A PROSODIC MODEL OF MANDARIN SPEECH AND ITS APPLICATION TO PITCH LEVEL GENERATION FOR TEXT-TO-SPEECH

Shaw-Hwa Hwang and Sin-Horng Chen

Department of Communication Engineering and
Center For Telecommunications Research
National Chiao Tung University
Hsinchu, Taiwan 300, Republic of China

ABSTRACT

A prosodic model of Mandarin speech is proposed to simulate human's pronunciation mechanism for exploring the hidden pronunciation states embedded in the input text. Parameters representing these pronunciation states are then used to assist prosody information generation. A multirate recurrent neural network (MRNN) is employed to realize the prosodic model. Two learning methods were proposed to train the MRNN. One is an indirect method which firstly uses an additional SRNN to track the dynamics of the prosody information of the utterance; and then takes the outputs of its hidden layer as desired targets to train the MRNN. The other is a direct training method which integrates the MRNN and the following MLP prosody synthesizers to directly learn the relation between the input linguistic features and the output prosody information. Simulation results confirmed the effectiveness of the approach. Most synthesized prosodic parameter sequences match quite well with their original counterparts.

1. INTRODUCTION

In a text-to-speech (TTS) system, prosody information is generated from linguistic features obtained by analyzing input text. Although many approaches[1-3] have been studied in the past, it remains a general difficulty to elegantly invoke high-level linguistic features in prosody generation. This is especially true for fundamental frequency (F0) generation. Usually, only simple intonation patterns to consider the syntactic structure of input sentence were used to count the effect from high-level linguistic features. The resulting synthesized speeches are in general not very natural.

In this paper, a novel approach to employ a prosodic model to assist the generation of proper prosody information for Mandarin TTS is proposed. The prosodic model tries to simulate human's pronunciation mechanism to explore the hidden pronunciation states, which control the prosody information generation, based on interpreting the linguistic features of the input text. In other words, the prosodic model attempts to track the hidden prosodic states embedded in the input text. With the help of such a prosodic model, we can expect that the synthesized prosody information is more nature to mimic that of human's speech.

The organization of the paper is described as follows. Section 2 discusses the proposed prosodic model in detail. The new approach of synthesizing prosody information in Mandarin TTS assisted with the prosodic model is presented in Section 3. Effectiveness of the proposed method is examined by simulations discussed in Section 4. Some conclusions are given in the last section.

2. THE PROSODIC MODEL

A simplified model of the pronunciation process in human's brain is shown in Fig.1. In the model, input text is first processed by a text analysis to extract some linguistic features. Then, these linguistic features are interpreted by a prosody pronunciation mechanism to generate prosody information. Basically, a variety of linguistic features ranging from low-level lexical features, such as the tone of a syllable, to high-level features like a syntactic boundary can be used by the pronunciation mechanism to affect the prosody information generation. Due to the fact that the affection of high-level linguistic features are more global, we suggest in this paper to conceptually divide the prosody pronunciation mechanism of human into two parts as shown in Fig.2. One is a prosodic model which uses high-level linguistic features to generate prosodic states which track the global trend of prosody inforamtion generation. The other is a prosody generator which uses low-level lexical features and prosodic states to produce prosody inforamtion. In this section, we will focus our discussions on the method to construct the prosodic model.

The basic idea of our approach is to realize the prosodic model by an artificial neural network to automatically learn from a large database containing nature utterances accompanying with their texts. A multirate recurrent neural network (MRNN) shown in Fig.3 is adopted in this study to implement the prosodic model. The MRNN has two hidden layers operating on different clocks. The first hidden layer works with a clock synchronized with the phrases of the in-

This work was supported by the National Science Council, ROC, under contract NSC83-0404-E-009-091. The authors want to thank Telecommunication Laboratories, MOTC, ROC for supporting the speech database.

0-7803-2431-5/95 $4.00 © 1995 IEEE

put text. The second hidden layer works with another clock synchronized with the syllables. Both hidden layers feedback their outputs to themselves with unit delay of their own clocks. Input features of the first layer include the part-of-speeches (POSs) and the lengths in syllable of the current and the following phrases, and punctuation mark (PM) indicators. For the second hidden layer, aside from accepting outputs of the first hidden layer, an additional feature is fed-in explicitly. It is the location of the processing syllable in the current phrase.

Two methods are proposed to train the MRNN. One is an indirect method which firstly uses an additional simple recurrent neural net (SRNN) to learn the dynamic states embedded in the prosody information of the utterance associated with the input text. The SRNN operates in synchronization with syllable-syllable boundaries of the utterance. Input features include the duration and the energy dip of the processing boundary, and the F0 means and the energy levels of the preceding and the following syllables. Output features are four binary indicators respectively to show whether the boundary is an inter-word, an intra-word, or the existence of two types of PM. The reasons of using only low-level linguistic features in the training are to suppress the affection from text's syntax and let the SRNN have enough flexibility to track the dynamics of the prosody information of the utterance. Then, outputs of all hidden nodes of the SRNN at the boundary immediately following the processing syllable are taken as desired output targets to train the MRNN. The other is a direct training method which integrates the MRNN and the following prosody generator to directly learn the relation between the input linguistic features and the output prosody information by the back propagation (BP) algorithm. We will discuss the integrated prosody synthesizer in more detail in the following section.

3. SYNTHESIS OF PROSODY INFORMATION

For Mandarin TTS, prosody information to synthesize include pitch (or F0) contour, energy contour and duration of syllable as well as pause duration between syllables. In this study, only the syntheses of pitch mean, energy level, and final duration of syllable, and pause duration between syllables are considered. These synthesis parameters are normalized to reduce their variabilities for suppressing the influences from local lexical features. This can make their syntheses more easy. To be more specific, both energy level and final duration of syllable are normalized for each of 39 final types. Pause duration, on the other hand, is normalized for each of 22 initial types of the following syllable.

The use of the MRNN prosodic model to assist a basic MLP prosody synthesizer[4] for improving the generations of these prosody parameters is now discussed. The basic MLP prosody synthesizer uses only local lexical features such as the tones of the processing syllable and its two neighbours, the initial types of the processing syllable and the following syllable, etc. Two methods to combine the MRNN prosodic model and the basic MLP prosody synthesizer to form a more powerful prosody synthesizer are proposed. The first method is to serially concatenate them by taking the outputs of the MRNN prosodic model as additional inputs of the basic MLP prosody synthesizer. Fig.4(a) shows the block diagram of the new prosody synthesizer. Training of the new synthesizer is stated as follows. First, the MRNN prosodic model is trained by the indirect training method discussed in the previous section. Then, the basic MLP prosody synthesizer is trained with all weights of the MRNN being forzen. An optional third step can be execute to fine tune the synthesizer by defrosting the MRNN and simultaneously retraining the MRNN and the MLP. The other method is to tightly incorporate the MRNN prosodic model and the MLP prosody generator to form a compact prosody synthesizer as shown in Fig.4(b). The second hidden layer of the MRNN is merged with the hidden layer of the MLP to make the whole system more compact. The synthesizer is trained by the direct training method discussed in the previous section.

4. SIMULATIONS

Performance of the new approach of prosody inforamtion synthesis for Mandarin TTS was examined by simulations. A continuous Mandarin speech database provided by Telecommunication Laboratories, MOTC, ROC was used. The database contains 655 sentential and paragraphic utterances and their texts. All utterances were generated by a single male speaker. They were all spoken naturally with a speed of 3.5 - 4.5 syllables per second. The database was separated into two parts: a training set and an outside test set. These two sets consist of 28191 and 7051 syllables, respectively.

All speech signals were digitally recorded in a 20 kHz sampling rate. They were then divided into 4-ms frames and manually segmented into silence, unvoiced and voiced parts based on observing some acoustic features including waveform, energy, zero crossing rate, LPC coefficients, cepstrum and delta-cepstrum. Prosody parameters to synthesis were then extracted from the down-sampled 10-kHz speech signals. They include pitch mean, maximal log energy, and final duration of each syllable, and pause duration in each syllable pair. Pitch period was detected by the SIFT algorithm[5] with manually error-correcting. The frame length for F0 analysis is 40 ms with 10-ms frame shift, and is 20 ms with 10-ms frame shift for log-energy analysis.

An automatic text analysis was used to transform all texts accompanying speech utterances into sequences of lexical words by using a Chinese lexicon containing approximately 80000 words. Each lexical word in the lexicon contains 1 - 5 syllables. After obtaining all proper word sequences, POSs of all words were manually determined. A set of 44 types of POSs was used in the study. Then, all linguistic features used in the system were extarcted.

Table 1 lists the mean squared errors of the synthesized prosody parameters for the two methods of combining the MRNN prosodic model and the basic MLP prosody synthesizer. Mean square errors of 0.7 ms/frame and 0.95 ms/frame were achieved by the first method for the synthesized pitch mean in the inside and the outside tests, respectively. They are 0.9 ms/frame and 0.84 ms/frame for the second method. Fig.5 shows typical examples of pitch mean synthesis by these two methods. Can be seen from these two figures that the trajectories of synthesized pitch

mean in both examples match quite well to their original counterparts. Fig.6 shows typical examples of energy level synthesis by these two methods. It can be found from the figure that reasonable good results were obtained for both examples. Experimental results shown above confirm that the MRNN prosodic model is very helpful to the prosody generation in Mandarin TTS. For further examining the characteristics of the prosodic model, we vector-quantized the outputs of the MRNN and calculated the state transition probabilities to form an 8-state automata. By examining some typical quantized state sequences of texts, we found that the automata roughly conforms to the syntax of Chinese language. Fig.7 shows some typical quantized state sequences. A sentence usually starts from State 1 or State 4 and ends at State 3 or State 7. A bisyllabic phrase always starts at State 8, 6, or 5 and ends at State 3. States 4 and 7 are associated with monosyllabic phrases. States 5 and 2 are associates with syllables of long phrases. From above discussions, we therefore conclude that the proposed prosodic model is an effective model to capture some high-level linguistic information of Chinese language for helping the synthesis of prosody information.

5. CONCLUSIONS

A neural network-based prosodic model has been proposed in this paper to explore the hidden pronunciation states embedded in the input text for assisting the prosody synthesis in Mandarin TTS. Experimental results have confirmed that it is an effective model to capture high-level linguistic information of Chinese language. With its help, promising results have been obtained in our study of synthesizing pitch mean, energy level, and final duration of syllable, and pause duration between syllables by a neural network-based prosody synthesizer.

6. REFERENCES

[1] Lin-Shan Lee, Chiu-Yu Tseng, Ming Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. ASSP. Vol-37, pp.1309-1320.1989

[2] Y. Sagisaka, "On the prediction of global F0 shape for Japanese text-to-speech," ICASSP, pp.325-328, 1990.

[3] Sin-Horng Chen, Su-Min Lee, and Saga Chang, "A statistical model based fundamental frequency synthesizer for Mandarin speech," J. Acoust. Soc. Am. pp.114-120, July. 1992.

[4] S. H. Hwang, and S. H. Chen, "A neural network based F0 synthesizer for Mandarin text-to-speech system," to appear on IEE Proceding-I

[5] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, No.5, pp.367-377, Dec. 1972.

Table 1. The MSEs of the synthesized prosody parameters for the two methods.

	Pitch	Max. Eng.	Pause	Dur.
Inside(I)	0.71ms	3.56dB		
Outside(I)	0.96ms	4.32dB		
Inside(II)	0.90ms	4.27dB	0.92ms	0.42ms
Outside(II)	0.84ms	3.73dB	0.93ms	0.39ms

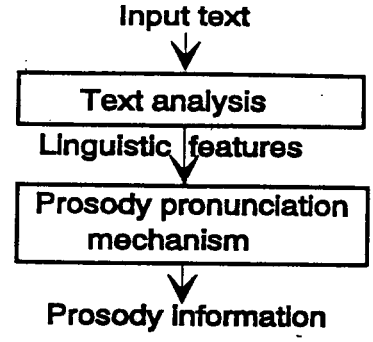

Fig.1. A simplified model of the pronunciation process in human's brain.

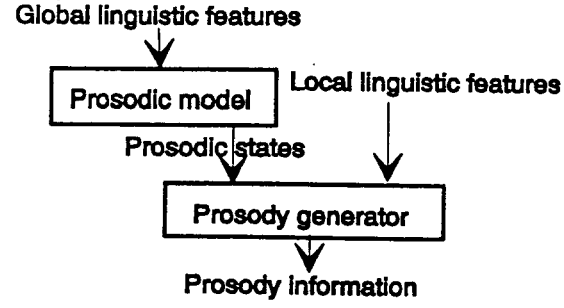

Fig.2. A conceptural model of the prosody pronunciation mechanism of human.

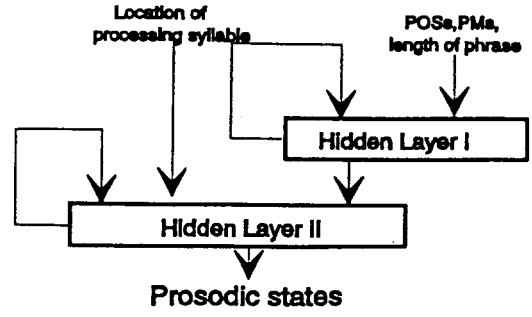

Fig.3. The MRNN prosodic model.

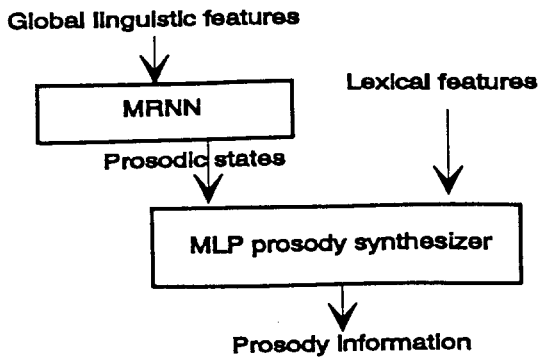

Fig.4a. Prosody synthesizer I.

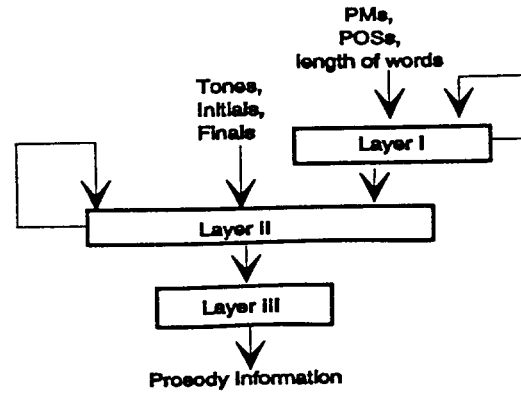

Fig.4b. Prosody synthesizer II.

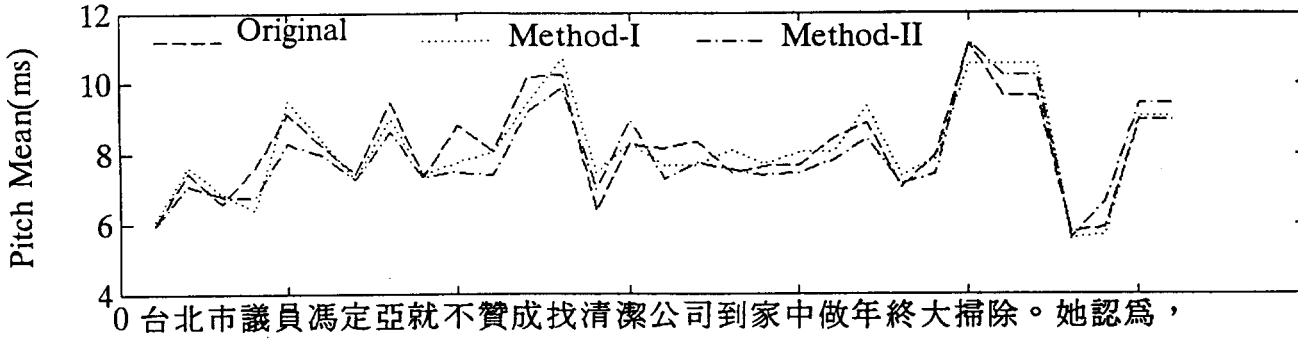

Fig.5 The original and synthesized pitch means with two methods.

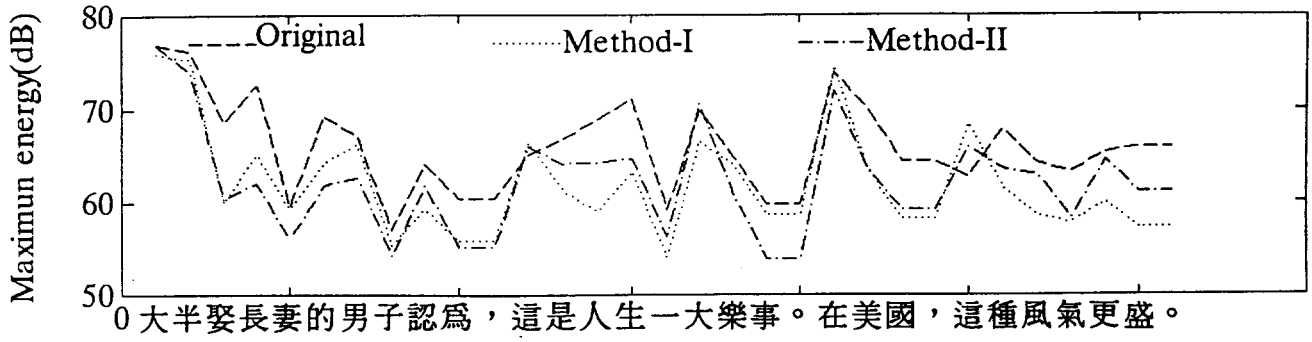

Fig.6. The original and synthesized energy with two methods.

請1把7這8籃5兔5子3送6走3。

大1半3娶7長8妻3的4男8子3認6爲3，這4是7人8生3一5大3樂6事3。

在4美8國3，這6種6風5氣3更4盛7。四8十5五7歲8的6歌3帶6韓2和5

三8十6九5歲8的5寇7特8羅6素2是7公8認8的4銀5色5佳6侶1；

Fig.4. The typical quantized state sequence of sentence.