A MANDARIN INTONATION PREDICTION MODEL THAT CAN OUTPUT REAL PITCH PATTERNS

Neng-Huang Pan, Ming-Shing Yu, and Ming-Jer Wu

Department of Applied Mathematics, National Chung-Hsing University, Taichung 40227, Taiwan.
E-mail: nhpan@amath.nchu.edu.tw, {msyu, mjwu}@dragon.nchu.edu.tw

ABSTRACT

In this paper we proposed an intonation prediction model for Mandarin TTS systems. Our model can output real pitch patterns by finding a suitable real pitch pattern from the training corpus. This method is a new experiment. The advantages of our model are as follows. (1) It can improve the naturalness of the synthesized speech. It gets higher scores in the subjective listening tests. (2) It has high accuracies. The average errors of 0.425ms and 0.457ms were obtained for the inside and outside tests, respectively. The pattern errors of 0.128ms and 0.129ms were obtained for the inside and outside tests, respectively. We found that pattern error measurement method compiles with human hearing. (3) The training corpus need not be very large. It can relieve the data sparsity problem.

1. INTRODUCTION

A Text-to-Speech (TTS) system can translate an arbitrary input sentence into its corresponding speech. A Mandarin TTS system is composed of three components, namely text analysis, prosody generation, and speech synthesis. The input sentence is first passed to the text analysis module. In this module, word segmentation and character-to-phoneme conversion are done. Some text analysis module also outputs the part-of-speech (POS) information and parsing results. Then the words and phonemes are passed to the prosody generation module. In this module, the prosodic information of each syllable (corresponding to a Chinese character) is obtained. They are duration, coarticulation or pause, volume, and pitch contours. Finally the speech synthesis module extracts the required synthesis units and does some signal processing to get the desired speech.

The goal of a prosody generator is to mimic the human rhythm, which includes syllable duration, pause, energy, and pitch contours produced by people. The prediction accuracy of prosody generator will affect intelligibility and naturalness of the synthesized speech. In general, the rule-based approach and data-driven approach are two major methods to generate prosodic information. Also some systems used the hybrid method — the Bell Laboratories Mandarin TTS system for example [13]. In Bell Laboratories Mandarin TTS system, the syllable duration prediction module used the data-driven approach, and the intonation prediction module used the rule-based approach. Besides, Wu and Chen [14] had proposed a word-prosody template tree method for a Mandarin TTS system. It needs linguistic experts to write rules in rule-based models [7,8,9]. It is difficult to create enough rules that cover all of the cases in a language. Moreover, some rules may contradict one another. The data-driven approach can automatically learn the relationship between input texts and labeled speech corpora, so it does not need a lot of linguistic knowledge. Neural nets [3,11,12], CART (Classification and Regression Tree) [1], and statistical models [6,10,15] are common methods in the data-driven approach.

This paper focuses on the intonation prediction module of prosody generator for Mandarin TTS systems. The framework of our model is a statistical model with hierarchical structure. In the past, there are many papers proposed to deal with the prosody generation problem [2,3,5-15]. The outputs of most prosody generators are not natural rhythms produced by people, so the naturalness of synthesized speech is still not good enough. In this paper we proposed an intonation prediction model that can output real pitch patterns. The block diagram of our model is show in Fig. 1. The input sentence is first passed to the text analysis module, and then the results of this module will be passed to the basic intonation prediction module. The outputs of this module are artificial pitch patterns. Finally, we use the predicted value of the basic intonation prediction module to find a suitable real pitch pattern from the training corpus. Most of outputs of our model are natural pitch patterns.

This paper is organized as follows. Section 2 describes the basic intonation prediction module. Section 3 presents how to find a suitable real pitch pattern from the training corpus. Section 4 focuses on the experimental results. Section 5 is the summary.

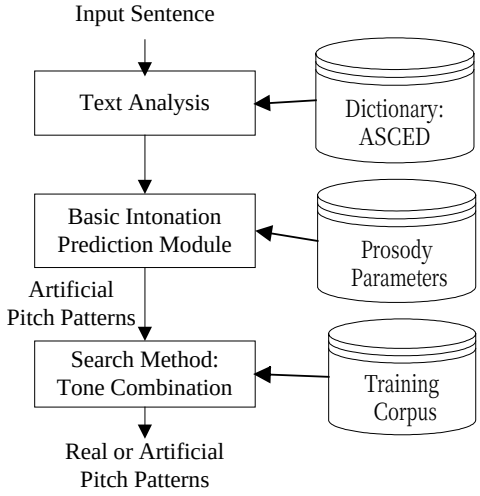

Fig. 1: Block diagram of our intonation prediction model.

2. BASIC INTONATION PREDICTIN MODULE

In this section we proposed a statistical intonation prediction module with hierarchical structure for Mandarin TTS systems. There are four levels in our module: syllable level, word level, prosodic phrase level, and utterance level. Here "hierarchy" means that each lower level is a subset of a higher level. The linguistic features that we used in each level are as follows:
Syllable level: consonant type, vowel type, and tone type.
Word level: word length and the position of the character in the word.
Prosodic phrase level: number of words in the phrase and the position of the word in the phrase.
Utterance level: number of phrases in the utterance and the position of the phrase in the utterance.

In each level, we calculate the means of syllables with the same condition. Finally, we combine the results of each level in our module by using a linear regression method to get a predicted value. Since there are only a few parameters in each level, the size of our training corpus need not be very large. Thus the data sparsity problem, which is often encountered in using some other models, can be relieved. The preparation of a corpus often needs lots of human work. The speech data must be correctly segmented and marked. These information can only be obtained semi-automatically up to now. Human checking is required to get precise information. Besides, smaller training corpus size can also save the training time and disk space. We use the orthogonal polynomial expansions to represent a pitch contour [4]. Each pitch contour can be represented as a four-coefficient vector, the first coefficient is the mean of the pitch contour, and the other three coefficients represent its shape. The distance measure between two pitch contours $A=(a_0, a_1, a_2, a_3)$ and $B=(b_0, b_1, b_2, b_3)$ can be defined as

$$Dis(A,B) = \sum_{j=0}^{3} (a_j - b_j)^2 \qquad (1)$$

3. SEARCH METHOD: TONE COMBINATION

In recent years, some approaches were proposed to choose the speech unit from a large speech corpus in order to improve the quality of the synthesized speech [5,14]. But there are no papers dealing with the prosodic information generation problem can output real pitch pattern by searching the speech corpus, as we know.

Mandarin is a tonal language. There are five tones in Mandarin. They are high level (tone 1), mid-rising (tone 2), mid-falling-rising (tone 3), high falling (tone 4), and neutral tone (tone 5). The pitch contour patterns of four lexical tones in Mandarin are show in Fig. 2 [4]. Tone is the most important feature that affects the pitch contours of a syllable. And word is the basic meaningful unit in Mandarin. Each word has one to several Chinese characters. Tone combination is the tone sequence of the characters in a word. We adopt three-syllable and two-syllable tone combination as our search feature, because most words in our training corpus are trisyllabic or disyllabic words. Besides, a multi-syllable word has strong inter-syllable coarticulations, which will affect the pitch contours.

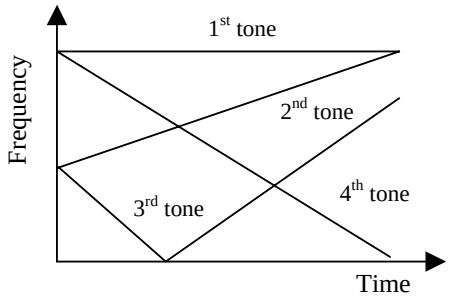

Fig. 2: The pitch contours of 4 lexical tones in Mandarin.

For each syllable in the test utterance, we use the following steps to output a real pitch pattern. First, we use the predicted value of the basic intonation prediction module to search a suitable three-syllable tone combination pattern in a word from the training corpus. If the search succeeds, we let it be the output. Secondly, if the search fails in the first step, we search the two-syllable tone combinations. If the search succeeds, we get the output. Third, if the search fails in the second step, we use the predicted value of the basic intonation prediction module as the output. The detail algorithm is show in Algorithm Natural Pitch. We first make some definitions.

Definitions*:* For syllable S, let P(S)and P'(S) be the original pitch contour and the predicted pitch contour from the basic intonation prediction module, respectively. N is the number of syllables in the training corpus. Consider a syllable S_i in a word $W=S_1S_2\ldots S_m$, where m is the word length of W. Let Ts_i be the tone of syllable S_i. Define $3TS(S_i)$= {Ts_{i-1} Ts_i Ts_{i+1} (S_{i-1}, S_i, and S_{i+1} are in the same word)} be the 3-syllable tone sequence of S_i. Define $2TS(S_i)$={Ts_i Ts_{i+1}, Ts_{i-1} Ts_i (S_i and S_{i+1} or S_{i-1} and S_i are in the same word)} be the 2-syllable tone sequence of S_i.

Algorithm Natural Pitch*:*
Input: A syllable S in the test set and P'(S), the predicted pitch contour by the basic intonation prediction module.
Output: A pitch pattern that is similar to P'(S), which may exist in the training corpus.
Steps:
1. Let M3 be the set of syllables S_i, i=1,2,…,N, in the training set such that $3TS(S_i)= 3TS(S)$. If M3 is non-empty, find the syllable S_k in M3 such that $Dis(P'(S),P(S_k))$ is minimum; output $P(S_k)$. Otherwise go to Step 2.
2. Let M2 be the set of syllable S_i, i=1,2,…,N, in the training set such that $2TS(S_i) \cap 2TS(S) \neq \varnothing$. If M2 is non-empty, find the syllable S_k in M2 such that $Dis(P'(S),P(S_k))$ is minimum; output $P(S_k)$. Otherwise go to Step 3.
3. Output P'(S) as the predicted pitch contour.

By Algorithm Natural Pitch, the predicted value for monosyllabic word is always the output of basic intonation prediction module, which is usually an artificial pitch pattern.

4. EXPERIMENTAL RESULTS

The texts in our corpus are articles adopted from daily newspapers. There are 11205 Chinese characters and 865 sentences in these articles. The ratio of training data and testing

data size are 84.51% (9469 Chinese characters) and 15.49% (1736 Chinese characters), respectively. A female speaker recorded the voice files. The average speaking rate is about three Chinese characters per second. All data were recorded at the sampling rate of 12kHz.

4.1. Prediction Errors

The average errors of the basic intonation prediction module (BIPM) and the basic intonation prediction module plus tone combination search method (BIPM+TC), i.e. our intonation prediction model, are show in Table 1.

Table 1: Average errors of our models. (Unit: ms/syllable)

Methods \ Tests	Inside	Outside
BIPM	0.43	0.446
BIPM + TC	0.435	0.451

For BIPM, average errors of 0.43ms and 0.446ms were obtained for the inside and outside tests, respectively. For BIPM+TC, average errors of 0.435ms and 0.451ms were obtained for the inside and outside tests, respectively. In our intonation prediction model, there are 144 (about 1.52%) and 0 syllables that can find the original pitch pattern from the training corpus for inside and outside test, respectively.

4.2. Subjective Listening Tests

From Table 1, BIPM seems better than BIPM+TC, because it has smaller errors. It contradicts our supposition that using real pitch pattern can improve the naturalness of the synthesized speech, so we designed some listening tests to check whether our supposition is right. In our experiments, the subjective MOS (Mean Opinion Score) evaluation has ten classes. Score 10 for the best (excellent), and score 1 for the worst (unsatisfactory). The participators of the first inside test, second inside test, and outside test are 25, 40, and 42 college students, respectively. The subjective MOS of the BIPM and BIPM+TC are show in Table 2.

Table 2: MOS of our models.

Methods \ Tests	Inside		Outside
	First time	Second time	
BIPM	6.88	6.45	6.81
BIPM + TC	7.03	6.81	7.36

For BIPM, MOSs of 6.88, 6.45, and 6.81 were obtained for the first inside, second inside, and outside listening tests, respectively. For BIPM+TC, MOSs of 7.03, 6.81, and 7.36 were obtained for the first inside, second inside, and outside listening tests, respectively. In the first inside listening test, the evaluation data are 30 items (words or sentences). There are 20 testing items in BIPM+TC that have equal or higher prediction error than those in BIPM. But all of these items get higher MOS than those in BIPM. And there are 27 items in BIPM+TC that get higher MOS than those in BIPM for first inside listening test. The results are summarized in Table 3.

Table 3: Some information for BIPM+TC in the listening tests.

Tests \ Info.	Info. A	Info. B	Info. C
First inside test	20	27	30
Second inside test	20	28	30
Outside test	13	12	15

Info. A: # of test items in BIPM+TC that have equal or higher prediction error than those in BIPM.
Info. B: # of test items in BIPM+TC that get higher MOS that those in BIPM.
Info. C: # of test items in the test.

The evaluation data of the second inside listening test are 30 items. There are 20 testing items in BIPM+TC that have equal or higher prediction error than those in BIPM. But all of these items get higher MOS than those in BIPM. And there are 28 items in BIPM+TC getting higher MOS than those in BIPM for the second inside listening test, as can be seen in Table 3. In the outside listening test, the evaluation data are 15 items. There are 9 testing items in BIPM+TC that have higher prediction error yet higher MOS than those in BIPM. There are 4 items in BIPM+TC that have equal prediction error with those in BIPM; 2 of these items get higher MOS, the others get equal MOS. There are 12 items in BIPM+TC getting higher MOS than those in BIPM for the outside listening test, as can be seen in Table 3.

4.3. Pattern Error Measurement

Experimental results show that using real pitch pattern can improve the naturalness of synthesized speech, while the distance measure method, Eq. (1), does not comply with human hearing. So we use the "Pattern Error" [10] to calculate the prediction error instead. The pattern error between two pitch contours $X=(x_1,x_2,...,x_n)$ and $Y=(y_1,y_2,...,y_n)$, where n is the number of pitches, is defined as

$$E_i=\left|x_i-y_i\right| \quad , \tag{2}$$

$$AE(X,Y)=\frac{1}{n}\sum_{i=1}^{n}E_i \quad , \tag{3}$$

$$PatternError(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\left|E_i-AE(X,Y)\right| \quad , \tag{4}$$

where $AE(X,Y)$ means average error (distance) between X and Y.

A characteristic of pattern error is that if the pitch contours X and Y are two parallel lines, the pattern error between X and Y will be zero. The pattern errors of our models are show in Table 4. For BIPM, pattern errors of 0.136ms and 0.137ms were obtained for the inside and outside tests, respectively. For BIPM+TC, pattern errors of 0.128ms and 0.129ms were obtained for the inside and outside tests, respectively.

Table 4: Pattern errors of our model. (Unit: ms/syllable)

Method \ Test	Inside	Outside
BIPM	0.136	0.137
BIPM+TC	0.128	0.129

From Table 4, we can see that BIPM+TC has better accuracy (using pattern error measurement method) than BIPM. And by the results of listening tests, we know BIPM+TC gets higher MOS than BIPM. So we think the pattern error measure method is closer to human hearing (compared with Eq. (1)), and the outputs of BIPM+TC have better pitch contour patterns.

5. SUMMARY

In this paper we proposed a statistical intonation prediction model with hierarchical structure for Mandarin TTS systems. There are 4 levels in our model; they are syllable level, word level, prosodic phrase level, and utterance level. Here "hierarchy" means that each lower level is a subset of a higher level. This model can relieve the data sparsity problem, so the size of the training corpus need not be very large. Our model can output real pitch patterns by finding a suitable real pitch pattern from the training corpus. We found that using real pitch pattern can improve the naturalness of the synthesized speech. We also found that the pattern error measurement method complies with human hearing. Our model has high accuracies; the average errors of 0.425ms and 0.457ms were obtained for the inside and outside tests, respectively. The pattern errors of 0.128ms and 0.129ms were obtained for the inside and outside tests, respectively.

6. ACKNOWLEDGEMENT

This work was partially supported by National Science Council, Taiwan under the grant number NSC91-2213-E-005-016. The authors wish to thank Academia Sinica, Taiwan for supplying the lexicon Academia Sinica Chinese Electronic Dictionary. The authors also want to thank the people who participated in our listening tests.

7. REFERENCES

[1] Leo Breiman, Jeronme H. Friedman, Richard A. Olshen, and Charles J. Stone, "Classification and Regression Trees", Wadsworth & Brooks, Pacific Grove CA, 1984.

[2] Sin-Horng Chen, Saga Chang, and Su-Min Lee, "A Statistical Model Based Fundamental Frequency Synthesizer for Mandarin Speech", The Journal of the Acoustical Society of America, Vol. 92, No. 1, pp. 114-120, 1992.

[3] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 3, pp. 226-239, 1998.

[4] Sin-Horng Chen and Yih-Ru Wang, "Vector Quantization of Pitch Information in Mandarin Speech", IEEE Transactions on Communications, Vol. 38, No. 9, pp.1317-1320, 1990.

[5] Fu-Chiang Chou, "Corpus-based Technologies for Chinese Text-to-Speech Synthesis", Ph. D. Thesis, Department of Electrical Engineering, National Taiwan University, 1999.

[6] Fu-Chiang Chou, Chin-Yu Tseng, and Lin-Shan Lee, "Automatic Generation of Prosody Structure for High Quality Mandarin Speech Synthesis", ICSLP'96, Vol. 3, pp. 1624-1627, 1996.

[7] Dennis H. Klatt, "Review of Text-to-Speech for English", The Journal of the Acoustical Society of America, Vol. 82, pp. 737-793, 1987.

[8] Lin-Shan Lee, Chiu-Yu Tseng, and Ching-Jiang Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System", IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 3, pp. 287-294, 1993.

[9] Lin-Shan Lee, Chiu-Yu Tseng, and M. Ouh-Young, "The Synthesis Rules in a Chinese Text-to-Speech System", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 9, pp. 1309-1320, 1989.

[10] Neng-Huang Pan, Wen-Tsai Jen, Shyr-Shen Yu, Ming-Shing Yu, Shyh-Yang Huang, and Ming-Jer Wu, "Prosody Model in a Mandarin Text-to-speech System Based on a Hierarchical Approach", IEEE International Conference on Multimedia and Expo 2000, Vol. 1, pp. 448-451, 2000.

[11] Yoshinori Sagisaka, "On The Prediction of Global F0 Shape for Japanese Text-to-Speech System", ICASSP'90, pp. 325-328, 1990.

[12] Michael S. Scordilis and John N. Gowdy, "Neural Network Based Generation of Fundamental Frequency Contours", ICASSP'89, Vol. 1, pp. 219-222, 1989.

[13] Chilin Shih and Richard Sproat, "Issues in Text-to-Speech Conversion for Mandarin", Computational Linguistics and Chinese Language Processing, Vol. 1, No. 1, pp. 37-86, 1996.

[14] Chung-Hsien Wu and Jau-Huang Chen, "Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis", Speech Communication, Vol. 35, pp. 219-237, 2001.

[15] Ming-Shing Yu, Neng-Huang Pan, and Ming-Jer Wu, "A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-Speech System", International Symposium on Chinese Spoken Language Processing (ISCSLP) 2002, pp. 21-24, August 2002.