TOWARDS A CHINESE TEXT-TO-SPEECH SYSTEM WITH HIGHER NATURALNESS

Reh-Hua Wang Qinfeng Liu Yongsheng Teng Deyu Xia

University of Science & Technology of China

P.O.Box 4, Hefei, 230027 P.R.CHINA

Email: rhw@ustc.edu.cn

ABSTRACT

This paper presents our research efforts on Chinese text-to-speech towards higher naturalness, the main results can be summarized as follows: 1. In the proposed TTS system the syllable-sized units were cut out from the real recorded speech, the synthetic speech was generated by concatenating these units back together. 2. The integration of units synthesized by rules with natural units was tested. A LMA filter based synthesizer was developed successfully to test and generate those units, which were difficult to be collected from the speech corpus. 3. A new efficient Chinese character coding scheme - "Yin Xu Code"(YX Code) has been developed to assist the GB Code. Based on above results, a Chinese text-to-speech system named as "KD-863" has been developed. In the national assessment of Chinese TTS systems held at the end of March 1998 in Beijing, the system achieved a first of the naturalness MOS (Mean Opinion Score).

1. INTRODUCTION

Using the waveform of mono-syllables as elemental synthesis units, the Chinese speech can be synthesized by concatenating the sequence of syllable-based units after appropriate prosodic modifications. In the conventional synthesis system, this kind of modifications are completed by the synthesizers, among which the concatenate of syllabic wavelet segments with the Pitch_Synchronous Overlap Add(PSOLA) technique becomes much popular, and with this method the quality of output speech of the synthesis system has been greatly improved[1]. However, experiments show that the speech quality generated by using the PSOLA will be sharply declined with the increased prosodic modification. Therefore , we prefer to use a simplest approach to bypass this problem, i.e. cutting out all required syllables from the real recorded speech, and during synthesis, concatenating these units back together based on the prosodic rules. The key point in this concatenating synthesis is to select the syllable-sized units in such a way as to minimize discrepancies between adjacent units, which come originally from completely different utterances. It would be theoretically possible to take all contextually determined syllable-sized units, but the number of such units would inevitably be huge. It is not only hard to obtain these units, but also complicated to sum up the prosodic rules. For this reason, the context-sensitive units must be merged, meanwhile the prosodic rules appeared in the continuous speech must be simplified and incorporated. In this paper a simplified prosodic pattern stemmed from the analysis of the digit strings was proposed, based on the concept of "perception-quantizing" a set of syllable-sized units were designed, how to obtain these units from the natural speech was presented.

Based on the proposed method, a Chinese text-to-speech system named as "KD-863" has been developed. In this system a new efficient Chinese character coding scheme (YX Code) was developed to contribute to the text preprocessing, the new Code contains the pronunciation information, and is much helpful for the words-segmentation, which will be introduced in the section 4. The performance evaluated by the national assessment project was given at last.

2. SIMPLIFIED PROSODIC RULES

In the natural speech, one sentence may have varied prosodic expressions, but each of which sounds fluent. Analysis by synthesis experiments made it clear that, when the prosodic modification of some parts in a sentence doesn't exceed a given range, the naturalness of synthesized speech will not be damaged. The prosodic features of the syllable in a given speech environment are also allowed to be changing in a limited range, so long as the prosodic parameters remain in this range, the perception will be natural. Based on this phenomenon, a concept called "perception-quantizing " was raised here, which is the base to form the simplified prosodic model. In order to carry out this "perception-quantizing" idea, we first summed up a simplified prosodic model. Which is based on the broadcasting declarative style. And the simplification and classification were mainly aimed at dealing with the tone and duration, because the two parameters have the largest influence on the naturalness and their rules are also the most complicated.

For the convenience to extract the prosodic rules, we choose the digit strings as analyzing target. Digit strings have such advantages: (1)Include all the Chinese tones; (2)Can compose various tonal connection and the corresponding string's naturalness is easily to be evaluated. (3)The rhythm of digit strings with the same length can be fixed at a unified mode(e.g. All the 4-syllable phrases can use 2-2 rhythm and all the 5-syllable phrases can use 2-3 rhythm) which make it convenient to carry out tests for rhythm and pause. (4) There is no emphasis,

no light tone in digit strings, so the prosodic rules are very standard, stable and easily to be summed up. Using the analyzing process of monosyllable->phrase->sentence, the tone and duration rules were summed as below:

2.1. The Tonal rules

(1) The tonal rules of short phrases (less than 4 syllables) have already been introduced in details in former articles[2] , synthesis experiments have shown the efficiency of these rules.

(2) Constructing long phrases (longer than 4 syllables) from the tonal-connection of short phrases . The processing steps are: (a)Cutting the long phrase into several short phrases based on the rhythm composition; (b) Giving out the internal tonal contours of short phrases; (c) Droping the tonal height of the short phrases in the tail part of big phrase in accordance with the gradual descendence principle. (d)Checking the tonal range of the whole long phrase, if the range beyonds 3 semitones, it should be reduced to 3 semitones while the middle value keeps unchanged. (e) Refering to the 4-syllable tonal rules to smooth the tonal contours at the connection positions of short phrases,.

(3) When the length of long phrases reaches 8, a breath group is formed. Connecting the tonal contours of long phrases(or breath groups) and inserting some gaps, the basic F0 shape of a whole sentence is constructed. Then adding a gradual descendent part to the basic shape, the F0 contour of continuous sentence is finnaly built.

2.2. The Duration Rules

(1) The influence of duration to naturalness is mainly shown on the length comparison of connected syllables. As long as the ratio is suitable, a high naturalness will be reached. The absolute length is only used to constitute the speed characteristics with the pause among phrases or breath groups.

(2) Excluding weakening and emphasis, the internal duration mode of phrase can be defined as below: bi-syllable word has a longer duration at the second syllable, the ratio is 1:1.1; tri-syllable word has a shorter duration at the middle syllable, the ratio is 1.1:1:1.2; four-syllable word has three different ratio modes depending on different rhythm, when the rhythm is 2-2 composition, the ratio is 1:1.1:1:1.2; when the rhythm is 1-3 composition ,the ratio is 1.2:1.1:1:1.2; when the rhythm is 3-1 composition, the ratio is 1.1:1:1.1:1.2.

(3) Big phrases of longer than 5 syllables can be seen to be constructed by small phrases of shorter than 4 syllables. To duration parameter, it is enough to compress the syllables at the joined positions to 0.9 ratio as before.

(4) The last syllable of a breath group or a sentence should have a longer duration, normally the ratio 1.2 can satisfy the need.

The above prosodic rules come from the analysis of digit strings. By adding weakening and emphasis rules to the prosodic model, we obtain all the tone and duration rules under the Chinese simplified mode. As for the tone and duration rules at the sentence level, we make the correspondence of the changes between words with the changes in the words as far as possible, so as to discover as many prosodic parameters and prosodic-units as possible.

The Fig.1 and Fig.2 give an example of concatenating the prosodic-units from quad-syllable words to form a five-syllable word “中国共产党(zhong1guo2gong4chan3dang3)”.

Five small phrases shown in Fig.1 are merely pronounced according to tone combination and certain rhythm respectively without any specific implications : 中州毁卡(zhong1 zhou1hui3ka3), 兵国群病(bing1guo2qun2bing4), 管共信(guan3gong4xin4), 酸兵产怎(suan1bing1chan3zhen3) and 隐甭拖党(ying3beng2tuo1dang3). By cutting out five prosodic-units “中zhong1”,”国guo2”,”共gong4”,”产chan3”,”党dang3” and connecting them together, we get the synthesized speech. Fig.2 gives the comparative waveforms of the synthesized speech and original speech. It is thus clear from the illustration that the tone and duration characteristics of the synthesized speech are so similar to that of original speech. Listening experiment also shows that the quality and naturalness of these two kinds of sentences are very close.

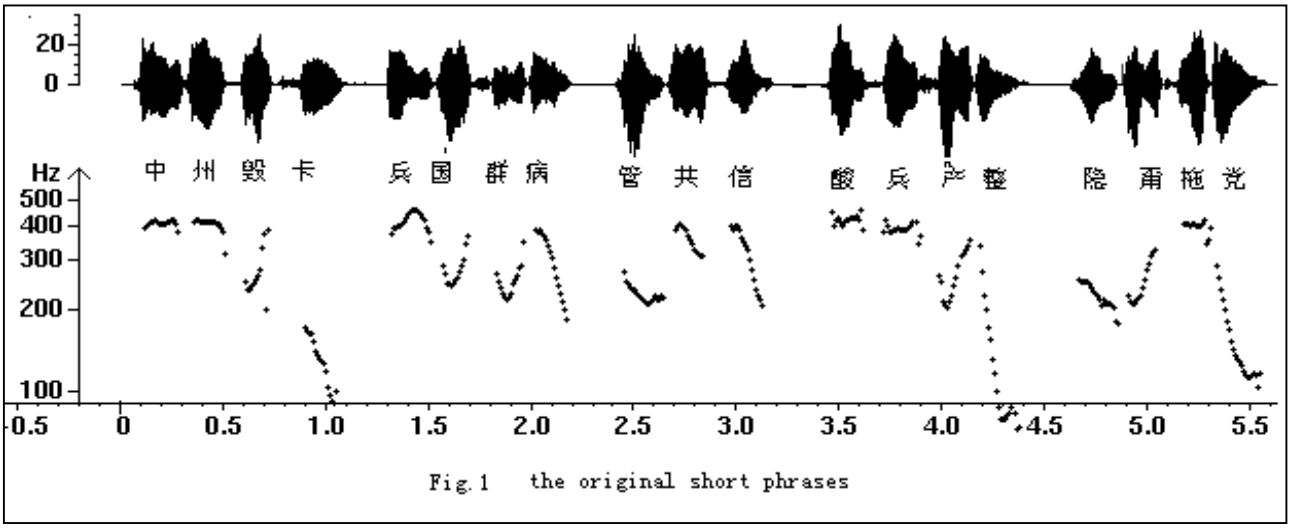

Fig.1 the original short phrases

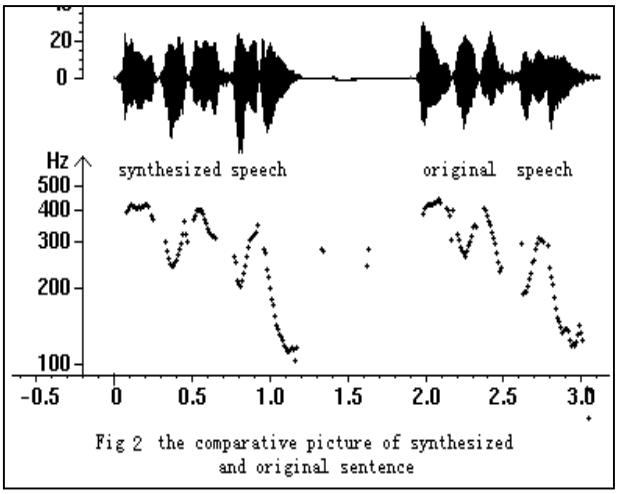

Fig 2 the comparative picture of synthesized and original sentence

3. PERCEPTION-QUANTIZING UNITS

The syllable-based concatenating system is still to come across the problem of cross-syllable co-articulation. Though the co-articulation phenomena do not have the same marked influence as tone and duration, sometimes it may also play important roles in the naturalness. On the basis of "perception-quantizing" A set of simplified rules was summed up to deal with the cross-syllable co-articulation problems .

3.1. The Simplification And Classification of Co-articulation Phenomena

Listening tests have shown that in a TTS system using monosyllables as synthesis units, the major influence to naturalness was caused by the transition pattern between connected syllables. As we know from the pronouncing physiology, the co-articulation phenomenon reflects the continuous change of the tongue position and vocal tract shape which leads to the continuous trend of the Formant track. Based on these knowledge, the co-articulation phenomena problems were simplified and classified as below:

(1)When the next syllable's head is an unvoiced consonant, the co-articulation can be ignored. So long as we insert a 30ms gap between the connected syllables, the un-successive influence caused by the un-continuous Fomant track can be eliminated.

(2)When the next syllable's head is a voiced consonant, the initial one's tail does not have perceptional difference with the syllable followed by an unvoiced consonant. The co-articulation influence to the initial syllable can also be ignored.

(3) When the next syllable's head is a voiced consonant, it will be influenced by the initial one's tail. According to the initial syllable's tongue position and the F2 track, the influence can be sorted into 3 main types: (a) Front, high, apical; (b) Front, high, blade-palatal; (c)Middle, lower, dorsal or back, high, velar.

3.2.Getting The Perception-Quantizing Units

With the above simplified prosodic mode, we can completely analyze all the possible cases of tone, duration, co-articulation of each syllable in continuous speech-streams and their collocation relationship. According to this kind of collocation relationship we get all the corresponding prosodic units. These units are once more simplified and combined and we get the needed "perception-quantizing units". The principle of simplification and combination is: (1)ensure that in any context environment, always can find a "unit" according with the demand of prosodic rules, so make the synthesis output having relatively high naturalness. (2) The redundancy of "unit base" should be as little as possible.

In order to conduct redundancy compression and naturalness tests, we ever devised a speech synthesizer based on LMA vocal tract model[3]. Synthesizing practice indicates that it is better than PSOLA and Formant synthesizer in the ability of adjusting prosodic parameters and synthesis quality, and basically can realize the demand of adjusting each parameter needed under simplified prosodic rules with high quality. To counter the prosodic parameters of each prosodic-unit under simplified prosodic mode, first synthesize the corresponding syllables with LMA synthesizer, and test them with the substitution. Relying upon the results, those syllables, which could be exchanged, would be substituted by one unit, also the corresponding rules would be incorporated. This way we obtain all the essential prosodic units and the corresponding calling rules , which are required in synthesizing the phrases from the syllable. Final result indicates that for all the syllables (about 1300 ones), we only need 13000 "units" in all.

After having analyzed all the "units" with LMA synthesizer, we then designed a recording vocabulary table according to corresponding simplified calling rules, and cut off these prosodic units from the original speech of the announcer. Because it is very hard for the announcer to accurately control the prosodic parameters of some units in the speech-stream when naturally speaking, we obtain those prosodic units which have relatively large deviation from the syllables that have similar rules of rhyming directly by utilizing LMA synthesizer. The excellent performance of LMA synthesizer enables that the synthesized unit has no difference in quality with the original unit.

Based on the set of "units" and the simplified prosodic rules, the developed KD-863 TTS system have remarkably improved its naturalness and quality, and make a big step forward the practicability.

4. YIN XU CODE

The Chinese text-to-speech usually accepts input from Chinese characters coding in GB Code, which is the standard coding scheme for the Chinese characters. However the GB Code does not contain the pronunciation information, so it is difficult to define the reading of a character according to the responding GB code word. There are a few hundred Chinese characters that have multiple pronunciations (around 891 polyphones from 6763 characters in the GB Code), on the other side, one pronunciation may be responding to several characters. In order to remedy the weaknesses of the GB Code, a new coding scheme, named Yin Xu Code (YX Code, coding by pronunciation and order), has been developed to assist the GB Code. The YX Code mixed the pronunciation together, and more fit for those systems, where pronunciation information is

wanted , such as in the text-to-speech or speech recognition systems.

As people know, the basic pronunciation units in the standard Chinese are the tonal syllables, which are labeled by the Pin Yin symbols. According to the statistics on the Modern Chinese Dictionary, the 6763 characters in GB Code are only responding to 1310 different tonal syllables except the er suffixation and light reading. So that the YX Code was designed like this:
1.As same as GB Code, each code have two bytes;
2.Each code of YX Code has one and only one grapheme and one pronunciation;
3.From the code, it is easy to get the only pronunciation;
4.The graphemes come from GB2312 and the pronunciations come from “Modern Chinese Dictionary”;
5.The scheme is unite and standard..

We use the high 11 bits denoting the pronunciations which represent the syllable code, and the low 5 bits distinguishing the graphemes which represent the grapheme code. That is enough to the syllables but not enough to the graphemes because there are 12 syllables, which have more than 32 homographs. They are bi4、fu2、jil、ji4、li4、qi2、shi4、xi4、yi4、yu2、yu4、zhi4. The resolved method is allocating one more syllable code for these syllables, it is reasonable because none has more than 64 homographs in the 12 syllables. The grapheme code mainly distinguishes different characters, and the homographs are arranged in the frequency of the use. Combining the syllable code with the grapheme code the YX Code is defined. There are total 7756 units in YX Code.

Based on the YX Code, a new dictionary system was designed. The proposed lexicon structure, not only supply with the pronunciation information, but also is helpful to implement the words-segmentation, it plays an important role in the KD-863 system. The details on the lexicon structure and it's application to the words-segmentation were introduced in another paper [4].

5. EVALUATION OF PERFOMANCE

The developed KD-863 Chinese text-to-speech system runs on the platform of common PC586 with 30 MB disk capacity for storage of the units base. The system can convert any Chinese text to speech output with very high naturalness in real time. In the end of March,1998, a national assessment of the performance of the Chinese text-to-speech systems was held in Beijing. The test was based on the subject evaluation, total four systems were evaluated.

First specification is on the syllable's clarity, word's and sentence's intelligibility. The mean opinion score for the KD-863 system is listed in the table 1.

Second is on the speech naturalness, which is most interesting specification. The naturalness was evaluated by the trained listeners using 5 point scale MOS after hearing the speech output of the tested Chinese text-to-speech system. The testing process is full automatic, manual operation was not allowed. The inputs for the tested system are 5 files, each one is a text file which contains a short article with several paragraphs. These articles were selected from the newspaper, and the content covers a variety of domains. The naturalness testing is a comprehensive results because it is really involved in the text-to-speech. The results were shown in figure . where the judgment on the 5 point scale is like that: 5 excellent, the same quality as natural speech; 4 good, willing to accept; 3 fair, can accept; 2 bad, not willing to accept; 1 very bad, can not accept.

Table 1 Intelligibility score of KD-863 system

	Syllable's Clarity	Word's intelligibility	Sentence's intelligibility
Score (%)	80.1	66.9	84.3

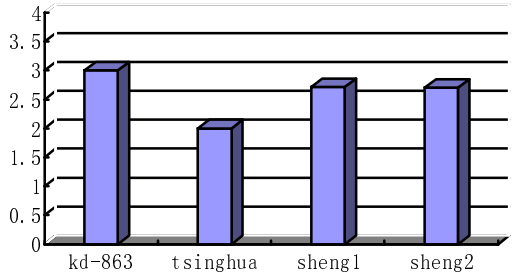

Figure 3 The naturalness in MOS of the 4 tested systems

From the Fig.3 it can be seen, even though the KD-863 system achieved the top of the score in the all tested systems, but it is just 3 points, that means only achieved the extent of "can accept" . As for the real nice Chinese TTS system there is still a long way to go.

Finally, it is worth explaining that although this kind of simplified prosodic rules and “units” were extracted and summarized from the utterances with broadcast style, which have comparatively single tone, it can be extended to obtain sigh, question, emphasis style by means of comparing the changes of tone patterns in the prosodic rules , and the expand and contract of duration. With the same group of units, we can realize varied changes in speed, tone and semantic characteristics through using different calling order and calling rules.

6. REFERENCES

1. Wang, R.H., Liu, Q., Tang, D. "A New Chinese Text-To-Speech System with High Naturalness”, *Proc. of ICSLP96, p1441-1444, 1996*.

2. Wu, Z., "Roles for processing Prosodic variations of Quadro-syllabic Groups in Standard Chinese", *Report of Phonetic Reserach, Vol 7, 1988, p12-27*

3. Liu, Q., Wang, R.H., "A New Speech Synthesis Method Based on The LMA Vocal Tract Model", *Chinese Journal Of Acoustics, Vol.17,1998, p153-162*

4. Teng, Y., et al " YX- Code and Word Parsing in Chinese TTS System" to be published at ISCSLP98.