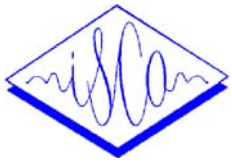

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

AN RNN-BASED SPECTRAL INFORMATION GENERATION FOR MANDARIN TEXT-TO-SPEECH

Shaw-Hwa Hwang, Sin-Horng Chen@, and Saga Chang**

*E000/CCL, Industrial Technology Research Institute, Chutung, Hsinchu, Taiwan, R.O.C
@Department of Communication Engineering, National Chiao Tung University, Taiwan, R.O.C
email:hsf@porsche.ccl.itri.org.tw Tel:+886-3-5917255, Fax:+886-3-5820098

ABSTRACT

In this paper, an RNN-based spectral model is proposed to generate spectral parameters for Mandarin text-to-speech(TTS). The RNN is employed to learn the relations between the linguistic features and the spectral parameters. The phoneme-to-spectral parameter rules and the coarticulation rules between each two adjacent phones are automatically learned and memorized into the weights of RNN. The synthesized speech sounds more fluent and smooth. The RNN is divided into two parts. The first part is synchronized with syllable and is expected to simulate the phoneme-to-spectral parameter rules. The second part is synchronized with frame and is expected to simulate the coarticulation rules between each two adjacent phones. The line spectrum pair(LSP) parameters and the normalized energy contour are taken as target value. Training with large database, the synthetic LSP and energy contour match to the original LSP and energy contours quite well. Moreover, an RNN-based prosodic model which was proposed in our previous study was combined to the spectral model to efficiently simulate the spectral and prosodic information generation. Lastly, the LPC-based Mandarin TTS is implemented to examine the performance of our spectral model. The synthetic speech sounds fluent and natural. The coarticulation effect between each two adjacent phones which makes synthesized speech sounds un-fluent and echo-like was improved. However, due to the simple structure of LPC-based synthesizer, the clarity of synthetic speech can be improved by using the other spectral parameter as target value. For example, the modify mel-cepstrum parameter[5, 6, 7] or the FFT-based spectral parameter can also be learned by RNN and synthesizes more clarity speech. This is a initial work on the RNN-based spectral model for text-to-speech. Some advantages of our spectral model can be found. First, large memory space of synthesis unit in traditional TTS is replaced by small memory space of RNN's weights. Second, the coarticulation effect can be alleviated and produces more fluent speech. Third, the RNN-based prosodic and spectral information generator[8, 9] can be easily combined to formed a more compact RNN-based TTS system.

The authors want to thank Telecommunication Laboratories, MOTC, ROC for supporting the speech database. We also want to thank Academia Sinica for supporting the lexicon.

1. INTRODUCTION

Speech is the most friendly interface between human and machine. Amoung many speech technologys, the text-to-speech(TTS), automatic convertion of stored text to synthetic speech, plays an important role in many applications of computer. Many attractive applications on the text-to-speech will be found in the future.

An generic TTS system can be functionally divided into four parts: text analysis, spectral information generation, prosodic information generation, and speech synthesizer. Input text is first analyzed in text analysis to extract some linguistic features. Then, the template sequence of synthesis unit and prosodic information can be assigned and generated according to the linguistic feature. The template sequence of synthesis unit is formed in the spectral information generation in order to make the synthetic speech sounds clear. Nevertheless, spectral smoothing method may also be performed on the template sequence in order to alleviate the coarticulation effect between each two adjacent synthesis units and make synhetic speech sounds fluent. Prosodic information generation assigns the pitch, timing, and stress pattern according to the linguistic features for the speech synthesizer to modify the template sequence in order to generate an intelligible and natural speech. Speech synthesizer receives prosodic and spectral parameter and products synthetic speech.

A clear, fluent, natural, and intelligent speech for unlimited text is the main goal to develop an TTS system. Prosodic and spectral informations which embed in the speech are the two major parameters to make speech sounds clear, fluent, natural, and intelligent. Thus, in the past, most researches were focused on the prosodic information generation and spectral information generation. In the prosodic model, the intelligence and naturalness of synthetic speech are the two main goals. Two main approaches which include rule-based and data-driven were used to generate prosodic information. Successful results were obtained by those approaches and natural speech was obtained. In the spectral model, the clarity and fluency of synthetic speech are the two main goals. The first goal can be easily achieved by increasing the data rate of synthesis unit. The PSOLA-based approach is a typical example. In order to achieve the second goal, smoothing the spectral parameters between each two adjacent synthesis units is very important. There are two major approaches were studied in the past. The first one is to directly use a

spectral smoothing technique[1]. In this approach, many coarticulation rules must be inferred by observing a large set of utterances with the help of linguists and acoustic expert. The other is to model each synthesis unit by using multiple templates and then choose a proper one in the synthesis according to the context[3, 4].

Recently, the HMM-based spectral model[5, 6, 7] is proposed to model the transition on spectral parameter between each two adjacent phones. In this approach, both static and dynamic features are taken into account when spectral parameters are generated from HMMs. Synthetic speech with quite smooth can be obtained. It avoided the difficulty of manually inffered coarticulation rule.

Motivated by the successes of our researches on prosodic information generation[8, 9] in the past, an RNN-based spectral parameter generation is proposed in this paper. The RNN is employed to learn the relations between the linguistic features and the corresponding spectral parameters. Two input features with different clocks are used to learn the phoneme-to-spectral information rule and coarticulation rule between each two adjacent phones individually. Then, the RNN can be taken as a mechanism of generating the spectral parameter from the given linguistic features. Experiment result shows that phoneme-to-spectral information and coarticulation rules are automatically learned. An LPC-based Mandarin text-to-speech system was used to examine the performance of the spectral model. The synthesized speech sounds smooth and fluent.

This paper is organized as follows. The proposed system of synthesizing spectral parameters is discussed in Section 2. Simulation results and discussions are listed in Section 3. Conclusions are given in the last section.

2. THE PROPOSED SYSTEM

Mandarin is a tonal and syllabic language. Each Chinese character is pronounced as a syllable. There are only about 1300 phonetically distinguishable syllables, which are the set of all legal combinations of 411 base-syllables and 5 tones. Each base-syllable is composed of an optional consonant initial and a vowel final. The continue speech is composed of syllabic sequences which are pronounced according to Chinese characters and are directly connected. In the global view on the continue speech, each syllabic waveform is static and maps to an Chinese character. In the local view on the running speech, each syllabic waveform is dynamic and smoothly or abruptly changes from one pattern to another. Moreover, the sylable is composed of an consonant initial and a vowel final, the speech waveform intra the syllable is also dynamic. Thus, producing the proper spectral information and synthsizing fluent and smooth speech is not an easy job.

In our spectral model, the multi-layer recurrent neural network(RNN) is adopted in this work to implement the spectral information generation. Fig.1 depicts the block diagram of the RNN. The RNN is composed of two hidden layers, one output layer, and two input layers which operate in different clock. It can be functionally decomposed into two parts. The first part consists of the first input layer and the first hidden layer with all outputs feeding back to the input of itself. It is regared as a mechanism of phoneme-to-spectral parameters. It operates on a clock synchronized with syllable to generate some outputs representing the steady state of spectral parameter at the current syllable. The input featues include the tone $T(S_j)$, the type of initial $I(S_j)$, and the type of final $F(S_j)$ of the processing syllable S_j; the tone $T(S_{j+1})$ and the initial type $I(S_{j+1})$ of the following syllable S_{j+1}; the tone $T(S_{j-1})$ and the final type $F(S_{j-1})$ of the preceding syllable S_{j-1}; the pause duration $P(S_j)$ preceding the processing syllable S_j; and the pause duration $P(S_{j+1})$ preceding the processing syllable S_{j+1}.

The second part of the RNN consists of the second input layer, the second hidden layer, and the output layer. It is the real spectral parameter generator. It operates on a clock synchronized with frame to generate spectral parameters need by a Mandarin TTS system by using some frame-level features, and the value generated from the first part. All outputs of the second hidden layer are fed back to the input of itself. Besides, the output spectral parameters are also fed back to the input of output layer. By this arrangement, the spectral parameter generator becomes a dynamic system to be able to predict these time-varying spectral parameters of real speech. The input features used in the second part include the initial indicator $II(F_j)$ and the final indicator $FI(F_j)$ of the processing frame F_j. It is used to indicate the position of current frame where it is in the processing syllable. This arragement is expected to simulate the transition effect intra the processing syllable and the coarticulation effect inter the processing syllable. In the output layer, the sigmoid function is replaced by a linear combination function. The spectral parameters which include 12-order LSP coefficients and 1 normalized energy contour are directly obtained from the output layer.

3. SIMULATION

The performance of this RNN-based spectral parameter generation is examined by simulation. A speech database provided by Telecommunication Laboratories(TL) was used in our simulation. The database contains 655 sentential and paragraphic utterances which contain the phonetic-balance sentence and the sentential text from newspaper. All utterances were generated by a single male speaker. The database was divided into two parts: a training set and an outside test set. These two sets consist of 31730 and 7832 syllables, respectively. Speech signal was sampled at 10k Hz and segmented into 10ms frames. Then, the syllable segmentation was first done manually by observing the speech waveform and with help of hearing. The 12-order LSP parameters and 1 normalized energy contour of each frame were calculated and taken as target features. These two sets consist of 590359 and 146360 frames, respectively. The input features with syllable-level and frame-level which are stated in previous section were extracted from context and syllable duration, respectively. The error back propogation(EBP) algorithm is employed to train the RNN. Over 100 training epoches were used to converged approximately during the training pro-

cess. The RNN contains two parts which operate with two different clocks. The first part receives syllable-level linguistic features and operates with syllable-synchronized clock. The second part receives frame-level linguistic features and the output of the first part. It operates with frame-synchronized clock and produces the 12-order LSP and energy contours. Fig.2 shows the original and synthesized LSP contours of an sentence for outside test. The solid lines are the original LSP contours and the dash lines are the synthesized LSP contours. It can be found that the synthesized LSP contours match to the original LSP contours well. Fig.3 shows the original and synthesized energy contours with outside test. The synthesized energy contour also matchs to the original energy contour well. Table 1 lists the mean, standard deviation, and root mean square error(RMSE) of the 12-order LSP as well as 1 normalized energy contour. All the RMSEs of the inside and outside test are small. It is proved that the relations between linguistic features and spectral parameters were automatically learned. An LPC-based Mandarin TTS system was employed to test the performance of our spectral model. In this system, an RNN-based prosodic information generation which is proposed in our previous study[8, 9] was also employed. The RNN-based prosodic information and spectral information generation combined to integrate a more compact model. The synthetic speech sounds fluent and natural. However, due to the LPC-based synthesizer, the clarity of synthesized speech must be improved in the future.

4. CONCLUSION

A RNN-based spectral parameter generation for Mandarin TTS is proposed. The phoneme-to-spectral parameter rule and the coarticulation rule are automatically learned and memorized into the weight of RNN. The coarticulation effect between each two adjacent synthesis units is alleviated by this approach. The RNN-based spectral model can be taken as a mechanism of spectral parameter generator. It can automatically generate spectral parameter from the input linguistic features. A more compact TTS system with RNN-based spectral and prosodic parameters generator can be achieved. However, the clarity of synthetic speech can be improved by using another spectral parameter as target values.

Table 1. The mean, standard deviation, and RMSE of the 12-order LSP as well as one normalized energy parameters.

	LSP-1	LSP-2	LSP-3	LSP-4	LSP-5	LSP-6	LSP-7
Mean	0.1800	0.3440	0.6183	0.9440	1.1187	1.3946	1.6646
Standard Deviation	0.0830	0.1255	0.1625	0.1975	0.2085	0.1535	0.1221
RMSE(Inside)	0.0452	0.0551	0.0784	0.0753	0.0738	0.0780	0.0697
RMSE(Outside)	0.0467	0.0564	0.0847	0.0822	0.0774	0.0858	0.0780

	LSP-8	LSP-9	LSP-10	LSP-11	LSP-12	Energy	
Mean	1.8312	2.0570	2.2518	2.4875	2.6987	0.8704	
Standard Deviation	0.1257	0.1289	0.1165	0.1303	0.1288	0.1117	
RMSE(Inside)	0.0634	0.0700	0.0635	0.0766	0.0840	0.0583	
RMSE(Outside)	0.0694	0.0755	0.0671	0.0823	0.0913	0.0604	

5. REFERENCES

[1] J. P. Olive "Rule Synthesis of Speeh from Diadic Units," ICASSP, pp.568-570, 1977.

[2] I. Mikuni, and K. Ohta, "Phoneme based text-to-speech synthesis system" ICASSP, pp.2435-2437, 1986.

[3] Nakajima, S. & Hamada, H. "Automatic generation of synthesis units based on context oriented clustering," ICASSP, Vol.1, pp.659-662, 1988.

[4] W. Wang, W. Campbell, N. Iwahashi and Y. Sagasaka, "Tree-Based Unit Selection for English Speech Synthesis," ICASSP, Vol.2, pp.191-194, 1993.

[5] T. Keiichi, K. Takao, and I. Satoshi, "Speech Parameter Generation from HMM Using Dynamic Features," ICASSP, vol.1, pp.660-663, 1995.

[6] F. Toshiaki, T. Keiichi, K. Takao, and I. Satoshi, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," ICASSP, vol.1, pp.137-140, 1992.

[7] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai "Speech Synthesis Using HMMs with Dynamic Features," ICASSP, Vol.1, pp.389-392, 1996.

[8] S. H. Chen and S. H. Hwang "A Prosodic Model of Mandarin Speech and Its Application to Pitch Level Generation for TTS," ICASSP, Vol.2, pp.45-48, 1995.

[9] S. H. Chen, S. H. Hwang, and Y. R. Wang "An RNN-based Prosodic Information Generation for Mandarin TTS,", to appear on IEEE Trans. on Speech and Audio Signal Processing.

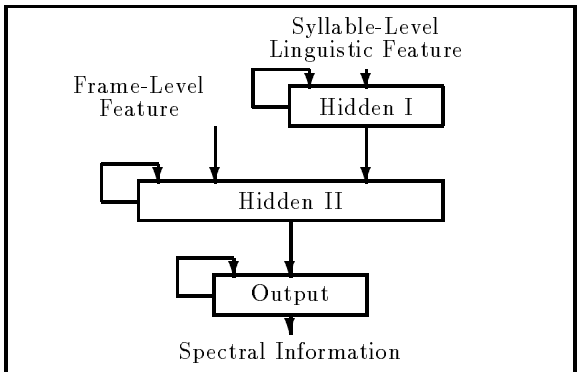

Fig.1 The block diagram of RNN-based spectral information generation.