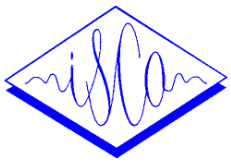

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A SCHEME FOR CHINESE SPEECH SYNTHESIS BY RULE BASED ON PITCH-SYNCHRONOUS MULTI-PULSE EXCITATION LP METHOD

Changfu Wang† *Wenshen Yue*† *Keikichi Hirose*†† and *Hiroya Fujisaki*†††

†Department of Electronic Engineering, University of Science and Technology of China
Hefei, Anhui, 230026 P. R. China
††Department of Electronic Engineering, University of Tokyo
7–3–1 Hongo, Bunkyo–ku, Tokyo, 113 Japan
†††Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

ABSTRACT

A pitch-synchronous analysis and synthesis method has been developed for Chinese speech synthesis by rule. This method is based on the multi-pulse excitation linear predictive (MELP) analysis. The wavelet transform is adopted to detect the instant of glottal closure (GCI) from the speech signal. Since the analysis is carried out pitch-synchronously within the interval of a pitch period bounded by two adjacent GCI's for a voiced speech segment, the major excitation occurring at a GCI is excluded from the analysis interval. Thus higher accuracy is achieved for vocal tract and excitation source parameters than that obtainable by conventional fixed frame methods. Since, in the synthesis process, the speech segments of one pitch period each are used as the units of synthesis for voiced speech, the pitch-synchronous scheme also simplifies the process of control parameter generation and increases the flexibility and controllability of the synthesizer.

1. INTRODUCTION

Because speech is expected to be the most efficient and flexible means of communication between man and machine, a large amount of effort has been devoted to develop talking machines. Since speech synthesis by rule is capable of producing unlimited number of sentences from a limited amount of stored data, it has been studied by many research groups. Many of the works on Chinese speech synthesis by rule have been based on concatenating syllable segments represented by linear predictive parameters or their equivalents[1]. Since in the conventional methods, the vocal tract parameters are extracted paying little attention to the influence of the excitation source, and the model of excitation source is over-simplified, it is rather difficult to synthesize speech with a high quality. Moreover, without good models for generating F_0 contours of sentences, the control of prosodic features becomes rather complex.

Chinese morphemes are monosyllabic. Even in continuous speech, the constituent syllables can be distinguished rather clearly. Therefore, concatenating syllables to produce sentences is especially suited for Chinese speech synthesis. The structure of Chinese syllables is relatively simple. A syllable consists of an initial consonant part (which may be absent) and the main part consisting of a vowel or a diphthong, sometimes terminated by /n/ or /ng/. In this paper, we define the initial consonant part as the portion of a syllable consisting of the initial consonant and the transition to the following vowel, and the main part as the portion starting with the quasi-stationary part of the vowel following the initial consonant and the transition. In Chinese syllables, the initial consonant is relatively short, and its duration is rather independent from the speech rate. On the other hand, the main part has a longer duration which depends heavily on the speech rate, and its acoustic characteristics are relatively stationary as compared with those of the initial consonant part.

Chinese is a typical tone language. Although the total number of phonologically allowed syllables in Chinese is about 1300 considering the tone types, the number is reduced to 412 if we disregard the tone types. The F_0 contours for different syllables with the same tone-type remain basically the same, and, therefore, we can produce syllables with different lexical tone-types from a stored syllable template by superposing F_0 contours of four tones on the segmental feature pattern of the template. Since the main part consists of three phonemes at most, it can roughly be represented by five segments including two transitions between the phonemes. Therefore, we select manually those five speech segments with typical pitch periods and use them to represent the main part of each syllable.

Based on these considerations, we propose a pitch-synchronous method for synthesis by rule of Chinese based on the multi-pulse excitation linear predictive method. Because of the pitch-synchronous analysis, higher accuracy can be achieved for extracted parameters. Combined with the use of a generative model for sentence F_0 contour generation, the proposed scheme assures high quality of synthetic speech.

2. DETECTION OF GCI

It is well known that at the instant of glottal closure (GCI), the major excitation of the vocal tract occurs causing sudden changes in the speech signal. These sudden changes corresponding to GCIs can be detected using the Dyadic Wavelet Transform (DyWT):

$$DyWTs(b,2^j) = \frac{1}{2^j}\int_{-\infty}^{+\infty} s(t)g(\frac{b-t}{2^j})dt \equiv s(t) * g_{2^j}(t),$$

where s(t) denotes the speech signal, $g_{2^j}(t) = 1/2^j * g(t/2^j)$ denotes the wavelet function. It has been shown that if we choose the first derivative of a smoothing function as a wavelet function and adopt an appropriate j, the local maxima of the DyWTs take place at points with sudden changes in the speech signal [2]. In other words, the locations of local maxima indicate the instants of glottal closures. Hence, we can detect GCI on the basis of the local maxima of DyWTs.

Using spline interpolation, we construct a wavelet function which satisfies the condition mentioned above. In Figure 1, panels (A), (B) and (C) respectively show the wavelet, the speech segment and the detected GCI. From panels (B) and (C) we can see that the GCI's are detected accurately. The discrepancy between the GCI detected using DyWTs and that estimated by visual inspection of the speech waveform is, in almost all the cases, less than three sampling points. Thus the information concerning GCI detected by DyWTs can be used to conduct pitch-synchronous analysis.

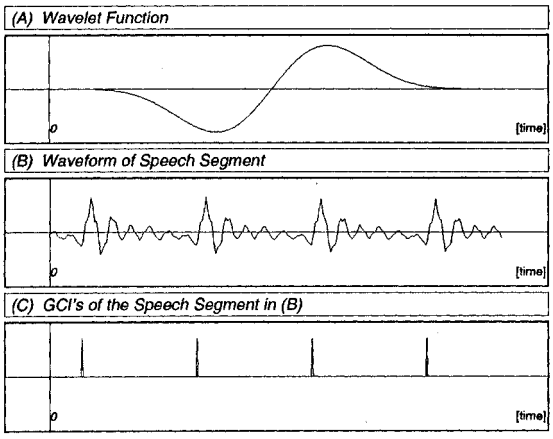

Fig. 1. Wavelet function, speech segment and its GCI.

3. PITCH SYNCHRONOUS ANALYSIS USING MELP METHOD

Atal et al. proposed a multi-pulse excitation model, in which an excitation signal is represented by multiple pulses [3]. Speech synthesis is conducted using a synthesis filter excited by the multiple pulses. In conventional methods, parameters of the synthesis filter are calculated from the original speech signal using a frame-by-frame LPC analysis. The multi-pulse excitation signal is determined using the analysis-by-synthesis method with a sequential pulse search procedure. A weighting filter is introduced to reduce the distortion

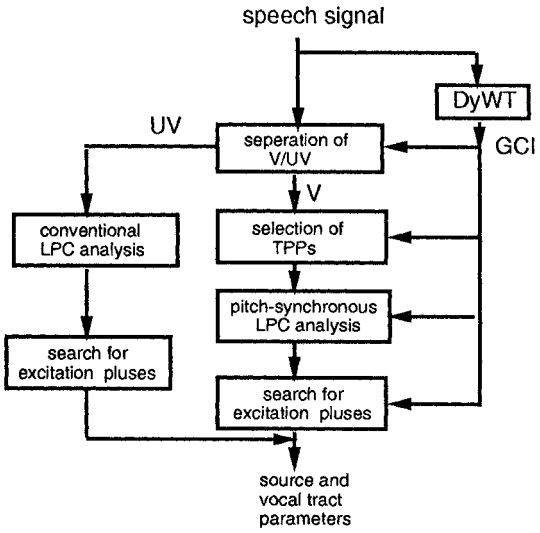

Fig. 2. Pitch-synchronous MELP analysis.

using the criterion of minimized mean square error between the original and the synthetic speech waveforms. In order to reduce the computational complexity, a simplified pulse search algorithm is adopted: the optimum pulse location n_i of the ith pulse is determined by searching for the maximum absolute value g_i:

$$g_i = \frac{R_{hx}(n_i) - \sum_{k=1}^{i-1} g_k * R_{hh}(n_k - n_i)}{R_{hh}(0)}, \quad 1 < n_i, n_k < N,$$

where N is the length of the analysis frame, $R_{hx}(n)$ is the crosscorrelation function between the weighted speech signal and the weighted impulse response, $R_{hh}(n)$ is the autocorrelation function of the weighted impulse response, and g_i is the amplitude of the ith pulse.

In the conventional MELP analysis, the LPC parameters of the vocal tract are extracted using a fixed frame length. Therefore, the influence of the excitation source on the vocal tract parameters is not fully taken into account, even though the inaccuracy of LPC parameters can be compensated for automatically by adjusting the locations and amplitudes of the pulses in the search process. This adjustment may affect the locations and amplitudes of the pulses, and may even destroy their periodicity. In order to overcome the weakness of the conventional MELP analysis, we propose a pitch-synchronous method for multi-pulse excitation LPC analysis, as shown in Figure 2.

In the proposed method, we first detect and separate voiced and unvoiced intervals of a syllable using the GCI information. For the voiced interval, LPC analysis is conducted on each pitch period bounded by two adjacent GCI's, while the conventional LPC analysis with a fixed frame length is adopted for the unvoiced interval. In order to avoid the inaccuracy in GCI detection, the actual interval for the LPC analysis is set from gci(k) + 5 to gci(k+1) − 5, where gci(k) denotes the instant of occurrence of the kth GCI. This assures that the major

excitation pulses are excluded from the analysis interval, so that it not only reduces the effect of excitation source on the LPC parameters, but also increases the accuracy of extraction of locations and amplitudes of excitation pulses for the voiced interval of the syllable. Since the main part of a syllable can be represented by five speech segments of typical pitch periods (TPP), the analysis of the main part can be replaced by that of these five typical speech segments for the purpose of speech synthesis. This scheme considerably reduces the computational time for extracting the LPC parameters and for searching for the excitation pulses as well as the memory size for the storage of the syllable templates. As a result, the following information is stored only for the first tone syllables:

1. Vocal tract and excitation parameters of the initial consonant part including its transition to the main part, together with the length and the number of analysis frames.

2. Positions, lengths, vocal tract and excitation source parameters of the five typical pitch periods of the main part of the syllable, together with the number of pitch intervals represented by each one of the five typical periods.

The difference between the position of the main pulse with the largest amplitude in a pitch period and that of GCI detected by DyWT is, in almost all the cases, not more than three sampling points. Since the acoustic characteristics of the initial consonant part varies more quickly, and contains more acoustic information than the quasi-stationary main part, compression is usually not conducted on the initial consonant part.

4. PITCH SYNCHRONOUS SYNTHESIS OF CHINESE MONOSYLLABLES

We first describe the reconstruction of the first tone syllables. The lattice synthesizer with k parameters is adopted. For the unvoiced interval, the speech segments are synthesized frame by frame using the conventional MELP synthesis method. For the voiced interval, the length of each pitch period is first calculated according to a given F_0 contour of the first tone, and then each interval between two adjacent pulses in a pitch period is compressed or expanded proportionately by inserting or deleting zero-point(s) in order to adjust the length of the pitch period of the stored syllable template to the desired length (see Figure **3**).

For the transition segment, the excitation pulses modified drive the synthesizer with corresponding k parameters to synthesize the speech segment for each pitch period. For the main part of a syllable, we first compress or expand linearly the length of each pitch period, and count the number of pitch periods of the segment which can be represented by the parameters of each one of the five typical pitch periods according to the difference between the original and given lengths of the first

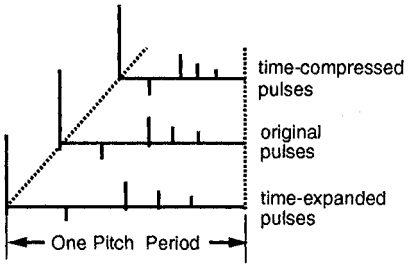

Fig. 3. Compression and expansion of excitation pulses.

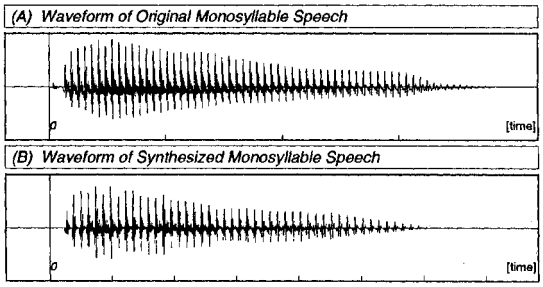

Fig. 4. Waveforms of original and synthetic speech for Chinese monosyllable /bao(1)/.

tone (T1) syllables. We then synthesize the speech segments represented by the parameters of the five typical pitch periods. Since the speech segment represented by one typical pitch period is long, the synthetic speech may contain noise caused by discontinuities of the vocal tract and excitation parameters. In order to reduce the noise, for the speech segments of the three pitch periods adjacent to the next speech segment represented by the parameters of another typical pitch period, the k parameters and amplitudes of excitation pulses are smoothed linearly toward those of the next segment. In synthesizing the speech segment represented by the parameters of the fifth typical period, the amplitudes of the excitation pulses are reduced linearly from the tenth pitch period counted from the end to the last pitch period of a syllable , while the k-parameters representing the fifth typical pitch period are left unchanged. In Figure **4**, panel (A) shows the original waveform for the first tone monosyllable, while panel (B) shows the synthesized waveform using the parameters extracted from the speech segment shown in the panel (A).

The synthesis procedure for syllables with tones other than T1 is almost the same as for the syllables with T1, except for the larger amount of adjustment of the pitch period length required by the tone-contour of the synthesized syllable.

It should be pointed out that the fifth tone called "neutral tone" or "light tone" is phonologically related to weak stress, and is usually described as flattened. Although its F_0 contour is flat as in the case of the first tone, its pitch period is longer than that of the first tone, and its duration is relatively short. Its manifestation heavily depends on the tone type of its preceding syllable. Therefore the fifth tone is treated as the first tone with reduced power and tone command amplitude

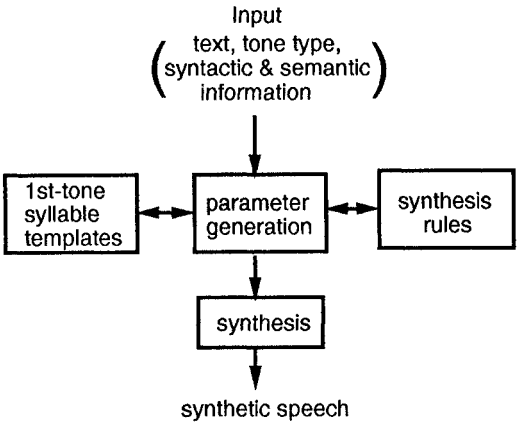

Fig. 5. Block diagram of the system for speech synthesis by rule for Chinese.

as well as shorter duration in our system. The fifth tone cannot exist in isolation, but can only follow syllables with other tone types.

5. SYNTHESIS OF CHINESE SPOKEN SENTENCES

The acoustic parameters for the synthesis of spoken sentences are obtained from syllable-unit parameters using synthesis rules for concatenation. These rules define how the parameters describing the acoustic properties of syllables should be adjusted when these syllables are concatenated to produce a sentence. They consist of rules for tone type, tone sandhi, tone enhancement/suppression, syllable duration and pause, based on the analysis of continuous speech and phonological knowledge, and are incorporated into the speech synthesis system. Figure **5** shows the block diagram for the synthesis by rule of Chinese sentences. Since, in Chinese, the F_0 contour is the major manifestation of the prosodic features, a functional F_0 contour model is adopted to realize flexible control of F_0 contours [4-6]. According to this model, the instantaneous fundamental frequency of the voiced interval of each syllable consists of contributions by the lexical tone and by the phrasal intonation, which can be derived respectively from the lexical tone of each syllable and from the syntactic structure of the sentence to be synthesized. In calculating the lexical tone component, the rules of tone sandhi and of tone enhancement/suppression have to be applied. In the synthesis process, speech segments of one pitch period are used as the units of synthesis. This scheme increases the flexibility and controllability of the synthesizer. Figure **6** shows the waveform of the synthesized speech for a Chinese sentence.

6. CONCLUSION

A prototype for text-to-speech conversion system of Chinese weather forecast has been constructed on a computer. The system has only 80 stored syllable templates with the first tone related to weather forecast sentences. Necessary inputs for the system are texts in

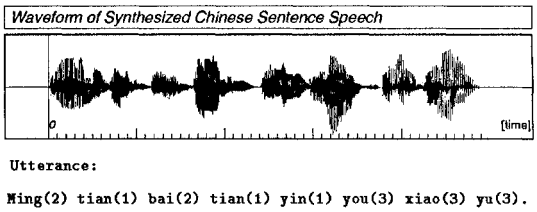

Fig. 6. Waveform of synthesized speech for a Chinese sentence.

phonetic alphabet, with tone type of each syllable, syntactic information and semantic information. The system is capable of reconstructing the syllables with the first tone using corresponding parameters stored as syllable templates. To synthesize the syllables with other tones using the parameters of the first tone syllables, pitch period is adjusted according to the F_0 contour of the corresponding tone-types. Synthetic sentences are generated by concatenating the synthesized syllables by rule. Synthetic speech of rather high quality is obtained by the system, indicating that the proposed method is effective.

Improving the method for extracting acoustic parameters and making the synthesis rules complete should be our further tasks.

ACKNOWLEDGEMENT

The authors gratefully thank the colleagues at the University of Tokyo and the University of Science & Technology of China for their help and support in this research.

REFERENCES

[1] L. Lee, C. Tseng and M. Ouh-Young, "*The synthesis rules in a Chinese text-to-speech system*," IEEE Trans. on ASSP, Vol. 37, No. 9, pp. 1309-1319, 1989.

[2] S. Mallat and W. Hwang, "*Singularity detection and processing with wavelets*," IEEE Trans. on Information Theory, Vol. 38, No. 2, pp. 617-637, March 1992.

[3] B. S. Atal and J. R. Remde, "*A new model of LPC excitation for producing natural-sounding speech at low bit rates*," Proc. ICASSP 82, Vol. 1, pp. 614-617, 1982.

[4] H. Fujisaki, K. Hirose, P. Halle and H. Lei, "*Analysis and modeling of tonal features in polysyllablic words and sentences of the standard Chinese*," Proc. ICSLP 90, Vol. 2, pp. 841-844, Kobe, Japan, 1990.

[5] H. Fujisaki, K. Hirose and H. Lei, "*Prosody and syntax in spoken sentences of standard Chinese*," Proc. ICSLP 92, Vol. 1, pp. 433-436, Banff, Canada, 1992.

[6] K. Hirose, H. Lei and H. Fujisaki, "*Analysis and formulation of prosodic features of speech in standard Chinese based on a model of generating fundamental frequency contour*," J. Acoust. Soc. Jpn., Vol. 50 No. 3 pp. 177-187, 1994.