Modeling Pitch Contour of Chinese Mandarin Sentence with PENTA Model*

PANG Hui[1], WU Zhiyong[2], CAI Lianhong[3]!
[1,2,3] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

Abstract: In continuous speech, it is believed that the pitch contour of the same syllable may vary a lot due to its different context information. To apply the Parallel Encoding and Target Approximation (PENTA) model to Mandarin speech synthesis and improve its prediction accuracy, this paper proposed a method to predict pitch contours for Chinese syllables with different contexts by combining the Classification And Regression Tree (CART) with the PENTA model. We first used CART to cluster syllables' normalized pitch contours according to the context information of syllables and the distances between pitch contours. For each cluster, we calculated the average pitch contour and trained the PENTA model with this average contour. The initial pitch value is required while using PENTA model to predict a continuous pitch contour. We further proposed a Pitch Discontinuity Model (PDM) to predict such initial pitch values at positions where voiceless consonants and prosodic boundaries are found. We first conducted experiments on a Chinese four-syllable word corpus containing 2048 items and then extended experiments to a continuous speech corpus containing 5445 sentences. The results were satisfactory in terms of the Root Mean Square Error (RMSE) values comparing the predicted pitch contour with the original contour. With this method, we can model pitch contours for Mandarin sentences of any text and apply the trained model parameters into speech synthesis.

Keywords: Speech synthesis; PENTA model; Prosody analysis; Prosody modeling
Chinese Library Classification: TN912.3; TN912.33

Introduction

Natural speech not only expresses the semantic meaning of the text, but also transfers characteristics of speaker's personality, emotional state, attitudes and speaking style[1]. These "implications" are often expressed through stress, rhythm, intonation and other prosodic features. Speech prosody plays a very important role in improving the naturalness of synthetic speech. A good prosody model is rather helpful to predict more accurate prosodic information according to the text analysis result in a text-to-speech (TTS) system. This paper focused on combining the CART algorithm with the PENTA model [2, 3] so that the PENTA model can be utilized for pitch (F0) modeling and prediction in a practical Mandarin speech synthesis system.

The PENTA model encodes emotional state, attitudes, speaking style, etc. simultaneously to get pitch targets, and uses the pitch targets to estimate a continuous pitch contour [2, 3]. Hence it can play a very important role in expressive or personalized speech synthesis.

However, the PENTA model itself is indeed a curve fitting method which only takes into account the pitch contour of several adjacent Chinese syllables. In continuous speech, it is believed that the pitch contour of the same syllable may vary a lot due to its different contextual information. Hence, we need to train several PENTA models for the syllable with several different contexts, and also find a method to associate different model with different context so that the PENTA model can be used in a practical TTS system.

*Supported by the National Natural Science Foundation of China (No. 60805008, 60928005, 61003094, 60931160443) and the Ph.D. Programs Foundation of Ministry of Education of China (No. 200800031015)
Wu Zhiyong. Tel:+86-755-26036389 Email:zywu@sz.tsinghua.edu.cn

Furthermore, in the current training process of PENTA model, a continuous pitch contour is required [2, 3]. While in Chinese, the pitch contour may break into several discontinuous segments due to the occurrence of voiceless consonants and prosodic boundary pauses. We must compute the initial pitch value of each segment while using PENTA model to estimate the pitch contour.

We attempted to address the above two practical problems while using PENTA model for pitch contour modeling and prediction in a Mandarin TTS system. To solve the first problem, we utilized the CART tree to associate the PENTA model with syllable's contextual information. As for the second problem, we proposed a Pitch Discontinuity Model (PDM) to predict initial pitch values at positions where voiceless consonants and prosodic boundaries are found. Finally, we combined the PDM and PENTA model to predict the complete pitch contour of Chinese sentence in a Mandarin TTS system.

1 Training Corpus and Pre-processing

1.1 Training Corpus

Two corpora were used in this work to train and evaluate the proposed method. The first one is a *word corpus* containing 2048 common Chinese four-syllable words. The second one is a *sentence corpus* containing 5445 selected phonetically and contextually balanced Chinese sentences; these sentences have been carefully designed for a corpus based concatenative TTS system. The speech recordings of the corpora were saved in Microsoft WAV format with 16 kHz sampling rate, 16 bits per sample and double channels: the left channel is the speech waveform and the right channel is the glottal wave.

1.2 Pre-processing of Pitch Contour

1) Syllable segmentation and manual refinement

The pitch contour was derived directly from the right channel of glottal wave of the speech recordings. A further refinement was conducted by following procedures of pre-processing: (1) syllable segmentation, (2) pitch contour smoothing at syllable onset.

We segmented the sentence into syllables with a homegrown forced alignment tool with further manual refinement. We also observed F0 frequency jitter at the onset section of some syllables. We hence adjusted the pitch value of outliers to ensure the smoothness of pitch contour.

2) Pitch contour normalization

To eliminate the dependence of syllable duration while training the PENTA model, we need to normalize the original pitch contour to the same length for syllables in the corpus. For each syllable, its pitch contour was re-sampled to M pitch (F0) points at equal distance. Linear interpolation was applied if the re-sampled F0 point is located between two adjacent F0 points of the original contour.

2 Analysis of Conventional PENTA Model

The PENTA model recognizes four melodic primitives (local pitch target, pitch range, articulatory strength and duration) and treats them as both basic encoding elements for the communicative functions and control parameters for the articulatory system that generates F0 contours. The PENTA model further assumes that the articulatory system generates F0 by successively approaching syllable-synchronized local pitch targets, across specific pitch ranges, and with specific articulatory strengths [3]. The basic mathematical fitting formulas are as follows [5]:

$$f_0(t) = x(t) + \left(c_1 + c_2 t + c_3 t^2\right) \times e^{-\lambda t} \quad (1)$$

$$x(t) = mt + b \quad (2)$$

$$c_1 = f_0(0) - b \tag{3}$$

$$c_2 = f_0^{'}(0) + c_1\lambda - m \tag{4}$$

$$c_3 = \left(f_0^{''}(0) + 2c_2\lambda - c_1\lambda^2\right)/2 \tag{5}$$

Where, *m* and *b* represents the slope and height of a syllable's pitch target; λ is the approximation rate to the pitch target; $f_0^{'}(0)$ and $f_0^{''}(0)$ represent the velocity and acceleration of the initial pitch value $f_0(0)$; $c_1\ c_2\ c_3$, as the constant items, are then computed with formula (3-5).

Given pitch contour of a syllable, the parameters of the PENTA model (*m*, *b*, λ) can be estimated with curve fitting method by minimizing the Root Mean Square Error (RMSE) between the estimated and the original pitch contour. During prediction stage, an estimated pitch contour can be calculated from the same set of parameters (*m*, *b*, λ).

As can be seen, the PENTA model only cares about the pitch contour itself by ignoring the fact that pitch contour of the same syllable may vary a lot due to its different contextual information. To apply PENTA model in a TTS system, we need to find a way to incorporate both context information and PENTA model while predicting pitch contour for generating continuous synthetic speech.

3 Modeling Pitch Contour with PENTA Model and CART

In this paper, we proposed to use Classification And Regress Tree (CART) [6] to classify pitch contours of syllables in the corpus into several different clusters according to syllables' contextual information. Then for each cluster, we trained a PENTA model so that the estimated curve can approach the cluster center (i.e. the average pitch contour of the cluster). As for the initial pitch value that is required for PENTA model to predict a continuous contour, we proposed a Pitch Discontinuity Model (PDM) to predict initial pitch values at positions where voiceless consonants and prosodic boundaries are found.

3.1 CART for Context based Pitch Contour Clustering

As a kind of decision tree, Classification And Regression Tree (CART) provides very useful way to map observations about an item to conclusions about the item's target value [6]. In this work, we used CART to map syllables' context information to their different variations of pitch contours in continuous speech.

The wagon program in Edinburgh speech tools [14] is adopted to train the CART. During training, the input parameters include a distance matrix that is composed of Euclidian distances of normalized pitch contours between every two syllables, and a set of features that represent different context information of syllables. After training, similar pitch contours (i.e. with small Euclidian distances between them) are grouped into the same leaf node. For each leaf node, a path from the root to current node can be tracked; and the features on the path represent the best context information related to the current pitch contour cluster.

3.2 Context Information for CART

For the context information used in CART, we considered: (1) initial, final and tone information of current, previous and next syllables; (2) prosodic structure information.

1) Initial, final and tone information

We considered the following context information related to initial, final and tone of current, previous and next syllable:

- *p.final, p.fType*: the name and the articulation way of the final of previous syllable. The finals (*p.final*) and the final types (*p.fType*) used in this paper are summarized in Table 1, where *F_NONPY* indicates that there is no

previous syllable. The finals are categorized into 6 types according to the manner of articulation of the finals.

- *n.initial, n.iType*: the name and the articulation way of the initial of next syllable. The initials (*n.initial*) and the initial types (*n.iType*) used in this paper are summarized in Table 2, where *I_NONPY* indicates there is no next syllable and *I_ZERO* represents that next syllable has zero initial (i.e. the Pinyin starts with final directly). The initials are categorized into 10 types according to the manner of articulation of the initials.
- *tone, p.tone, n.tone*: the tone type of current, previous and next syllable. Five standard Mandarin tone types are used including *H* (tone 1), *R* (tone 2), *L* (tone 3), *F* (tone 4) and *N* (tone 0, neutral tone). *X* denotes no previous or next syllable.

Table 1. Final classification according to articulation way

Final Type	Finals	Manner of Articulation
F_NS	*ian in iang ing iong uan uen uang ueng ong an en ang eng van vn m n ng*	*Final ends with Nasal*
F_OP	*a ia ua o io uo ao iao ou iou e E er ie ve*	*Final ends with Open*
F_PT	*v*	*Final ends with Protruded*
F_RO	*u*	*Final ends with Round*
F_ST	*i -i -I ai uai uei ei*	*Final ends with Stretched*
F_NO	*F_NONPY*	*No previous syllable*

Table 2. Initial classification according to articulation way

Initial Type	Initials	Manner of Articulation
I_AA	*c, ch, q*	*Aspirated affricate*
I_AU	*z, zh, j*	*Unaspirated affricate*
I_FN	*f, s, sh, x, h*	*Voiceless fricative*
I_FV	*r*	*Voiced fricative*
I_LR	*l*	*Lateral*
I_NS	*m, n*	*Nasal*
I_PA	*p, t, k*	*Aspirated plosive*
I_PU	*b, d, g*	*Unaspirated plosive*
I_ZR	*I_ZERO*	*Initial is NULL (Zero initial)*
I_NO	*I_NONPY*	*No next syllable*

2) Prosodic structure information

According to prosodic hierarchical structure, the prosodic boundary may occur at the level of single syllable (B0), prosodic word (B1), prosodic phrase (B2), and intonation phrase group (B3, corresponding to the end of a sentence) [10].

We further took into account the context information that is related to the position at different prosodic levels:

- *pbound, nbound*: the boundary type before and after current syllable, which takes 4 values (*B_SYL, B_PWD, B_PPH, B_UTT*) representing syllable boundary, prosodic word boundary, prosodic phrase boundary and the sentence (utterance) boundary.
- *sylTpwdF, sylTpphF, sylTuttF*: the position of the current syllable in prosodic word, prosodic phrase and sentence, which takes 4 values (*S, T, H, M*) representing single syllable, at the end (of word, phrase or sentence), at the beginning (of word, phrase or sentence), and in the middle (of word, phrase or sentence).
- *sylTpwdP, sylTpphP, sylTuttP*: the relative position of the current syllable in prosodic word, prosodic phrase and sentence. Here the relative position means the percentage value of the position of current syllable over the whole length of prosodic word, prosodic phrase or sentence, which takes a continuous float value in [0, 1].
- *pwdTpphF, pwdTpphP, pwdTuttF, pwdTuttP*: the position and the relative position of the current prosodic word in prosodic phrase and sentence. The values of these features are the same as those of syllable's position.
- *pphTuttF, pphTuttP*: the position and the relative position of the current prosodic phrase in sentence, which also take the same values as those of syllable's position.

3.3 PENTA Model with Context Information

After training the CART, similar pitch contours are clustered into the same leaf node; while different variations of pitch contours are grouped into different leaf nodes.

Hence, each node of the CART represents a cluster of pitch contours. For each leaf node (i.e. each cluster),

we computed the average pitch contour of this cluster. The PENTA model was then trained for this average pitch contour to get model parameters (m, b, and λ) related to current cluster.

Moreover, the context information for each leaf node of the trained CART can be retrieved by tracing the path from root to current leaf node.

In this way, CART maps syllables' context information to their different variations of pitch contours, and further to different PENTA models represented by parameters (m, b, and λ).

3.4 Pitch Discontinuous Model

As shown in Equations (3-5), the initial pitch value $f_0(0)$ is required while using PENTA model to predict a continuous pitch contour.

We also proposed to use CART to predict the initial value of pitch contour at positions where voiceless consonants and prosodic boundaries are found. We called this the Pitch Discontinuity Model (PDM).

For training the PDM, the original initial pitch values at positions where voiceless consonants and prosodic boundaries are extracted from the corpus, together with the context information of the syllables to which the initial pitch values are attached. The CART is then trained to find the relationship between the context information and the initial pitch value in the regression manner. Such trained CART (i.e the PDM) is then utilized in predicting the initial pitch value given the context information of a syllable.

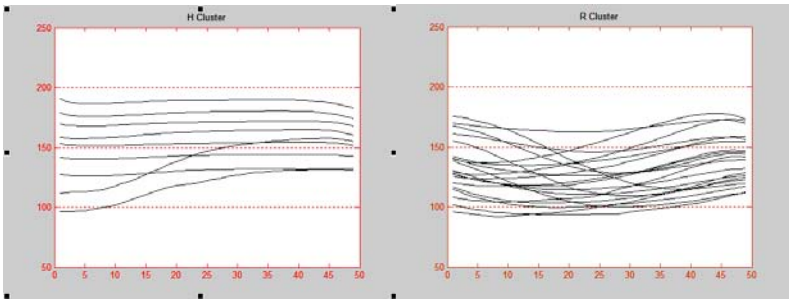
4 Experiments

In this paper, we conducted experiments with a four-syllable word corpus containing 2048 items and a sentence corpus containing 5445 items including: PENTA training experiment, pitch discontinuity modeling experiment, and pitch contour prediction experiment.

4.1 PENTA Training Experiment with CART

We first conducted the experiment to validate the efficiency of the PENTA model considering context information with CART.

The normalized pitch contours in each corpus were classified into four classes according to syllable's tone type: *H*, *R*, *L* and *F*. For *word corpus*, each class contains 2049, 2419, 1675 and 2049 items respectively. And for *sentence corpus*, each class contains 20777, 26947, 19689 and 31284 items respectively. The pitch contours in each class were further grouped into different clusters respectively with *k-means* algorithm. For *word corpus*, the number of clusters is 9, 20, 10 and 20 for tone type *H*, *R*, *L* and *F* respectively; while for *sentence corpus*, the number of clusters is 200, 150, 200 and 150 for tone type *H*, *R*, *L* and *F* respectively. Figure 1 depicts the *average pitch contour* of the four-syllable *word corpus* for the clustering result for each tone type.



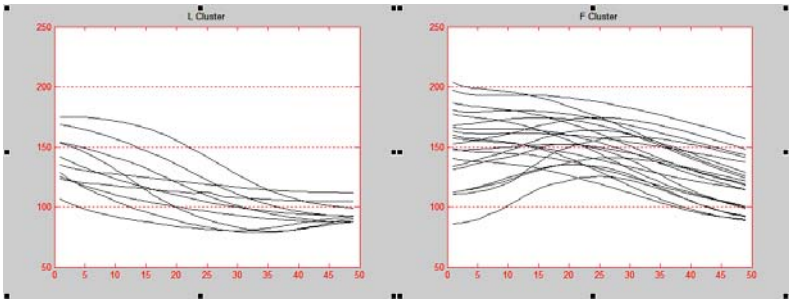


Figure 1. Pitch contour clustering result for *H R L F* tone in four-syllable corpus.

The PENTA models were then trained using the *average pitch contour* data for each cluster *considering the context information*. For prediction, the model parameters *predicted by CART tree*, the original initial pitch value and syllable duration were used to predict the pitch contour. RMSE values were then calculated between the predicted and original pitch contour. Results are shown in Figure 2. The left histogram in Figure 2 is the RMSE distribution of the four-syllable *word corpus* and the right histogram is the distribution of the *sentence corpus*. The results show that most of the RMSE is below 50Hz, which indicates that our method can achieve good performance of pitch contour prediction with PENTA considering context information.

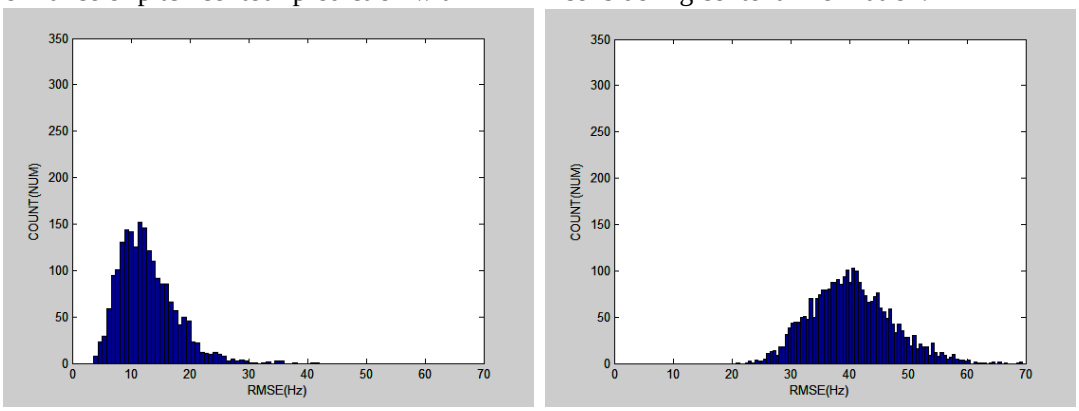


Figure 2. RMSE distribution between original pitch contours and predicted pitch contour with CART. The left histogram is for the four-syllable word corpus. The right one is for the sentence corpus.

4.2 Pitch Discontinuity Modeling Experiment

CART tree was used to train the pitch discontinuity model. In the four-syllable *word corpus*, 5121 items with either voiceless consonants or prosodic boundaries were selected as the data set; and in the *sentence corpus*, 85120 items were selected. Regression function of CART was used to predict the initial pitch value for each pitch contour. Table 3 shows the RMSE and correlation between the predicted and original initial pitch value for both corpora. The results indicate that our PDM can achieve good performance in predicting the initial pitch value from context information in terms of RMSE and correlation coefficients for both *word corpus* and *sentence corpus*. The result for sentence corpus is worse than that for word corpus. This might be due to the performance degradation of the model at the positions of prosodic phrase boundaries in the sentence corpus where large pitch reset values might occur.

Table 3. RMSE and correlation in two corpus

Corpus Name	RMSE	Correlation
Four-syllable word corpus	11.35	0.91
Sentence corpus	49.19	0.52

4.3 Pitch Contour Prediction Experiment

To evaluate the performance of the proposed method of combing PENTA model with CART and the PDM model, we conducted another experiment to predict pitch contour for text-to-speech (TTS) synthesis. In this experiment, the parameters of the PENTA model were predicted from the context information of a syllable with CART. The initial pitch value was also predicted from the context information with PDM model. While the

original syllable duration was used directly.

We first conducted the experiments in the training set, where the speech recordings of the whole corpus are used. We computed the RMSE between the original and the predicted pitch contours. The results of the RMSE distribution for both *word corpus* and *sentence corpus* are shown as follows in Figure 3. Most of RMSE values are below 40 Hz and centered at around 20 Hz for the four-syllable word corpus (shown in the left histogram in Figure 3); while for the *sentence corpus*, most of RMSE values are below 60 Hz and centered at around 50Hz (shown in the right histogram in Figure 3). Again, the performance of the method in the sentence corpus is a bit worse than in the word corpus.

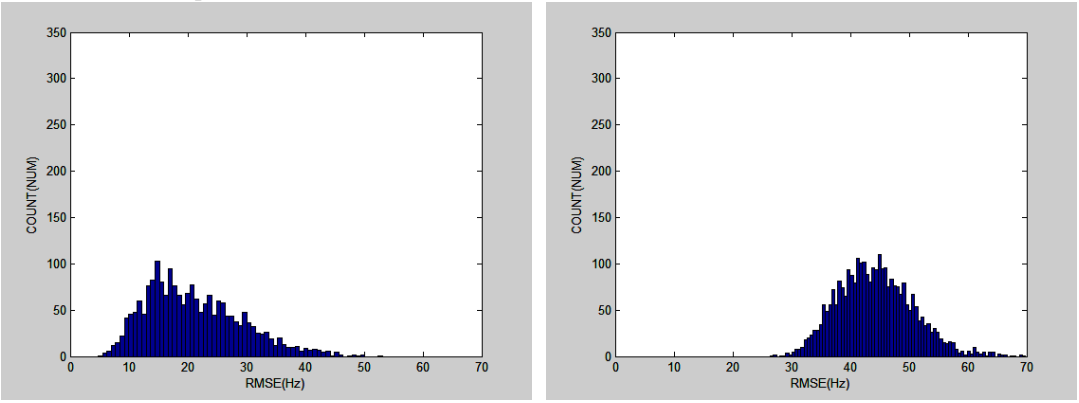


Figure 3. RMSE Distribution between the original and the predicted pitch contour with CART and PDM. The left histogram is for the four-syllable word corpus and the right one is for the sentence corpus.

We then conducted the experiments in two new test sets, where the speech recordings are set aside from the training set for training the CART, PENTA and PDM. The first test set contains 100 four-syllable words and the second test set contains 100 sentences.

For the four-syllable word corpus, an example is shown in Figure 4, which depicts the original pitch contour (red dot curve) and the predicted contour (blue + curve) of a Chinese four-syllable word “wu2 yuan1 wu2 chou2 (无怨无仇)”. The voiceless consonant “ch” in syllable “chou2” leads to the discontinuity of pitch contour. RMSE value between the original and predicted pitch contour is 6.15Hz. The average RMSE value of the whole test set of 100 four-syllable words is 10.52 Hz.

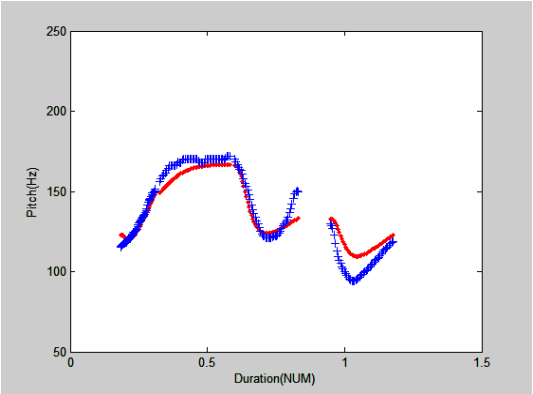


Figure 4. An example of pitch contour prediction result comparing the original (red dot curve) and the predicted (blue + curve) pitch contour. The PENTA model parameters and the initial pitch values are predicted by CART and PDM respectively from context information.

As for the sentence corpus, an example is shown in Figure 5. The sentence is “li2 wei3 piao2 jiao4 lian4 mei3 tian1 wei4 wo3 men2 shang4 xia4 gang4 ling2 pian4 ji3 shi2 wan4 gong1 jin1 (/李伟朴|教练/每天|为我们|上下|杠铃片/几十万|公斤/)” with 20 syllables, where “|” indicates the boundary of prosodic word, and “/” indicates the boundary of prosodic phrase. The average RMSE value of the whole test set of 100 sentences is 40.12 Hz, which indicates that our proposed method can accurately predict the pitch contour of sentence by combining the CART algorithm with the PENTA model. In this way, the PENTA model can be utilized for pitch modeling and prediction in a practical Mandarin TTS system.

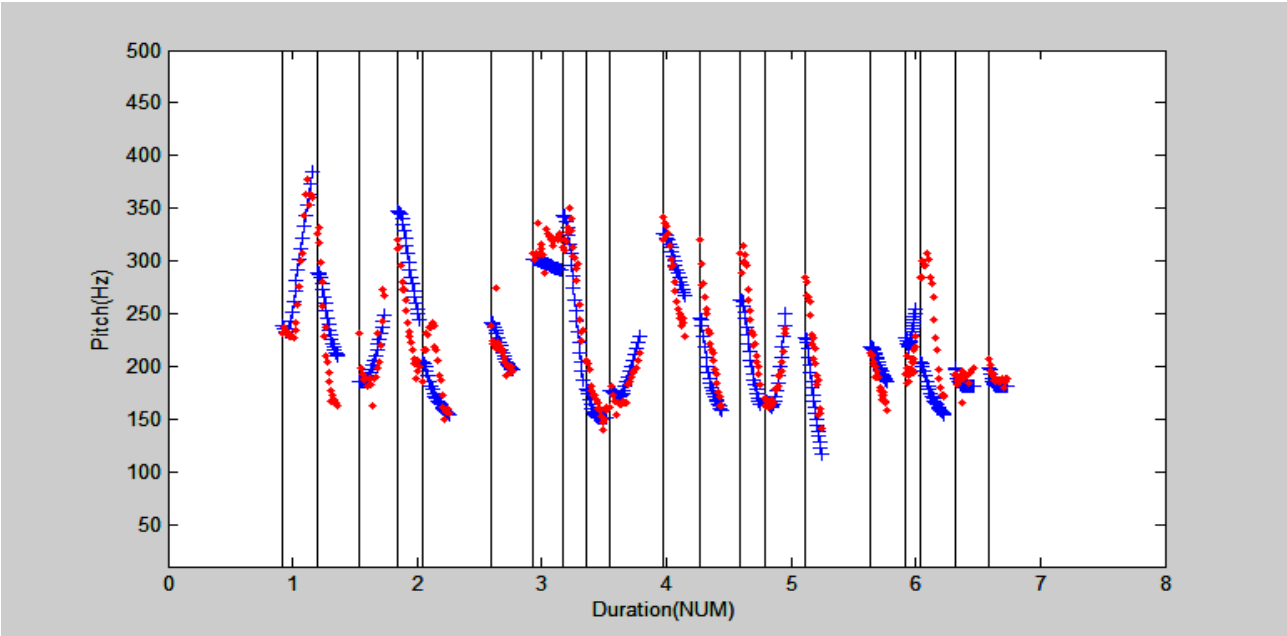


Figure 5. An example of pitch contour prediction result of a sentence comparing the original (red dot curve) and the predicted (blue + curve) pitch contour. The black vertical line represents the start position of the syllables.

5 Conclusion

This paper proposed a method to predict pitch contours for Chinese syllables with different context information using the Parallel ENcoding and Target Approximation (PENTA) model and the Classification And Regression Tree (CART).

In continuous speech, it is believed that the pitch contour of the same syllable may vary a lot due to its different context information. We used CART tree to cluster syllables' normalized pitch contours according to the context information of syllables and the distances between pitch contours. For each cluster, we calculated the average pitch contour and trained the PENTA model with this average contour. With this method, different set of parameters of the PENTA model could be finely tuned according different context information of syllables, leading to better performance in predicting pitch contours for continuous speech.

The initial pitch value is required while using PENTA model to predict pitch contour. To deal with this, we further proposed a Pitch Discontinuity Model (PDM) to predict such initial pitch value at positions where voiceless consonants and prosodic boundaries are found. The PDM is also based on CART by considering context information.

Experiments on a Chinese four-syllable word corpus and a Chinese sentence corpus for building TTS system indicate that this method can achieve good performance in predicting pitch contours by taking into account the context information with PENTA and CART.

In the future, we will make further research to model the syllable duration by statistical method with corpus. Only getting the duration and pitch information, can we get a complete prosody model of a sentence and apply it to speech synthesis.

References

[1] H Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations [A].the 4th FASE Symposium on Acoustics and Speech [C].1981.

[2] Xu Y. Speech melody as articulatorily implemented communicative functions [J]. Speech Communication. 2005. vol 46: p. 220-251.

[3] Prom-on S, Xu Y, Thipakorn B. Quantitative target approximation model: simulating underlying mechanisms of tones and intonations [A]. Acoustics, Speech and Signal. 2006

[4] Xu Ching X, Xu Y, Luo Li-Shi. A pitch target approximation model for F0 contours in Mandarin [A]. 1999.

[5] Prom-On S, Xu Y, Thipakorn B. Modeling tone and intonation in Mandarin and English as a process of target approximation [J]. The Journal of the Acoustical Society of America. 2009. 125: p. 405.

[6] The Centre for Speech Technology Research. Classification and Regression Trees [EB]. http:// www.cstr.ed.ac.uk/projects/speech_tools/

[7] Fujisaki,H, Ohno S, Gu W. Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their F0 contours [A], Proc. TAL: with emphasis on tone languages. 2004: p. 61-64.

[8] Zemlin WR. Speech and hearing science anatomy and physiology [J]. Otology & Neurotology, 1982. 4(2): p. 186

[9] Liu Tao, Cai Lianhong. The clustering research based on fundamental frequency of the syllable [J]. Journal of Chinese Computer Systems, 2004. 25(007): p. 1145-1150.

[10] Shen Weijun, Lin Fuzong, Li Jianmin, et al. A CART-based Prosodic Phrasing Method for Chinese Text-to-Speech [J]. Computer Sciences, 2002(04):p. 50-52.

[11] Xu Y. Contextual tonal variations in Mandarin [J]. Phonetics. 1997: p. 61--83.

[12] Feng Yongqiang, Chu Min, He Lin, et al. The statistical analysis of Chinese syllable duration [A], Modern Phonetics of New Century. Fifth National Conference on Modern Phonetics Proceedings. 2001.

[13] Hu Weixiang, Xu Bo, Huang Taiyi. The Acoustic Research about Prosodic Boundary [J]. Journal of Chinese Information. 2002. 16(001): p. 43-48.

[14] http://www.cstr.ed.ac.uk/projects/speech_tools/