IIIGH INTELLIGIBILITY AND NATURALNESS CHINESE TTS SYSTEM AND PROSODIC RULES

Chu Min and Lu Shinan
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 are 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 the pitch contour and the duration, which are main factors that affect the naturalness of Chinese, often change greatly in continues 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 lexical items, duration model for words, stress patterns for words and sentence intonation models, are built up according to acoustic analyses of broadcast-style speech. By scanning the input text on the word level and the sentence level, the prosodic rules assign pitch contours and durations to syllables of a sentence comprehensively.

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

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 .

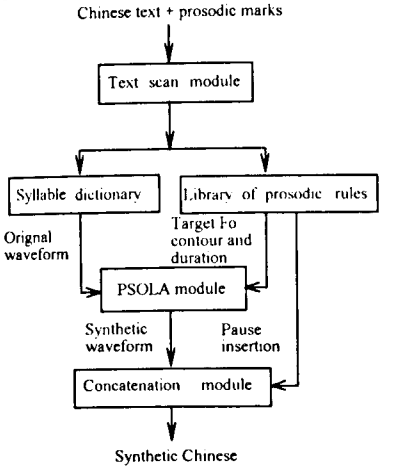


Fig.1. Flowchart of the TTS system

3.1 Tone and sandhi patterns of lexical items

Though the four Chinese tones have rather stable tone patterns in the case of isolated syllables, they undergo various modifications in continues speech due to coarticulations. Yet, for words or phrases with the same tone combination, the overall tone-sandhi patterns are almost unchanged.

In the library of prosodic rules, tone-sandhi patterns for disyllabic, trisyllabic and polysyllabic items are stored. Since the majority of Chinese lexical items are disyllabic or trisyllabic ones, there are 20 tone-sandhi patterns for disyllabic items and 120 patterns for trisyllabic ones. For items having more than three syllables, tone-sandhi patterns are formed by concatenating monosyllabic tone patterns 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 decrease. The syllable duration is also affected by its position in sentence. The last syllble of a breathing group, a subsentence or a sentencee 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 subsentence 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 tonic 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

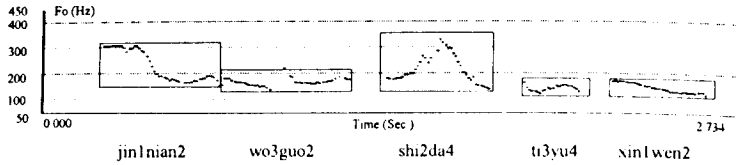


Fig.2.Tonal pitch range of the sentence of "Jin1nian2 wo3guo2 shi2da4 ti3yu4 xin1wen2"

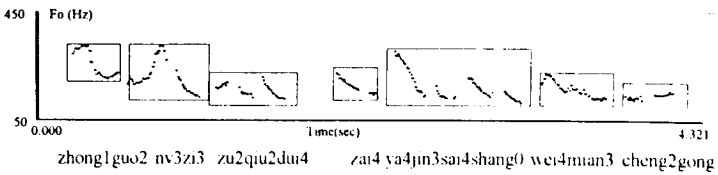


Fig.3. Movements of the up-line and the base-line of sentence of "Zhong1guo2 nv2zi3 zu2qiu2dui4 zai4 ya4jun3sai4shang0 wei4mian3 cheng2gong1."

reflectes the rhythm. At the end of sentences with different mood, such as indicative mood, interrogative mood and exclamatory mood, the position of the two line are different.

In acoustic studying, the tonal pitch ranges of words are found to be the basic blocks of the intonation of a sentence. The up-line of the stressed word always moves upwards to make the size of the stressed tonal pitch range larger than the size of the not stressed one. In Fig 2, the stressed words are "jin1nian2" and "shi2da4" . The base-line of tonal pitch ranges have a trend of moving downwards in a breathing group, and they rise at the beginning of a new breathing group. In Fig.3, there are two breathing groups in the sentence. The base-line falls in the first group, rises at "zai4" and falls again in the second breathing group. The two-line intonation model is used in our TTS system.

There are altogether five stress levels for words and phrases, which are heavy, secondary, normal, weak, and light. The up-line moves up and down according to the stress levels of words. Sometimes Chinese expresses stressing by expanding the duration. The system proposes a series of symbols for users to modify the durations of words when necessary.

The rising and falling of base-lines and inserting proper pauses in speech are corresponding to the overall rhythm. Breathing groups usually express relatively independent meanings, and are units of rhythm. The base-line falls in a breathing group until reaching the end of the group. After a little period of pausing, a new breathing group begins. The base-line rises and another falling period starts Movements of the upline and the base-line at the end of a sentence are also the main means for expressing the mood of the speech. At the end of indicative speech, the base-line falls and the up-line falls more greatly; At the end of interrogative speech, the base-line rises and the up-line almost unchanges; At the end of exclamatory speech, the base-line falls and the up-line rises.

The system inserts 10ms pause between words or phrases, 100ms pause between breathing groups, 300ms pause between subsentences. At the end of indicative sentences there are 500ms pauses and at the end of interrogative and exclamatory ones the pauses are 700ms.

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.

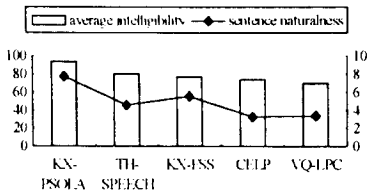
4. RESULT OF EVALUATION

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 a 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 7.8 (the naturalness of natural speech is 10.) respectively.

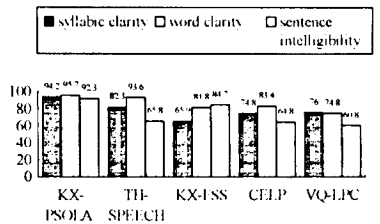
In Fig.4 (a), the average intelligibility of the two waveform concatenation systems KX-PSOLA and TH-SPEECH are higher than others, and the sentence naturalness of KX-PSOLA and KX-FSS, which have a good prosodic model, are higher. Both the average intelligibility and the naturalness of KX-PSOLA are the 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



(a)



(b)

Fig.4. Some results of evaluation

controlling of speech prosody is the key point for improveing the speech quality. Doing more reseach work on the prosody of Chinese and improving the prosodic rules are our main work of next stage.

REFERENCES

[1] Lu Shinan, Qi Shiqian and Zhang Jialu (1993), "Experimental study on the naturalness of synthetic speech", *Chinese Journal of Acoustics*, Vol.12, No.3, P.256-264

[2]F. Charpentier, M. Stella (1986), "Diphone ynthesis using an overlap_add technique for speech waveforms concatenation", *Proc. Int. Conf. ASSP*, P.2015-2018

[3]Y. R. Zhao (1933) ,"Tone and intonation in Chinese", *Shi Yu Suo Ji Kan*, Vol.4, No.2, P.121-134

[4]Shen Jiong (1983), "Pitch range of tone and intonation in Beijing dialect", Working papers in experrimental phonetics, P. 73-120 (In Chinese)