HIGH INTELLIGIBILITY AND NATURALNESS CHINESE TTS SYSTEM AND PROSODIC RULES

Chu Min and Lu Shinnan
Institute of Acoustics, Academia Sinica, Beijing, China, 100080

ABSTRACT
This paper presents a new waveform concatenation Chinese TTS system based on TD-PSOLA method. It can produce clear and natural Chinese speech. The intelligibility and naturalness of the above system are 94.1% and 7.8 respectively. The flowchart of the system is given. The prosodic rules of this TTS system, which are based on the acoustic analyses of broadcast-speech, are discussed in details.

1. INTRODUCTION
Researches on synthetic Chinese disclose that only when both the segmental and supra-segmental features of the synthetic speech are similar to those of the natural one, the synthetic speech will sound intelligible and natural[1]. Segmental features of Chinese syllables are relatively stable, but prosodic features such as pitch, contour and duration, which are main factors that affect the naturalness of Chinese, often change greatly in continuous speech. The TD-PSOLA[2] method can modify the pitch, duration and intensity of waveforms with little distortion of dynamic spectrum, so it is very suitable for synthesis Chinese. The system discussed in this paper are based on the TD-PSOLA method and utilizes mono-syllables as synthetic units.

There are many kinds of Chinese dialects in use today. Even for Standard Chinese, people from different age groups or from different educational backgrounds speak differently. The speech uttered in CCTV news broadcast is the first prototype of our TTS system. A library for prosodic rules, which includes tone and sandhi patterns of chosen texts reading by a male and a female, both of whom are senior students studying announcing arts in the Institute of Broadcast at Beijing. On the basis of the work already done, a library of prosodic rules, which includes tone and sandhi patterns of lexical items, duration distribution model for words and phrases, stress patterns for words and phrases and sentence intonation models, are built up.

2. FLOWCHART OF THE SYSTEM

The flowchart of the system is in Fig.1. The system starts from Text Scan Module, where the input text is decomposed into sentences, breathing groups and words, pronunciations of constituent Chinese characters are decided, prosodic markers are separated from the text and the order of syllables, words or phrases, breathing groups and sentences are registered. Then, pitch contours and durations of syllables of a sentence are assigned. After that, syllable waveforms and their pitch marks fetched from the syllable library are conveyed with the corresponding target pitch contour of the syllable to the PSOLA module. At last, the system concatenates the synthetic waveforms, inserts pauses in proper position, and conveys them to the D/A converter. Then, very clear and natural continues synthetic speech can be heard. The system showed in Fig.1 is carried out on a 386 PC computer. All the additional hardwares needed are a sound blaster and a speaker. The system can run on real time.

3. LIBRARY OF PROSODIC RULES

We recorded six-hour broadcast-speech and analyzed them in the length of about half an hour. We also analyzed half an hour's recording materials of several by special rules. The library stores three monosyllabic tone patterns for each of the four Chinese tones according to its position (at beginning of a word, at end of a word and others). All tone patterns have been normalized into the same tonal pitch range of normal stressed words.

3.2 Syllable duration distribution model for words and phrases
It is very helpful for improving the naturalness of synthetic speech to study the distribution of syllable durations in words or phrases. It is found that there are two kinds of duration patterns for disyllabic word. If the accent is on the first syllable, the duration of the first syllable is longer, otherwise, the last syllable is longer. In trisyllabic words, the middle syllable is the shortest, while the last syllable is the longest. In words with more than three syllables, the syllable duration is often alternated between long and short. Besides when the number of syllables in word increases, the duration of each syllable decreases. The syllable duration is also affected by its position in sentence. The last syllable of a breathing group, a subentence or a sentence is always lengthened.

3.3 Stress patterns for words and phrases and the sentence intonation model

The system turns Chinese text to speech sentence by sentence. A sentence is usually decomposed into several subsentences and a subentence into breathing groups, a breathing group into words or phrases. Y. R. Zhao first used the concept of the range of pitch to describe the tone and the intonation in Chinese[3]. Shen Jiong proposes the term of the tonal pitch range[4], which is adopted in this paper. It has many advantages to describe the intonation of a tonal language such as Chinese by the movements of the up-line and the base-line of tonal pitch ranges. The movement of the up-line indicates the stress levels of words and the movement of base-line
When there is no prosodic symbols, the system can decide the tonal pitch range and the duration automatically. By using a few prosodic symbols to change the stress level or the duration of some words in texts, the synthetic speech will be improved in naturalness.

4. RESULT OF EVALUATION

The system took part in a formal evaluation of speech quality of synthetic Chinese which is held by the specialist group of the State High Technology Development Project of China in May, 1994. The evaluation includes intelligibility and naturalness evaluations. Here gives out some results in Fig.4. KX-PSOLA is the proposed system in this paper. KX-FSS is another TTS system of our laboratory, which is a formant synthesizer and uses the same prosodic rule library as KX-PSOLA. TH-SPEECH is another waveform concatenation system, and CELP and VQ-LPC are two systems using LPC method.

There are altogether 16 listeners. The average intelligibility and the naturalness of the synthetic speech of KX-PSOLA are 94.1% and 78 (the naturalness of natural speech is 10) respectively.

In Fig 4 (a), the average intelligibility of the two waveform concatenation systems KX-PSOLA and TH-SPEECH are higher than others, and the sentence naturalness of KX-PSOLA and KX-FSS, which have a good prosodic model, are higher. Both the average intelligibility and the naturalness of KX-PSOLA are the highest. In Fig.4 (b), though the syllable and word clarity of KX-FSS is lower, the sentence intelligibility of KX-FSS is higher. This is due to the contribution of the prosodic rules.

5. CONCLUSION

Waveform concatenation synthetic technique based on TD-PSOLA method is suitable for synthesis Chinese. Proper controlling of speech prosody is the key point for improving the speech quality. Doing more research work on the prosody of Chinese and improving the prosodic rules are our main work of next stage.

REFERENCES