AUDITIVE LEARNING BASED CHINESE F0 PREDICTION

Tao Jianhua[(1)] *Ni Xing*[(2)]

[(1)]National Lab of Pattern Recognition, Chinese Academy of Science, Beijing China
[(2)]Tsinghua University, Beijing, China
[(1)]jh_tao@yahoo.com [(2)]nx01@mails.tsinghua.edu.cn

ABSTRACT

The paper describes a new F0 model based on auditive learning (AL) method. Being focused on the notion of prosody templates, we confirmed that F0 patterns for a syllable can be extracted from various anamorphosis of F0 contours in spontaneous speech. It is much suitable to use F0 templates selection method for Chinese F0 prediction with prosody cost function (PCF). Furthermore, an AL method is used to adjust the weight of PCF dynamically in application. Unlike other methods, the approach may give feedback as to exactly what are the crucial parameters determining the successful choice of patterns. The paper also analyzes the error distribution of the F0 predicting results. Both smoothing testing and F0 range testing show that the synthesis results are much closed to human being.

1. INTRODUCTION

During the last several years, the speech synthesis quality has been highly improved, and corpus based unit selection method has become the most popular method used all around, but the production of a natural prosody still remains a difficult and challenging problem. Normally, corpus based system requires a very large database which limits the usage of the system quite a lot, such as, in embedded operation systems, chips, etc. Currently, there are many algorithms tried for F0 prediction, such as decision trees [2], neural networks [3], and HMMs [4]. They resulted in noticeably better synthetic speech than the traditional rule-based approach. Nevertheless, without any help of human interposition, the automatic learning methods try to get the average prosody effect according to different training set. Human cannot revise the synthesis results in application, even they find the notable errors. The paper develops a new method for F0 prediction based on F0 template selection method with prosody cost function (PCF). The method integrates both automatic training method and auditive learning (AL) method in F0 model, which can be revised by human dynamically with the help of listening test in application.

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, but the contours will be transformed from the isolated ones while they appear in the spontaneous speech in accordance with different context information. 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. Then a clustering method is adopted to classify the F0 contours of each tonal syllable into several patterns, which are used to generate prosody templates. In the phase of F0 prediction, PCF is used to select the F0 templates and concatenate them into the intonation. The paper describes detailedly in how to assign the value for the weights of PCF. Furthermore, to make human be able to revise the synthesis results in application, an AL method was generated, which is more like a interaction procedure between the speech synthesis system and human being. If human found the synthesis errors in application, they are able to change and to find the better candidate among the F0 templates, and to make the system trained on the new changes.

The model has been tried in speech synthesis system successfully, which shows good synthesis results, meanwhile, the other prosodic parameters, duration and energy, are generated from a statistical database directly. The full paper is organized into four main sections. In section 2, the typical Chinese prosody feature, tone and stress, are analyzed. With the help of the analysis of the behaviors of F0 contours in various context, the idea of prosody template generated by a clustering method. In section3, the paper establishes a model to select the F0 template for each syllable with PCF, both automatic training method and AL method are generated here. Context parameters, which are sensitive to prosody features, are also described in this section. In section 4, acoustic validation and listening testing results are analyzed. The results show the good naturalness of the synthesis speech.

2. PROSODY TEMPLATE

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. 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. The tone shapes often deviate from the expected canonical one in spontaneous speech, such as the tonal variations unexpected tone shape is associated with weak prosodic strength, and a weak tone accommodates the shapes of neighboring strong tones [5]. The F0 movement of stressed syllable in Chinese 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 F0 movement of syllable stress is realized by

shifting up of the pitch with relatively constant F0 contours. When it is stressed, the pitch range increased.

According above analysis, the F0 template for each tonal syllable can be generated by a fuzzy clustering method, which is described detailedly in [1]. In the paper, there are 110 F0 templates generated in total. Those are, 20 for tone 1, 20 for tone 2, 20 for tone 3, 20 for tone 4 and 30 for the neutral tone.

3. F0 PREDICTION AND AUDITIVE LEARNING

3.1. Prosody cost function (PCF)

The trends in modification of syllable F0 contours are various in different sentences. Research of corpus based speech synthesis has shown that the modification of the syllable contours is related to the syntactic structure of the sentence and speaking surroundings [3]. The same element may have different contours, being in different position of the sentence or the phrase. In Chinese, the context information can be composed into four levels, which are syllable level, prosody word level, prosody phrase level and sentence level.

- **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.

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. The kernel idea of processing the intonation for Chinese is how to select appropriate F0 template for each syllable in accordance with the context information and concatenate them into the whole sentence. To perform the prosody selection, a Prosody Cost Function (PCF) is used here, which is shown as below,

$$S_{n,m} = \sum_i \gamma_i V(a_{n,m,i}) \text{, where, } \gamma_i = f(\omega_i) \quad (1)$$

Here $a_{n,m,i}$ means the context parameter i of F0 template candidate m in syllable n within the sentence. $a_{n,m,i}$ is a non-negative integer. $V(a_{n,m,i})$ 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 to 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 context parameter i of syllable n in synthesized speech.

The result of PCF is the sum among the context parameters with the weight vector. The template which makes the largest PCF result will be taken as the most appropriate F0 parameter for the syllable. The procedure is shown as following,

$$\vec{Y}_n = \arg\max_m (S_{n,m}) = \arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] \quad (2)$$

3.2. Weights assigning

Actually, different weights of PCF reflect 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, which includes syllable location, prosody word boundary and prosody phrase boundary, tone, stress and POS [10].

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.

Let's 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,

$$\sum_{i=1}^{p} \omega_i^j = 1 \quad (3)$$

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

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

$$\omega_i^{j+1} = \omega_i^j + \eta^j \cdot d_i^j \quad (4)$$

Here, η^j determines the learning rate 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,

$$d_i^j = [1 - \frac{(\Delta\vec{Y}_n)_{\min}}{\Delta\vec{Y}_n}][\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s))] + C \quad (5)$$

- $\Delta\vec{Y}_n$ is the F0 error between synthesis result and the target of the syllable n, which is
  $$\Delta\vec{Y}_n = |\arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] - \hat{\vec{Y}}_n| \quad (6)$$
- $\Delta(V(a_{n,i}^s))$ 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 of F0 templates 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 (3), 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,

$$C = \frac{1}{P}[\frac{(\Delta \vec{Y}_n)_{\min}}{\Delta \vec{Y}_n} - 1] \sum_{i=1}^{P} [\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^{s}))] \qquad (7)$$

Then, the whole automatic training procedure will be finished with the combination of (4), (5) and (7). The block diagram of the approach for prosody prediction and training is given in Figure 1.

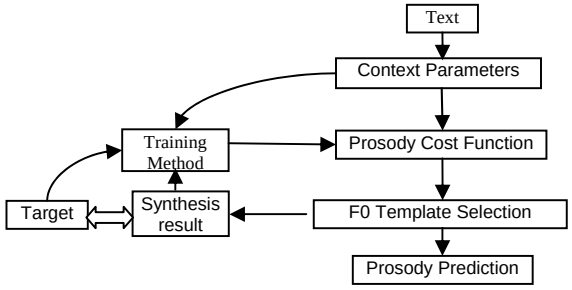

Figure 1, Diagram of F0 template selection and weight assigning

3.3. Auditive Learning

There is always some deviation between subjective listening and the automatic training results, since the apperceive can not be fully represented by acoustic simulation. How can we revise the results while we meet error in application? It will be very useful to ask the users correct the candidate of F0 templates manually and re-train the system with the new results.

All of the candidates of the prosody templates can be classified into two categories, suitable candidate and unsuitable candidate according to auditive testing. The PCF score related to them can be defined as,

$$S_\alpha^j = \sum_i \omega_i^j V(a_{\alpha,i})$$

$$S_\alpha^{j+1} = \sum_i \omega_i^{j+1} V(a_{\alpha,i})$$

$$S_\beta^j = \sum_i \omega_i^j V(a_{\beta,i})$$

$$S_\beta^{j+1} = \sum_i \omega_i^{j+1} V(a_{\beta,i})$$

Here, j means learning time step. $a \in A$, and A is the set of suitable candidates, the amount of candidates in A is $|A|$. $\beta \in B$, B is the set of unsuitable candidates. The learning procedure should ensure that weight vector related to suitable candidate makes better PCF score than that related to unsuitable candidate. Thus, learning procedure should make $S_\alpha^{j+1} > S_\alpha^j$ and $S_\beta^{j+1} < S_\beta^j$. After time step J-1, the final results should be $S_\alpha^J > S_\beta^J$.

Since there is no training data in application anymore, the target could be assigned as the center of suitable candidates. That is,

$$\vec{M} = \sum_{\alpha=1}^{|A|} \vec{Y}_\alpha$$

Then the learning direction of (4) can be replaced by

$$d_i^j = [1 - \frac{(\Delta \vec{M}_n)_{\min}}{\Delta \vec{M}_n}][\Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^{s}))] + C \qquad (8)$$

- $\Delta \vec{M}_n$ is the F0 error between the templates selected result and the center of suitable candidate $\vec{M}$, which is

$$\Delta \vec{M}_n = |\arg\max_m \left[\sum_i \gamma_i V(a_{n,m,i}) \right] - \vec{M}| \qquad (9)$$

The whole auditive learning procedure can be composed by (4), (7) and (8). It is described as following,

- Find the syllable whose synthesis result sounds unnatural.
- Change the F0 template candidates of the syllable to find a group of the suitable templates to make better F0 output.
- Calculate the center of all suitable templates.
- Re-train the model with (4), (7) and (8).

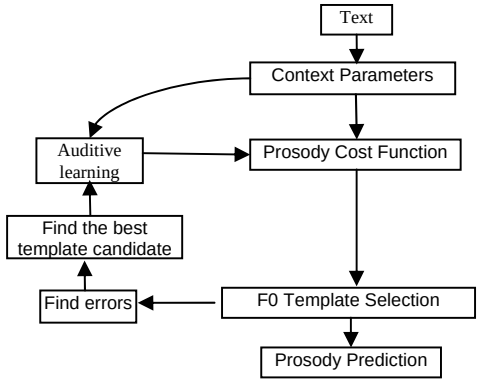

Figure 2, Diagram of AL method

4. TESTING AND EVALUATION

Speech database used in this evaluation is the continuous male speech database of 3000 phoneme balanced 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, and 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 listening 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 to F0 contours of the target automatically. The comparison is composed in two phases: smoothing testing and F0 range testing.

4.2. Smoothing testing

Smoothing testing 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,

$$ASB = \frac{1}{2N}\sum_{n}(|\log F_n^T - \log F_n| + |\log E_n^T - \log E_n|) \quad (10)$$

$$ASER = \frac{ASB}{Average\ \log F0\ Value} \times 100\% \quad (11)$$

F and E are initial and final F0 value of the syllable. From the results shown in figure 3, 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.

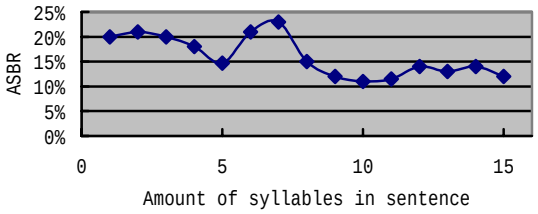

Figure 3, Results of smoothing testing

4.3. F0 range testing

As mentioned above, F0 range is very important to stress perception [12]. F0 range testing is used to assess how much the maximum and minimum F0 value deviates from the target. 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 as below,

$$AFRB = \frac{1}{2N}\sum_{n}(|\log B_n^T - \log B_n| + |\log H_n^T - \log H_n|) \quad (12)$$

$$AFRER = \frac{AFRB}{Average\ \log F0\ Value} \times 100\% \quad (13)$$

B and H are the bottom and top F0 values of the syllable. From the results shown in figure 4, a similar phenomenon can be found as smoothing testing. The deviation of F0 range shows low error rates, which comes from 7% to 18%, and becomes smaller with more syllables in the sentence. It means the longer of the sentence, the better synthesis results can be generated by AL method. It confirms opinion that there is interaction between prosody features in spontaneous speech [1].

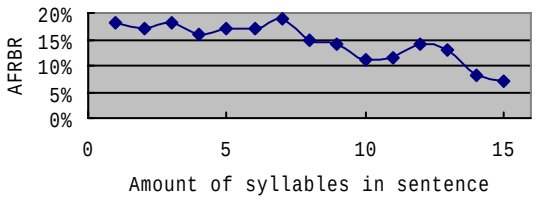

Figure 4, Results of F0 range testing

Using the above results, we also conducted a MOS 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. The final result is 4.3. It is much closed to human voice in general, nevertheless there are few words which sound unnatural.

5. CONCLUSION

The paper establishes a new method in generating a F0 model for Chinese speech synthesis with both automatic training method and AL method. The model trained on actual speech and revised manually may learn subtler nuances of variation in speech than traditional rule-based or corpus based text-to-speech system can do. It not only makes the speech synthesis system trainable and flexible, but also improves the naturalness of synthesized speech. The model has been integrated into speech synthesis system successfully, which shows good synthesis results, and has been used widely in applications.

6. REFERENCES

[1] Tao Jianhua, etc, “Clustering and feature learning based F0 prediction for Chinese speech synthesis”, ICSLP2002, Denver

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

[3] Tao Jianhua, etc, “Trainable prosodic model for standard Chinese Text-to-Speech system”, Chinese Journal of Acoustic, Vol.20, 2001, P257-265

[4] 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.

[5] Chilin Shih and Greg P. Kochanski, “Chinese Tone Modeling with Stem-ML”, ICSLP2000

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

[7] 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

[8] 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

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

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

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

[12] Tao Jianhua etc, “Automatic stress prediction of Chinese speech synthesis”, International symposium on Chinese spoken language processing, 2002, 8. Taipei