The Context-based Method of Creating Chinese Prosodic Model

Jian-Hua Tao Lian-Hong Cai
Information group, Department of Computer Science,
Tsinghua University, Beijing, 100084
FAX: 0861062772001
tjh@tts.cs.tsinghua.edu.cn clh-dcs@mail.tsinghua.edu.cn

ABSTRACT

The paper studies the linguistic characteristics of the context and moves a normalized model of word tone. We use fuzzy clustering model to classify and state the word tones. Finally, the paper describes a prosodic model, which uses a neural network to perform linguistic characteristics to speech intonation mapping. The neural network system requires less memory than a concatenation system, and performed well in tests comparing it to commercial systems using other technologies.

KEYWORDS: Prosodic Model, Neural Network, Linguistic Characteristic, Fuzzy Clustering

1. INTRODUCTION

Text-to-speech conversion has traditionally been performed either by concatenating short samples of speech or by using rule-based systems to convert a phonetic representation of speech into an acoustic representation, which is then converted into speech. Chinese is a tonal language, where four lexical tones exit for a syllable: namely, tone 1 characterized by a high-flat F0 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 F0. Prosody of spoken Chinese has two major factors, i.e. tone and intonation in sentence which is produced to express emotion. The relation of tone and intonation in the Standard Chinese had been described:" the small ripples riding on large waves" (Y.R.Chao), the large waves represents emotion changing, and the small ripples represents word tone. Each Chinese tonal model must solve word contour and analysis the tone-sandhi rules firstly. The sentence intonation can be gotten from adding word contour to emotion changing contour.

In this paper, we design a new Chinese prosodic model, in which the model is including three levels, first is word tone, second is bias of pitch and third is emotion changing [2].

Being word tone do a very important role in the Chinese prosody. We analyzed a large number of words of di-syllables and tri-syllables, and classified the words using fuzzy clustering method. From which, we can get some important rules, and create the word tone model [3].

The results of modern phonetic research show that Chinese prosody is related nearly to syllable position in words, types of accent, surroundings of syntax etc. In most synthesizers the task of generating a prosodic tune consists of two sub-tasks, the prediction of intonation labels(accents, tones, etc) from text and the generation of a pitch contour from those labels (and possibly other information). It is difficult to analyze all prosodic characteristics from them directly. This paper

Proceedings of the 1998 symposium on Image, Speech Signal processing and Robotics, 271-276, Sep. 1998, Hongkong

deals with a lot of text, makes the labeling information (both syntactical labeling and prosodic labeling) of them, and creates the relation between the labeling information of the texts and the intonation of speech [4], using Neural-Network.

Finally, based on the above discussions, we designed a high- quality prosodic model for the Chinese text-to-speech system [6], in which we use PSOLA technology.

2. TONE MODEL

The tone contours of the Standard Chinese (SC) sentences can be segmented into frames with different ranges on the basis of their semantic contents. Each unit contour can be getten from the word tone model. Thus, the intonation contour of a sentence in SC can be considerd being composed of a number of word tone-sandhi contours which correspond to individual sense groups and help to increase their intelligibility (Z.J.Wu).

The word tone in the sentence can be described as following equation,

$$F_n(t) = \log^{-1}\left(V_n + D_n(t) \times S_n(L \times t)\right) \qquad (1)$$

Where, $F_n(t)$ stands for word contours in the sentence, and $S_n(t)$ is word tone model which is gotten from the sole di-syllables or sole tri-syllables, it's a logarithmic coordinates value of the pitch contour]. L represents the duration changing of the words, it is generated by NN model, which will be detailed in section [5], and $D_n(t)$ will be changed synchronously with $S_n(t)$ according to the grammatical structures or prosodic labeling [3]. V_n is the bias of the mid-value of the word pitch contour. In this formula, V_n and $D_n(t)$ together represent the emotion changing of the speaker. By the formulation, we can get the maximum of the frequency domain of word contour $V_n + \max[D_n(t) \times S_n(t)]$, and the minimum of it $V_n + \min[D_n(t) \times S_n(t)]$.

In order to generate high-quality word contour, it needs four pre-processed input parameters: word tone of the sole di- syllables or sole tri-syllables $S_n(t)$, which will be discussed below [4], duration changing of the word L, bias of pitch V_n and emotion changing $D_n(t)$. We must have settled the above parameters firstly before generating sentence intonation.

Thus, sentence intonation can be gotten from composing a serials of word contours, which is shown as following,

$$S(t) = F_1(t) + F_2(t) + \cdots \qquad (2)$$

Where, $S(t)$ stands for "sentence intonation".

3. SENTENCE LABLING

The results of modern phonetic research show that the word contours are modified at a large range in deferent position or surroundings of the sentence. Chinese prosody characters relies on the syntactical surroundings or grammatical structure nearly, it is important to get the relationship between them. To find that, we selected 265 sentences (with both text and speech) to be analyzed, and labeled them with syntactical information and prosodic information. The mapping model between the labeling information and prosodic characters will be discussed in [5].

3.1 Syntactical Labeling

The aim of the syntactical labeling is to plot the syntactical component in the sentence's text and label them with syntactical information which will be much helpful to reflect the emotion changing of the word contours in deferent surroundings of the sentence, and to supply the training units for the neural network created for the prosodic model [5].

All of the deferent sentences can be plotted into six parts, alone the method of

Section-Sentence-Clause-Phrase-Word-Syllable. Every part is also including several attributes listed in the table 1. Among the parameters of each part, the properties of the phrases, words and syllables do the most important role in the whole sentence.

So, each syllable is correspond to a syntactical labeling vector:

$$\vec{X} = (T_1, S_1, S_2, S_3, S_4, S_5, C_1, C_2, P_1, P_2, P_3, P_4, W_1, W_2, W_3, W_4, W_5, Y_1, Y_2) \quad (3)$$

Where, every ponderance of the vector $\vec{X}$ stands for an attribute of the syntactical component.

3.2 Prosodic Labeling

Currently, speech synthesizers are also controlled by a multitude of proprietary tag sets. These tag sets vary substantially across synthesizers and are an inhibitor to the adoption of speech synthesis technology by developers. User can control the synthesizer to produce speech with the special manner by using serials tag sets inside the text. Now, there are several synthesis markup languages in the world, each of which supplies a serial of tag sets to control the synthesizer, such as SABLE, SSML, STML, JSML etc. There is also another popular tone labeling system named TOBI, which is used usually to describe the characteristics of the prosody of the speech.

Because Chinese has special characteristics in

	SECTION (T)	SENTENCE (S)	CLAUSE (C)	PHRASE (P)	WORD (W)	SYLLABLE (Y)
Labeling Information	1.Kind of the section	1. Kind of the sentence 2. Style of the text 3. Mood 4. Speed 5. Accent in sentence	1. Syntactic style in the sentence 2. Accent in the clause	1. Part of speech 2. Structure 3. The number of the syllables 4. Accent in the phrase	1. Part of speech 2. The number of the syllables 3. The function in the emotion changing 4. Accent in the word 5. The kind of accent	1. The kind of accent 2. The code of phonetic

Tab 1. Syntactic labeling information of the sentence

syllable tone, word tone, phrasal tone, intonation, and pause etc. Consulting several other languages, we developed a labeling language named AMCL ourselves, which is proved to be suitable for Chinese prosodic characteristics and can be used easily. The AMCL markup language is being developed with the following goals in mind:

- Enable markup of speech synthesis text input.
- Internationalized: appropriate to a large number of languages.
- Easy to learn and use: AMCL should not require specialized knowledge of speech synthesis, linguistics or markup languages, though users with such experience should be able to apply their knowledge. Portability: provide application developers with a consistent mechanism for controlling Synthesizers from different companies and on different platforms.
- Tools: enable the creation of tools for use and control of speech synthesis: for example, software that generates AMCL text, AMCL editing tools, pronunciation and lexicon tools, AMCL parsers and verifiers.
- Extensibility: AMCL should be able to evolve to support new features in future releases. AMCL should allow individual

synthesizers to provide enhanced features without compromising the portability of AMCL text.

The prosodic labeling vector can be described as $\vec{A}$, which is also the training data for the neural network created for the prosodic model [5]. The high-natural speech will be generated with AMCL text.

4. WORD TONE CLUSTERING

As we know, word tone plays very important role in the intonation of Standard Chinese and most important in the word tone is the tone coarticulation between two adjacent syllables.

In this paper, 641 words of sole di-syllable are selected carefully for clustering and statistic, in order to find the rules of the tone coarticulation and generate the word tone model of sole di-syllables. Word tone of tri-syllable and four-syllable can be extended from di-syllable.

The words are analyzed by speech-analysis tool, which is developed by us for calculating the speech pitch contours automatically. Ten points in the syllable pitch contour is selected equably, $t_0, t_1, \sim t_9$, at which the pitch value $f(t_i)$ can be thought to represent the contour characteristic of the pitch mainly. Assign the pitch value of previous syllable is $fr_j(t_i)$, next is $fn_j(t_i)$. Where, t_i stands for the i'th point of the pitch contour of the syllable. And j represents j'th word.

For the same tone combination of di-syllables, the average pitch contour can be calculate by,

$$\overline{f_j(t_i)} = \frac{1}{20}\left[\sum_{i=0}^{9}\left(fr_j(t_i) + fn_j(t_i)\right)\right] \qquad (4)$$

Assign: $fr_{ji}' = fr_j(t_i) - \overline{f_j(t_i)}$ (5)

We can get the relation coefficient between two word pitch contour,

$$R_{jk} = \frac{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ki}' + fn_{ji}' \times fn_{ki}'\right]}{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ji}' + fn_{ji}' \times fn_{ji}'\right] + \sum_{i=0}^{9}\left[fr_{ki}' \times fr_{ki}' + fn_{ki}' \times fn_{ki}'\right]} \qquad (6)$$

For all the words, a similar matrix ($M \times M$) can be get,

$$A = \begin{bmatrix} R_{00} & R_{01} & \cdots & R_{0M} \\ R_{10} & R_{11} & \cdots & R_{1M} \\ \cdots & \cdots & \cdots & \cdots \\ R_{M0} & R_{M1} & \cdots & R_{MM} \end{bmatrix} \qquad (7)$$

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

Thus, the word pitch contours can be classified accurately, using max-tree method from the similar matrix. The words with same tone combination will comprise a serial of deferent word tone models $s_{ij}(t)$ by clustering, Where, i stands for deferent tone combination, and j is the index of the deferent word tone model in the same tone combinaion. $s_{ij}(t)$ can be gotten by calculating the average value of the most similar pitch contour in the same tone combination. For the reason of that B curve has the perfect curve-smooth function, and can be controlled at the edge of the segment accurately and easily, it has been used in image processing widely. In this paper, it is selected to smooth the value points which is generated by tone model, and generates high-quality pitch contour from them.

The word tone model of sole tri-syllable and sole quad-syllable can be developed using the same method as above.

Because the word pitch contour in the sentence is deferent from that is in sole word. It will be influenced by complicated surroundings. Next part of the paper will develop a model to improve the quality of word pitch contour in intonation.

5. NEURAL NETWORK FOR PROSODIC MODEL

The notion of using a neural network, or other machine learning system, 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 can presently be incorporated fully into rule-based or concatenation text-to-speech system. System can also suit deferent styles of users. Though there has been some neural network models used in the TTS system now, most of them have been used to select the current synthetical units or generate a series of coder parameter vectors. They often require a large number of training data. And the training processing of them is usually finished slowly.

In this paper, a neural network with time delay in input data is developed. Which is shown as Figure 1. The system can expediently convert the linguistic and prosodic description into a series of high-natural word pitch contours, and form the high-quality intonation finally. The train databases of it are the linguistic labeling and prosodic labeling, output will be the emotion changing of the words in deferent sentences or in deferent surroundings. The system doesn't need very large training sentence, but the grammatical structure of the them must have been designed carefully. The data storage requirements in this system are also smaller than other's. It should also be easier to be trained on a new language than to determine a rule set for that language.

5.1 Output of Neural Network

The output of the NN model includes both emotion changing and duration changing of the words. If the output of the model is syllable tone, it may lead to low-natural pitch contours at the parts of tone-sandhi in the sentence. Based on this experience, we use emotion changing of the words for the minion output parameter. It can at least insure that speech will sound naturally at the part of tone coordination. We have got the deferent word tone model $S_{ij}(t)$ in the section 2. The average value of it is,

$$\bar{S} = \frac{1}{N}\sum_{n\in N} S_i(t_n) \qquad (8)$$

The pitch contours of the words in the sentence can be assigned as, $f_i(t)$, the average value of it can also be gotten,

$$\bar{F} = \frac{1}{N}\sum_{n\in N} f_i(t_n) \qquad (9)$$

Thus, the emotion changing of the words in the sentence can be described as following:

The bias of the pitch value,

$$V = \bar{F} - \bar{S} \qquad (10)$$

and,

$$D_i(t) = \frac{f_i(t) - \bar{F}}{S_i(t) - \bar{S}} \qquad (11)$$

Therefor, the output vector can be described as,

$$\vec{O} = (D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, V, L) \qquad (12)$$

D_1~D_{10} and V represent the emotion changing characteristic, L represent the duration changing of the words.

5.2 Training Data

In order to train a neural network to perform the text-to-prosody mapping, it was necessary to prepare an appropriate database. This database, consisting of most of the deferent Chinese sentence structures, and complex surroundings, was then labeled syntactically, and prosodically. Syntactic labeling vector $\vec{X}$, and prosodic labeling vector $\vec{A}$ can be gotten from the text. The neural network model actually represent the relationship between the emotion changing vector $\vec{O}$ and the $\vec{X}$, $\vec{A}$, the

mapping function can be described as,

$$\vec{O} = \Phi\left(\vec{X}, \vec{A}\right) \quad (13)$$

Because Chinese prosodic characteristic is relate to context nearly, the neural network must consider the affections of the adjacent syllables or words, and the input data must consider the adjacent labeling vector.

Let's assign the previous labeling vector as, $\overrightarrow{XR}$, $\overrightarrow{AR}$, and the next is $\overrightarrow{XN}$, $\overrightarrow{AN}$. The mapping function of the NN model can be revised as

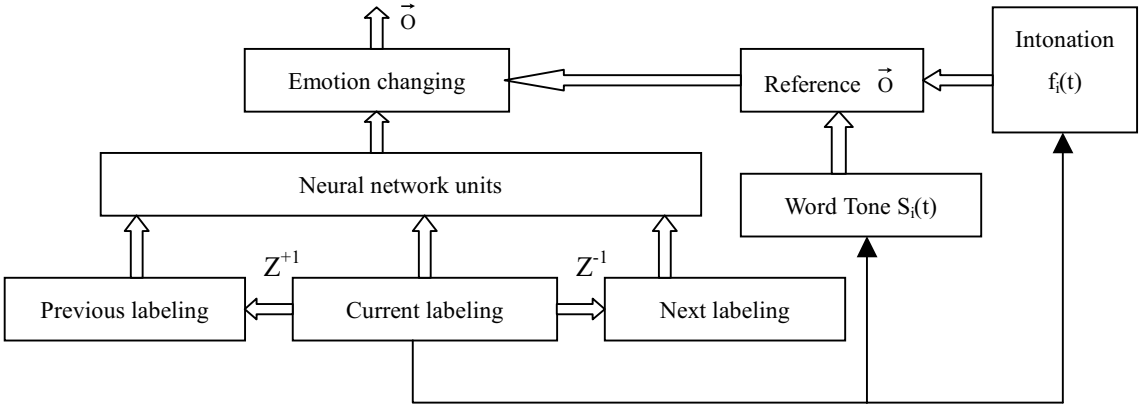

Fig 1. The model of Neural Network

following,

$$\vec{O} = \Phi\left(\vec{X}, \vec{A}, \overrightarrow{XR}, \overrightarrow{AR}, \overrightarrow{XN}, \overrightarrow{AN}\right) \quad (14)$$

5.3 Neural Network Model

Shown as Figure 1, the NN model includes three levels: Input level, hidden level and output level. The function of the unit is RBF function in hidden level, and liner function in output level. The training method includes both LBG and BP. Result show that this NN model has the perfect astringency and distribution.

6. PROSODIC MODEL

We have gotten the word tone contours with word tone model, emotion changing and duration changing with NN model above. Thus, the sentence intonation can be generated with formula 1 and 2. Total prosodic model is shown in Figure 2, which includes four subsystems, a text analysis system used to analyze the text and labeling it automatically(includes both syntactic labeling and prosodic labeling), a word tone model system used to generate word pitch-contours, a neural network used to convert the labeling information into a series of emotion changing parameters. At last the tonal model and duration adjusting part, which use formula 1 and 2, syncretize the word pitch contours with emotion changing to generate high-quality intonation pitch contours.

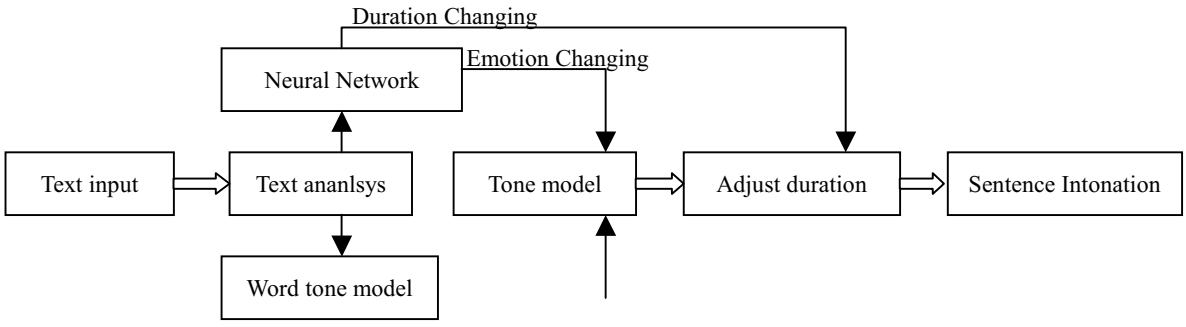

Fig 2. Prosodic Model for Synthesis system

7. CONCLUSION

Chinese prosodic characteristics are related to context nearly. Based on this theory, the paper develops a new Chinese prosodic model, which uses NN model and includes four parts. Each part of them is described detailedly. The prosodic model performs very good results in TTS system.

REFERENCE

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

[2] Z.J.Wu, " A new method of intonation analysis for Standard Chinese: frequency transposition processing of phrasal contours", Analysis, Perception and Processing of Spoken Language, pp 255-268 (1996)

[3] 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

[4] O. Karaali, G. Corrigan, and I. Gerson, " Speech Synthesis with Neural Networks", Word Congress on Neural Networks, pp 45-50 (1996)

[5] Jin-Fu. Ni etc, " Quantitative Analysis and Formulation of Tone Concatenation in Chinese F0 Contours", Europe Speech 96

[6] H.Fujisaki, et al, " Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese," ICSLP90, pp 841-844