Pitch Prediction for Mandarin TTS with Mutual Prosodic Constraint

[1]Jian Yu, [1]Jianhua Tao, [2]Xia Wang
[1]National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences
[2]Multimedia Technologies Laboratory, Nokia Research Centre, China
[1]{jyu, jhtao}@nlpr.ia.ac.cn, [2]Xia.S.wang@nokia.com

Abstract. Most of current pitch prediction methods for mandarin TTS try to get pitch contours from the contextual information with a group of weights assigning. Without a good method in prosody concatenation constraint, the predicted pitch contours are not always stable because of the incomplete accordance between prosody information and text information. The paper presents a new mandarin pitch prediction method with mutual prosodic constraint between syllables. The idea of this mutual constraint is first inspired by lots of observations on corpus, but then it has been strictly proved with performance comparison and feature contribution analysis of CART-Based prosodic parameter prediction. Based on this, a reasonable definition of prosody concatenation cost is presented to measure the naturalness of pitch contours between two adjacent syllables. By minimizing this cost, the model can generate fluent pitch contours, which has been proved to be able to make the TTS system more natural than traditional systems.

Keywords: Speech synthesis, TTS, Mandarin, prosody model, pitch generation, mutual constraint.

1 Introduction

With the popularity of corpus-based technology in Text-to-Speech, the synthesis quality has been highly improved. Many statistical models, including decision tree, neural network, GMM and HMM [3][4][5][6], are used to describe the relationship between prosody information and text information, which results in better synthetic speech than the traditional rule-based approach. However, all of these studies fail to consider the fact that there is no complete accordance between prosody information and text information. For example, even the same sentence uttered by one person, the appearances of pitch contours may be different at different time and different locations. This fact leads to unsatisfactory results generated by previous pitch prediction models.

Recently, we realize that the reason for this incomplete accordance is that there is strong mutual prosodic constraint between adjacent units, which are syllables in Mandarin particularly. The idea of this mutual constraint is first inspired by lots of

The work is supported by National Natural Science Foundation of China (No. 60575032)

observations on large corpus, but then it has been strictly proved with performance comparison and feature contribution analysis of CART-Based prosodic parameter prediction in this paper. Based on these analyses, our new pitch prediction model makes full use of the mutual prosodic constraint between syllables: In the prosodic parameter prediction process, not only frequently used text features, but some prosodic features, such as adjacent syllables' pitch contours and initial length, are involved, which greatly reduces the predicting error; In pitch generation process, a new concatenation cost is defined to measure the naturalness of pitch contours between two adjacent syllables. This cost is minimized at every concatenation place, which makes the output pitch contour similar to that of the natural sentence on overall trend.

The structure of this paper is organized as follows: Section two introduces the method to extract prosodic parameters. Section three describes the meaning of mutual prosodic constraint using an example from recorded corpus, and then the existence of this constraint is strictly proved with performance comparison and feature contribution analysis of CART-based prosodic parameter prediction. Section four shows how to make use of this constraint in detail. Base on a new definition of prosody concatenation cost, the whole pitch contour is generated using a nonlinear estimation algorithm. Section five makes an evaluation to show the good performance of this model. Lastly, section six brings a conclusion and lists the deficiency of this model and the future work.

2 Pitch Contour Parameterization

In Mandarin, the pitch contour within a syllable has its own standard tone pattern, which is subject to various modifications in continuous speech. To construct a pitch prediction model, this continuous curve must be converted to a pattern represented by a series of parameters. In our pitch contour parameterization method, the pitch contour within a syllable is parameterized into five parameters, as Fig 1 shows. Among these, $F0_M$ is the mean value of pitch contour, which reflects the pitch register of current syllable, while $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$ are the starting value, the ending value, the starting derivative value, and the ending derivative value of the pitch contour, respectively. These four parameters are considered as boundary parameters, which can be used to measure the naturalness of pitch contours at concatenation places. All in all, the pitch contour within a syllable can be noted as a vector:

$$PitchContourWihinASyllable = (F0_M, F0_S, F0_E, F0_{SD}, F0_{ED})$$

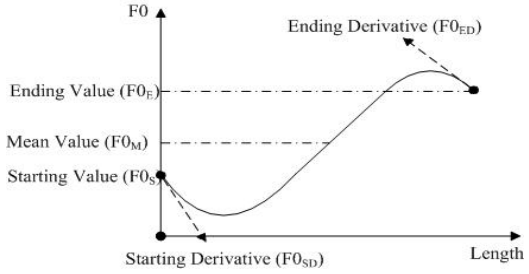

Fig. 1. Pitch contour parameterization

3. The Mutual Prosodic Constraint

The mutual prosodic constraint is the basic framework of the new pitch model. In this section, the meaning of mutual prosodic constraint is expatiated, and then the existence of this strong mutual constraint is strictly analyzed and proved from two aspects: the performance comparison and the feature contribution analysis of CART-based prosodic parameter prediction.

3.1 The meaning of mutual prosodic constraint

The meaning of mutual prosodic constraint between syllables is that adjacent syllables' pitch contours have great impacts on the current one. This viewpoint is inspired by lots of observations on large corpus. Fig 2 lists a typical pitch contour from recorded corpus. It seems that the pitch contour is virtually connected across the silence and voiceless initial. Both $F0_S$ and $F0_{SD}$ values are greatly impacted by the previous syllable's pitch contour; and $F0_E$ and $F0_{ED}$ are mostly impacted by the next syllable's pitch contour. In addition, we also notice that the initial category and its length have some great impacts on the boundary prosodic parameters. For example, when current initial is nasal or non-initial, the current syllable's starting pitch value is exactly the same as the previous syllable's ending value; in other situations, it seems that the pitch value continues to vary as same speed and same direction. All of these facts are caused by the mutual prosodic constraint between syllables.

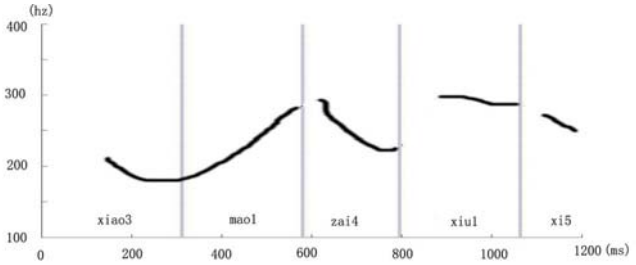

Fig. 2. An example showing the mutual prosodic constraint

3.2 The performance comparison

Maybe just an illustration from one figure is too impressionistic and not enough as a basis for a whole argument. Therefore, much effort should be done to prove the existence of strong mutual prosodic constraint between syllables. If this constraint is strong enough, adjacent syllables' prosodic parameters must be very useful for the prediction of current ones. Therefore, a performance comparison is made between two different kinds of prediction methods, in which one only predicts target prosodic parameters from frequently used text features, named as rough prediction method; while another involves more adjacent prosodic features in its input feature set(the detailed information of involved prosodic features are listed in Table 2), named as

precise prediction method. CART (Classification and Regression Tree), which can make use of both continuous features and categorical features at the same time, is adopted as the training model. For each feature, two kinds of regression trees are constructed, and their performance results are compared, listed in Table 1.

In table 1, RMSE depicts the predicting error and correlation depicts the relationship between predicted values and target values. These two terms are different measures to describe how well the CART model performs. From this table, we can see that the inclusion of prosodic features has greatly improved the performance of CART in terms of both RMSE and correlation. Based on the analysis in section 1, the imprecise results of rough prediction method are induced by the incomplete accordance between prosody information and text information. In addition, we also can see that the influence of mutual prosodic constraint focuses on boundary prosodic features including $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$, while the predicting result of $F0_M$ does not get much improvement.

Table 1. Comparison of predicted results of $F0_S, F0_E, F0_{SD}$ and $F0_{ED}$ by CART

	Precise Prediction Method (Prosodic features included)		Rough Prediction Method (Prosodic features excluded)	
	RMSE	Correlation	RMSE	Correlation
$F0_S$	24.7hz	0.92	32.8hz	0.84
$F0_E$	23.4hz	0.91	35.2hz	0.81
$F0_{SD}$	0.36hz/ms	0.75	0.45hz/ms	0.61
$F0_{ED}$	0.34hz/ms	0.78	0.49hz/ms	0.63
$F0_M$	22.1hz	0.91	25.5hz	0.89

3.3 The feature contribution analysis

For getting more evidences of the existence of mutual prosodic constraint, another experiment is done to show the feature contribution in the precise CART-Based boundary prosodic parameter prediction, in which the performances of CARTs are evidently improved.

As we know, a general splitting criteria for regression tree is least-squares deviation. For every node, the best splitting feature is the one which maximize this equation:

$$S_{feature\_name} = R_{parent} - (R_{left\_node} * N_{Left\_node} + R_{right\_node} * N_{right\_node}) / N_{parent}$$

Where, $R = \sum_{i=1}^{N} \left(x_i - \sum_{j=1}^{N} x_j \right)^2$, N is the number of input vectors in particular node.

$S_{feature\_name}$ reflects this feature's contribution in current node. By going through all nodes, the contribution for one particular feature can be calculated, the formula is:

$$Contribution_{FeatureI} = \sum_{i=FeatureI} S_i$$

Table 2 lists the contributions of most input features in CART-based prediction and shows several remarkable points: (1) All in all, prosodic features make significant contributions in the prediction of all four boundary prosodic parameters, which

further proves the existence of strong mutual prosodic constraint. (2) The degree of mutual prosodic constraint is not uniform. For instance, the contribution of previous syllable's prosodic parameters in $F0_S$ and $F0_{SD}$ prediction is much larger than the contribution of next syllable's prosodic parameters in $F0_E$ and $F0_{ED}$ prediction. This fact shows that the constraint of first syllable on second syllable is stronger than that of second syllable on first syllable, which is consistent with previous research [11]. (3) Actually, some prosodic features can be considered as more precise descriptions of corresponding text features. Let's take $F0_S$ prediction as an example, as our priori knowledge, the previous tone should plays an important role in this prediction, but from this table the contribution of previous tone is not very large, just 3.08%. In our opinion, the reason for this phenomenon is that previous $F0_E$ and $F0_{ED}$ actually cover more precise information than previous tone. For example, if previous tone is tone 4, its ending value is between 150 Hz and 200 Hz, which influences current syllable's $F0_S$ value. But now, we can exactly know the value of previous ending value, so it can get more precise results just with little contribution from previous tone. (4) Some duration information also make significant contributions in the CART-based prediction, the explanation of this fact can be got from Fig.2. When the initial is voiceless, the pitch deviation across the voiceless initial and pause is increased as the length increased.

Table2. The feature contribution analysis in CART-based prediction

	Feature Contribution			
	$F0_S$	$F0_{SD}$	$F0_E$	$F0_{ED}$
All prosodic features	**43.4%**	**40.5%**	**19.4%**	**16.7%**
Previous syllable's $F0_E$	27.3%	15.45%	*	*
Previous syllable's $F0_{ED}$	1.25%	5.66%	*	*
Next syllable's $F0_S$	*	*	12.94%	5.80%
Next syllable's $F0_{SD}$	*	*	1.23%	4.50%
Pause length before current syllable	3.16%	1.68%	*	*
Pause length after current syllable	*	*	1.29%	3.11%
Current syllable's initial length	11.67%	17.69%	*	*
Next syllable's initial length	*	*	3.87%	3.21%
All frequently used text features	**56.6%**	**59.5%**	**80.6%**	**83.3%**
Current syllable's tone	39.68%	33.39%	53.72%	61.7%
Previous syllable's tone	3.08%	7.99%	*	*
Next syllable's tone	*	*	6.15%	3.58%
Syllable ID(include initial and final)	1.43%	6.4%	0.04%	1.2%
Prosodic structure(includes position and length of word, phrase and intonation phrase)	12.37%	11.65%	20.68%	16.82%

4. Pitch Prediction with mutual prosodic constraint

4.1 The Definition of Concatenation Cost

The existence of strong mutual prosodic constraint has been strictly proved, but there is still a big problem remaining to be resolved before this constraint can be made use of in pitch prediction: how to make adjacent syllables' prosodic parameters be the input features of CART while predicting them at the same time. In the previous work [2], a prosody template based method was proposed to solve this problem. In this work, a new target cost and concatenation cost was defined to measure the naturalness of pitch contours between syllables, and then the viterbi algorithm was used to select the best prosody template sequence which minimizes the cost.

In our new pitch model, the definition of concatenation cost is still valuable. But now because no template could be used, different prosodic parameters of one syllable can be treated separately, while in the previous method, once the template is selected, all of the prosodic parameters are fixed. Therefore, the concatenation cost can be minimized separately at each concatenation place. Currently, the best in local is the best in whole.

Before presenting the new algorithm in detail, it's necessary to explain the definition of concatenation cost. A reasonable concatenation cost could be applied to the measurement of the naturalness of pitch contours between adjacent syllables. The prosodic parameter values predicted by precise prediction method with adjacent syllables' prosodic features involved can be considered as the expected values by adjacent syllables, thus the differences between the predicted values and the real values reflect the naturalness of pitch contours between these two adjacent syllables. Therefore, the value of concatenation cost is defined as the weighted sum of differences between predicted values and real values of the four boundary prosodic parameters, noted as $DF0_S, DF0_E, DF0_{SD}$ and $DF0_{ED}$ respectively. The formula is:

$$concatenation\_cost = w_1 * DF0_S + w_2 * DF0_E + w_3 * DF0_{SD} + w_4 * DF0_{ED}$$

Where w_i is the weight of corresponding parameter, which is assigned according to expert knowledge. Fig 3 schematically illustrates the definition of concatenation cost. This definition is much reasonable because it makes full use of the mutual prosodic constraint between syllables.

4.2 Prosodic parameters prediction algorithm

Based on this definition of cost, our algorithm tries to minimize the concatenation cost separately at every concatenation place, which is described in detail as follows:

Step 1: Range Estimation. Based on the rough prediction method described in section 3.2, the rough range of each prosodic parameter can be obtained, noted as [Avg-r*Dev, Avg+r*Dev].

Where Avg and Dev are the predicted value and RMSE of the rough prediction method, respectively, and r is a coefficient to modify the range, whose default value is 1. The final value of each prosodic parameter should be in this range and the next step

is to find the precise position based on the mutual prosodic constraint between syllables.

Step 2: Precise Search. For each parameter, the range is equally divided into N parts. That is to say, there are N candidate values for each parameter, as Fig 4 shows. Then, the next algorithm tries to choose the most proper values which minimize the concatenation cost. As showed in Fig 3, among all input features of CART, the contextual information can be got through text analysis module and the initial length and pause length can be predicted by duration model. Therefore, once all the four parameters values in Fig 4 are fixed, the value of concatenation cost at current concatenation place can be calculated. Obviously, the most direct and simplest method to minimize the concatenation cost is to enumerate all possible values of prosodic parameters at concatenation place. Suppose the number of syllables in current sentence is M , the computation complexity is $N^4 * (M-1)$.

From the feature contribution analysis on the precise prediction method, it can be seen that the derivative values $F0_{SD}$ and $F0_{ED}$ do not play a very important role in the prediction for pitch values $F0_S$ and $F0_E$. Therefore, the pitch values and the pitch derivative values can be treated separately in turn to simplify the computation:

Step 2.1: assign the values of $F0_{SD}$ and $F0_{ED}$ as the mean values of their rough ranges, respectively. Then enumerate all possible values of $F0_S$ and $F0_E$ and select the values which minimize current concatenation cost.

Step 2.2: based on the values of $F0_S$ and $F0_E$ obtained at step 2.1, enumerate all possible values of $F0_{SD}$ and $F0_{ED}$, and select the values which minimize current concatenation cost.

By this way, the computation complexity is reduced to $(N^2 + N^2) * (M-1)$, which does not cause imprecision.

Such search algorithm is carried out at every concatenation place, which updates the values of all boundary prosodic parameters at concatenation places. Meanwhile, all the F0 mean values, the starting value of first syllable and the ending value of last syllable are kept as the rough prediction results. Until now, all prosodic parameters in the whole sentence are precisely obtained.

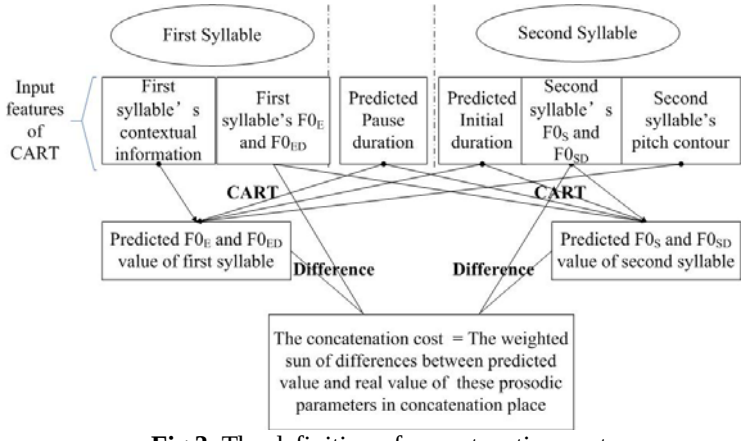

Fig.3. The definition of concatenation cost

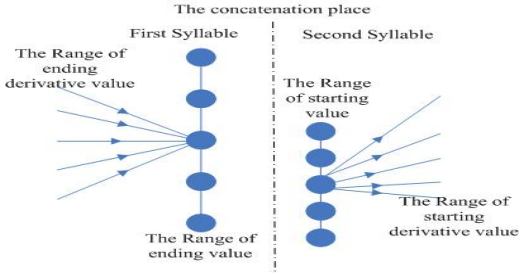

Fig.4. The precise search (N=5)

4.3. Pitch generation based on prosodic parameters

After precise search, the last step is the generation of pitch contour based on the predicted prosodic parameters. Suppose the pitch contour within a syllable can be precisely depicted by a function f(x), an equation group can be listed based on the values of prosodic parameters. That is:

$$f(s) = F0_S \tag{1}$$

$$f(e) = F0_E \tag{2}$$

$$f'(s) = F0_{SD} \tag{3}$$

$$f'(e) = F0_{ED} \tag{4}$$

$$\int_s^e f(x)dx = (e-s) * F0_M \tag{5}$$

Where e and s are the starting time and ending time of the pitch contour within a syllable, respectively.

Two kinds of formulas are supposed to represent the function f(x): third order polynomial function and exponential function, as follows:

$$1: \quad f(x) = ax^3 + bx^2 + cx + d$$

$$2: \quad f(x) = a * e^{bx} + cx + d$$

Among these two formulas, the second is based on the theory of PENTA model [7], and it is much easier to explain in phonetics. However, it can not precisely depict the pitch contour with more than one polar, just like the pitch contour plotted in Fig 1. Therefore, the first one is adopted in our pitch model.

In addition, both two functions only have four coefficients, but the number of equations is five. Actually, this equation group has no solutions. To solve this problem, a nonlinear least squares minimization algorithm named Levenberg-Marquardt [12] is used to calculate similar solutions. Fig 5 shows a comparison between a natural pitch contour sketched as gray and a generated one sketched as black, which shows the advantages of the new pitch model.

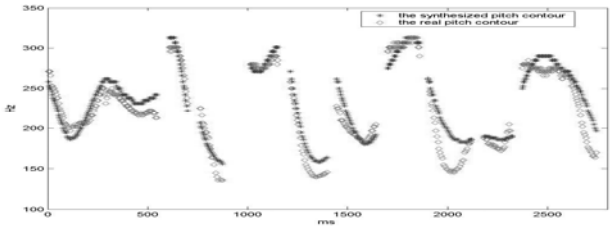

Fig.5. A comparison between a natural pitch contour and a synthesized one

5. Evaluation

How to evaluate a TTS system is a very difficult problem for a long time. On one hand, unlike speech recognition and machine translation, there is no perfect objective evaluation target for TTS system. On the other hand, the most frequently used subjective evaluation method-Mean Opinion Score(MOS), may not be able to provide much information in the performance comparison of TTS systems based on different corpuses and designed for different domains.

Even with above difficulties, making an informal evaluation for our new pitch model is essential. In this paper, the evaluation task is done by two ways: one subjective test with MOS which shows human's perception feelings and another objective test with correlation and RMSE between real and synthesized pitch contours, in which correlation indicates the similarity in shape and RMSE indicates the characteristic divergence. Two models are compared in this evaluation, one is the new model presented in this paper, and the other is a primitive one which predicts prosodic parameters only from text information, just like the rough prediction method introduced in section 3.2.

One more thing need mentioned, in the MOS test, the evaluation target is the naturalness of pitch contour. To abandon the influences of other acoustic parameters, an HMM-based speech synthesis system [5] is used to generate the spectrum of synthesized speech, and then after coordination with different pitch contours generated by two systems, STRAIGHT-based vocoding algorithm is used to generate speech [10]. By this way, it focuses on the difference of pitch contour.

The corpus used in the paper contains 6000 sentences, in which 5000 sentences are used for training, and others are used for open test, among these sentences, only 200 are used in MOS scoring for simplicity. Table 3 shows the comparison result, which reveals that the new model generates much more natural pitch contours than the primitive one does.

Table 3. The comparison result of pitch contours generated by two pitch models

	Objective evaluation		Subjective evaluation MOS
	RMSE	Correlation	
Primitive Model	51hz	0.66	2.9
New Model	22hz	0.87	4.0

6. Conclusions

This paper presents a new pitch prediction model for mandarin TTS system based on the mutual prosodic constraint between syllables. By concentrating on the mutual constraint, the model can make sure that there is no unnatural pitch contours between every two adjacent syllables, which leads to very natural pitch contours on overall trend.

However, this model still has some shortcomings. Firstly, the supposition of function f(x) in pitch generation can only depict the main trend and register of the pitch contour within a syllable, but can not depict small changes near syllable boundaries, especially in tone 2 and tone 3. Secondly, the computation complexity is still high when the value of N is large. How to design an iterative algorithm to further simplify computation is another part to be improved.

References

[1]. Jianhua Tao, “F0 Prediction Model of Speech Synthesis Based on Template and Statistical Method”, Lecture Notes of Artificial Intelligence, Springer, 2004

[2]. Jian Yu, Wanzhi Zhang, Jianhua Tao, “A New Pitch Generation Model Based on Internal Dependence of Pitch Contour for Mandarin TTS System”, ICASSP 2006, Toulouse, France

[3]. Jianhua Tao, etcl, “Trainable Prosodic Model for Standard Chinese Text-to-Speech System.”, Chinese Journal of Acoustic, Vol.20, 2001, p257-265

[4]. Sin-horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, “An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech”, IEEE Transaction on Speech and Audio Processing, VOL.6, No.3, May 1998

[5]. Tomoki Toda, Keiichi Tokuda, “Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis”, InterSpeech 2005, Lisboa, Portugal.

[6]. Fu-Chiang Chou, Chiu-Yu Tseng, and Lin-Shan Lee, “A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese”, IEEE Transaction on Speech and Audio Processing, VOL.10, No.7, October 2002

[7]. Yi Xu "Speech Melody as Articulatorily Implemented Communicative Functions", Speech Communication 46 (2005) 220-251.

[8]. Chen-Yu Chiang, Yih-Ru Wang and Sin-Horng Chen " On the Inter-syllable Coarticulation Effect of Pitch Modeling for Mandarin Speech" INTERSPEECH 2005, Lisboa, Portugal.

[9]. Yongguo Kang, Jianhua Tao and Bo XU, “Applying Pitch Target Model to Convert F0 Contour for Expressive Mandarin Speech Synthesis”, ICASSP 2006, France

[10]. H.Kawahara, I. Masuda-Katsuse and A. deCheveigne, “Restructuring speech representations using pitch-adaptive timefrequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187-207, 1999

[11]. Yi Xu and Q. Emily Wang, "Pitch Targets and Their Realization: Evidence from Mandarin Chinese", Speech Communication 33, 2001.

[12]. http://www.ics.forth.gr/~lourakis/levmar/