CLUSTERING AND FEATURE LEARNING BASED F0 PREDICTION FOR CHINESE SPEECH SYNTHESIS *

(1) Tao Jianhua *(2) Cai Lianhong*

Department of Computer Science and Technology
Tsinghua University, Beijing, China, 100084
{(1)jhtao, (2)clh-dcs}@tsinghua.edu.cn

ABSTRACT

The paper describes a Chinese prosody model based on clustering and feature learning method. As the tonal language, the features of Chinese prosody are analyzed in accordance with the various context information. Focusing on the notion of prosody templates, we confirmed that a F0 pattern can be extracted based on various context parameters for each syllable. Statistic algorithm was used for template selection and training. Finally, the paper analyzes the error distribution of the F0 predicting results. Unlike other methods, the approach may give feedback as to exactly what the crucial parameters are that determine the successful choice of patterns. Both acoustic validation test and listening test show that the synthesis results are much closed to human being. And the system has been used widely in applications.

1. INTRODUCTION

With the development of the technology in speech processing, the speech synthesis system has been made rapid progress during the last few years, and has been used in various places successfully. While the synthesis quality has been highly improved, the production of a natural prosody still remains a difficult and challenging problem. Many automatic prediction methods have already been tried for this topic, including decision trees [1], neural networks [2], and HMMs [3]. They resulted in noticeably better synthetic speech than the traditional rule-based approach. As we know, Chinese is a tonal language and syllable is normally assigned as the basic prosody element in processing. Each syllable has a tone, and has a relatively steady F0 contour. And the contours will be transformed from the isolated ones while they appear in the spontaneous speech in accordance with different context information. Normally, it's a complicated problem to deal with lots of the parameters of the surroundings in sentence, which can affect the F0 contours.

This paper describes a clustering and feature learning approach which can be thought to solve the problem very well. With the method, the F0 contours of each syllable with tone are classified into several patterns, which are used to generate prosody templates. During the prosody prediction in the phase of speech synthesis, prosody cost function (PCF) is used to select the prosody template and concatenate them into the intonation. Furthermore, a statistic algorithm was generated for training the model.

* Supported by Hi-Tech Research and Development Program of China (863 program) (2001AA114072)

The full paper is organized into four main sections. In section 2, the typical Chinese prosody feature, tone and stress, are analyzed. The Syllable Pitch Stylized (SPiS) parameters are generated to describe the shape of F0 contour for each syllable. With the help of the analysis of the behaviors of F0 contours in various context, the idea of prosody template generated by clustering method. In section3, the paper establishes a model to select the prosody template for each syllable with prosody cost function, and a statistic method is generated for the automatic training. Context parameters, which are sensitive to prosody features, are also described here. In section 4, acoustic validation and listening testing results are analyzed. The results show the good naturalness of the synthesis speech.

The model has been integrated into our speech synthesis system successfully, which shows good synthesis results, meanwhile, the other prosodic parameters, duration and energy, are generated from a statistical database directly.

2. CHINESE PROSODY FEATURES AND TEMPLATE GENERATION

Tone is the most important prosody feature in Chinese. It is much complicated in how to process the intonation with tone, and how to perceive and process the stress with tone.

2.1. Syllable Pitch Stylized (SPiS) Model

There are five lexical tones exist in Chinese: 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, tone 4 characterized by a failing contour form high F_0, and neutral tone without steady F0 contours.

Studies show that tone shapes often deviate from the expected canonical one in spontaneous speech. The tonal variations unexpected tone shape is associated with weak prosodic strength, and a weak tone accommodates the shapes of neighboring strong tones. The distorted tone shapes both occur on weak syllables, and the observed distortion conforms with the neighbor's influence [4].

In our work, we use SPiS parameters to describe the shape of the syllabic F0 contours. In Figure 1, the F0 contour of the syllable is normalized into six SPiS parameters $\vec{P} = (B, H, N_1, N_2, F, E)$. They denote the minimum and the maximum point, beginning and ending point of the contour.

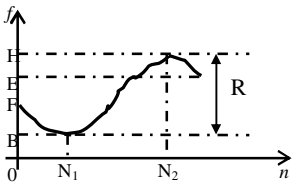
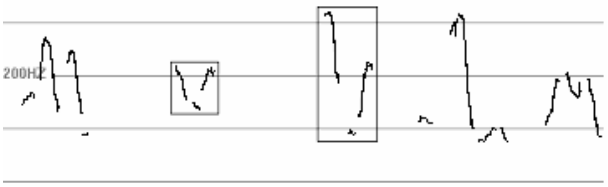


Figure 1, The features of the syllabic tone

F and E normally reflect the smoothing characters between the syllables, which are mainly influenced by phoneme characters such as initial and final types, and various reading speed. H and B denote the F0 range, which are normally related to the stress of the syllable. If tone level doesn't change too much, the shape of F0 maintains better than others.

2.2. F0 transform according to stresses

The F0 movement of stressed syllable in Chinese is complicated that it cannot be described as one line intonation model. The modification of the range is somewhat as a graph drawn on an elastic band would be magnified when stretched (Chao, 1933). The pitch range of tones can be described as top-line correlates to the stressed word while the declination of the bottom-line keeps constant to make the rhythm of the sentence, which results in the widening of the pitch range.



"ni3 gei3 wo3 guo4 lai2! gan4 shen2 me5? gan4 shen2 me5! ni3 xin1 li3 ming2 bai5 bie2 zhuang1 suan4."
(You come here! What? What! You know what I mean, don't feign deafness.)

Figure 2, F0 contour of the sentence

The F0 movement of syllable stress is realized by shifting up of the pitch with relatively constant F0 contours. Figure 2 shows that the F0 range of a syllable has been changed in accordance with the different stresses. When it is stressed, the pitch range increased. Further experiments also show that syllable F0 contours can also be influenced by the stress degree of the neighboring syllables.

With the character of relatively steady syllable F0 contours and various transformations in spontaneous speech, we found that quantifying of the stress with the different syllabic prosody template is an efficient way to solve the problem of prosody modeling for Chinese.

2.3. Generate Syllable prosody templates

With the SPiS parameters, the average of the normalized pitch values can be acquired by,

$$f_{ji}^{'} = f_j(t_i) - \overline{f_j(t_i)} \quad (1)$$

Here, j is the index of the syllables.

Generate the correlation between two normalized pitch contour R_{jk}, then we can draw a similar matrix ($M \times M$) for all syllables within the same tone,

$$A = \begin{bmatrix} R_{00} & R_{01} & \dots & R_{0M} \\ R_{10} & R_{11} & \dots & R_{1M} \\ \dots & \dots & \dots & \dots \\ R_{M0} & R_{M1} & \dots & R_{MM} \end{bmatrix} \quad (2)$$

With the max-tree method

$$FR_i = \frac{1}{M}\sum_j f_j(t_i) \quad (3)$$

Where $f_j(t_i)$ stands for pitch point of the same contour type, and $f_j(t_i) - \overline{f_i} \le \alpha$. Then mean value of F0 contours in each type is assigned as one F0 pattern. To reduce the computing cost, 20 templates of each tone are selected finally.

3. F0 PREDICTION AND TRAINING

3.1. Prosody template selection based on PCF

Research show that the modification of the syllable contours is related to the syntactic structure of the sentence and speaking surroundings. In the paper, the context information is composed into four levels, which are syllable level, prosody word level, prosody phrase level and sentence level. 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,

- **the current syllabic information**
  the initial and final types, syllabic tone, the location in the prosody word, the preceding and succeeding boundary type, the stress and the duration of the syllable.
- **the preceding syllabic information**
  the tone and final type of the preceding syllable.
- **the succeeding syllabic information**
  the tone and initial type of the succeeding syllable.
- **the prosody word level information**
  POS, the amount of syllables in prosody word, the location of the word in the prosody phrase
- **the prosody phrasal level information**
  the amount of words in group, the location of the phrase in the sentence.
- **sentence level information**
  the type of sentence and the amount of phrases inside.

The kernel idea of processing the intonation for Chinese is how to select appropriate prosody template for each syllable in accordance with the context information and concatenate them into the whole sentence. To perform the prosody selection, Prosody Cost Function (PCF) is used here, which is shown as below,

$$S_{n,m} = \sum_i \gamma_i V(a_{n,m,i}) \text{ , where, } \gamma_i = f(\omega_i) \quad (4)$$

Here $a_{n,m,i}$ means the i'th context parameter of m'th prosody template candidate in syllable n in the sentence. $a_{n,m,i}$ is an integer with non-negative value. $V(a_{n,m,i})$ denotes the similarity of the context information between the candidate template and target unit. It is normalized into 0 to 1. Here, we classify the context parameters into two kinds, grad-numerical parameters and non-grad-numerical parameters according to their numerical features. Grad-numerical parameters include stress, boundary features, location information and distant information etc, which are comparable with other parameters. However,

non-grad-numerical parameters cannot reflect the hierarchy of the parameters. They denote different classification, such as POS, initial and final types, and so on.

With the non-grad-numerical parameters, $V(a_{n,m,i})$ will be replaced as $\begin{cases} 0 & if\ a_{n,m,i} = \hat{a}_{n,i} \\ 1 & if\ a_{n,m,i} \neq \hat{a}_{n,i} \end{cases}$. And, $V(a_{n,m,i})$ will be described as $1 - \frac{|a_{n,m,i} - \hat{a}_{n,i}|}{\max_i(a_{n,m,i})}$, if $a_{n,m,i}$ belongs to grad-numerical parameters.

Here, $\hat{a}_{n,i}$ denotes the i'th context parameter of syllable n in synthesized speech.

The result of PCF is a sum result among the context parameters with a weight vector. The prosody template which makes the largest PCF result will be taken as the most appropriate prosody parameter for the syllable. The procedure is shown as following,

$$\vec{Y}_n = \arg\max_m (S_{n,m}) = \arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] \quad (5)$$

If we take the whole sentence into account, we will find function (5) cannot reflect the interaction among the prosody features. As mentioned in 2.1 and 2.2, the shape and the range of F0 contour of the syllable will be distorted conforming with the neighbor's influence. In spontaneous speech, it is influenced not only by the neighbor's tones, but also by the stresses in other syllables. Then, (5) should be changed to

$$\vec{Y}_n = \arg\max_m \left[\sum_g \sum_i [\omega_i V(a_{n,m,i}) P(\vec{Y}_{n,m} \mid \vec{Y}_{n-1,g}) P(\vec{Y}_{n-1,g})] \right] \quad (6)$$

$P(\vec{Y}_{n,m} \mid \vec{Y}_{n-1,g})$ is the transition probability from one syllabic prosody feature to another, and reflects the interaction between two prosody features.

3.2. Weights assigning and Training

The most important parameter of PCF is weight vector corresponding to the context parameters. They have the key influence to the synthesis results. Actually, the value of the weights reflects the different sensitive of the context parameters to prosody features. If the context parameter leads a rapid change in prosody features, it always needs a relatively large weight value. Experiments show that some context information, such as syllable location, prosody word boundary and prosody phrase boundary, tone, stress and POS, make the large influence for prosody features. The weights related to them are normally be assigned as large value.

Though the initial weights can be assigned manually according to the researcher's experience, further training method is still necessary to adapt the model to different speech corpus.

Suppose the initial weight vector of PCF is,

$\vec{\omega}^0 = \{\omega_1^0, \omega_2^0, \cdots \omega_p^0\}$.

The training set and synthesis outputs are $\{\hat{\vec{Y}}_1, \hat{\vec{Y}}_2, \cdots, \hat{\vec{Y}}_N\}$ and $\{\vec{Y}_1, \vec{Y}_2, \cdots, \vec{Y}_N\}$,

After time step j-1 of learning, the weight vector will be

$\vec{\omega}^j = \{\omega_1^j, \omega_2^j, \cdots \omega_p^j\}$

Here, p is the length of weight vector. Weight vector is restricted by the following condition,

$$\sum_{i=1}^{p} \omega_i^j = 1 \quad (7)$$

The condition ensures that weights will converge to steady ones, and ensures the balance of the weights in the whole vector space.

After the training at time step j, the new weight vector is acquired with,

$$\omega_i^{j+1} = \omega_i^j + \eta^j \cdot d_i^j \quad (8)$$

Here, η^j determines the rate of learning at time step j. d^j denotes the learning direction. Since all context parameters were normalized from 0 to 1, d^j can be got from,

$$d_i^j = [1 - \frac{(\Delta\vec{Y}_n)_{\min}}{\Delta\vec{Y}_n}][\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s))] + C \quad (9)$$

- $\Delta\vec{Y}_n$ is the F0 error between synthesis result to the target of the syllable n, which is
$$\Delta\vec{Y}_n = |\arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] - \hat{\vec{Y}}_n | \quad (10)$$
- $\Delta(V(a_{n,i}^s))$ is the corresponding error of context parameters between synthesis results and targets,
$\Delta(V(a_{n,i}^s)) = [V(a_{n,i}^s) - V(\hat{a}_{n,i})]$
- $\Delta(V(a_{n,i}^{\min}))$ is the error of context parameters corresponding to the candidate F0 patterns which makes minimum F0 error between synthesis results and targets.
$\Delta(V(a_{n,i}^{\min})) = [V(a_{n,i}^{\min}) - V(\hat{a}_{n,i})]$

With the condition (7), then we get,

$$\sum_{i=1}^{P} \omega_i^{j+1} = \sum_{i=1}^{P} \omega_i^j + \sum_{i=1}^{P} \eta^j \cdot d_i^j = 1 \quad \text{, thus, } \sum_{i=1}^{P} \eta^j \cdot d_i^j = 0$$

To reduce the computing cost η^j normally is assigned as a constant value, we get,

$$C = \tfrac{1}{P}[\frac{(\Delta\vec{Y}_n)_{\min}}{\Delta\vec{Y}_n} - 1] \sum_{i=1}^{P} [\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s))] \quad (11)$$

Then, the whole training procedure will be finished with the cooperation of (8), (9) and (11).

4. TESTING AND EVALUATION

Speech database used in this evaluation is the continuous male speech database of phoneme balanced 3000 Chinese sentences. All of the sentences were sampled in 22050HZ, were labeled automatically and checked manually. The speaker was asked to read the sentences in neutral mood. 2200 sentences are used for training and the rests are used for validation.

4.1. Acoustic validation test

An acoustic analysis of the corpus was carried out in parallel to listen test method mentioned below. The results of the perception experiment were analyzed under the light of this acoustical information. In acoustic validation test, all of the synthesis results are compared in F0 contours with the target one's automatically. The comparison is composed in two phases: smoothing validation and F0 range validation.

4.1.1. Smoothing validation

Smoothing validation is used to assess if the F0 is smoothed or not during the transition part of two syllables. Here, we define two new parameters to perform this test, average smoothing

bias (ASB) and average smoothing error rate (ASER), which are described below,

$$ASB = \frac{1}{2N}\sum_{n}(|F_n^T - F_n| + |E_n^T - E_n|)$$

$$ASER = \frac{ASB}{Average\ F0\ Value} \times 100\%$$

F and E are two of SPiS parameters described in Fig 1. From the results shown in table 1, the bias and error rate in the conjunction part of the speech are not high and decreased accompanying the increasing syllable amount in the sentence.

Sentence Length (syllable number)	*5*	*10*	*15*
ASB	22.1HZ	12.5HZ	8.2HZ
ASER	14.7%	8.3%	5.4%

Table 1, Results of smoothing validation

4.1.2. F0 range validation

As mention above, F0 range is a very important to stress perception. F0 range validation is used to assess how much the maximum and minimum F0 value deviates from the target ones or not. Testing results will open out if the synthesis speech sounds naturally or not. Here, we define average F0 range bias (AFRB) and average F0 range error rate (AFRER) to perform the assessment, which are described below,

$$AFRB = \frac{1}{2N}\sum_{n}(|B_n^T - B_n| + |H_n^T - H_n|)$$

$$AFRER = \frac{AFRB}{Average\ F0\ Value} \times 100\%$$

B and H are also two of SPiS parameters. From the results shown in table 2, we found the similar phenomenon mentioned in smoothing validation. It shows the low error rates during the testing.

Sentence Length (syllable number)	*5*	*10*	*15*
AFRB	10.8HZ	8.2HZ	7.8HZ
AFRER	7.2%	5.5%	5.2%

Table 2, results of F0 range validation

4.2. Listening test

Using the above results, we conducted a listening test. On the basis of the 50 utterance corpus, a dissociation experiment is performed. The aim of this experiment is to assess the naturalness of the synthesis speech in general, and the 20 listeners was asked to concentrate on the stress, rhythm and other prosody features. We thus present to the listeners pairs of stimuli constructed from one reiterant sentence. For each sentence, listeners were asked "*How natural do you think for this sentence?",* The answer should be one of the following, *"absolutely unacceptable", "unacceptable but some of them sound natural", "acceptable but some of them sound bad", "sounds good but cannot be compared to human's voice", "very good, it is much closed to human voice".* The whole corpus is presented twice randomly, and the final result is got from the average score of evaluation among the listeners. Experiment shows that the quality of synthesis voice is between the best two levels. That means the synthesis results are much closed to human voice in general, nevertheless there are few words which sound not good.

4.3. Analysis of context information

The trends in modification of syllable F0 contours are various in different sentences. The same element may have different contours, being in different position of the sentence or the phrase. In General, 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. Besides the factors, which are pointed out above, there are also some others, which can effect on the syllable contours, such as silence between the syllables, etc. Their contributions to the pitch contours are in a large range.

5. CONCLUSION

The notion of using a prosody template to implement components in a Chinese text-to-speech system is an attractive one. The 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 establishes a method in how to generate a model for predicting the F0 contours of Chinese, and integrates the model into the TTS system successfully. It not only makes the system trainable and flexible, but also improves the naturalness of synthesized speech. Now, our speech synthesis system has been used widely in China Telecom, China Mobile and other network applications.

6. REFERENCES

[1] Ross, K., Modeling of intonation for speech synthesis, PhD. Thesis, College of Engineering, Boston University, 1995.

[2] Tao Jianhua, etcl, "Trainable prosodic model for standard Chinese Text-to-Speech system", Chinese Journal of Acoustic, Vol.20, 2001, P257-265

[3] Jensen, U., Moore, R.K., Dalsgaard, P., and Lindberg, B., Modeling intonation contours at the phrase level using continuous density hidden Markov models, Computer Speech and Language, Vol. 8: 247–260, 1994.

[4] Chilin Shih and Greg P. Kochanski, "Chinese Tone Modeling with Stem-ML", ICSLP2000

[5] Andrew J. Hunt and Alan W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", ICASSP 96

[6] Fujisaki,H.et al., "Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese", ICSLP'90 Vol.2, pp841-844

[7] 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

[8] Selkirk, E. (1984) Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT press.

[9] Achim Mueller, Jianhua Tao, Ruediger Hoffmann, "Data-driven importance analysis of linguistic and phonetic information", ICSLP2000.

[10] Wu,Z.J., "Tone-sandhi in sentences in Standard Chinese", Chinese of China, No.6,pp.439-450

[11] Yang Shunan, "Synthesis technology of the Mandarin Speech", Publishing of the Social Science, 1994, 4