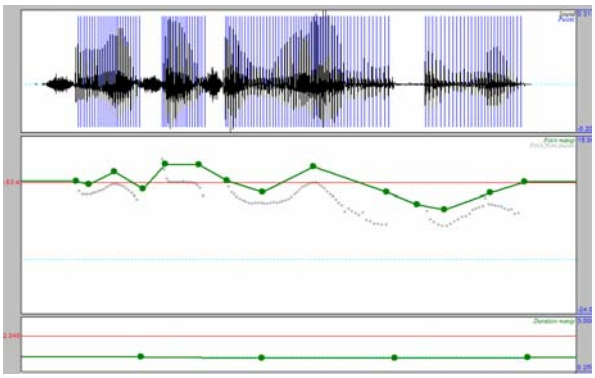

Fig. 4: An interrogative sentence 请问先生您要买房吗？(Would you like to buy a house, sir?)

3.3 Perceptual method and results

5 subjects were involved in the experiment. They listened to the 5 stimuli of each sentence randomly as many times as they wanted. Then they rated the friendliness of each stimulus in 5 points scale according to their perception.

Table 5 shows the mean scores for 5 kinds of stimuli, and Fig. 5 depicts the score distributions of 5 kinds of stimuli in declarative and interrogative sentences.

The statistic results tell us that pitch is the most important feature that contributes to friendly speech; modifying duration can't solely result in friendly speech. Duration and pitch combination have the same effect to friendly speech as pitch alone. Interrogative sentences got higher scores than declarative sentences.

Table 5: Means and 4 groups in homogeneous subsets

	Subset for alpha = .05		
stimuli	1	2	3
N	2.0154		
D	2.0154		
PD		2.8000	
P		2.8308	
F			4.8308
Sig.	1.000	.829	1.000

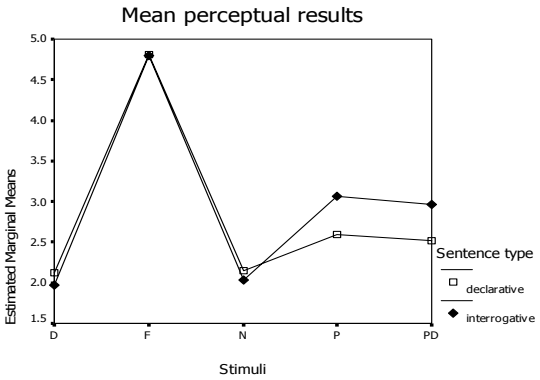

Fig. 5: Mean perceptual results for declarative and interrogative sentences in 5 kinds of stimuli

4. Discussion and conclusions

This research gives some acoustic parameters for synthesizing friendly speech. We conclude for the perceptual experiments 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; (3) Only adjusting duration is no use in produicng expressive friendly speech;. (4) Interrogative sentences got higher perceptual results than declarative sentences; (5) A high boundary tone for interrogative sentence was usually used by speakers to express friendly speech.

A higher perceptual score for interrogative sentence confirm that intonation of friendly speech has high register feature. But for the lack of phonemic duration analysis, the duration of each syllable in the prosodic words could not be well controlled. Another work is being done but has not been finished is the analysis on the detailed tonal pattern of prosodic words in friendly speech which is also important for speech synthesis.

Many thanks to Dr. Wang Wei for his editing, and Dr. Xiong Ziyu for his help in data preparation.

5. References

[1] Fangxin Chen, Aijun Li, Haibo Wang, Tianqing Wang and Qiang Fang, "Acoustic Analysis of Friendliness", to appear in *proceedings of ICASSP2004*.

[2] Aijun Li, Fangxin Chen, Haibo Wang and Tianqing Wang, "Perception on Synthesized Friendly Standard Chinese Speech", *TAL2004,* Beijing.

[3] Aijun Li, "Chinese prosody and prosodic labeling of spontaneous speech", *Proceedings of speech prosody 2002*.

[4] E.Douglas-Cowie, R. Cowie and N. Campbell ed., Special Issue on Speech and Emotion, *Speech Communication*, 40 (2003).

[5] Nick Campbell, "Listening between the lines: study of paralinguistic information carried by tone-of-voice", *TAL2004*, Beijing.

[6] Nick Campbell, "Voice Quality, the 4th prosodic dimension", *ICSPHS2003*, Barcelona.

[7] Jianhua Tao, Emotion Control of Chinese Speech Synthesis in Natural Environment, *Eurospeech2003*.