A NEW CHINESE TEXT-TO-SPEECH SYSTEM WITH HIGH NATURALNESS

Ren-Hua Wang Qinfeng Liu Difei Tang

University of Science & Technology of China

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

Email: rhw@ustc.edu.cn

ABSTRACT

This paper introduces a new Chinese text-to-speech system that produces far more natural and intelligible synthesized speech than existing systems. There are two distinguishing features in this system. One is the perfect prosodic rules that were summed up from the linguistic knowledge and statistical results made on a standard Chinese database. These rules are successfully used to modify the elemental synthesis units to get high naturalness while concatenate them to a sentence. The other feature is that, the Log Magnitude Approximate(LMA) filter is used as the synthesis filter in the introduced system. With the LMA filter, the prosody of the synthesized speech can be modified at a wide range while maintain high intelligence and naturalness. In this paper the formulated prosodic rules are presented, and the LMA filter based speech synthesis is described in details.

1. INTRODUCTION

In the past five years Chinese text-to-speech synthesis has made a rapid progress. Using the original waveforms of the mono-syllables as elemental synthesis units, the speech can be synthesized by concatenating the sequence of acoustical units after appropriate modification. With this method the quality of output speech of the synthesis system has been greatly improved. It is well known that if we just concatenate the mono-syllables to generate sentences without any prosody control, they may be intelligible, but sound much unnatural. In general, there are two key points in developing a high-quality speech synthesis system : one is how to find out and sum up the prosodic rules, the other is to have a flexible speech synthesizer, which permits to modify the prosodic features effectively based on the rules. The first problem is language dependent. Differing from English or other western languages, Chinese is a tonal language. The tone and intonation play the most important roles in the prosodic features in standard Chinese. So more attentions are put on the tone-sandhi in multi-syllable words in this paper. The second problem is related to the synthesis methods, among which the concatenate of syllabic wavelet segments with the Pitch_Synchronous Overlap Add(PSOLA) technique becomes much popular in recent years. However, experiments show that the speech quality generated by using the waveform concatenate will be sharp decline as the increased prosodic modification. Therefore we prefer to take the source/filter model in this system. Instead of the Formant or LPC based model, a novel LMA filter is used as the vocal tract model. Since the log magnitude of the LMA filter's transfer function can optimally approximate the log spectral envelope of the required speech, high-quality speech can be generated from the LMA filter with a small set of parameters. Using the LMA filter based synthesizer and the elaborately designed prosodic rules, a Chinese text-to-speech system with high naturalness has been developed .

2. THE PROSODIC RULES

Prosody, or supra-segment features in continuous speech is mainly presented by the variation of the pitch, duration, pause and amplitude, and play an important role on naturalness and fluency of synthesized sentences. Many valuable results on prosodic features have been achieved by phoneticians[1]. The acoustic analysis and synthesis experiments also have shown that the pitch contour and duration are the two most important ones among the prosodic features in Chinese.

2.1. Tone and intonation

Being a tonal language, the Chinese syllables are distinguished by tone, and the intonation in sentence is produced to express emotion. The relation of tone and intonation in the Standard Chinese had been described: "the small ripples riding on large waves", so that a sentence intonation might be treated firstly in phrasal contour, and analyzed by tone-sandhi rules, then the remained coarticulation contour are considered separately. On the basis of this theory, the tone contours of the mono-syllables will keep the basic form in sentences, at the phrase level the variation of the tone contour is mainly reflected in the range and integrity of the contour. This point is approved by the analysis and statistics on a database of two thousands words, which is designed specially to capture the possible co-articulation effects in Standard Chinese, and composed of 434 two-syllable words, 534 three-syllable words, and 1032 four-syllable words[2].

In the rule library of our system, we first extend the monotonemes set from four tonemes of mono-syllable words to 16 tonemes , that is, from the primary four tones H, R, L and F (High, Raising, Low and Falling) to H^1, H^2, H^3, R^1, R^2, R^3 and L^1, L^2, L^3, F^1, F^2, F^3, F^4, F^5, together with two kinds of light tone. The H^n indicates the nth derivations of H. Secondly, the

tone patterns of multi-syllable words are formed by the combinations of basic units of proposed monotonemes set with tone-sandhi rules . The tone-sandhi rules expressed by formula as follows:

Two-syllable words:

$H \longrightarrow H^1 / (H|F) + \underline{H}$
$H \longrightarrow H^2 / (R|L) + \underline{H}$
$R \longrightarrow R^1 / \underline{R} + H$
$R \longrightarrow R^2 / X + \underline{R}$
$L \longrightarrow L^1 / \underline{L} + (H|R|F)$
$L \longrightarrow L^1 / H + \underline{L}$
$L \longrightarrow L^2 / (R|L|F) + \underline{L}$
$L \longrightarrow R / \underline{L} + L$
$F \longrightarrow F^2 / \underline{F} + X$
$F \longrightarrow F^3 / X + \underline{F}$

Three-syllable words:

$H \longrightarrow H^1 / H + \underline{H} + X$
$H \longrightarrow H^1 / (H|R|F) + \underline{H} + H$
$H \longrightarrow H^1 / R + H + \underline{H}$
$H \longrightarrow H^2 / (H|L) + H + \underline{H}$
$H \longrightarrow H^2 / (R|F) + (R|L|F) + \underline{H}$
$L \longrightarrow R^1 / (H|R) + \underline{L} + L$
$F \longrightarrow F^1 / \underline{F} + F + F$
$F \longrightarrow F^2 / H + \underline{F} + X$
$F \longrightarrow F^2 / \underline{F} + F + (H|R|L)$
$F \longrightarrow F^2 / (H|R|L) + \underline{F} + F$

$R \longrightarrow R^1 / H + \underline{R} + (H|R|L)$
$R \longrightarrow R^1 / R + \underline{R} + X$
$R \longrightarrow R^1 / (L|F) + \underline{R} + (H|F)$
$R \longrightarrow R^2 / X + X + \underline{R}$
$R \longrightarrow R^2 / H + \underline{R} + F$
$L \longrightarrow L^1 / X + \underline{L} + (H|R|F)$
$L \longrightarrow L^1 / X + H + \underline{L}$
$L \longrightarrow L^2 / X + (R|L|F) + \underline{L}$
$L \longrightarrow R / \underline{L} + L + (H|R|F)$
$L \longrightarrow R / (L|F) + \underline{L} + L$
$\underline{L} + \underline{L} + L \longrightarrow \underline{L^1} + \underline{R} + L$
$F \longrightarrow F^2 / (R|L) + H + \underline{F}$
$F \longrightarrow F^3 / (R|L|F) + \underline{F} + (H|R|L)$
$F \longrightarrow F^3 / (H|F) + X + \underline{F}$
$F \longrightarrow F^3 / (R|L) + (R|L|F) + \underline{F}$
$F \longrightarrow F^4 / F + \underline{F} + F$
$R \longrightarrow R^1 / H + \underline{R} + (H|R|L)$
$R \longrightarrow R^1 / R + \underline{R} + X$
$R \longrightarrow R^1 / (L|F) + \underline{R} + (H|F)$
$R \longrightarrow R^2 / X + X + \underline{R}$
$R \longrightarrow R^2 / H + \underline{R} + F$

Where mono-syllables are connected with "+", X indicates arbitrary tone, "|" means "or", "/"means different combinations. For example, formula

$H \longrightarrow H^1 / (H|F) + \underline{H}$

indicates "In two-syllable words, if the first syllable's tone is high or falling, and second syllable's tone is high, then the second syllable's tone changes to the first derivation of high tone".

As for the four-syllable words, the tone-sndhi is much complicated. We first devide the four syllables into two two-syllable words and change the tone contours according to two-syllable rules,then regard the second and the third syllable as a new tow-syllable word to change their tone contours.

Light tone is another kind of tone-sandhi, its contour pattern changes following the precedent syllable.We devide the patterns of light tone into two types: the type of low if following the high, rasing and falling tone, and the type of high if following the low tone. For convenience some common used mono-syllables with both types of light tone can be stored in the library in advance.

Finally, the sentence intonation can be obtained by the mixture of an amount of local patterns and global modifications. There are plenty of variations for the intonation in the spoken speech, so that how to synthesize sentence intonation is still an undertaking problem. In genaral, for the declaration sentence the range of tone contour is reduced gradually, and a big step lower when the syllable is at the end of sentence. If one phrase is emphasized in the sentence the range will be increased clearly. For the interrogative sentence, just the upper threshold of the tone contour is raised at the last syllable. These principles are used to modify the sentence intonation temporarily.

2.2. Duration

The duration feature is another important factor to present the prosodic features in Standard Chinese. We take the same strategy as tone processing. First it is modified at word level. Experiments on the words database also show that, the first syllable's duration is longer than the second one in the two-syllable words, the ratio is about 0.95 for male speaker and 0.93 for female speaker, in three -syllable words the last syllable is the longest, the first syllable takes second place, the middle one is shortest. In the four-syllable words the order of syllable duration is : the last syllable is longest, next is the first syllable, then the third one, the second syllable is the shortest. Based on that a set of rules for the syllable's duration modification are proposed. When the sentence level is taken into consideration, the syllable's duration is relative to the stress positions and grammar structure. the duration feature becomes more complicated. A practical method is to shorten the syllable's duration as the word's position moves apart from the beginning of the sentence, usually moving one word step the syllable duration shorten by 2.5%. The duration of the stressed syllable or the ended syllable in the sentence will lengthen relatively. The duration of light tone reduces to 70% to 80% of original.

The pause and amplitude are also prosodic features. Since the effect to speech naturalness by the amplitude is much less than by the tone and duration, no much attentions are paid in this time. The length of pause between the syllables is reduced in proper order like as: in words, between words, between sentences.

3. A LMA FILTER BASED SYNTHESIZER

The LMA filter[3], composed by a set of cascade rational function filter, is constructed by the following formula :

$$H(z) = \exp\left(\sum_{m=0}^{M} C_m z^{-m}\right) \qquad (1)$$

Where C_m are the cepstrum coefficients of the undertaking signal.

In the realization of the filter, the formula (1) can be simplified to a rational function form by the Pade approximation.

It can be easily deduced from the definition of the cepstrum that , when M is large enough, the logarithmic amplitude spectrum of H(z) can optimally approximate the logarithmic envelope of the analyzed signal under the criterion of least mean square. With the LMA filter as the vocal tract model a proper glottal excitation wave will generate high quality speech[4]. Especially, our experiments indicate that, by controlling the parameters of LMA filter and the glottal excitation waveform the prosody of the

synthesized speech can be modified at will while retaining high level of naturalness..

3.1. The excitation model

In the LMA filter based synthesizer, the excitation waveform is generated by the excitation model according to different sounds:

(1) Voiced.

When a voiced sound is asked, the model generates quasi-periodic quasi-triangular glottal wave. Experiments of speech synthesis show that, when the quasi-periodic characteristics are kept, the form of the glottal wave are not harsh. Here we construct the triangular wave with a negative exponential function. The diagram of the excitation model is shown below:

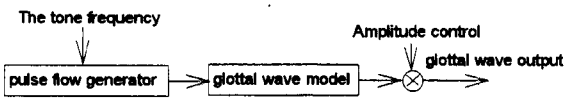

Fig.1 The voiced glottal excitation model

(2)Unvoiced

The excitation for the unvoiced sounds is comparatively simple. Usually it is simulated by putting a Gaussian white noise through a LP filter. In this system, the residual signal is taken as the excitation, which comes from the output of the corresponding inverse LMA filter with the original unvoiced speech as input, its duration and energy are modified depending on the practical demands of the synthesized speech. The diagram of the excitation model is shown below:

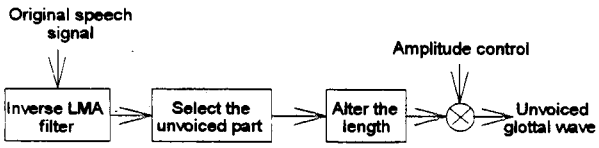

Fig.2 The unvoiced glottal excitation model

The following figure shows a wave section in the beginning part of the generated excitation waveform of "天" (sky), which includes the transition from the unvoiced part to the voiced part.

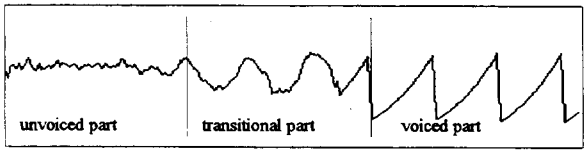

Fig.3 The excitation wave of the beginning part of "天"(sky)

3.2. The adjustment of the prosodic features

In this system, the speech library stores the acoustical parameters of the comparatively small speech units. During the synthesis, the system adjusts the parameters by a set of rules. The objects of these adjustments include the fundamental frequency (F0), duration and intensity of the units as well as the LMA filter parameters at the connection part of different units. The above prosodic features can easily be modified using the synthesis method reported in this paper.

(1) Pitch updating

The pitch contour is namely the tone contour. Since the pitch contour of a certain speech section is reflected exactly by the quasi-periodicity in its glottal excitation wave, when updating the unit's pitch contour we only need to build a new glottal excitation wave with the required F0 contour.

(2) Duration updating

Experimental phonetics shows that when we speak, from the beginning to the end, the vocal tract always changes continuously. It can be naturally supposed that when we produce a complete single syllable, the vocal tract changes its shape slowly and slightly, and that when we produce the same syllable in a sentence or a word, in order to keep the speech coherent, it will be shorter and quicker, and the corresponding change in the vocal tract shape will also be faster, but the universal change tendency should be the same. Based on the above assumption, in our speech synthesizer, if the vocal tract change tendency of the original speech unit is calculated out in advance, whenever the synthesis rules tell us that its duration must be changed, we can modify the speed of the change tendency to accord with the new duration while the universal characteristics remain. If the duration becomes shorter, we will delete some frames of the part in which the vocal tract changes slowly; if the duration becomes longer, we will add some frames to the part in which the vocal tract changes quickly.

(3) The co-articulation problem.

In the continuous speech, due to the organs of speech such as the tongue and the oral cavity always change gradually from one syllable to another, the beginning and the end of a syllable will be influenced by the adjoining syllables in a word or sentence, especially by the next syllable. Experiments show that the co-articulation problem in Chinese is much prominent when a vowel is followed by a zero-initial syllable. The problem is solved quite successfully with the LMA-Filter vocal tract model, where we just add the information of the second syllable's initial into the vocal tract parameters of the first one's tail gradually, so that the transition would become smooth. Synthesis experiments show that this method solves the co-articulation problem very well. The retroflexing is a special phenomenon in Beijing dialect, we can easily synthesize the corresponding retroflexed syllable from the original syllable using the same method.

As for the intensity, it only needs to normalize the synthesized waveform's amplitude directly.

4. A CHINESE TEXT-TO-SPEECH SYSTEM

On the basis of the LMA filter and the proposed prosodic rules, a Chinese text-to-speech system has been developed. The system can translate any Chinese text(including English alphabet and Arab digits) to continuous speech. The synthesized speech is clear and understandable, and sounds natural. The system is composed of three main modules: linguistic processing module,

prosody rule module, and speech synthesis module. The system frame is given in Fig.4.

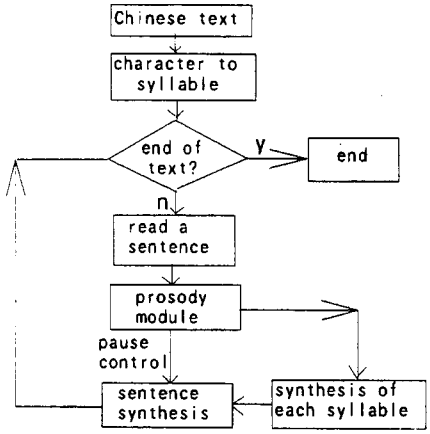

Fig.4 system frame

In the linguistic processing module the Chinese characters are converted to monosyllables, at the same time, the words are auto-segmented with the aid of a dictionary and lexicon, the syllables with light tone are found out. In the prosody controlling module, each syllable is endowed with a set of prosodic parameters (such as pitch contour, duration, amplitude and pause etc.) by rules. According to the prosodic parameters the syllables are synthesized by using the LMA filter based model, and then concatenated to sentences.

5. EVALUATION OF PERFORMANCE

Last year, a formal evaluation of speech quality for the Chinese text-to-speech systems was held in Beijing by the office of the State High Technology Development Project of China. The test was based on the subject evaluation of the syllable's clarity, word's and sentence's intelligibility. The mean opinion scores are shown in Figure.5.

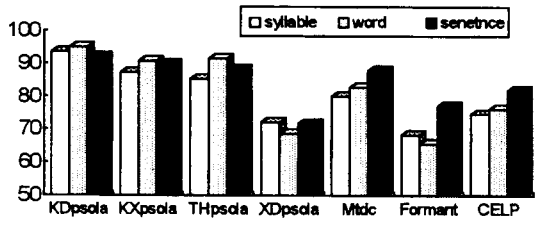

Fig.5 Evaluation results of speech quality

Where most systems use the method of waveform concatenation with the PSOLA technique. The results show that the intelligibility of those systems is much higher than of the other systems, such as Formant synthesis, CELP coding, and even the Mtdc--waveform concatenate directly with the natural syllables.

Though the PSOLA method improves the quality of synthesized speech greatly, it can not solve the co-articulation problems at all, especially, when the changes of F0 contour are too large, the output quality will be injured seriously. That is why we prefer to try the source/filter model. The LMA filter based synthesizer is essentially a kind of parameter's synthesis, which keeps the flexibility to control tone changing and the co-articulation. Meanwhile , the excellent performance of the LMA filter lets the synthesizer possible to output high quality speech. With the good strategies to control the excitation waveform and the duration, the introduced system manifests tremendous advantages for Chinese text-to-speech synthesis. An informal evaluation was carried out for testing the system performance recently. The KDpsola system, the top of the evaluated PSOLA based systems in Fig.5, is used as comparative object. Testing method is as same as the national one in last year. The results is given in Figure 6.

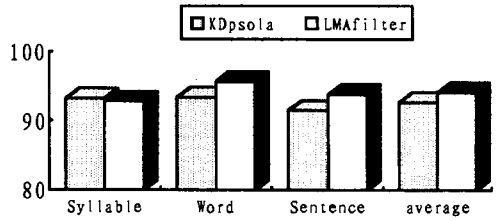

Fig.6 Intelligibility comparing of two systems

Fig.6 shows that the average intelligibility of the LMA filter based system already achieves 94.1%, and is higher than the 92.7% of KDpsola.

Synthesis experiments have shown that the 2-, 3- and 4-syllable words synthesized by our system are very close to their original speech now. Since the variation of sentence intonation is much more complicated, how to synthesize sentence intonation still remains a challenging project. Due to the quite successful of this system, we can foresee that: with the further perfecting of the prosodic rules, the system will be much hopeful to resolve the Chinese text-to-speech project.

6. REFERENCE

1, Zong-ji Wu, "further experiments on spatial distribution of phrasal contours under different range registers in Chinese intonation", International Symposium on Prosody, 18, September 1994, Yokoham, Japan

2, Ren-Hua wang, Deyu Xia, Jinfu Ni, Bicheng Liu, "USTC95--A Putonghua Corpus", Proc of ICSLP 96, 1996.

3, Satoshi, "Log Magnitude Approximation(LMA) Filter", Trans. IEICE, 80/12 Vol.J63-A No.12

4, Satoshi, Tadashi, "Speech Analysis Synthesis System Using the Log Magnitude Approximation Filter", Trans. IEICE, 78/6 Vol.J61-A No.6