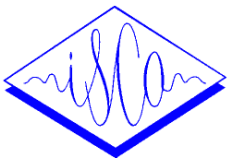

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PROSODIC CONTROL IN CHINESE TTS SYSTEM

*LU Shinan** *HE Lin** *YANG Yufang*+ *CAO Jianfen*×

*Institute of Acoustics, Academia Sinica
+Institute of Psychology, Academia Sinica
×Institute of Linguistics, Chinese Academy of Social Sciences

17 Zhongguancun Street, 100080 Beijing, China

lusn@infor.unet.net.cn

Tel: 010/62324061 Fax: 010/62553898

ABSTRACT

In this paper, the prosodic control strategy is discussed under the collectivity of Chinese TTS system design. A four level (syllable, prosodic word, prosodic phrase and sentence) pitch modification and multiplicative duration model are suggested. Although the prototype of models was formed in 1994, the subsequent results of concerned research based on large speech databases are also represented, which effectively advance to perfect the prosody control mode of the Chinese TTS system.

1. INTRODUCTION

With an effort in the past 15 years, the Chinese TTS has started in application in some communication system. When it is applied in a simple system with small vocabulary, such as a computer weather forecast system, Chinese TTS realized by concatenation of non-uniform units is with high naturalness. But for unlimited vocabulary Chinese synthesis up to today the improving on naturalness of synthetic speech is still a great challenge. The data driving speech synthesis system is in face of data cover, whereas the rule driving speech synthesis system is in face of knowledge cover. Although in data driving system the function of rule is indirect, but the rules are the key for improving naturalness for both synthesis systems. In principle, synthesis rule should include segmental and supra-segmental rules. In daily Chinese it is hard to say which diversification of pronunciation is dominating, segmental or supra-segmental diversification? But in broadcast the Chinese is pronounced clearly for every syllable, we found out the targets of formant of syllable are quite steady. Therefore the segmental diversification is not dominating for broadcast Chinese. At present, the announcer is the pattern of Chinese TTS system. It is the reason why the TD-PSOLA is compatible for Chinese TTS. Moreover TANG etc.[1] found out the perception for aberrancy of supra-segmental characteristics is more sensitive than of segmental characteristics., therefore we pay great attention to prosodic control.

The prosodic rule aggregation for Chinese TTS system was formed in 1994, after then a series psychological, phonetic and acoustical experiment researches of Chinese prosody had been done. These researches based on large speech databases effectively advance to perfect the prosody control model of Chinese TTS system.

Section 2 describes the collectivity of Chinese TTS system design, under which pitch and duration control strategy are introduced in Section 3 and 4 respectively.

2. GENERAL SURVEY OF SYSTEM

The first Chinese TTS system based on TD-PSOLA was successfully developed in Institute of Acoustics, Academia Sinica, in 1994[2]. Because in that time we were short of prosodic knowledge of Chinese and did not find out a practicable linguistic process method for input text, the prosodic control of output speech from our TTS was unsubstantial. As a result the naturalness of synthesis speech could not be well accepted. Therefore the new Chinese TTS system illustrated in Fig.1 will consist of three parts, linguistic process, prosodic design and acoustic process. Linguistic process and prosodic design module are prominent in the system

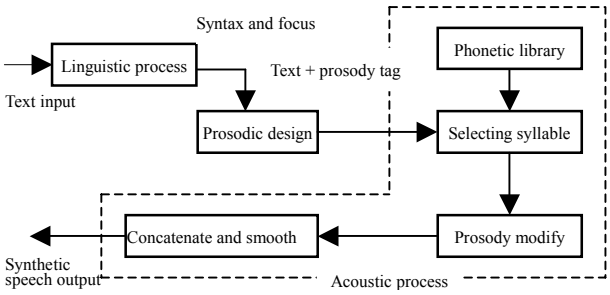

Fig. 1. Block diagram of Chinese TTS System

Linguistic process The aim of linguistic process is to get the clues which will help us to design prosody of synthesis

speech. YAO[3] developed a Chinese parser for their Chinese-foreign language transducer. By use of this parser the syntax of input text, as well as the semantic focus may be gained.

Prosodic design The task of prosodic design module is to decide where and which kind of break should be inserted in the synthesis speech, and where should be accented. Although grammatical structure is not completely consistent with prosodic structure of input text, but at present the syntax of input Chinese text is the one and only practical clue for designing the prosody of synthesis speech. The differences of both structures are mainly in word level. Generally speaking, prosodic word may be lager than grammatical word. It means that a prosodic word may include several grammatical words in Chinese. CHU[4] proposed a method how combine grammatical words to a prosodic word. We also consider creating a prosodic lexicon for word segmentation in Chinese TTS system. The prosodic phrase is usually consistent with syntactic phrase; but sometimes it is smaller than syntactic phrase.

Although the sentence stress depends on semantic expression and one can speak same sentence with diverse sentence stress, the common accent, or grammatical accent is regular. ZHEN etc.[5] validated following rules by psychological experiment research.

Rule 1 The predicate is the regular accent in the subject-predicate sentence.

Rule 2 The object is the regular accent in the subject-predicate-object sentence.

Rule 3 The object is the regular accent in the predicate-object sentence.

Rule 4 The subject is the regular accent when it is the interrogative pronoun.

Rule 5 The attribute is the regular accent mostly.

The outputs of prosodic design module are not only text, but also the additional prosody tag, which will be used to control the prosody of synthesis speech.

Acoustic process The text with prosody tag pass through acoustic process module, and convert to speech by the use of TD-PSOLA method.

Because the syllables are selected as synthesis unites, the duration and pitch contour of syllable are the two important control parameters. Currently we don't change the intensity of original syllables when they are concatenated into a sentence, because it is proved by ZHONG[6] that the aberrancy of intensity has a less sensitive than of duration and pitch on perception. The pitch and duration control strategy are introduced respectively.

3. PITCH CONTROL

The pitch is controlled in four levels as syllable, prosodic word, prosodic phrase and sentence, in Chinese TTS system.

Syllable and word levels It is well-know, that Chinese is a tone language. Every syllable must be with one of four tones, namely high level, rising, low dipping and falling tone. But some time the syllable loses its inherent tone, the pitch contour will dependent on the tone of foregoing syllable, usually called neutral tone. Besides the general tone sandihi as pointed by WU[7] the pitch contour is modified by the stress, when syllable is combined into a word. For example the combination of initial, final and tone for Chinese word "技术 ji4shu4 (technique)" and "计数 ji4shu4 (take count of)" are uniform, but the stress at first syllable for frontal word, at secondary syllable for latter word, the difference of their pitch contour is remarkable as shown in Fig. 2.

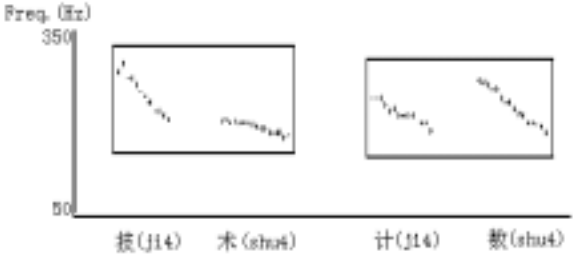

Fig. 2. F0 contour of "技术" and "计数"

When the Chinese tone is increased from four to fourteen, see Fig. 3, it is just about sufficient for to deal with the modification of pitch contour at the word level. For example, "技术" will use the combination of secondary (去声 2) and fourth falling tone（去声 4）; whereas "计数" use the combination of secondary（去声 2）and first falling tone（去声 1）. KONG etc.[8] studied on pitch models of disyllable in Chinese by vector quantization. The 39 sub-classes summarized in their study can also combine by these increased tone models.

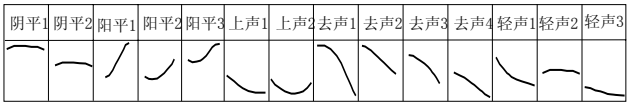

Fig.3. 14 tone models used in the Chinese TTS system

Taking one with another, if a syllable is stressed, the pitch contour will be intact and with higher top, such as high level 1 (阴平 1), rising 1 （阳平 1）, low dipping 2 （上声 2） and falling 1（去声 1）.

WU[9] pointed out the middle syllable of a Chinese tri-syllabic word is usually un-stressed, the pitch contour will be the transition from the end of first pitch contour to the begin of last pitch contour, especially when the speech rate is rapid. For example " 西 红 柿 xi1hong2shi4 (tomato)" usually is pronounced as "西轰柿 xi1hong1shi4". In our system the right tone combinations of a word are checked in the word library,

and a little modification of tone and duration of syllable are potential by the use of technically designed prosodic mark[10], which are also checked in the word library.

Prosodic phrase level At the prosodic phrase level, the effect of accent and declination on pitch are considered. According to SHEN's[11] standpoint the top line of Chinese pitch is modified by accent and intonation of sentence, while the base line is modified by rhythm. If we draw the top line and base line of pitch for each prosodic word in Fig. 4, the boundary of prosodic phase "今年我国" and "十大体育新闻" is shown very clearly by the declination and reset of base line. The elevations of top line at word "今年" and"十大"come down to the accent. The Chinese TTS is designed to a tardily going down base line and an unfixed top line to express the accent in phrase level. It means the base line will be modified according to the position of word in phrase; and the top line modified according to the accent.

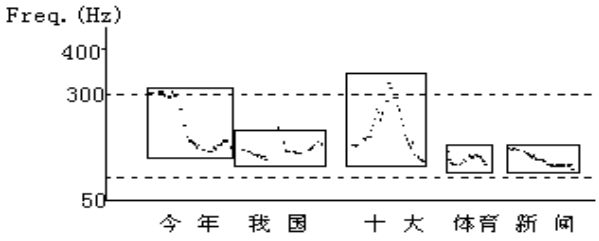

Fig. 4 The pitch contour of "今年我国十大体育新闻"

WANG etc. [12] found out that, although the Chinese falling tone with fixed range moves up, or the top of falling tone moves up going with expanded range, both can apperceive as a stressed syllable, the effect of the top of falling tone on stress is ultimate. The acoustic representation of stress as well as SHEN's intonation standpoint has been validated firstly by the normal psychological experiment based on the large speech database.

Sentence level When several prosodic phrases are combined a sentence, the going down base line will be reset at phrase boundary, and the maximum and minimum of base line of following phrase will be a little reduce respectively. The last phrase of sentence the top line and base line will be modified by intonation as described by SHEN[11]. Fig.5 is an example of modified pitch in system compared with the natural speech.

4. DURATION CONTROL

A multiplicative duration model is suggested to modify the syllable duration according to the context:

$$D = D_i \times f_w \times f_b \times f_s$$

Here D_i stands for intrinsic duration of syllable, which adopts the duration of syllable previously stored in phonetic dictionary. f_w, f_b and f_s are modification factors respectively for the

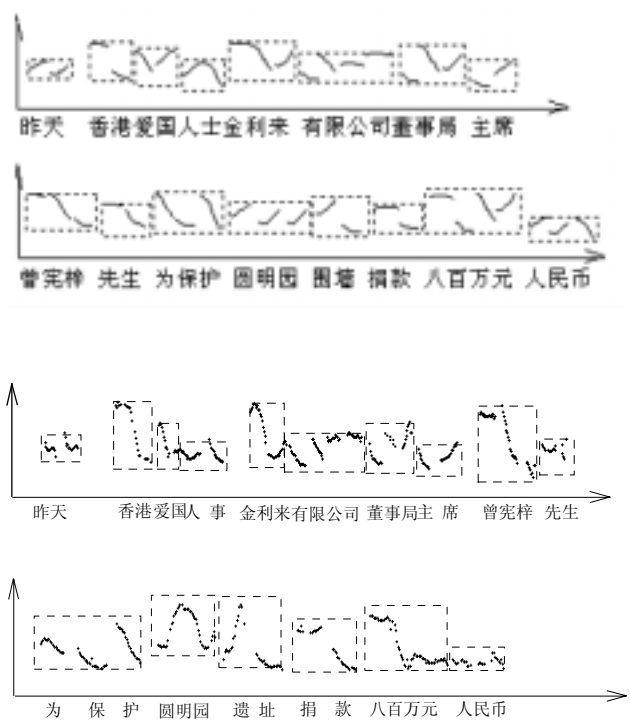

Fig.5. An example of pitch modification for synthesizing a Chinese sentence (upper) compared with pitch curves measured from same sentence of natural speech (under).

syllable position in word, the effect of boundary and of stress. Because the pronunciation of syllable samples in phonetic dictionary are pronounced in the same carrier sentence, "wo3 fa1 'X' zhe4ge zi2 (I pronounce the syllable 'X'.)". It means the target syllable has the same context, or the same effect of context on the duration. Therefore we believe D_i only depends on the components of syllable, namely the initial, final and tone, so it is defined as intrinsic duration.

The effect of syllable position in word on duration was described as a set of experiential modification coefficient suggested by CHU[4] as followings:

Table 1. Syllable duration modification coefficients at word level

Syllable position in	1st	2nd	3rd	4th
Monosyllabic word	1.0			
Disyllabic word	0.90	0.95		
Trisyllabic word	0.85	0.80	0.90	
Quadrisyllabic word	0.85	0.75	0.80	0.90

CAO etc.[13] pointed out that, if a syllable is before the boundary of prosodic phrase, its duration will be extended, expect the last syllable of the sentence. f_b = 1.3 for the last

syllable of prosodic phrase; $f_b \leq 1.0$ for the last syllable of sentences were suggested by her.

In a general way the accented syllable will be extended. But XU etc.[14] fount out an interesting phenomena. The distributions of the duration for stressed, normal and weak stressed (light) Chinese word are respectively shown in Fig. 6. There is a common peak closed to normalized duration value 0 in different sentence stress. And except "Normal", the distribution curve represents as two peaks, one is close to the peak of "Normal"; another moves to the right in "Stress" and to the left in "Light". It means there are at least two kinds of Chinese words in "Un-Normal" sentence stress, for one of them the duration is dependent on sentence stress, the duration will be lengthen in "Stress" and shorten in "Light"; for another the duration is independent on sentence stress.

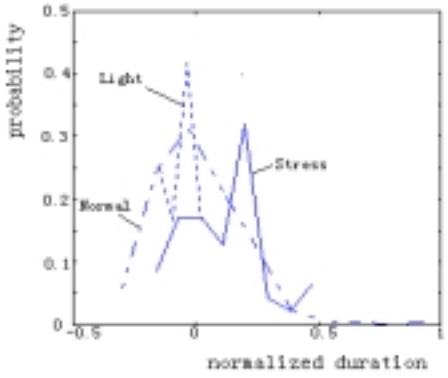

Fig. 6 The duration distribution of stressed, normal and light Chinese word

This primary result is based on only 40 sentences or 228 words, and the judgment of stress level is not regular. A formal psychological experiment for the judgment of word stress level of 1000 sentences and their acoustical analysis are going on, in order to deeply study the relationship between the word duration and sentence stress. Because the primary result does not be validated yet, we could not confirm which kind of Chinese word will be affected by stress and how to modify its duration. Therefore in above expressions the stress modification factor f_s is vacant yet (set $f_s = 1$).

Moreover, if the syllable with neutral tone, the duration is reduced to about 60%.

The silence are inserted to different prosody boundary with different duration. CHU[4] suggested the default values are 10ms, 200ms, 400ms, 500-700ms and 1000ms for the boundary of prosodic word, prosodic phrase, sub-sentence, sentence and paragraph respectively. The values will be also modified by the tempo.

5. CONCLUSION

Some bread-and-butter pitch and duration modification models for Chinese TTS are integrated in this paper. A four level pitch control strategy and multiplicative duration model are propounded. It is far from a perfect Chinese prosody rule system. The technique obstruction comes from insufficiency of knowledge about Chinese prosody. We hope the cooperative research of psychology, phonetics and acoustics will enrich knowledge about Chinese prosody, and finish the Chinese prosody rule system in not far future.

6. References

[1] 唐涤飞，吕士楠，周同春和王仁华，(1993)，"汉语语音合成协同发音规则研究"，**第六届全国语音图象通讯信号处理学术会议论文集**，四川南平, p75-78。

[2] CHU Min and LU Shinan， (1996)， "A Text-to-Speech System with High Intelligibility and Naturalness for Chinese"， **CHINESE JOURNAL OF ACOUSTICS**， Vol.15， No.1， p81-90.

[3] 姚天顺（1995），《自然语言理解》，清华大学出版社。

[4] CHU Min (1995) Research on Chinese TTS system with high intelligibility and naturalness, **Ph. D Thesis**, Institute of Acoustics, Academia Sinica.

[5] ZHENG Bo, WANG Bei, YANG Yufang, LU Shinan and CAO Jianfen (2000), The Distribution Rules of the Regular Focus Accent in Chinese Sentences, **Proc. ICSLP2000**, Beijing, October.

[6] ZHONG Xiaobo (2000), The perception of prominences in Mandarin and their Acoustic Correlates, **Ph. D. Thesis**, Institute of Psychology, Academia Sinica.

[7] 吴宗济（1980），普通话语句中的声调变化，**中国语文**，第 6 期。

[8] KONG Jiangping and LU Shinan (2000), A VQ study on pitch models of disyllable in Mandarin, **CATA ACUSTICA,**Vol.25, No.2

[9] 吴宗济（1984），普通话三字组变调规律，**中国语言学报**第二期。

[10] 吕士楠，初敏，贺琳，陆亚民和李晓光（1996），计算机汉语口语输出系统的设计与实现，**软件学报**，863 专刊。

[11] 沈炯（1999），汉语音高载信系统模型，《**中国语言学的新拓展**》石锋、潘悟云主编，香港城市大学出版社。

[12] WANG Bei, ZHENG Bo, LU Shinan, YANG Yufang, CAO Jianfen (2000), The Pitch Movement of Stressed Word, **Proc. ICSLP2000**, Beijing, October.

[13] CAO Jianfen（2000），Rhythmic grouping and timing, **Proc. ICSLP2000**, Beijing, October.

[14] XU jianpin, CHU Min, HE Lin and LU Shinan (2000), The influence of Chinese sentence stress on pitch and duration, **ACTA ACUSTICA**, Vol. 25, No. 4.