TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/12 pp218-224
Volume 17, Number 2, April 2012

Modeling Pitch Contour of Chinese Mandarin Sentences with the PENTA Model*

Hui Pang, Zhiyong Wu**, Lianhong Cai

Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

Abstract: In continuous speech, the pitch contour of the same syllable may vary much due to its contextual information. The Parallel Encoding and Target Approximation (PENTA) model is applied here to Mandarin speech synthesis with a method to predict pitch contours for Chinese syllables with different contexts by combining the Classification And Regression Tree (CART) with the PENTA model to improve its prediction accuracy. CART was first used to cluster the syllables' normalized pitch contours according to the syllables contextual information and the distances between pitch contours. The average pitch contour was used to train the PENTA model with the average contour for each cluster. The initial pitch is required with the PENTA model to predict a continuous pitch contour. A Pitch Discontinuity Model (PDM) was used to predict the initial pitches at positions with voiceless consonants and prosodic boundaries. Initial tests on a Chinese four-syllable word corpus containing 2048 words were extended to tests with a continuous speech corpus containing 5445 sentences. The results are satisfactory in terms of the Root Mean Square Error (RMSE) comparing the predicted pitch contour with the original contour. This method can model pitch contours for Mandarin sentences with any text for speech synthesis.

Key words: speech synthesis; PENTA model; prosody analysis; prosody modeling

Introduction

Natural speech not only expresses the semantic meaning of the text, but also transfers characteristics of the 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 will predict more accurate prosodic information according to textual analysis results in a Text-To-Speech (TTS) system. This paper focuses on combining the Classification And Regression Tree (CART) algorithm with the Parallel Encoding and Target Approximation (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 the emotional state, attitudes, speaking style, etc. simultaneously to get pitch targets, and uses the pitch targets to estimate a continuous pitch contour. Hence, the model plays a very important role in expressive or personalized speech synthesis[2,3].

However, the PENTA model itself is a curve fitting method which only takes into account the pitch contours of several adjacent Chinese syllables[4]. In

Received: 2011-06-30; revised: 2011-09-29

* Supported by the National Natural Science Foundation of China (Nos. 60805008, 60928005, and 61003094) and the Ph.D. Programs Foundation of the Ministry of Education of China (No. 200800031015)

** To whom correspondence should be addressed.
E-mail: zywu@sz.tsinghua.edu.cn

continuous speech, the pitch contour of the same syllable will vary due to the contextual information. Hence, several PENTA models must be trained for each syllable in different contexts and a method is needed to associate different models with different contexts so that the PENTA model can be used in practical TTS systems.

Furthermore, current PENTA model training needs a continuous pitch contour[2,3]. However in Chinese, the pitch contour may break into several discontinuous segments due to the occurrence of voiceless consonants and prosodic boundary pauses. Thus, the initial pitch of each segment is needed when using the PENTA model to estimate the pitch contour[5].

This paper addresses these two problems when using the PENTA model for pitch contour modeling and prediction in Mandarin TTS systems. For the first problem, the CART tree is used to associate the PENTA model with the syllable's contextual information. For the second problem, a Pitch Discontinuity Model (PDM) was designed to predict the initial pitches at positions with voiceless consonants and prosodic boundaries. The PDM and PENTA model were then combined to predict the complete pitch contour of Chinese sentences in Mandarin TTS systems.

1 Training Corpus and Pre-processing

1.1 Training corpus

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

1.2 Pitch contour pre-processing

1.2.1 Syllable segmentation and manual refinement

The pitch contour was derived directly from the right channel of the glottal wave of the speech recordings. The result was further refined by pre-processing with syllable segmentation and pitch contour smoothing at the syllable onset.

The sentence was segmented into syllables with an inhouse forced alignment tool with further manual refinement. F0 frequency jitter was observed at the onset of some syllables, hence, the pitch of outliers was adjusted to ensure pitch contour smoothness.

1.2.2 Pitch contour normalization

The dependence of the PENTA model training on the syllable duration was eliminated by normalizing the original pitch contour to the same length for all syllables in the corpus. The pitch contours of each syllable were re-sampled to M pitch (F0) points at equal distances. Linear interpolation was used if the re-sampled F0 point was located between two adjacent F0 points of the original contour.

2 Analysis of the 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 the F0 contours[6]. 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. The basic mathematical fitting formulas[5] are:

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

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

$$c_1 = f_0(0) - b \quad (3)$$

$$c_2 = f_0'(0) + c_1\lambda - m \quad (4)$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \quad (5)$$

Where m and b represent the slope and height of a syllable pitch target, λ is the approximation rate to the pitch target, and $f_0'(0)$ and $f_0''(0)$ represent the velocity and acceleration of the initial pitch value $f_0(0)$. c_1, c_2 and c_3 are constants computed from Eqs. (3)-(5).

Given the syllable pitch contour, the PENTA model parameters, m, b and λ can be estimated using the curve fitting method by minimizing the Root Mean Square Error (RMSE) between the estimated and the original pitch contours. During the prediction stage, the estimated pitch contours is then calculated using these

parameters.

The PENTA model only considers the pitch contour itself while ignoring the fact that the pitch contour of the same syllable may vary in different contexts. Thus, the PENTA model in a TTS system needs a way to incorporate context information with the PENTA model to predict pitch contours for continuous synthetic speech.

3 Modeling Pitch Contours with the PENTA Model and CART

A CART[7] was used to classify the pitch contours of syllables in the corpus into several different clusters according to the syllable contextual information. Then, a PENTA model was used for each cluster so that the estimated curve approaches the cluster center (i.e., the average cluster pitch contour). The initial pitch required for the PENTA model to predict a continuous contour was given by a PDM that predicts the initial pitches at positions with voiceless consonants and prosodic boundaries.

3.1 CART for context based pitch contour clustering

The CART is a decision tree that provides a very useful way to map observations about an item to conclusions about the item's target value. In this work, CART is used to map the syllable contextual information to their different pitch contour variations in continuous speech.

The wagon program in the Edinburgh speech tools[8] was used to train the CART. The training input parameters include a distance matrix composed of the Euclidian distances of the normalized pitch contours between each pair of syllables and a set of features representing the different syllable contextual information. After training, similar pitch contours (i.e., with small Euclidian distances) were grouped into the same leaf node. A path from the root to the current node was tracked for each leaf node with the features on the path representing the best contextual information related to the current pitch contour cluster.

3.2 CART contextual information

The CART contextual information used the initial, final and tone information of the current, previous and next syllables and the prosodic structure information[9].

3.2.1 Initial, final and tone information

The contextual information related to initial, final and tone of the current, previous and next syllables was described by:

(1) p.final, p.fType: the name and articulation of the final of the 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 articulation of the finals;

(2) n.initial, n.iType: the name and articulation of the initial of the 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 the next syllable has zero initial (i.e., the Pinyin starts directly with the final). The initials are categorized into 10 types according to the articulation of the initials;

(3) tone, p.tone, n.tone: the tone types of the current, previous and next syllables. 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 articulation classification

Final type	Finals	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 articulation classification

Initial type	Initials	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

3.2.2 Prosodic structural information

According to the prosodic hierarchical structure, the prosodic boundary may occur at the level of a single syllable (B0), a prosodic word (B1), a prosodic phrase (B2), or an intonation phrase group (B3, corresponding to the end of a sentence)[10,11].

The contextual information related to the position at different prosodic levels is taken into account using:

(1) pbound, nbound: the boundary type before and after the current syllable, which takes 4 values (B_SYL, B_PWD, B_PPH, B_UTT) representing the syllable boundary, prosodic word boundary, prosodic phrase boundary or the sentence (utterance) boundary;

(2) sylTpwdF, sylTpphF, sylTuttF: the position of the current syllable in the prosodic word, prosodic phrase or sentence, which takes 4 values (*S*, *T*, *H*, *M*) representing a single syllable, at the end (of a word, phrase or sentence), at the beginning (of a word, phrase or sentence), or in the middle (of a word, phrase or sentence);

(3) sylTpwdP, sylTpphP, sylTuttP: the relative position of the current syllable in the prosodic word, prosodic phrase or sentence. Here, the relative position means the percentage of the position of the current syllable over the whole length of the prosodic word, prosodic phrase or sentence, which is a continuous floating value in [0, 1];

(4) pwdTpphF, pwdTpphP, pwdTuttF, pwdTuttP: the position and the relative position of the current prosodic word in the prosodic phrase or sentence. The values of these features are the same as those for the syllable position;

(5) pphTuttF, pphTuttP: the position and the relative position of the current prosodic phrase in the sentence, which also takes the same values as those of the syllable position.

3.3 PENTA model with contextual information

After training the CART, similar pitch contours are clustered into the same leaf nodes, while different variations of pitch contours are grouped into different leaf nodes[12]. Hence, each node of the CART represents a cluster of pitch contours. The average pitch contour of the cluster is calculated for each leaf node (i.e., each cluster). The PENTA model was then trained with this average pitch contour to get the model parameters (m, b, and λ) related to the current cluster. The contextual information for each leaf node of the trained CART can be retrieved by tracing the path from the root to the current leaf node.

In this way, CART maps the syllable contextual information to their different pitch contours variations and then to different PENTA models represented by the parameters m, b, and λ.

3.4 Discontinuous pitch model

As shown in Eqs. (3)-(5), the initial pitch $f_0(0)$ is required by the PENTA model to predict a continuous pitch contour. The CART is then used to predict the initial pitch at voiceless consonants and prosodic boundaries in the PDM.

The PDM is trained using the original initial pitches at voiceless consonants and prosodic boundaries extracted from the corpus, together with the contextual information of the syllables requiring the initial pitches. The CART is then trained to find the relationship between the contextual information and the initial pitch using linear regression. The trained CART (i.e., the PDM) is then used to predict the initial pitch given the syllable contextual information.

4 Tests

Tests were conducted with a four-syllable word corpus containing 2048 words and a sentence corpus containing 5445 sentences in a PENTA training experiment, a pitch discontinuity modeling experiment, and a pitch contour prediction experiment.

4.1 PENTA training test with CART

Tests were conducted to validate the efficiency of the PENTA model with the contextual information with CART.

The normalized pitch contours in each corpus were classified into four classes according to syllable tone type: *H*, *R*, *L* and *F*. For the *word corpus,* the classes contained 2049, 2419, 1675 and 2049 syllables. For the *sentence corpus*, the classes contained 20 777, 26 947, 19 689 and 31 284 syllables. The pitch contours in each class were further grouped into different clusters using the *k*-means algorithm. For the *word corpus*, the four tone types had 9, 20, 10 and 20 clusters. For the *sentence corpus*, the four tone types had 200, 150, 200 and 150 clusters. Figure 1 shows the *average pitch contours* of the four-syllable *word corpus* for the clustering result for each tone type.

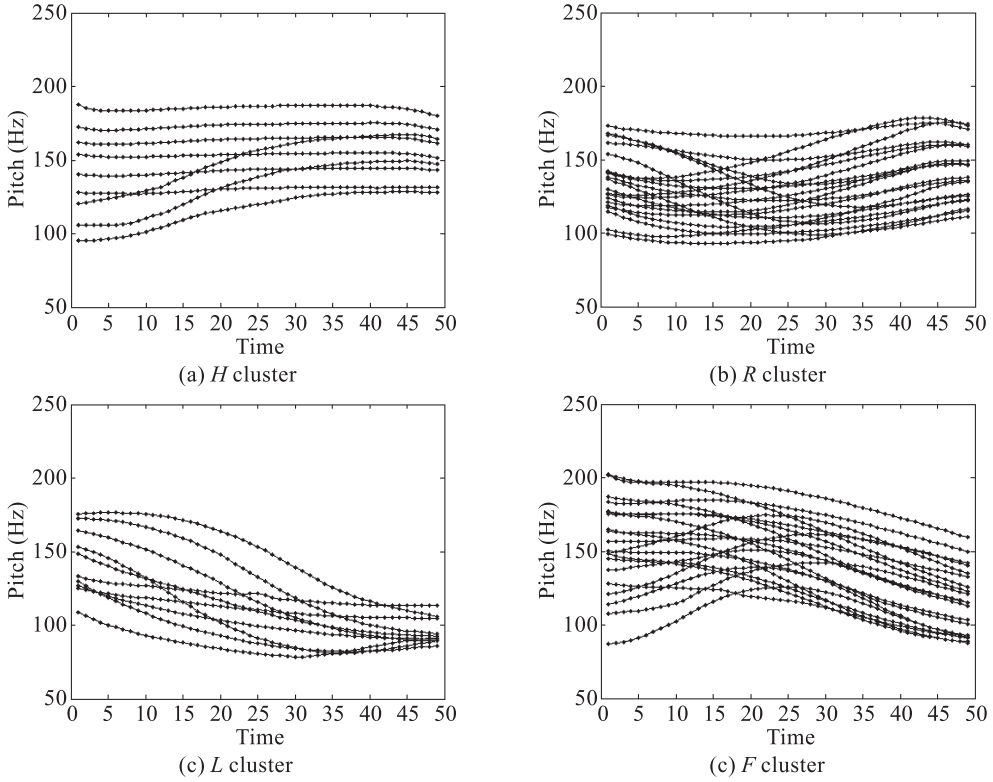


Fig. 1 Pitch contour clustering result for the *H*, *R*, *L*, and *F* tones in the four-syllable corpus

The PENTA models were then trained using the *average pitch contour* data for each cluster *considering the contextual information*. The model parameters *predicted by the CART tree*, the original initial pitches and the syllable durations were used to predict the pitch contours. The RMSE between the predicted and original pitch contours were then calculated. Figure 2a shows the RMSE distribution for the four-syllable *word corpus* while Fig. 2b shows the distribution for the *sentence corpus*. The results show that most of the errors are below 50 Hz, which indicates that the method accurately predicts the pitch contours with PENTA and the contextual information.

4.2 Pitch discontinuity modeling test

The CART tree was used to train the pitch discontinuity model with the four-syllable *word corpus* with 5121 syllables with either voiceless consonants or prosodic boundaries selected as the data set. In the *sentence corpus*, 85 120 syllables were selected. The CART regression function was used to predict the initial pitches for each pitch contour. Table 3 shows the RMSE and correlation rates between the predicted and

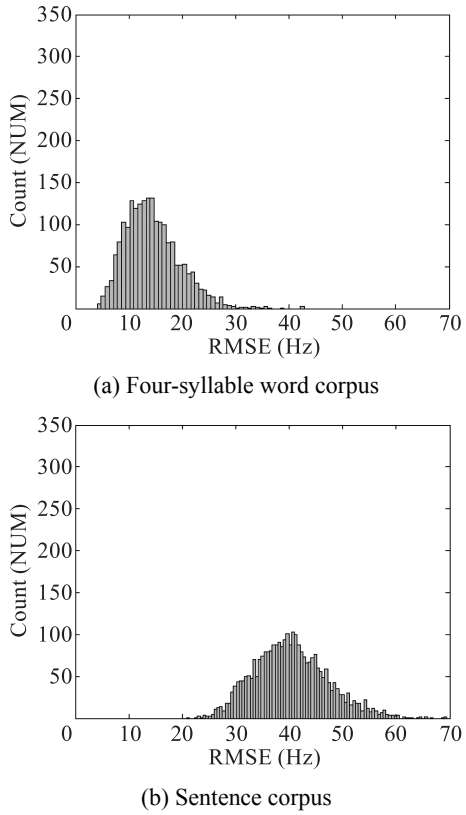


Fig. 2 RMSE difference between the original pitch contours and the predicted pitch contours with CART

Table 3 RMSE and correlation rates for the two corpora

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

original initial pitches for both corpora. The results indicate that the PDM accurately predicts the initial pitches from the contextual information in terms of the RMSE and correlation coefficients for both the *word corpus* and the *sentence corpus*. The result for the *sentence corpus* was worse than for the *word corpus* perhaps due to the performance degradation at the prosodic phrase boundaries in the sentence corpus where large pitch resets can occur.

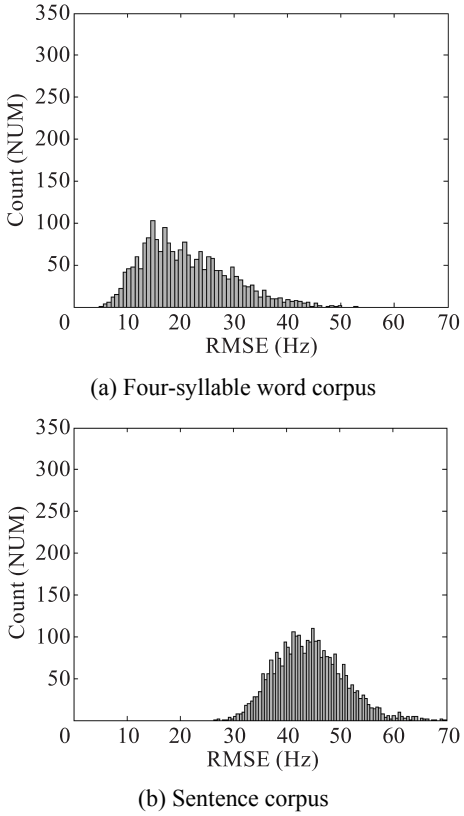
4.3 Pitch contour prediction test

The combined PENTA, CART and PDM model was evaluated in another test to predict pitch contours for TTS synthesis. In this test, the PENTA model parameters were predicted from the syllable contextual information by CART. The initial pitch was predicted from the contextual information by the PDM model. The original syllable duration was used directly.

The initial tests were conducted with the training sets, using the speech recordings of the entire corpus. The RMSE between the original and the predicted pitch contour for both the *word corpus* and the *sentence corpus* are shown in Fig. 3. Most of RMSE were below 40 Hz and centered around 20 Hz for the four-syllable *word corpus* (shown in Fig. 3a); while for the *sentence corpus*, most of the RMSE was below 60 Hz and centered around 50 Hz (shown in Fig. 3b). Again, the performance with the *sentence corpus* is slightly worse than with the *word corpus*.

Tests were then conducted with two new test sets with selected speech recordings set aside in the training set for training the CART, PENTA and PDM. The first test set contained 100 four-syllable words while the second test set contained 100 sentences.

An example for the four-syllable word corpus is shown in Fig. 4 which shows the original pitch contour (red dotted curve) and the predicted contour (blue plus curve) of a Chinese four-syllable word “wu2 yuan1 wu2 chou2 (无冤无仇)”. The voiceless consonant “ch” in the syllable “chou2” leads to a pitch contour discontinuity. The RMSE between the original and the predicted pitch contour is 6.15 Hz. The average RMSE for the whole test set of 100 four-syllable words was 10.52 Hz.



(a) Four-syllable word corpus

(b) Sentence corpus

Fig. 3 RMSE distribution between the original and the predicted pitch contours with CART and PDM

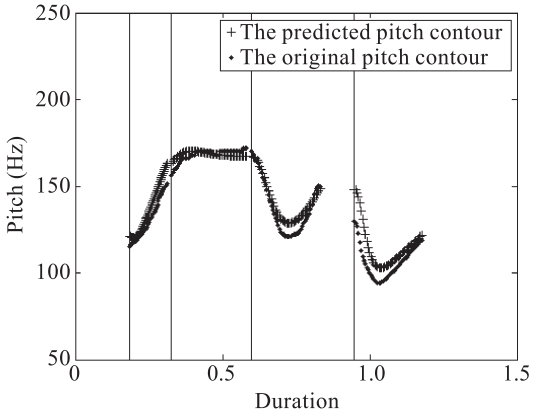


Fig. 4 Pitch contour prediction result of a word combining the original and the predicted pitch contours. The PENTA model parameters and the initial pitches were predicted by CART and PDM from contextural information.

An example for the sentence corpus is shown in Fig. 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 prosodic word boundaries, and “/” indicates the prosodic phrase boundaries. The average RMSE of the whole test set of 100 sentences was

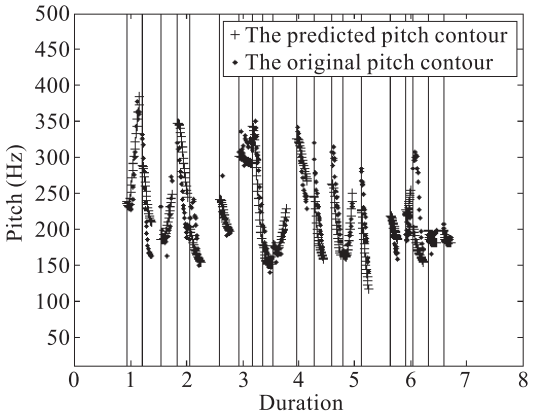


Fig. 5 Pitch contour prediction of a sentence comparing the original and the predicted pitch contours. The vertical lines represent the syllable's start positions.

30.12 Hz, which indicates that the method can accurately predict sentence pitch contours by combining the CART algorithm with the PENTA model. In this way, the PENTA model can be utilized for pitch modeling and prediction in practical Mandarin TTS systems.

5 Conclusions

This paper presents a method to predict pitch contours for Chinese syllables with different contextual information using the PENTA model and the CART.

In continuous speech, the pitch contour of the same syllable will vary depending on the context. The CART is used to cluster the syllables' normalized pitch contours according to the contextual information and the distances between pitch contours. The average pitch contour was calculated for each cluster for training the PENTA model. With this method, the PENTA model parameters can be finely tuned according to the syllable contextual information leading to better performance in predicting pitch contours for continuous speech.

The initial pitch is required by the PENTA model to predict the pitch contours. A PDM was then developed to predict the initial pitches at voiceless consonants or prosodic boundaries. The PDM is also based on CART and is based on contextual information.

Tests on a Chinese four-syllable word corpus and a Chinese sentence corpus indicate that this method can accurately predict pitch contours by taking into account the contextual information with PENTA and CART.

Further research will model the syllable duration in the corpus statistically. Only the duration and pitch information will then be used to get a complete prosody model of a sentence for speech synthesis.

References

[1] Fujisaki H. Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations. *STL-QPSR*, 1981, **22**(1): 1-20.

[2] Xu Yi. Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 2005, **46**(3-4): 220-251.

[3] Prom-on S, Xu Y, Thipakorn B. Quantitative target approximation model: Simulating underlying mechanisms of tones and intonations. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France, 2006: I.

[4] Xu C X, Xu Y, Luo L S. A pitch target approximation model for F0 contours in Mandarin. In: Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco, USA, 1999: 2359-2362.

[5] Prom-on S, Xu Y, Thipakorn B. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 2009, **125**(1): 405-424.

[6] Fujisaki H, Ohno S, Gu Wentao. Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command-response model for generation of their F0 contours. In: Proceedings of International Symposium on Tonal Aspects of Languages with Emphasis on Tone Language. Beijing, China, 2004: 61-64.

[7] The Centre for Speech Technology Research. Classification and Regression Trees. http://www.cstr.ed.ac.uk/projects/speech_tools/, 2011.

[8] Edinburgh Speech Tools Library. http://www.cstr.ed.ac.uk/projects/speech_tools/, 2011.

[9] Liu T, Cai L. The clustering research based on fundamental frequency of the syllable. *Journal of Chinese Computer Systems*, 2004, **25**(7): 1145-1150.

[10] Hu W, Xu B, Huang T. The acoustic research about prosodic boundary. *Journal of Chinese Information*, 2002, **16**(1): 43-48

[11] Shen W, Lin F, Li J, et al. A CART-based prosodic phrasing method for Chinese text-to-speech. *Computer Sciences*, 2002, **34**(4): 50-52.

[12] Xu Y. Contextual tonal variations in Mandarin. *Phonetics*, 1997, **25**: 61-83.