ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2012
Sixth International Conference
Shanghai, China
May 22-25, 2012

Tone Generation by Maximizing Joint Likelihood of Syllabic HMMs for Mandarin Speech Synthesis

Xingyu Na [1], *Chaomin Wang* [1], *Xiang Xie* [1], *Jingming Kuang* [1], *Yaling He* [2]

[1] School of Information and Electronics, Beijing Institute of Technology, China
[2] Eastel Corporation, Beijing, China
{naxy, wangchaomin, xiexiang}@bit.edu.cn, heyl@eastel.cn

Abstract

A tone generation method by maximizing the joint likelihood of syllabic HMMs is proposed to improve the Mandarin speech synthesis. F0 sequence is generated by jointly maximizing the likelihood of the state-level F0 model and syllable-level tone model under the constraint of mean F0 of the adjacent units. The optimal weight of the tone component is searched in terms of the parameter generation error and correlation coefficients. Objective and subjective evaluations both prove the positive effects of this method. The generation error is reduced by 26.7%, the correlation coefficient is increased by 6.5%, and the prosody perception is significantly improved.

Index Terms: speech synthesis, F0 contour, tone generation, speech prosody

10.21437/SpeechProsody.2012-9

1. Introduction

Tonal information is important for the understanding and prosodic perception of tonal languages, such as Mandarin, Cantonese and Thai, etc. From acoustic aspect, it is directly represented by the shape of fundamental frequency (F0) contour. HMM-based speech synthesis (HTS) algorithm has been proven to generate natural and intelligent speech [1]. In a typical HTS system, F0 contour is generated in a state-based maximum likelihood sense given the sentence HMM sequence of a labeled text [2]. Though F0 values could be modeled and generated under the constraint between static and dynamic features, known as the trajectory HMM [3], the long term tendencies of the F0 contours are not properly considered. Therefore, the prosody of generated speech sounds flattened.

There have been a lot of efforts towards improving the prosody of the generated speech using better tone modeling methods. Parameter generation considering global variance of the generated parameters was proposed to address the parameter flatten problem [4]. Consistent voicing condition modeling for dynamic F0 features [5] and continuous F0 modeling [6] are proposed to improve the conventional multi-space distribution modeling [7] in HTS. However, the state-based HMM is inadequate in modeling supra-segmental features, such as F0 contour. Another drawback is the decision-tree-based state tying is difficult to capture the underlying additive structure of features. Tonal labeling using F0 symbol is employed to improve the decision tree based state clustering [8]. F0 symbols of quantized values of segmented F0 contour are used in the state tying in order to capture the additive information. The generated tone of the professional speaker, which is considered to be the standard tone predictor, will highly affect the prosody of the target speaker. To address the issue of additive structure, a multi-layer F0 model was proposed to capture the underlying pitch patterns of different prosodic layers [9]. A modified prosody generation based on F0 contour parameterization was proposed to model the additive structure explicitly [10]. Compared with the piecewise modeling methods, F0 contours are hierarchically parameterized and modeled as a long term prosody feature in this method. Phone-based HMMs are used and syllabic tone and phrase intonation are jointly modeled to improve the generated prosody. In the training stage, F0 values are normalized using the mean of the voiced segments of an utterance. In all these methods, tone generation errors are reduced and prosody perceptions are improved.

However, seldom investigation has been implemented for improving tone modeling of syllable-based tonal languages synthesis. Among phoneme, syllable and word, syllable is the longest unit of a continuous, stable articulation phase, as well as the shortest unit in the hierarchical prosody structure in most tonal languages. In previous researches, syllable has been proven to be a more robust modeling unit for Mandarin speech synthesis in large data sets [11]. Unlike phonemes, syllables carry meanings as well as prosody in mono-syllabically paced tonal language [12], which suggests it as a fine modeling unit in speech synthesis.

In this paper, we propose a tone generation method for improving prosody perception of syllable based Mandarin speech synthesis. In addition to the syllabic state models, F0 contours are jointly generated under the constraint between the state-based F0 model and the syllabic tone model, which is estimated using the parameterization of F0 contour.

The rest of this paper is organized as follows. In Section 2 and 3, the tone generation methods in syllable based Mandarin HTS are described. Experiments and analysis of the methods are presented in Section 4. Section 5 concludes the paper.

2. Tone modeling of syllabic HTS

In statistical parametric speech synthesis, speech waveforms are parameterized and reconstructed using acoustic parameters. Vocal tract and vocal cord oscillation are firstly represented by spectrum envelope and fundamental oscillation frequencies respectively. These raw features are then parameterized, such as mel-cepstral coefficients, linear prediction coefficients and logarithm F0. Tone, as a long-term feature, is more sensitive to F0 variations than frame based pitch values. Therefore, F0 should be modeled considering the joint likelihood of the contours of different time spans.

2.1. State level F0 model

In typical HTS systems, F0 is modeled by frame logarithm observations. Each state of the HMM is supposed to emit one F0 value at each occupation, of which the count is firstly determined by the duration model. Static and dynamic features are composed and trained jointly using the multi-space distribution (MSD) HMMs to discriminately modeling the voiced and unvoiced components of F0 [1].

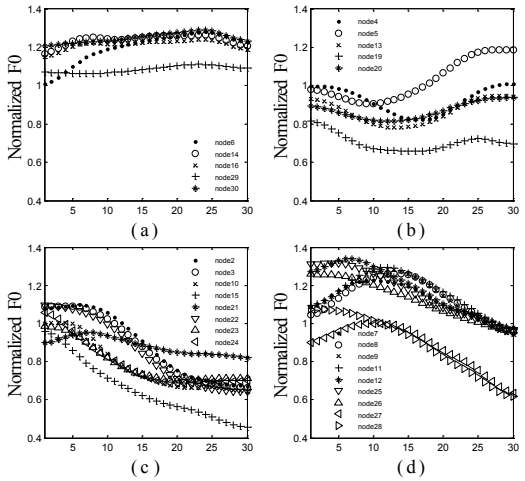

Figure 1: *Tone clustering result using CART. (a), (b), (c) and (d) are reconstructed F0 contours of the tree nodes of tones level, rising, falling-rising and falling respectively.*

However, the voicing condition of dynamic and static features on the boarders of voicing segments are inconsistent, which reduces the accuracy of F0 modeling. Consistent voicing condition modeling [5] was applied in state level F0 models to address this issue.

2.2. Syllabic level tone model

Considering the need for time warping and dimensional reduction, tone model of syllabic level should be trained using parametric representation of contours other than logarithm frame values. Discrete cosine transformation (DCT) has been proven to be a robust parameter for F0 contour modeling for its fitting performance and inverse transformation complexity [10]. The DCT of F0 contour is

$$c_d = \frac{2}{T}\sum_{t=0}^{T-1} f_t \cos\left[\frac{\pi}{T} d\left(t+\frac{1}{2}\right)\right], \quad d = 0,\ldots,D-1 \tag{1}$$

where $f_0, \ldots, f_{T-1}$ are F0 values of length T and represented by D coefficients of DCT , $c_0, \ldots, c_{D-1}$. The F0 values to be transformed need to be divided by the mean of the F0 values of the voiced frames through the whole sentence. The inverse DCT is defined as

$$f_t = \frac{1}{2}c_0 + \sum_{d=1}^{D-1} c_d \cos\left[\frac{\pi}{T} d(t+\frac{1}{2})\right], t = 0,\cdots,T-1 \tag{2}$$

Context dependent HMM with single Gaussian component is used to model the syllabic level DCTs. To capture the pitch variation between adjacent syllables, dynamic features of the first order DCT coefficients are calculated because they represent the mean values of the normalized F0s. We use a classification and regression tree (CART) to cluster the context dependent models. The questions for splitting the nodes are automatically chosen in a maximum likelihood (ML) sense. Minimum description length (MDL) criterion is used as the stopping criterion for balancing the modeling error and the complexity [13].

We use nearly 19000 syllables from 1000 utterances of a female speaker to check the tone clustering performance of CART. In Mandarin, tone patterns are characterized by their F0 trajectory shapes of the final part with the four names of level, rising, falling-rising and falling. The patterns are used as a part of the question set in decision tree growing. Notably, tone patterns are automatically selected as the questions at the first four splitting nodes of the decision tree, so we can easily figure out which tone a node belongs to. As shown in Figure 1, there are totally 28 leaf nodes, of which 5, 5, 8 and 9 nodes are clustered as level, rising, falling-rising and falling respectively and there is one leaf for silence and short pause whose DCT coefficients are zero. The sensitivity to context of rising, falling-rising and falling tones are captured by CART according to the differences of the starting value, ending value and turning-point position of the contours.

3. Tone generation of syllabic HTS

Sentence HMM is concatenated given the labeled text in HTS. State-based acoustic parameters aligned by the duration model are generated under constraint between static and dynamic features to physically simulate vocal cord and vocal tract. Synthesis filter is stimulated by the excitation generator to output speech waveforms.

3.1. State level F0 generation

In the synthesis part of typical HTS, logarithm F0 values are generated based on ML under the constraint of dynamic features of logarithm F0 models. Given the state level F0 model λ_s and state sequence $\mathbf{q}_s$, observation probability is defined as

$$\begin{aligned} P(\mathbf{f} \mid \lambda_s) &= P(\mathbf{W}_s\mathbf{f} \mid \mathbf{q}_s, \lambda_s) \\ &= N(\mathbf{W}_s\mathbf{f} \mid \boldsymbol{\mu}_{\mathbf{q}_s}, \boldsymbol{\Sigma}_{\mathbf{q}_s}) \end{aligned} \tag{3}$$

where $\mathbf{W}_s$ is the window matrix of dynamic features and $\boldsymbol{\mu}_{\mathbf{q}_s}$ and $\boldsymbol{\Sigma}_{\mathbf{q}_s}$ are the means vectors and covariance matrices of the sentence HMM. To calculate $\mathbf{f}$, we set

$$\frac{\partial \log P(\mathbf{f} \mid \lambda_s)}{\partial \mathbf{f}} = 0 \tag{4}$$

and get

$$\mathbf{W}_s^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_s}^{-1}\mathbf{W}_s \cdot \mathbf{f} = \mathbf{W}_s^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_s}^{-1}\boldsymbol{\mu}_{\mathbf{q}_s} \tag{5}$$

By solving this linear formula, we can obtain state level F0s.

3.2. Syllabic level tone generation

Tone contour is generated by the DCT model with dynamic feature constraint in the ML sense. For a given syllabic DCT model λ_y and syllable sequence $\mathbf{q}_y$, observation probability is defined as

$$\begin{aligned} P(\mathbf{f} \mid \lambda_y) &= P(\mathbf{W}_y\mathbf{D}_y\mathbf{N}\mathbf{f} \mid \mathbf{q}_y, \lambda_y) \\ &= N(\mathbf{W}_y\mathbf{D}_y\mathbf{N}\mathbf{f} \mid \boldsymbol{\mu}_{\mathbf{q}_y}, \boldsymbol{\Sigma}_{\mathbf{q}_y}) \end{aligned} \tag{6}$$

where $\mathbf{W}_y$ is the window matrix of dynamic features and $\boldsymbol{\mu}_{\mathbf{q}_y}$ and $\boldsymbol{\Sigma}_{\mathbf{q}_y}$ are the means vectors and covariance matrices of the syllabic models. The bases of DCT transformations are smooth orthogonal functions, which makes it possible to be represented as the transformation matrix $\mathbf{D}_y$. $\mathbf{N}$ is the

normalization matrix of syllabic F0 contours based on mean of sentence F0 values.

The probability of F0 contour is the joint probability of state level F0 and syllabic level DCT, i.e.,

$$P(\mathbf{f} \mid \lambda_s, \lambda_y) = P(\mathbf{W}_s\mathbf{f} \mid \mathbf{q}_s, \lambda_s) \cdot P(\mathbf{W}_y\mathbf{D}_y\mathbf{N}\mathbf{f} \mid \mathbf{q}_y, \lambda_y)^{\alpha} \quad (7)$$

where α is the weight of syllabic tone model which works as a control factor in synthesis phase. Setting the derivation of the joint likelihood as

$$\frac{\partial \log P(\mathbf{f} \mid \lambda_s, \lambda_y)}{\partial \mathbf{f}} = 0 \quad (8)$$

we get

$$[\mathbf{W}_s^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_s}^{-1}\mathbf{W}_s + \alpha \cdot (\mathbf{W}_y\mathbf{D}_y\mathbf{N})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_y}^{-1}\mathbf{W}_y\mathbf{D}_y\mathbf{N}] \cdot \mathbf{f} = \mathbf{W}_s^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_s}^{-1}\boldsymbol{\mu}_{\mathbf{q}_s} + \alpha \cdot (\mathbf{W}_y\mathbf{D}_y\mathbf{N})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{q}_y}^{-1}\boldsymbol{\mu}_{\mathbf{q}_y} \quad (9)$$

In the synthesis phase a previous joint modeling method, the generated F0 values are multiplied by the mean of F0s in the training data [10]. This inconsistency between the scale factor used in normalization and inverse normalization would increase the generated pitch error and cause discontinuity between two voiced units. In the proposed syllabic level tone modeling and generation method, normalization matrix **N** is calculated using the means of the utterance F0 values. Therefore, the variation of pitch level and the prosody of the generated speech would be improved.

4. Experiments

4.1. Experimental setup

A female Mandarin speech database containing 5200 syllable-balanced utterances are used in our experiments. Sampling rate of recorded speech waveforms is 16 kHz. 5000 sentences were selected from the database for model training, and 100 sentences are used as developing set. The remains were used as testing data.

The spectral analysis is performed at 5-ms shift using the spectral envelopes estimated by STRAIGHT [14] and represented by 18 dimensional linear spectrum pair (LSP) coefficients. F0 values were extracted using the one best selection of the results of a robust algorithm for pitch tracking (RAPT) [15] and the STRAIGHT pitch tracker without manual check, and then passed through a five-point median filter. Considering the length of F0 real contour of one syllable is usually less than 50 frames, seven-point DCT could guarantee a fine approximation [10]. Ten-state left-to-right syllabic HMMs are adopted in our baseline system. Static and dynamic features of F0 are aligned using voicing conditions and modeled in a single stream to keep a consistency with the state level F0 model described in Section 2.1.

The rich phonetic and prosodic contexts used in growing decision trees include: tri-syllable, tri-tone; type of initial and final, method and place of articulation [16]; the position of syllable and word in phrase and sentence; and the length of word, phrase and sentence in number of syllable, word and phrase. The same question set is used for LSP, F0 and tone modeling. For the tone model, MDL factor is set to 0.5 and the occupancy floor of each leaf node is set to 5 to obtain a more elaborate tone model. The numbers of leaf nodes in the decision trees are shown in Table 1. The MDL factor and occupancy floor of the state-level models are set as 1.0 and 10. The baseline system uses the state-level models only.

4.2. Evaluation results and analysis

Objective and subjective evaluations are implemented to test the performance of the proposed method. In the objective measure, RMSE and correlation coefficients between the original and generated F0 sequences are calculated over all voiced frames aligned to the syllabic duration of the developing and the testing data. Voicing decisions are made by the MSD weight of the baseline system and modified using a GMM-based voicing condition optimization method [17].

The developing set was used to find the optimal α for maximizing the joint probability of state and syllabic models. In Figure 2 (a) and (b), RMSE and correlation are independently considered, in which the optimal values are 0.7 and 0.6 respectively. Meanwhile, it can be observed that the error increases when a larger α is used. We find that the course for the increase of RMSE is the F0 discontinuity between different syllables. Although dynamic features of the mean F0 value are considered in the generation of tones, they are modeled independently to the static features. Besides, higher order DCT coefficients reflect the absolute value of the F0 contours as well. The correlation coefficients also decrease when using a large α, and this is because the averaged tone patterns cannot retain the micro-prosody inside a syllable, especially at the boundaries of the voiced segments. In Figure 2 (c), the two metrics are jointly considered, which results in a joint optimal α of 0.6.

Table 1. *Number of leaf nodes of different feature decision trees.*

Feature (level)	Number of leaf nodes
LSP (state)	7689
F0 (state)	17054
DCT (syllable)	309

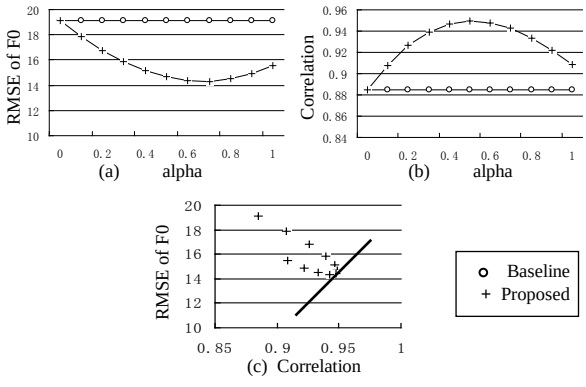

Figure 2: *Optimal value search of α for maximizing the joint likelihood of tone DCT and state F0. (a) RMSE of F0. (b) Correlation. (c) Joint search using RMSE and correlation. The bold solid line represents the joint threshold.*

Table 2. *Objective evaluation results in terms of RMSE of F0 and correlation.*

Measure	System	Results
RMSE of F0 [Hz]	Baseline	21.745
	Proposed	15.934
Correlation	Baseline	0.881
	Proposed	0.938

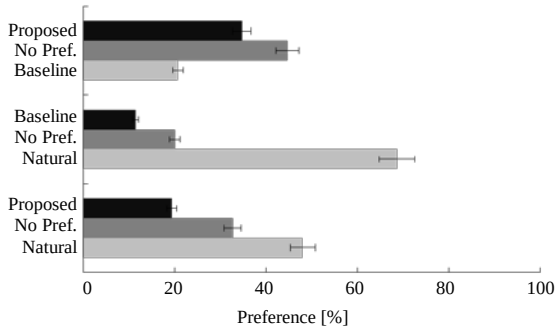

Figure 3: *Preference evaluation result.*

By fixing the tone model weight α to be 0.6, objective and subjective evaluations are performed. RMSE of F0 and contour correlation are measured on the testing set. Table 2 shows the objective test results comparing the baseline system and the proposed method. The proposed method reduces the RMSE of generated F0 by 26.7% and improves the correlation by 6.5%. The perception performance is further evaluated by an AB preference test participated by 6 listeners. 50 utterances randomly chosen from the testing set and synthesized by the baseline and proposed systems are cross evaluated and the subjects are asked for a choice according to their prosody perception among three options 1) the former is better; 2) the latter is better; 3) no preferences. As shown in Figure 3, the proposed method is perceptually preferred than the baseline system. Additionally, the generated prosody is evaluated compared to natural prosody. To avoid the influence of spectral distortion, original LSP sequences extracted from the testing data are used together with generated F0 sequences aligned to original LSP. It is shown in the lower of Figure 3 that the prosody of generated speeches is significantly improved using the proposed method.

5. Conclusions

In this paper, we propose a tone modeling and generation method for syllable-based Mandarin HTS. The likelihoods of conventional F0 model and syllabic tone model are jointly maximized in the F0 generation procedure. In our experiments, we firstly search the optimal weight of the tone component in joint likelihood in terms of RMSE of F0 and correlation coefficients, which can be seen as two main measures of the generated prosody. Objective and subjective evaluations show that the tone is a crucial factor for prosody perception of Mandarin speech synthesis. The subjective perception is magnificently improved by the proposed method. Further, we intend to integrate phrase level intonation modeling into the syllable-based Mandarin speech synthesis.

6. Acknowledgements

This work was supported by the National Science and Technology Major Projects (Grant No.2010ZX03004-003-01), National Natural Science Foundation of China (Grant No. 11161140319, No. 90920304 and No. 91120015) and Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20101101110020).

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, “Statistical Parametric Speech Synthesis”, Speech Communication, 51(11): 1039-1064, 2009.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, in ICASSP, 2000.

[3] H. Zen, K. Tokuda, and T. Kitamura. “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”, Comput. Speech Lang., 21(1):153–173, 2006.

[4] T. Toda and K. Tokuda. “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE Trans. Inf. Syst., E90-D(5):816–824, 2007.

[5] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A pitch pattern modeling technique using dynamic features on the border of voiced and unvoiced segments”, Technical report of IEICE, 101(325): 53–58, 2001.

[6] K. Yu and S. Young, “Continuous F0 modelling for HMM based statistical speech synthesis”, IEEE Transactions on Audio, Speech and Language Processing, 19(5):1071-1079, 2011.

[7] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, “Multispace probability distribution HMM”, IEICE Trans. Inf. & Syst., E85-D(3):455–464, 2002.

[8] V. Chunwijitra, T. Nose, and T. Kobayashi, “Tonal context labeling using quantized F0 symbols for improving tone correctness in average-voice-based speech synthesis”, in ICASSP, 2011.

[9] M Lei., Y. Wu, Z. Ling and L. Dai, "Investigation of Prosodic F0 Layers in Hierarchical F0 Modeling for HMM-based Speech Synthesis", in ICSP, 2010.

[10] Y Qian, Z. Wu, B. Gao and F. K. Soong, "Improved Prosody Generation by Maximizing Joint Probability of State and Longer Units", IEEE Audio, Speech, and Language Processing, 19(6):1702-1710, 2011.

[11] Q. Duan, S. Kang, Z. Wu, L. Cai., Z. Shuang and Y. Qin., "Comparison of Syllable/Phone HMM Based Mandarin TTS", in ICPR, 2010.

[12] Y. Li, T. Lee, and Y. Qian, “F0 analysis and modeling for Cantonese text-to-speech”, in Speech Prosody, 2004.

[13] K. Shinoda and T. Watanable, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn(E), 21(2): 79–86, 2000.

[14] H. Kawahara, I. M. Katsuse, and and A. D. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds”, Speech Communication, 27(3–4), pp. 187–207, 1999.

[15] A. D. Talkin, “chapter A robust algorithm for pitch tracking (RAPT),”in Speech Coding and Synthesis. Amsterdam, The Netherlands: Elservier, 1995.

[16] X. Chen, A. Li., G. Sun., W. Hua and Z. Yin, "An Application of SAMPA-C for standard Chinese". in ICSLP, 2000.

[17] S. Kang, Z. Shuang, Q. Duan, Y. Qin and L. Cai, “Voiced/ Unvoiced Decision Algorithm for HMM-based Speech Synthesis”, in Proc. of InterSpeech, 2009.