

A New Pitch Modeling Approach for Mandarin Speech

Wen-Hsing Lai^{1,2}, Yih-Ru Wang¹ and Sin-Horng Chen¹

¹Dept. of Communication Engineering, National Chiao Tung University, Taiwan

²Chunghwa Telecommunication Laboratories, Taiwan

Abstract

In this paper, a new approach to model syllable pitch contour for Mandarin speech is proposed. It takes the mean and shape of syllable pitch contour as two basic modeling units and considers several affecting factors that contribute to their variations. Parameters of the two models are automatically estimated by the EM algorithm. Experimental results showed that RMSEs of 0.551 ms and 0.614 ms in the reconstructed pitch were obtained for the closed and open tests, respectively. All inferred values of those affecting factors agreed well with our prior linguistic knowledge. Besides, the prosodic states automatically labeled by the pitch mean model provided useful cues to determine the prosodic phrase boundaries occurred at inter-syllable locations without punctuation marks. So it is a promising pitch modeling approach.

1. Introduction

Prosody is an inherent supra-segmental feature of human's speech. It carries stress, intonation pattern, and timing structure of continuous speech which, in turn, decide naturalness and understandability of the utterance. How to automatically generate, analyze and recognize the prosody in speech is one of the unresolved problems confronting many speech synthesis and recognition researches. Although it is known that prosody can be affected by many factors such as sentence type, syntactical structure, semantics and emotional status of speaker, the relationships between prosodic features and those affecting factors are not totally understood.

Among all features known to carry prosodic information, pitch is the most important one. It was reported that F0 contour characterizes speaking style and speaker [1]. An adequate pitch control is very important for synthetic speech to be natural in text-to-speech (TTS) [2]. The method of using the concept of separating an utterance's pitch contour into a global trend and a locally variational term is a popular pitch modeling approach, e.g., superpositional modeling [3] and two-stage modeling [4]. Pitch modeling is even more important for Mandarin speech. This is because Mandarin is a tone language and the information of the tonality of a syllable mainly appears on its pitch contour. Although there are only five lexical tones, syllable pitch contour patterns vary dramatically from their canonical forms (i.e., high-level tone, mid-rising tone, low-falling tone, high-falling tone, and low-energy tone) in continuous speech. So pitch modeling is a nontrivial research issue for Mandarin speech processing.

In this paper, a new pitch modeling approach for Mandarin speech is proposed. It takes the mean and shape of syllable pitch contour as two basic modeling units and models them separately by using statistical approaches to consider several affecting

factors that control their variations. These affecting factors include speaker, prosodic state, tone, and syllable initial and final classes. Here prosodic state is conceptually defined as the state in a prosodic phrase and is treated as a hidden variable. Due to the fact that prosodic states are not explicitly given, an expectation-maximization (EM) algorithm is derived to estimate all parameters of each model from a large training set. A by-product of the EM algorithm is the determinations of the hidden prosodic states of all syllables in the training set. This is an additional advantage because prosodic labeling has become a popular research topic recently [5].

The paper is organized as follows. Section 2 discusses the proposed pitch modeling approach for Mandarin speech in details. Section 3 describes the experimental results. An application to pitch prediction in TTS is given in Section 4. Some conclusions and possible future work are given in the last section.

2. The Proposed Approach

In the proposed pitch modeling approach, we first separate each syllable log-pitch contour into two parts, mean and shape, by a 3-rd order orthogonal polynomial expansion [2] with the zero-th order coefficient represents mean and the other three higher order coefficients represent shape. Note that log-pitch period instead of pitch period itself is chosen here because the dynamic range of log-pitch is almost the same for male and female speakers [6]. We then take syllable's pitch mean and shape as basic modeling units and employ two separate statistical models for them to consider several major affecting factors. We discuss them in detail as follows.

In the pitch mean model, it first considers the affecting factor of speaker by

$$Z_n = (Y_n + \beta_{s_n})\gamma_{s_n}, \quad (1)$$

where Z_n is the observed log-pitch mean of the n th syllable;

β_{s_n} and γ_{s_n} are two companding (compressing-expanding) factors (CFs) of the speaker affecting factor representing respectively the effects of speaker's level shift and dynamic range on Z_n ; and Y_n is the speaker-compensated log-pitch mean. The model then further considers other affecting factors by

$$Y_n = X_n + \beta_{t_n} + \beta_{p_n} + \beta_{f_n} + \beta_{i_n} + \beta_{f_n} + \beta_{p_n}, \quad (2)$$

where X_n is the normalized log-pitch mean of the n th (current) syllable and is modeled as a normal distribution with mean μ

and variance V ; β_r is the CF for affecting factor r ; t_n , pt_n and ft_n represent respectively the lexical tones of the current, previous and following syllables; i_n , f_n and p_n represent respectively the initial class, final class and prosodic state of the current syllable. Note that t_n ranges from 1 to 5 while both pt_n and ft_n ranges from 0 to 5 with 0 denoting the cases of punctuation mark and the non-existence of the previous or following syllable. The affecting factors for $pt_n = 0$ and $ft_n = 0$ are set to zeros because we don't want the affection of tone across punctuation mark.

To estimate the parameters of the model, an EM algorithm based on the ML criterion is adopted. The EM algorithm is derived based on incomplete training data with prosodic state being treated as hidden or unknown. To cure the drawback of resulting in a non-unique solution, we modify each optimization procedure in the maximization step (M-step) to a constrained optimization one via introducing a global constraint and the auxiliary function becomes

$$Q(\bar{\lambda}, \lambda) = \sum_{n=1}^N \sum_{p_n=1}^P p(p_n | Z_n, \bar{\lambda}) \log p(Z_n, p_n | \lambda) + \eta \left(\sum_{n=1}^N (\mu + \beta_{t_n} + \beta_{pt_n} + \beta_{ft_n} + \beta_{i_n} + \beta_{f_n} + \beta_{p_n}) \gamma_{s_n} - N \mu_Z \right), \quad (3)$$

where N is the total number of training syllables; P is the total number of prosodic states; $p(p_n | Z_n, \bar{\lambda})$ and $p(Z_n, p_n | \lambda)$ are conditional probabilities; $\lambda = \{\mu, \nu, \beta_t, \beta_{pt}, \beta_{ft}, \beta_i, \beta_f, \beta_p, \beta_s, \gamma_s\}$ is the set of parameters to be estimated; and λ and $\bar{\lambda}$ are, respectively, the new and old parameter sets. η is a Lagrange multiplier. The constrained optimization is finally solved by the Newton-Raphson method. Based on the assumption that the normalized log-pitch mean X_n is normally distributed, $p(Z_n, p_n | \lambda)$ can be easily derived from the assumed model given in Eqs. (1) and (2). Then, sequential optimizations of these parameters can be performed in the M-step.

To execute the EM algorithm, initializations of the parameter set $\bar{\lambda}$ are needed. This can be done by estimating each individual parameter independently. Specifically, the initial multiplicative/additive CF for a specific value of an affecting factor is assigned to be the ratio/difference of the mean of Z_n with the affecting factor equaling the value to the mean of all Z_n . Notice that, in the initializations of affecting factors of prosodic state, each syllable is pre-assigned a prosodic state by vector quantization. After initialization, all parameters are sequentially updated in each iteration. The iterative procedure is continued until a convergence is reached. The prosodic state can finally be assigned by

$$p_n^* = \max_{y_n} p(p_n | Z_n, \lambda), \quad (4)$$

The log-pitch shape model is expressed by

$$Z_n = X_n + \mathbf{b}_{tc_n} + \mathbf{b}_{p_n} + \mathbf{b}_{s_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n}, \quad (5)$$

where Z_n is the observed vector of orthogonal-expansion coefficients $[a_1 \ a_2 \ a_3]^T$ of the n th syllable log-pitch contour; X_n is the normalized vector of pitch shape coefficients of the n th syllable and is modeled as a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix \mathbf{R} ; \mathbf{b}_r is the CF vector for the affecting factor r ; tc_n represents the lexical tone combinations. When the coupling effect between syllables is tight, three-tone combination (previous, current and following tone) is considered. If the coupling is loose, only one or two-tone combination is considered. The strength of coupling effect is simply determined by pause duration.

A similar EM algorithm is used to solve the problem. An auxiliary function is defined by

$$Q(\bar{\lambda}, \lambda) = \sum_{n=1}^N \sum_{p_n=1}^P p(p_n | Z_n, \bar{\lambda}) \log p(Z_n, p_n | \lambda) + \mathbf{L}^T \left(\sum_{n=1}^N (\boldsymbol{\mu} + \mathbf{b}_{tc_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n} + \mathbf{b}_{p_n} + \mathbf{b}_{s_n}) - N \boldsymbol{\mu}_Z \right), \quad (6)$$

where \mathbf{L} is an $n \times 1$ Lagrange multiplier vector and $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \mathbf{R}, \mathbf{b}_{tc_n}, \mathbf{b}_{i_n}, \mathbf{b}_{f_n}, \mathbf{b}_{p_n}, \mathbf{b}_{s_n}\}$ is the set of parameters to be estimated. Again, $p(Z_n, p_n | \lambda)$ can be easily derived based on the normal distribution assumption of X_n . By minimizing the auxiliary function, we can get the optimal parameter set.

3. Experimental Results

Effectiveness of the proposed log-pitch modeling approach was examined by simulation using a multi-speaker microphone-speech database. The database was divided into two sets, one for training and the other for testing. The training set contained utterances of 455 sentences and 200 paragraphs uttered by four speakers including two males and two females. It contained, in total, 99,232 syllables. The test set contained utterances of 100 different paragraphs read by another female. The total number of syllables was 21980.

We first examined the effect of the log-pitch modeling with the number of prosodic states being set to 8. Assigning the best prosodic state to each syllable by Eq. (4), the RMSEs of the estimated and observed pitch are 0.551 and 0.614 ms for closed and open tests respectively. Notice that the RMSEs from the orthogonal transformation are 0.214 and 0.202 ms.

We then made some analyses to the inferred model for better understanding the effects of some important affecting factors. Before discussing the effects of the estimated factors, we give a brief introduction to the priori knowledge of tone pattern in the following. As reported in [7], a high-level tone, or Tone 1, starts in a speaker's high F0 range and remains high. A mid-rising tone, or Tone2, starts at the speaker's mid F0 range, remains level or

Table 1: The estimated CFs for the affecting factors of the current, previous and following tones in the pitch mean model.

tone	1	2	3	4	5
β_t	-0.150	0.052	0.160	-0.036	0.118
β_{pt}	-0.010	-0.011	0.006	0.009	0.011
β_{ft}	0.008	-0.004	-0.019	0.007	0.006

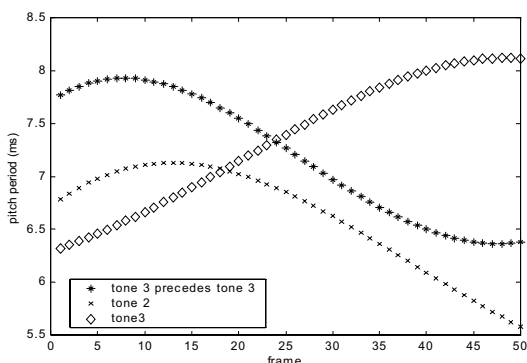


Figure 1: Compare a Tone 3 precedes another Tone 3 with canonical Tone 2 and 3.

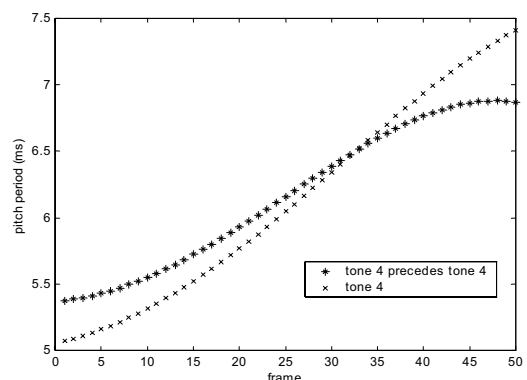


Figure 2: Compare a Tone 4 precedes another Tone 4 with canonical Tone 4.

drops slightly during the first half of the vowel, and rises up to high at the end. A low-falling tone, or Tone 3, starts at the speaker's mid range and falls to the low range. A high-falling tone, or Tone 4, usually peaks around the vowel onset, then falls to the low F0 range at the end. A low-energy tone, or Tone 5, has a relatively arbitrary pitch contour pattern. Table 1 shows the CFs of the affecting factors of the current, previous and following tones in the pitch mean model. It can be found from Table 1 that the CFs of the affecting factors of the current syllable are negative valued for Tones 1 and 4 and positive valued for the other 3 tones. Due to the fact that the effect of a positive (negative) CF is to decrease (increase) the F0 mean, the CFs of the affecting factors of the current tone matched well with the priori phonetic knowledge discussed above. It was also

Table 2: The estimated CFs for affecting factors of 7 initial/final classes in the pitch mean model.

class	0	1	2	3	4	5	6
β_i	-0.007	0.002	0.013	-0.015	0.003	-0.014	0.003
β_f	0.011	0.006	-0.0004	0.005	-0.007	-0.015	-0.001

reported in [8] that all tones before a Tone 3 have a much higher F0 level than when they precede other tones, and all tones are slightly lowered before a Tone 1. Besides, all tones after Tone 1 and Tone 2 have a higher F0 level than after Tone 3 and Tone 4. Those phenomena match with the results shown in Table 1 that a negative affecting factor was obtained when the following tone is a Tone 3, and a positive affecting factor was obtained when the following tone is a Tone 1. It is also observed that when the previous tone is a Tone 1 or 2, the negative affecting factor increases the F0 mean, and when the previous tone is a Tone 3 or 4, the positive affecting factor decreases the F0 mean. Combined with the pitch shape model, we also found out very promising results from the CFs of some tone combination affecting factors. Some famous sandhi rule [9] can be observed easily from those CFs. For examples, we know when a Tone 3 precedes another Tone 3, it is pronounced approximately as a Tone 2; and when a Tone 4 precedes another Tone 4, its pitch slope will be decreased [9]. Those two phenomena can be clearly observed from Figs. 1 and 2 in which tone patterns are drawn based on our model without counting factors other than tone.

Table 2 shows the estimated CFs for the affecting factors of 7 initial/final classes in the pitch mean model. Initial classes include 0: null initial, 1: {b, d, g}, 2: {f, s, sh, shi, h}, 3: {m, n, l, r}, 4: {ts, ch, chi}, 5: {p, t, k}, 6: {tz, j, ji}. Final classes include 0: low vowel, 1: middle vowel, 2: high vowel, 3: compound vowel, 4: nasal ending, 5: retroflexion, 6: null vowel. After deconfounding the effects of other factors, it can be found from Table 2 that positive CFs for {b, d, g}, {f, s, sh, shi, h}, {ts, ch, chi} and {tz, j, ji} lower syllable F0 mean while all others raise syllable F0 mean. Besides, positive CFs of low vowel, middle vowel and compound vowel lower syllable F0 mean, and negative CFs of high vowel, nasal ending, retroflexion, and null vowel raise syllable F0 mean.

Another very interesting characteristics of our model is the determination of prosodic states which are linguistically meaningful. It is well-known that a global downtrend tendency of F0 is to decline over the course of an utterance [10]. We also know that a slight pitch reset of the bottom line of intonation will occur at the prosodic word boundaries and a significant pitch reset of the bottom line of intonation will occur at the prosodic phrase boundaries and intonational phrase boundaries [11]. In our model, we found that the CFs of prosodic states in a prosodic phrase were generally varies from small to large values, and got reset when crossing prosodic phrase boundaries. Therefore, the change of states' CFs from large to small may indicate the possible prosodic phrase boundaries. In Fig. 3, examples showing the possible minor, normal, and major prosodic phrase boundaries via judging from the states' CF change are presented. The results match well with our general knowledge. Hence, the

這位約翰霍普金斯大學名譽教授*在第一屆國際&性高潮會議中說*，他對這一始於&一九八〇年代的性趨勢&感到難過*。

這場比賽#將於今日下午2時&在&台北&市立棒球場舉行*，黑鷹組織&所屬&三級棒球隊#，包括台南六信*、台東農工#、屏東&鶴聲國中#、台東鹿野國中&及台南善化國小等隊*，將各著球隊服裝到場加油*，預計人數有近千人以上*。黑鷹兩位教練*黃永裕及江泰權*，對於&此場比賽*不敢掉以輕心*，除了排出鑽石陣容外#，也要親自上場*。

商人非法囤積&大量爆竹*，萬一發生爆炸事件*，不但會造成&死傷慘劇*，自己也可能成為&受害最大的當事人*。

證管會將於&二月二十日邀請台灣&證券交易所*、證券商公會*、發行公司*、承銷商&、會計師&以及&證券市場發展基金會等單位代表#舉行公聽會#，將針對&修正草案*，進行可行性研討*。

Figure 3: Examples showing the possible minor (&), normal (#), and major (*) prosodic phrase boundaries.

prosodic states automatically labeled by the pitch mean model provided useful cues to determine the prosodic phrase boundaries occurred at inter-syllable locations without punctuation marks.

4. An Application to Pitch Prediction for TTS

We apply the above model to predict pitch period for Mandarin TTS. A hybrid statistical/regression approach to synthesizing pitch is suggested. Instead of direct predicting pitch from the input features, it first estimates the CFs of prosodic state from the linguistic features by the linear regression technique, and then predicts the pitch by the statistical model. Here linguistic features are extracted via analyzing the input text by an automatic word tokenization algorithm with an 80000-word lexicon. The linguistic features used include some sentence-level features, such as sentence length and position in sentence, some word-level features, such as word length and position in word, punctuation-mark indicators, and part of speech categories. RMSEs of 1.10 and 1.52 ms were obtained for the closed and open tests, respectively. The results are better than those of 2.55 and 2.05 ms achieved by the conventional regressive prediction method. It is noted that the RMSEs caused from orthogonal transformation are 0.214 and 0.202 ms for training and testing sets, respectively.

5. Conclusions

A new statistical-based syllable log-pitch modeling approach for Mandarin speech has been discussed in the paper. Experimental results have confirmed its effectiveness on isolating several main factors that seriously affect the pitch mean and shape of Mandarin utterances. The inferred CFs of affecting factors conformed well to the prior linguistic knowledge. Besides, the prosodic-state labels produced by the EM algorithm were linguistically meaningful. By taking the benefit of pitch modeling

using only acoustic features and simple phonetic features, we can apply the pitch model to applications, like speech recognition and prosodic labeling, which do not need priori high-level linguistic information such as word tokenization and syntactic features.

ACKNOWLEDGEMENT

This work was supported in part by MOE under contract EX-91-E-FA06-4-4 and in part by NSC.

REFERENCES

- [1] A. I. C. Monaghan and D. R. Ladd, "Manipulating Synthetic Intonation for Speaker Characterisation," *ICASSP*, S7.11, pp. 453 – 456, 1991.
- [2] S. H. Chen, S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech and Audio processing*, vol. 6, no.3, pp.226-239, 1998.
- [3] Jerome R. Bellegarda, Kim E. A. Silverman, Kevin Lenzo and Victoria Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Trans. Speech and Audio processing*, vol. 9, no.1, pp.52-66, January 2001.
- [4] Masanobu ABE and Hirokazu SATO, "Two-stage F0 Control Model Using Syllable Based F0 Units," *ICASSP*, pp. II-53 – II-56, 1992.
- [5] Colin W. Wightman, Mari Ostendorf, "Automatic Labeling of Prosodic Patterns", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, October 1994, pp. 469 – 481.
- [6] Thomas Eriksson and Hong-Goo Kang, "Pitch Quantization in Low Bit-Rate Speech Coding," *ICASSP*, pp. 489 – 492, 1999.
- [7] Chi-lin Shih, "Tone and Intonation in Mandarin," *Working Papers of the Cornell Phonetics Laboratory*, No. 3, pp.83 – 109, June 1988.
- [8] Chao Wang and Stephanie Seneff, "Improved Tone Recognition by Normalizing for Coarticulation and Intonation Effects," *ICSLP 2000*.
- [9] Lin-shan Lee, Chiu-yu Tseng, Ming Ouh-young, "The Synthesis Rules in a Chinese Text-to-speech System," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 9, September 1989, pp. 1309 – 1319.
- [10] Chilin Shih, "Declination in Mandarin," *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, Athens, Greece, pp. 293-296, 1997.
- [11] Yang Yufang and Wang Bei, "Acoustic Correlates of Hierarchical Prosodic Boundary in Mandarin," *Speech Prosody 2002*.