A NEURAL NETWORK BASED PROSODIC MODEL OF MANDARIN TTS SYSTEM

Tao Jianhua *Cai Lianhong*

Department of Computer Science and Technology, Tsinghua University,
Beijing, China, 100084
tjh@tts.cs.tsinghua.edu.cn clh-dcs@mail.tsinghua.edu.cn

Martin Holzapfel *Herbert Tropf*

ZTIK 5, Siemens AG, Otto-Hahn-Ring 6, Munich, Germany, D-81730
martin.holzapfel@mchp.siemens.de herbert.tropf@mchp.siemens.de

ABSTRACT

To generate pitch contour in high quality is a very important issue for each TTS system. Until now, the naturalness of it is still far from being satisfactory. In this paper, a trainable prosodic model, based on neural network, is described for Mandarin TTS system. Extensive tests show that the structure of the neural network characterizes the Mandarin prosody more accurately than traditional models. The naturalness of the result has been improved a lot and the system performs more flexible in practice. Furthermore, personal and task specific characteristics are also maintained.
The paper adopts a fuzzy clustering algorithm in classifying the pitch contours of the Mandarin syllables. The algorithm has been proved much useful to optimize the neural network and make it suitable to deal with the pitch contours of Mandarin.

KEYWORDS: Prosodic Model, Neural Network, Fuzzy Clustering.

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 neural network in prosodic prediction. Some attempts have been made on the research of it for some western language [2][3]. 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 styles of 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, where four lexical tones exist for syllables: namely, tone 1 characterized by a high-flat pitch contour, tone 2 characterized by a rising contour, tone 3 characterized by a low-dip contour, and tone 4 characterized by a failing contour form high F_0. Much different to other western languages, it has its own method in the description of the prosodic features and in the prosodic processing.
In this paper, a neural network is described and applied to construct the Mandarin prosodic model in TTS system. Extensive tests show that the architecture of the neural network characterizes the features of Mandarin prosody. To make the neural network more suitable to tonal language, a fuzzy clustering algorithm is adopted in the paper. With this method, the pitch contours of different syllables within continuous speech are classified into several classes. The classes got from the speech are used as parts of the neural network's outputs.
The full paper consists of five main items. In section 2, the SPiS (Syllable-Pitch-Stylized) parameters are described. The parameters are used to control the pitch contours of synthetic speech, and also used as the parts of the outputs of neural network. The parameters can be got from the speech with less time consumed. Furthermore, the paper offers the algorithm with which the pitch contours of the synthesized syllable are modified by SPiS parameters. In section 3, the paper produces 20 linguistic features, which are divided into three levels and are used as input parameters of neural network. In section 4, the architecture and the learning algorithm of the neural network are described in detail. In section 5, the training method is brought forward. With this method, learning rate is speeded up and computational precision is improved. In section 6, some testing results and analysis of them are described.
The model has been integrated into multilingual Siemens TTS-system papageno successfully. The prosodic parameters phone, duration and energy are generated from a statistical database. The basic speech unit of the system is tri-phone, and the wavelet based PSOLA algorithm is used to concatenate the speech units. The system has been proved to be able to generate the speech with high quality.

2. SYLLABLE PITCH STYLIZED PARAMETERS

As we know, the pitch contours of the syllables do a very important role in the Mandarin prosodic processing. Thus, both input parameters and output parameters are organized on syllable level. Mainly, the figures of the Mandarin syllabic pitch contours can be divided into three groups according to different syllable tone, that are the flat shape, the curving shape and the conjunct shape, and they also consist of three parts, head, core and tail. Some algorithms and formulas, such as TOBI, etc, have been developed for the description of pitch contours in various ways (Yangsunan, 1990 [5], Fujisakii, 1984 [6], Xuyi, 1999 [7]). In order to make neural network calculable and more efficient, the Syllable-Pitch-Stylized (SPiS) parameters are brought forward in the paper. With the parameters, each syllabic pitch contour in continuous speech is normalized into six parameters

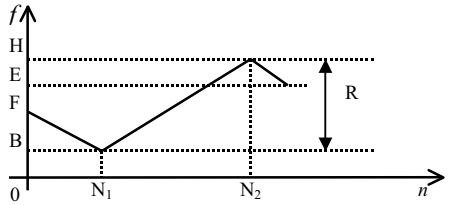

Figure 1: The features of the syllabic tone

$\vec{P} = (B, H, N_1, N_2, F, E)$, which are shown in Figure 1. Here, B and H are the minimum value and the maximum value of the pitch contour. F and E are the initial and the final pitch values. N_1 and N_2 are the positions of the minimum value and the maximum value. In prosodic model, the SPiS parameters denoted the contour intuitively and are used as outputs of the neural network.

3. LINGUISTIC PARAMETERS (INPUT PARAMETERS OF NN)

The input parameters of the prosodic model (neural network) are the results generated by the linguistic analysis module. They are chosen by the method, with which the most of the features in speech could be matched and balanced. The parameters are classified into phonology and linguistic features, which are detailed as following.

3.1 Phonology features

The phonology parameters mainly contain the tone information and stress information.

Tone Information

The tone information includes the lexical tone types of the current analyzed syllable, preceding syllable and succeeding syllable.

Stress Information

The stress degree of the current syllables and two adjacent syllables are covered in the parameters.

Duration of Speech

The duration of the syllables is also taken into account.

3.2 Linguistic features

The linguistic parameters are composed of three levels, which are syllable level, phrase level and sentence level. The syllabic information denotes the internal and external structures of the syllables, and also the relationship between the syllables.

Internal Syllabic Information

Internal syllabic information includes the initial and final types of the current analyzed syllable. Here, the initials are characterized into four classes: liquids, fricatives, vowels and nasals, and the finals are classified into five types according to four different calls in Mandarin vowel classification.

External Syllabic Information

External syllabic information includes the final type of preceding syllable and the initial type of the succeeding syllable. The types of the initials and the finals are classified in the same way mentioned above. The relational information between syllables is also described here. They are the tight degrees, which represent the mutual influence between the syllables. In the paper, both the tight degree between current syllable to preceding syllable and the tight degree between the current syllable to succeeding syllable are taken into account. The tight degrees are normalized into 3 levels in the paper according to the syntactic structure, which are syllabic boundary, phrasal boundary and sentence boundary.

Ultra-Syllabic Information

The ultra-syllabic information contains the phrasal information, sentence information and position of the syllables and phrases.

In the phrasal information, the number of syllables in the phrase, the position of the syllable in the phrase and the position of the phrase in the sentence are covered.

The sentence information contains the sentence type, and the number of phrases in the sentence.

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 syllables.

4. PITCH CONTOUR CLUSTERING

A corpus containing 2200 sentences is used for clustering to get the classification of the syllabic pitch contours. The pitch contours of the syllables are normalized into ten points, which are represented as, $f(t_0)$, $f(t_1)$,～ $f(t_9)$. Where, t_i stands for the i'th point of the pitch contour of the syllable in equal spaces. The algorithm of clustering is described as follows,

ALGORITHM 1:

A) The average of the normalized pitch values is calculated by,

$$\overline{f_j(t_i)} = \frac{1}{10}\sum_i f_j(t_i) \tag{1}$$

Assign $$f_{ji}' = f_j(t_i) - \overline{f_j(t_i)} \tag{2}$$

Here, j is the index of the syllables.

B) The relation coefficient between two syllable pitch contour can be got,

$$R_{jk} = \frac{\sum_{i=0}^{9}\left(f_{ji}' \times f_{ki}'\right)}{\sum_{i=0}^{9}\left(f_{ji}' \times f_{ji}'\right) + \sum_{i=0}^{9}\left(f_{ki}' \times f_{ki}'\right)} \tag{3}$$

C) Thus, we can draw a similar matrix ($M \times M$) for all syllables within the same tone,

$$A = \begin{bmatrix} R_{00} & R_{01} & \dots & R_{0M} \\ R_{10} & R_{11} & \dots & R_{1M} \\ \dots & \dots & \dots & \dots \\ R_{M0} & R_{M1} & \dots & R_{MM} \end{bmatrix} \tag{4}$$

Where, M is the number of syllables, which are being analyzed.

With the matrix A, the syllable pitch contours can be classified by max-tree method from.

D) For each classification, we got the mean values the pitch contours and the standard square root error in them.

$$\overline{f_i} = \frac{1}{M}\sum_j f_j(t_i)$$

$$\sigma = \frac{1}{N}\sum_{j=0}^{N-1}\sqrt{\frac{1}{10}\sum_{i=0}^{9}\left\{\left[f_j(t_i) - \overline{f_i}\right]^2\right\}}$$

E) Finally, the mean values are calculated through the values, which are inside the standard square root error.

The model can be gotten by calculating the average pitch contours of the word. Because the points, which exceed the bias of the square root error, may lead large warp in the average pitch contours, they must be dismissed from the calculating firstly, which is shown as following.
The average pitch can be gotten,

$$FR_i = \frac{1}{M}\sum_j f_j(t_i) \qquad (5)$$

Where $f_j(t_i)$ stands for pitch point of the same contour type, and $f_j(t_i) - \overline{f_i} \le a$.

The classifications of other tones are got in the same way. In the paper, 20 types of the pitch contours are selected for each tone, at last.

5. NEURAL NETWORK

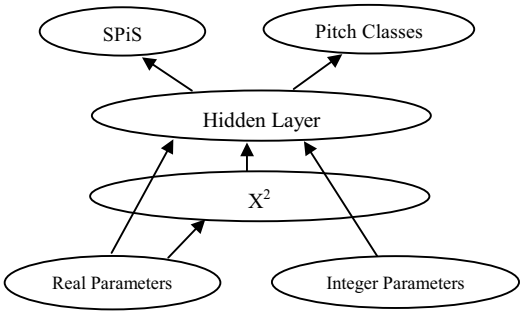

Figure 2: The Architecture of Neural Network

The architecture of neural network is shown in figure 2. Compared to a standard feed-forward network, a squaring layer is inserted into the topology, which is connected to the real valued inputs by an identity matrix. The activation function of it is,

$$y = x^2$$

The input parameters are divided into two groups with respect to the different actions they imposed on prosody, which are real parameters $\vec{X}_1$ (including 7 parameters)and integer parameters $\vec{X}_2$ (including 13 parameters). The hidden layer consists of 27 nodes with an activation function tanh(x). The network can be written as,

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

Using linear and squared input parameters easily combines two different classification properties. The direct use of the input parameters results in a linear separation of the feature space. The squared input parameters perform a weighted distance classification by radial basis functions. The first one is capturing global while the second one well localized patterns.

6. TRAINING

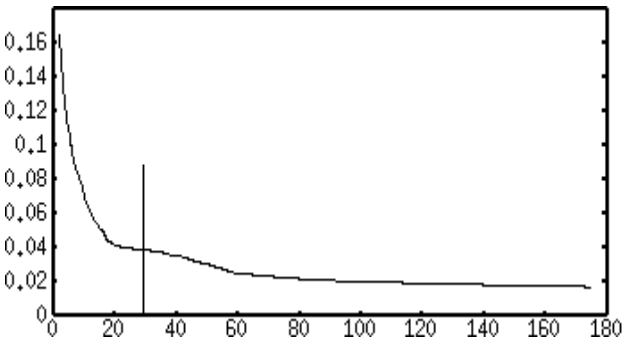

Figure 3: The training procedure

To make the weights of neural network converge quickly, at first, the full batch procedure is chosen to training the neural network. And the patterns of the inputs are trained sequentially. To avoid inconsistent evaluation of the weight, it is reasonable to perform a relatively small weight adjustment. This idea is implemented by introduction of a weight-specific factor

$$\beta_k = \frac{1}{\sqrt{\sum_{t=1}^{T}(\frac{\partial E^t}{\partial w_k} - \hat{E})^2}} \quad \text{with } \hat{E} = \tfrac{1}{T}\sum_{t=1}^{T}\frac{\partial E^t}{\partial w_k} \qquad (6)$$

Then the searching direction can be got by,

$$d^i = -\begin{bmatrix} \beta_1 & 0 & 0 \\ 0 & \cdot & 0 \\ 0 & 0 & \beta_p \end{bmatrix} \cdot \nabla E^i \qquad (7)$$

The weight modification function is $w^i = \eta \cdot d^i$. (8)

Figure 3 shows the rapid decline of error rate from the beginning. To make the neural network fall in the minimum , the training method is changed to LineSearch procedure after 30 steps. In contrast to a fixed training step length for weight modification, an optimal step length can be got by LineSearch procedure within an iterative procedure. It tries several possible values repeatedly to get the best step for training. And furthermore the Low-Memory-BFGS method is used to determine the searching direction. This idea can be taken further by approximating the error function by a quadratic polynomial and seeking the direction in which this quadratic function decreases most rapidly. From the figure 3, it shows a rapid decline in training result from 30 steps to 60 steps, and then it falls slowly again. After that, the stable weights of neural network are generated.

7. PITCH CONTOUR GENERATION

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 2:

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 $p(t)$:

$$\lambda_1 = \frac{N_1}{\hat{N}_1},\ \lambda_2 = \frac{|N_2 - N_1|}{|\hat{N}_2 - \hat{N}_1|} \text{ and } \lambda_3 = \frac{|T - N_2|}{|T - \hat{N}_2|}. \quad (9)$$

The functions of the pitch range modification of syllable:

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

$$\eta_E = \frac{|\hat{E} - \hat{B}|}{|\eta_{BH}(E - B)|}. \quad (10)$$

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

$$p'(t) = \begin{cases} \eta_{BH}[p(\lambda_1 t) - B] + \hat{B} & t \in (0, N_1) \\ \eta_{BH}[p(\lambda_2 t) - B] + \hat{B} & t \in (N_1, N_2) \\ \eta_{BH}[p(\lambda_3 t) - B] + \hat{B} & t \in (N_2, T) \end{cases} \quad (11)$$

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,

$$\hat{p}(t) = \begin{cases} \frac{\eta_F (N_1 - t)[p'(t) - \hat{B}]}{N_1} + \hat{B} & t \in (0, N_1) \\ p'(t) & t \in (N_1, N_2) \\ \frac{\eta_E (t - N_2)[p'(t) - \hat{B}]}{T - N_2} + \hat{B} & t \in (N_2, T) \end{cases} \quad (12)$$

The result coming from (12) 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.

8. TESTING AND EVALUATION

In Figure 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 declinedly 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

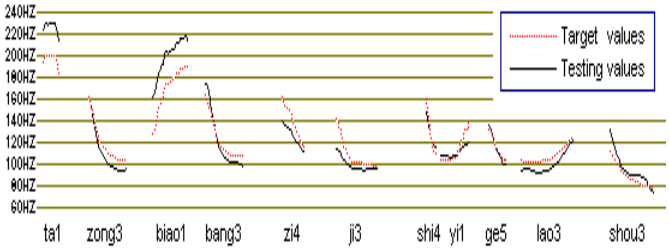

Figure 5: The f0 contours generated by the neural network

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'.

9. CONCLUSION

The notion of using a neural network to implement components in a text-to-speech system is an attractive one. A system trained on actual speech may learn subtler nuances of variation in speech than traditional rule-based or concatenation text-to-speech system can do. The paper establish a method in how to generate a neural network model to predict the pitch contours of Mandarin. The model has been integrated into the Mandarin TTS system successfully. It not only makes the system trainable and flexible, but also improves the naturalness of synthesized speech. Furthermore, System can also suit deferent styles of users.

10. REFERENCE

1. Tao Jianhua, Cai Lianhong, Zhong Yuzuo, "The Context-based Method of Creating Chinese Prosodic Model", *ISSPR'98,* pp271-276
2. R. Haury, M. Holzapfel, "Optimization of a Neural Network for Speaker and Task dependent F0-Generation", *ICASSP, 1998*
3. O.Karaali etc., "Text-to-Speech Conversion with Neural Networks: A Recurrent TDNN Approach". *Proc. Eurospeech, 1997*
4. 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.
5. Shun-an. Yang, " A Tonal Model For Synthesizing Polysyllabic Words and Phrases in Standard Chinese", *Essays on Linguistics*, pp 65-79 (1990)
6. 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
7. Ching X. Xu, Yi Xu, and Li-Shi Luo, "A Pitch Target Approximation Model for F0 Contours in Mandarin", *ICPHS99*, San Francisco, pp2359-2362