Trainable Prosodic Model for Mandarin TTS System *

Tao Jianhua Cai Lianhong Zhao Shixia
Department of Computer Science and Technology
Tsinghua University, Beijing 100084

ABSTRACT

Mandarin prosody is characterized by its hierarchical structure when influenced by linguistic environments. Based on this, a neural network, with specially weighted factors and optimizing outputs, is described and applied to construct the Mandarin prosodic model in TTS system. Extensive tests show that the structure of the neural network characterizes the Mandarin prosody more exactly than traditional models. Learning rate is speeded up and computational precision is improved, which makes the whole prosodic model more efficient. Furthermore, the paper also stylizes the Mandarin syllable pitch contours with SPiS parameters (Syllable Pitch Stylized Parameters), and analyzes them in adjusting the syllable pitch. It shows that the SPiS parameters effectively characterize the Mandarin syllable pitch contours, and facilitate the establishment of the network model and the prosodic controlling.

KEYWORDS Prosodic Model, Neural Network (NN), SPiS Parameters, TTS System

1. INTRODUCTION

With the development of the technology in speech processing, the Mandarin speech synthesis system (TTS system) has been made rapid progress during last few years, and has been used in various places successfully. But, the results of it is still far away from the high naturalness compared with humans, being lack of the good algorithm of prosodic processing. In recent years, with the increasing power of modern computers, there has been a growing interest on artificial neural network in prosodic processing. Some attempts have been made on the research of it for some western language, (O.Karaali etc. 1997) [2]. They resulted in noticeably better synthetic speech than the traditional rule-based approach. Not only the performance was improved but the system could also be easily configured for different persons by learning from existing database automatically. But, as for Mandarin, it is still in primary status to process the prosody with the neural network. As we know, Mandarin is a tonal language. Much different to other western languages, it has its own method in the description of the prosodic features and in the prosodic processing.

Chengxinhong (1998) [3] proposed an RNN-based model to generate prosodic parameters for Mandarin TTS system. The RNN was employed to learn the relations between the linguistic features and the prosodic parameters. It was divided into two parts. The first part explored the prosodic phrase structure of the synthetic speech, and the second part generated prosodic information in syllabic level. It represented the characters of the Mandarin prosody, and delivered relatively good synthetic speech quality. Furthermore, Hongyan [7], Zutingshao, Xujiepin, have also done some studies in different parts of the Mandarin prosody with neural network, such as, the prediction of syllable duration, the generation of pitch contours of the di- syllables, and the classification of stress

* Supported by 863 national high technology project and national natural scientific fund (69875008).

types. Though their works are limited in some special region of the Mandarin prosody, their experiences are also valuable.

In this paper, a neural network with specially weighted factors and optimizing outputs, is described and applied to construct the Mandarin prosodic model in TTS system. Extensive tests show that the structure of the neural network characterizes the Mandarin prosody more exactly than traditional models. Learning rate is speeded up and computational precision is improved, which makes the whole prosodic model more efficient. The paper consists of five main items. In section 2, the situations that the Mandarin prosody influenced by the linguistic environments are analyzed, in order to get the linguistic features for the input of neural network. In section 3, the SPiS parameters are developed. The SPiS parameters are much suitable to be used in neural network and can be got from the speech easily. The paper also offers the algorithm on how the SPiS parameters are used in adjusting the pitch contours of the syllable. In section 4, a neural network with specially weighted factors is brought forward. The architecture and the learning algorithm of it are detailed. In section 5, A scattering and reunion method is adopted in the optimizing of the output with Gauss function. Each neuron in output layer is divided into ten parts. The precision of the outputs is improved more by this way. In section 6, some testing results and analysis of them are described.

2. LIGUISTIC PARAMETERS FOR PROSODY

Normally, Mandarin prosodic model can be described as following,

$$\vec{P} = F(\vec{A}_1, \vec{A}_2, \vec{A}_3) \qquad [1]$$

Where, $\vec{P}$ denotes the prosodic features, $\vec{A}_1$, $\vec{A}_2$ and $\vec{A}_3$ denote the linguistic parameters in different levels respectively. Thus, All of these come to three main basic problems, which are mining the most influential linguistic parameters to prosodic features, developing an algorithm to describe the prosodic features perfectly, and finding out a way to construct the model for mapping the parameters.

As for the technology of neural network, it is more important to find out the substantial linguistic parameters, since the performance and the learning rate of the neural network are influenced by them deeply. In this paper, 17 linguistic parameters are selected. In terms of the syntactic structure of the text and the habits for the pronunciation, they are classified into 5 groups, which are the current syllabic parameters $\vec{C}$ (includes the initial type c_1, the final type c_2, tone type c_3, the position in the current word c_4, the tight degree between the preceding syllable c_5 and the tight degree between the next syllable c_6), the preceding syllabic parameters $\vec{L}$ (includes the final type l_1 and the tone type l_2), the next syllabic parameters $\vec{N}$ (includes the initial type n_1 and the tone type n_2), the phrasal parameters $\vec{W}$ (includes the number of syllables in group w_1, the position in the sentence w_2, the stress degree w_3, the distance between the previous stress w_4 and the distance between the next stress w_6) and sentence parameters $\vec{S}$ (includes the sentence type S_1 and the number of phrases in the sentence S_2).

In current syllabic parameters, the tight degree denotes the different silence level between the adjacent syllables. According to the syntactic structure, it's normalized into 4 levels in the paper, that is syllable boundary, word boundary, phrase boundary and sentence boundary. As a rule, the selected

linguistic parameters are classified into of 3 levels, which are sentence level, phrase level and syllable level. The parameters in sentence and phrasal levels usually determine the tendency of the prosody and stress modification of the whole sentence, while the others are mainly reflect the coarticulation of the prosody between the adjacent syllables.

3. SYLLABLE PITCH STYLIZED PARAMETERS

Some work has been done, and some algorithms and formulas have been developed for the description of the Mandarin pitch contours in various ways (Yangsunan, 1990 [4], Fujisakii, 1984 [5], Xuyi, 1999 [6] and Chumin), but it is still hard to get the perfect pitch parameters for speech synthesis system. Moreover, the parameters, have been used, are difficult to be drawn from the data directly, and are inconvenient for the neural network. Here, the SPiS parameters (with six pitch parameters) are brought out to meet this problem. With these parameters, the syllabic pitch contours in continuos speech are normalized and are denoted intuitively. Most of all, they can be much helpful for the design of the topology of the neural network. Being easily drawn from the speech, it also saves much time during the training period.

Mainly, the figures of the Mandarin syllabic pitch can be divided into three groups, that are the flat shape, the curving shape and the conjunct shape, and they also consist of three parts, head, core and tail. In Fig.1, it shows the SPiS parameters, $\vec{P} = (B, H, N_1, N_2, F, E)$. Here, B and H denote the minimum value and the maximum value of the pitch respectively. F and E are the initial and the final values of the pitch. N_1 and N_2 are the positions of the minimum value and the maximum value.

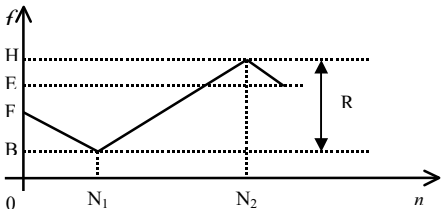

Fig.1, The features of the syllabic tone

Through the regulation of the speech by the SPiS parameters, various features of the mandarin prosody can be simulated, such as weakness, emphasis, slowness and speediness. In the process of prosodic generation for TTS system, the pitch parameters of the synthetic speech are generated by the combination of the SPiS parameters and the speech units in speech library. In this way, the pitch contours of the speech units are modified by the SPiS parameters to form the final pitch contours of the synthetic speech. The advantage is the method preserves the detailed pitch information of the speech.

The algorithm of generating the pitch contours with SPiS parameters is described as follows,

ALGORITHM 1:

A) The SPiS parameters used for generating the pitch contours are defined as, $\hat{B}$, $\hat{H}$, $\hat{N}_1$, $\hat{N}_2$, $\hat{F}$ and $\hat{E}$. And the pitch contour of the speech unit is assumed as $p(t)$, which is normalized into B, H, N_1, N_2, F and E.

B) With the condition $N_1 < N_2$ assumed, the following functions can be got,

The functions of the slope modification in pitch contour $p(t)$:

$$\lambda_1 = \frac{N_1}{\hat{N}_1},\ \lambda_2 = \frac{|N_2 - N_1|}{|\hat{N}_2 - \hat{N}_1|} \text{ and } \lambda_3 = \frac{|T - N_2|}{|T - \hat{N}_2|}.$$

The functions of the pitch range modification of syllable:

$$\eta_{BH} = \frac{|\hat{H} - \hat{B}|}{|H - B|}, \quad \eta_F = \frac{|\hat{F} - \hat{B}|}{|\eta_{BH}(F - B)|} \quad \text{and} \quad \eta_E = \frac{|\hat{E} - \hat{B}|}{|\eta_{BH}(E - B)|}.$$

C) By combining the regulation of the pitch range and the slope of the pitch contour, the following function is got, namely,

$$p'(t) = \begin{cases} \eta_{BH}[p(\lambda_1 t) - B] + \hat{B} & t \in (0, N_1) \\ \eta_{BH}[p(\lambda_2 t) - B] + \hat{B} & t \in (N_1, N_2) \\ \eta_{BH}[p(\lambda_3 t) - B] + \hat{B} & t \in (N_2, T) \end{cases}$$

D) Thus, the final pitch contour for the synthesis system will be got by formula [2], with the additional action on the regulation of the initial and the final value of the pitch contour,

$$\hat{p}(t) = \begin{cases} \dfrac{\eta_F (N_1 - t)[p'(t) - \hat{B}]}{N_1} + \hat{B} & t \in (0, N_1) \\ p'(t) & t \in (N_1, N_2) \\ \dfrac{\eta_E (t - N_2)[p'(t) - \hat{B}]}{T - N_2} + \hat{B} & t \in (N_2, T) \end{cases} \qquad [2]$$

The result coming from [2] is based on the condition $N_1 < N_2$. As for $N_1 > N_2$, the synthetic pitch contours can be got in similar way. But, it must be noted that the step C should be changed into the whole transition of the pitch contour, if the syllable is marked as tone 1.

4. THE ARCHITECTURE OF ARTIFICIAL NEURAL NETWORK AND THE TRAINING ALGORITHM

Accordingly, the input parameters of the neural network are linguistic parameters $\vec{X} = (\vec{C}, \vec{L}, \vec{N}, \vec{W}, \vec{S})$, described above, and output parameters are prosodic controlling parameters, including both the SPiS parameters $\vec{P}$ and the syllabic duration *L*. Thus, the whole output parameters of the neural network are defined in $\vec{Y} = (\vec{P}, L)$. Fig 2 shows the topology of the neural network, from which it is divided into three main layers, input layer (linguistic labeling layer), output layer (prosodic controlling layer) and hidden layer. The whole structure is represented by the formula,

$$\vec{Y} = F(\vec{X}) \qquad [3]$$

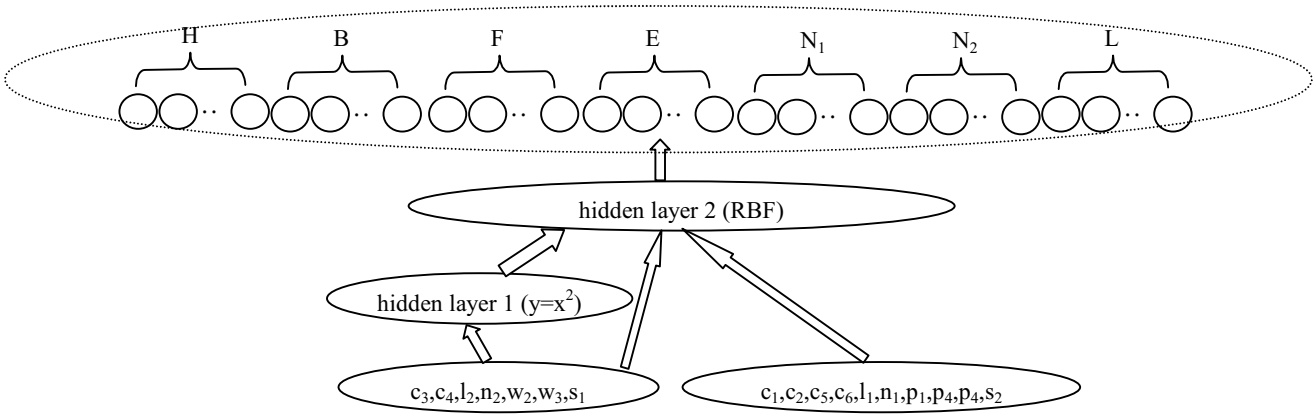

Fig.2, Topology of the neuron network

Being special in prosodic processing of Mandarin, the traditional neural network is difficult to converge at an ideal status, and also be trained slowly, if all of the linguistic parameters are identically treated. In the paper, the input parameters are divided into two groups with respect to the different actions they imposed on prosody, which are,

$\vec{X}_1 = (c_3, c_4, l_2, n_2, w_2, w_3, s_1)$, $\vec{X}_2 = (c_1, c_2, c_5, c_6, l_1, n_1, w_1, w_4, w_5, s_5)$ 。

Where, $\vec{X}_1$ mainly arises the global trend of the prosodic transition in sentence, and $\vec{X}_2$ implies the information of the coarticulation between the adjacent syllables. The studying results of the linguistics have proved that the stress of the syllables and the prosodic transition trend of sentence are more important than the other parameters in Mandarin. Based on this, a particular weighted hidden layer (named hidden layer 1) is imposed on the neural network in order to stand out the action of $\vec{X}_1$. It is inserted between the input layer $\vec{X}_1$ and hidden layer 2. The activation function of it is,

$$y = x^2$$

The learning rate is improved about 18%, compared to the normal back-propagation neural network, by this way. In application, to simplify the architecture of the neural network, the hidden layer 1 is usually mapped into the input layer. The network will be,

$$\vec{Y} = F(\vec{X}_1^2, \vec{X}_1, \vec{X}_2) \qquad [4]$$

Correspondingly, the output of the neural network is influenced by three vectors, $\vec{X}_1$, $\vec{X}_2$ and $\vec{X}_1^2$.

In order to enhance the fault tolerant of the neural network, a Radial-Base-Function (RBF) is introduced for the activation function of the hidden layer 2. The function is expressed as,

$$x_i^{(2)} = f\left[I_i^{(2)}\right] = \left[\sqrt{2\pi}\sigma_i^{(2)}\right]^{-1} e^{-I_i^{(2)}} \qquad [5]$$

where $\sigma_i^2 > 0$, $x_i^{(2)}$ is the output of the neuron i, $\sigma_i^{(2)}$ is the threshold, and $I_i^{(2)}$ is the corresponding input vector, that is

$$I_i^{(2)} = \varphi\left[\overrightarrow{X}^{(1)}, \overrightarrow{W}_i^{(2)}\right] = \frac{1}{2}\frac{\left\|\overrightarrow{X}^{(1)} - \overrightarrow{W}_i^{(2)}\right\|^2}{\left(\sigma_i^{(2)}\right)^2} \qquad [6]$$

here, $\overrightarrow{X}^{(1)}$ is the input patterns, acquired from training or testing data, $\overrightarrow{W}_i^{(2)}$ is the synaptic weight connecting the input patterns and the neuron i of the hidden layer 2.

As for the output layer, the activation function of it is,

$$y_i = x_i^{(3)} = f\left[I_i^{(3)}\right] = I_i^{(3)} \qquad [7]$$

where $I_i^{(3)} = \varphi\left[\overrightarrow{X}^{(2)}, \overrightarrow{W}_i^{(3)}, \theta_i^{(3)}\right] = \overrightarrow{X}^{(2)} \bullet \overrightarrow{W}_i^{(3)} + \theta_i^{(3)}$ [8]

$\overrightarrow{X}^{(2)}$ is got from the output of the hidden layer 2, $\overrightarrow{W}_i^{(3)}$ is the synaptic weight and $\theta_i^{(3)}$ is the threshold.

The whole learning algorithm of the neural network is described as follows,

ALGORITHM 2:

$\left(\hat{\vec{X}}_{(1)}^{(3)} = \hat{\vec{Y}}_{(1)}, \vec{X}_{(1)}^{(1)} = \vec{X}_{(1)}\right), \cdots, \left(\hat{\vec{X}}_{(M)}^{(3)} = \hat{\vec{Y}}_{(M)}, \vec{X}_{(M)}^{(1)} = \vec{X}_{(M)}\right)$ are assumed as the training dataset, where, *M* is the number of the total training patterns. The learning method of the neural network consists of three steps,

A) With respect to formula [4], the hidden layer 1 is mapped into input layer by adding an additional input vector $\vec{X}_1^2$.

B) The synaptic weights $\overrightarrow{W}_i^{(2)}$, between the input layer and the hidden layer 2, is trained by LBG method (unsupervised learning).

C) Finally, the remnant weights ($\vec{W}_i^{(3)}$， i = 1 ~ M，) are got by back-propagation method, in terms of making $\vec{W}_i^{(2)}$ and $\sigma_i^{(2)}$ invariable (supervised learning).

5. THE OPTIMIZATION OF THE OUTPUT

A potential problem, which exists in the application of neural network for prosodic model, is that the speakers are easily affected by some artificial factors. For the reason, the prosodic features of the continuous speech are usually unstable in mapping with the linguistic parameters. It's very harmful for the training of neural network. To make the results of output more stable, the paper made use of a statistic method, in which the output parameters are separated, trained and reunited in centroid again. The method was much useful to optimize the output of the neural network. With the method, each parameter of the output is separated into ten neurons in equal interval with the Gauss function during the training period. Furthermore, the target values of the outputs are got by reuniting these neurons in centroid (shown in Fig 3).

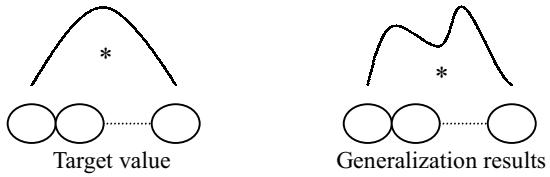

Fig.3, Each output parameter be separated into ten neuron in equal interval by Gauss method

The algorithm for the output optimization can be described as following,

ALGORITHM 3:

A) Each neuron of the output of the neural network is splitted into ten neurons in equal interval with Gauss function $f(x)=\frac{E}{\sqrt{2\pi}\sigma}e^{\frac{x^2}{2\sigma^2}}$ （Here, $\sigma=1$）.

Thus, the target value $\hat{Y}$ can be divided into following ten parameters,

$$\hat{Y}_1=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}},\quad \hat{Y}_2=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}},\quad \ldots,\quad \hat{Y}_{10}=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_{10}^2}{2}} \qquad [9]$$

Here, $\hat{Y}$ is the centroid of these ten values, that is $\hat{Y}=\sum_{i=1}^{10}\hat{Y}_i$.

B) The neural network is trained with $\hat{Y}_1,\cdots,\hat{Y}_{10}$.

C) During the generation period, $Y_1,\cdots,Y_{10}$ can be got from the output neurons of the neural network. Thus, the final results will be got by the function, $Y=\sum_{i=1}^{10}Y_i$

[10]

The optimized topology of neural network is shown in the dotted rectangle of Fig 2. Some tests show the precision of the neural network is improved about 7% with such scattering and reuniting method. The stability of the network is enhanced a lot, and the error ratio of the generation, which is caused by the randomicity of the input and output in neural network, is also furthest limited.

6. PROSODIC MODEL AND RESULTS

Based on above knowledge, the whole prosodic model is constructed as follow,

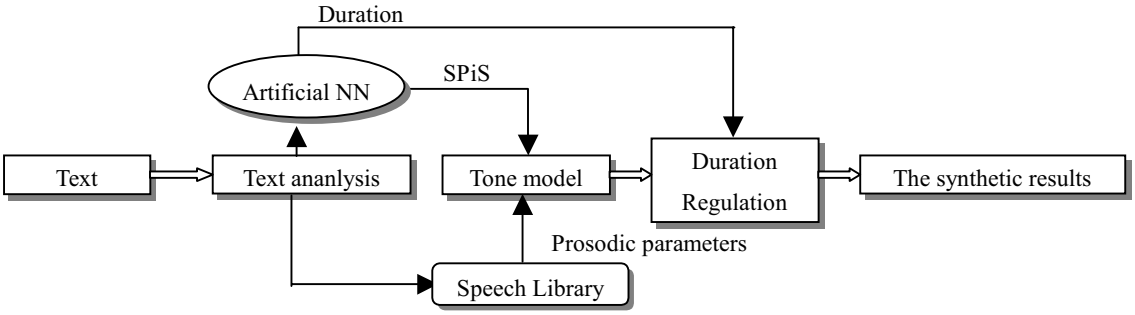

Fig.4, Whole Prosodic Model of TTS system

In the paper, 1000 sentences were selected deliberately to train and test the model. The contents of the sentences covered most of the situations appeared in Chinese, such as, sentence types, all of the phonemes and their combinations, coarticulation features in the context, all of the tone combination and various stress information, etc. Totally, 10157 syllables were contained. The sound was recorded at 16KHz-sample rate, and was processed by pitch normalizing and duration quantitative method. 75% of the database was for training, and the rest was for testing.

In application, there is strong relationship between the adjacent syllables. All of syllables in a

sentence should be treated as a whole unit for training, that make the neural network reflect the features of the Mandarin prosody more effectively. Some testing results of pitch contours and durations, and the analysis of them are offered below.

A) The testing result of pitch contours

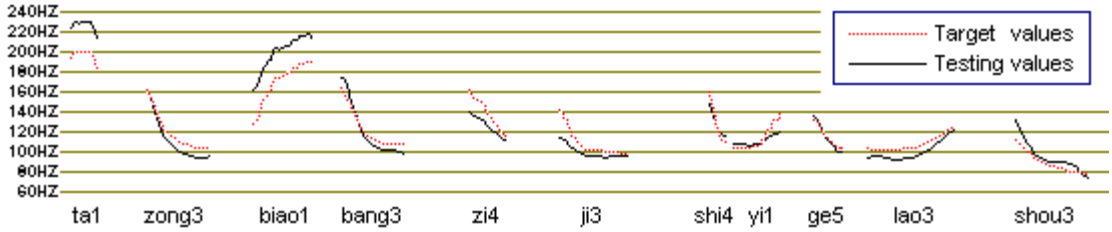

Fig.5, a testing result SPiS parameters of declarative sentence

In Fig.5, The dots denote the target values and the rectangles denote the outputs from the neural network. The results show that the output is so closed to the target value and the features of the Mandarin prosody incorporate into the neural network successfully. The pitch ranges of the syllables in both of them move decliningly in declarative sentence. And there are also the some other similar features found. For example, both the pitch of the syllable 'shi4' is lifted, being in the first place of the prosodic chunk and a long preceding silence in front of it. Moreover, the tone-sandi phenomenon is simulated by the neural network successfully. Here, the tone type of the syllable 'lao3' changes into 2, being affected by the following syllable 'shou3'.

B) The testing result of syllable durations

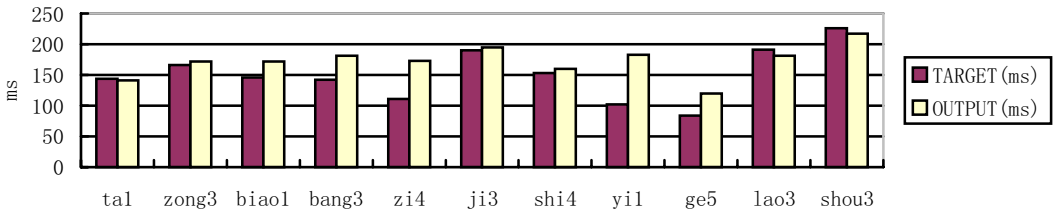

Fig.6, the duration results of a declarative sentence

The results also show that the transition tendency of the syllable durations is generated perfectly by the neural network. A statistical result of the distributing of the duration error is calculated, which shows: the 81% syllables' duration errors are below 50ms, the 14% are between 50ms to 120ms, and only the 5% exceed 120ms。If scaled by the percentage notation, that will be: the 89.8% syllables' errors are within 20% of target values, the 9% are between 20% to 50%, and only the 1.2% are over 50%. The syllable duration is one of the most important parameters in determining the stress of the Mandarin syllable, the results generated by neural network satisfy the need for the high quality prosodic controlling.

7. CONCLUSION

The traditional rule-based method of creating the Mandarin TTS system has been changed by the use of the neural network. The naturalness of the system is improved by the new way. And the TTS system becomes trainable and flexible. After the training, the system can mimic different person's

features of the prosody. Lots of listening tests prove that the architecture of the neural network and the corresponding optimizing method, developed in the paper, are much suitable to the processing of the Mandarin prosody. Now, the prosodic model has been integrated into the TTS system designed by Tsinghua University. The system produced speech, which was natural sounding, fluent and highly intelligible. Especially in the word of di-syllables, tri-syllables or qudr-syllables, it can be hard to distinguish between the synthesized sound and the recorded sound.

REFERENCE

[1] Tao Jianhua, Cai Lianhong, Zhong Yuzuo, “The Context-based Method of Creating Chinese Prosodic Model”, ISSPR’98, pp271-276

[2] O.Karaali etc., “Text-to-Speech Conversion with Neural Networks: A Recurrent TDNN Approach”. Proc. Eurospeech, 1997

[3] Sin-Horng Chen et al., “An RNN-Based Prosodic information Synthesizer for Mandarin Text-to-Speech”, IEEE Transcations on Speech and Audio Processing, VOL.6, NO.3, 1998,5.

[4] Shun-an. Yang, “ A Tonal Model For Synthesizing Polysyllabic Words and Phrases in Standard Chinese”, Essays on Linguistics, pp 65-79 (1990)

[5] H. Fujisaki and K. Hirose, “ Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese”, J. Acoust. Soc. Jpn.(E), Vol.5, No.4, pp 233-242, 1984

[6] Ching X. Xu, Yi Xu, and Li-Shi Luo, “A Pitch Target Approximation Model for F0 Contours in Mandarin”, ICPHS99, San Francisco, pp2359-2362

[7] Huang Yan, Huang Taiyi, “A Neural Learning Approach for Duration Parameter Generation in Mandarin Speech Synthesis”, ISCSLP’98, pp118-121