IMPROVED PROSODY GENERATION BY MAXIMIZING JOINT LIKELIHOOD OF STATE AND LONGER UNITS

Yao Qian[1], *Zhizheng Wu*[1,2*], *Frank K. Soong*[1]

[1]Microsoft Research Asia, Beijing, China, [2]College of Software, Nankai University, China

{yaoqian,frankkps}@microsoft.com

ABSTRACT

The current state-of-art HMM-bsed TTS can produce highly intelligible output speech and deliver a decent segmental quality. However, its prosody, especially at the phrase or sentence level, tends to be bland. The blandness of synthesized prosody is partially due to the fact that a state-based HMM is rather inadequate in modeling a global, hierarchical prosodic structure at a sentence or phrase level. In this study, the prosody of longer units are first modeled explicitly by appropriate parametric distributions. The resultant models are then integrated with the state-level baseline models to generate an optimal prosody by maximizing the joint likelihood of all, from state to longer, units. Experimental results in both Mandarin and English show consistent improvements over the state-based baseline system. The improvements are both objectively measurable and subjectively perceivable.

Index Terms— HMM-based TTS, Duration modeling, Pitch Modeling, Gamma distribution, DCT

1. INTRODUCTION

HMM-based TTS models spectral envelop, fundamental frequency, and duration simultaneously by the corresponding HMMs. For a given text sequence, speech parameter trajectories can then be generated from trained HMMs in the Maximum Likelihood (ML) sense [1].The speech generated from it is fairly smooth and exhibits no apparent glitches. However, overly-smoothed parameter trajectories tend to make synthesized speech sound less lively than natural.

Many research attempts have been tried to reduce over-smoothing of trajectory model and the resultant degraded synthesized speech quality. In [2], a parameter generation algorithm is proposed by considering the global variance (GV) of generated parameters. The probability of GV is used as a penalty for the reduced variance of generated trajectory. An extension which applies Gaussian mixture model to the GV term is used to improve the quality of an HMM-based polyglot speech synthesizer [3]. A trajectory model by imposing the explicit relationship between static and dynamic features was also proposed [4]. Minimum generation error is used as an alternative criterion in HMM parameter optimization [5]. It tries to adjust HMM parameters trained by the conventional EM algorithm to minimize the generation error between synthesized and original parameter trajectories in training data.

With the above improvements for overly-smoothed parameter trajectories, the segmental quality of synthesized speech is quite acceptable, while prosody, especially at the phrase and sentence level, still tend to be somewhat bland. As we know, Prosodic features are suprasegmental and have hierarchical structure. A state-based HMM is rather inadequate in modeling a global, hierarchical prosodic structure at a sentence or phrase level. Additionally, the prosody of long-term units should be modeled with more appropriate parametric distribution. Gamma distribution, which can model random variables with only positive semi-definite distributions, is more appropriate to duration modeling [6, 7, 8]. Discrete cosine transform (DCT), which expresses a finite signal in terms of a sum of cosine functions oscillating at different frequencies, is a good parametric representation to characterize F0 curves for F0 modeling [9, 10, 11].

In this paper, we use Gamma distribution to model durations on phone and syllable levels, and DCT to parameterize the F0 curves on syllable and phrase levels. The high-level models are integrated with state-level model in the generation procedure where their joint likelihood are maximized. The High-level model integration is similar to GV constraint [2] and utterance length constraint to parameter generation[12].

2. PROSODY MODELING AND GENERATION IN CONVENTIONAL HMM-BASED TTS SYSTEM

In conventional HMM-based TTS system, the state duration of a standard HMM is explicitly modeled with a single Gaussian distribution which is estimated by using state occupancy counts in the Baum-Welch re-estimation procedure [12]. F0 features are modeled by MSD-HMM [13], which can model the piece-wise continuous F0 trajectory stochastically. MSD models two, discrete and continuous probability spaces: discrete for unvoiced regions and continuous for voiced F0 contours. It models F0 in a different stream separated from the spectral feature stream.

In the synthesis part, the parameters are generated based on maximum likelihood (ML). The state duration is the mean of the corresponding Gaussian distribution. F0 trajectory is generated with dynamic feature constraint. For a given HMM model λ, it determines a sequence of F0, $F = f_0, ..., f_{T-1}$, which maximizes $\log P(O|\lambda)$ with respect to $O = WF$. W is static, delta and delta-delta coefficient matrix. If the state sequence Q is given by state duration, we set

$$\frac{\partial \log P(WF|Q,\lambda)}{\partial F} = 0 \tag{1}$$

and obtain

$$W^T U^{-1} W F = W^T U^{-1} M \tag{2}$$

where U and M are covariance matrix and mean vector of F0.

* Work performed as an intern in the Speech Group, Microsoft Research Asia

3. PROSODY MODELING FOR LONGER UNITS

Richer prosodic contexts are used to capture prosody co-articulation effects in HMM modeling. In practice, limited by insufficient training data, we usually have to cluster models of long contexts into generalized ones to predict unseen contexts in test robustly. State tying via a clustered classification and regression tree (CART) is commonly used in conventional HMM-based TTS. CART is an effective data mining tool which can efficiently handle messy data, missing values, or predictor variables measured in different scales. However, it has some limitations, e.g. difficulty in capturing underlying additive structure of the data [14]. The additive structure of prosody is commonly observed across different languages. Multi-layer models should be used to model the hierarchical structure of prosody components. The decision tree-based state tying is inappropriate to model hierarchical prosodic structure at a sentence or phrase level, even the questions about these high-level prosody components are already included in the question set to split nodes in decision tree growing. Therefore, we propose to model the prosody of longer units explicitly and integrate them with state-level model in the parameter generation procedure. Additionally, the prosody of longer units should be modeled with more appropriate parametric distribution.

In [8], we compared gamma and Gaussian distributions in their model fitting to the duration distributions of longer units. The distributions (histograms) of phone and syllable durations resemble gamma more than Gaussian distributions. The Chi-Square test of goodness-of-fit for duration distributions in each leaf node of decision tree also shows most of leaf nodes which gamma fit better. We use Gamma distribution for modeling phone and syllable durations. It has the form of

$$p(\boldsymbol{x}) = \frac{1}{\Gamma(a)b^a}\boldsymbol{x}^{a-1}e^{-\boldsymbol{x}/b} \tag{3}$$

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$. The expectation and variance of random variable $\boldsymbol{x}$ under gamma distribution are $E(\boldsymbol{x}) = ab$ and $Var(\boldsymbol{x}) = ab^2$, where a and b, $a = \mu^2/\sigma^2$, $b = \sigma^2/\mu$, are functions of μ and σ^2, the mean and variance of duration for a leaf node.

In [10], we investigated the parametric representations for F0 contours. The fitting errors in term of the number of coefficients show discrete cosine transform (DCT) is a more appropriate parametric representation to the F0 curves. We use DCT to parameterize the F0 contours on both syllable and phrase levels. The most common DCT definition is

$$c_n = \frac{2}{M}\sum_{m=0}^{M-1} s_m \cos[\frac{\pi}{M}n(m+\frac{1}{2})], n = 0, ..., N-1 \tag{4}$$

where $s_0, ..., s_{M-1}$ is a finite signal of length M and represented by N coefficients of DCT, $c_0, ..., c_{N-1}$. Similarly, the inverse transformation is defined as

$$s_m = \frac{1}{2}c_0 + \sum_{n=1}^{N-1} c_n \cos[\frac{\pi}{M}n(m+\frac{1}{2})], m = 0, ..., M-1 \tag{5}$$

4. PROSODY GENERATION BY MAXIMIZING THE JOINT LIKELIHOOD OF DIFFERENT UNITS

State durations can be estimated more precisely if they can be regulated by the durations of longer and higher level units like phone and syllables. The likelihood of state durations is jointly maximized in conjunction with the weighted likelihood of phone and syllable durations. The log likelihood of duration $L(D)$ is defined as

$$L(D) = \sum_j \Bigg[\sum_n \Bigg[\sum_k \log p_{j,n,k}(d_{j,n,k}) + \alpha \log p_{j,n}(d_{j,n})\Bigg] + \beta \log p_j(d_j)\Bigg] \tag{6}$$

subject to

$$\sum_k d_{j,n,k} = d_{j,n} \quad \text{and} \quad \sum_n d_{j,n} = d_j$$

where $d_{j,n,k}$ is the duration of state k, phone n, and syllable j. Correspondingly, $p_{j,n,k}(\cdot)$ is the probability density function of $d_{j,n,k}$. $p_{j,n}(\cdot)$ and $p_j(\cdot)$ are similarly defined. Two parameters, α and β, are to weight phone and syllable durations likelihood. We use gamma distributions for modeling phone and syllable durations to refine Gaussian model of state duration. To maximize likelihood $L(D)$, we set

$$\frac{\partial L(D)}{\partial d_{j,n,k}} = 0 \tag{7}$$

Limited by space, the detailed solutions to maximize likelihood $L(D)$ is given in [8]. If we set β =0 and use Gaussian distribution to model phone duration, the solution is the same as in [15].

Similar to duration, the likelihood of the F0 trajectory is jointly maximized in conjunction with weighted likelihoods of syllable and phrase contours. The log likelihood of F0 trajectory $L(F)$ is defined as

$$L(F) = \log P(O_s|Q_s, \lambda_s) + \alpha \log P(O_y|\lambda_y) + \beta \log P(O_h|\lambda_h) \tag{8}$$

with respect to

$$O_s = W_s F, \quad O_y = W_y D_y F \quad \text{and} \quad O_h = D_h A F$$

where λ_s is HMM parameters on state-level, λ_y and λ_h are DCT model parameters on syllable-level and phrase-level; Q_s is state sequence given by duration model. Voiced/unvoiced decision for each frame is given by state-level MSD. D_y is the DCT matrix for F0 contour on voiced part of syllable. On the phrase level, DCT matrix D_h is performed on the F0 mean of each consistent syllable. A is the matrix to get the mean of F0s on each syllable; To make F0 trajectory local continuous, W_s, the static, delta and delta-delta coefficient matrix, is used to calculate frame-level dynamic feature; To capture the phrase intonation and make neighboring syllable-level F0 contours globally continuous, W_y is the matrix to get the dynamic features of DCT first coefficient, which represents the mean of F0 curve on syllable; α and β are the parameters to weight the likelihood of syllable-level and phrase-level DCT models. When we set α=0 and β=0, only state-level is considered.

To maximize likelihood $L(F)$, we set

$$\frac{\partial L(F)}{\partial F} = 0 \tag{9}$$

and obtain the solution as

$$\begin{aligned}\{W_s^T U_s^{-1} W_s + \alpha[(W_y D_y)^T U_y^{-1} W_y D_y] \\ +\beta[(D_h A)^T U_h^{-1} D_h A]\} F \\ = W_s^T U_s^{-1} M_s + \alpha[(W_y D_y)^T U_y^{-1} M_y] \\ +\beta[(D_h A)^T U_h^{-1} M_h]\end{aligned} \tag{10}$$

where U_s and M_s are covariance matrix and mean vector of F0s on state-level; U_y, U_h, M_y, M_h are covariance matrices and mean vectors of DCT coefficients on syllable and phrase levels.

5. EXPERIMENTS AND RESULTS

5.1. Experimental setup

Two phonetically and prosodically rich, speaker-dependent continuous speech corpora in American English and Mandarin Chinese are used in our experiments. Both corpora were recorded by professional female speakers in broadcast news style. The two corpora are divided into three sets: training, developing and testing parts with corresponding sizes in number of sentences given in Table 1. The training set is used for training the prosody model. Developing set is used to determine the appropriate weights (α, β) in the Equations (6) and (8). The testing set is used to measure the performance of our improved prosody generation algorithm.

Table 1. *Training, developing and testing sets (# sentences) of the two speech corpora.*

# sentences	training	developing	testing
English	900	100	100
Mandarin	1,000	100	100

Speech signals are sampled at 16kHz, windowed by a 25-ms window with a 5-ms shift. The 40th order LPC spectral features are transformed into static LSPs and their dynamic counterparts. F0 is extracted on a short-time basis by applying the robust algorithm for pitch tracking (RAPT) and normalized by the mean of sentence F0 contours. Five-state, left-to-right HMM phone models are adopted in our baseline system.

Gamma distribution is used to model durations on phone and syllable levels. When gamma distribution is employed instead of Gaussian to model the durations of state, the formulas have too complicated form for practical implementation. In addition, five-state HMM phone model is used in our system and the state duration in terms of the number of frames per state ranges from 1 to 5 observed in 90% of states. It is difficult to tell the difference between Gaussian and gamma distributions on such a small scale. Therefore, we still use Gaussian distribution to model state durations.

F0 contours from voiced parts of syllable are used to F0 modeling on syllable-level. To reflect true tone contours, no artificial F0 values are interpolated for unvoiced parts. Considering the length of voiced part in some syllables can be less than 50ms, 7 DCT coefficients, delta and delta-delta features of the first DCT coefficient of proceeding, current and following syllables are used to represent syllable-level F0 contour. Our previous analysis also show the DCT with 7 coefficients can achieve the balance between the fitting error and the number of coefficients [10]. On phrase-level, 2 and 3 DCT coefficients are used to represent a contour that passes through the F0 mean of each constituent syllable. On state-level, no parametric representation is used for F0s.

Rich phonetic and prosodic contexts are used as a question set in growing decision trees. They include tones and breaks for Mandarin; stress, TOBI labels and POS for English; quin-phone, the position of phone, syllable and word in phrase and sentence, and the length of word and phrase for both Mandarin and English. The same question set is used for prosody modeling on different levels. The questions for splitting the nodes of tree are automatically selected in ML sense. Minimum description length (MDL) criterion for balancing model complexity and training data size is used as a stopping criterion for state clustering in decision tree growing.

5.2. Evaluation Results and Analysis

Objective and subjective measures are used to evaluate the performance of the proposed approach in testing data. Since the predicted phone durations of generated utterances are in general not the same as those of original speech, we first measure the root mean squared error (RMSE) of phone durations of synthesized speech. F0 distortions are then measured by RMSE and correlation coefficient between the original and synthesized F0 trajectories over all aligned voiced frames where the state durations of the original speech (obtained by forced alignment) are used for speech generation. Subjectively, a preference test is conducted to compare speech sentence pairs synthesized by our approach and the baseline. The duration and F0 in the baseline system are from the state-level model.

The RMSE results of phone duration predicted by the duration models on state (baseline) and integrated with models on phone and syllable levels are shown in Table 2. It shows that integrating phone and syllable duration models can reduce RMSE of 23.46 ms of English baseline to 21.35ms, and 30.1ms of Mandarin baseline to 26.78ms, i.e, the relative improvements of 9.9% and 11.1% are obtained for English and Mandarin corpora, respectively.

Table 2. *RMSE for baseline and improved duration generation with models on phone and syllable levels*

RMSE (ms)	English	Mandarin
state(baseline)	23.46	30.10
state+ph	21.44	29.86
state+ph+syl	21.35	26.78

Table 3 shows the RMSE and correlation coefficients between original and generated F0 trajectories for baseline and integrated with models on syllable and phrase levels. RMSE improvements of 0.87Hz and 0.67Hz, are obtained in English and Mandarin, respectively. The correlation coefficient is improved from 0.70 to 0.75 for English and 0.91 to 0.92 for Mandarin. A high correlation coefficient of 0.91 achieved by the baseline Mandarin TTS prevents it from being further improved significantly.

Table 3. *RMSE and correlation of F0 for baseline and improved F0 generation with the models on syllable and phrase levels*

	English		Mandarin	
	RMSE(Hz)	correlation	RMSE(Hz)	correlation
state(baseline)	13.46	0.70	21.39	0.91
state+syl	12.60	0.75	20.88	0.92
state+syl+phr	12.59	0.75	20.72	0.92

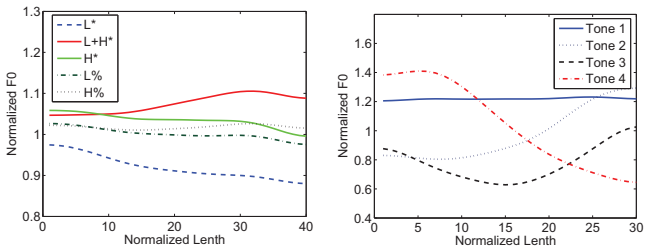
The improved prosody generation is further evaluated by a perceptual test. 50 Mandarin and 50 English sentences, which are selected from the testing set and synthesized by the baseline and the improved prosody generation, are evaluated in an AB preference test. 6 subjects participate in the preference test. There are three preference choices: 1) the former is better; 2) the latter is better; 3) no preference (The difference between the paired sentences can not be perceived or the difference can be perceived but it is difficult to choose which one is better). The preference scores between the

baseline and the improved prosody generation are shown in Table 4. It shows that the speech synthesized by the improved prosody generation outperforms the baseline system perceptually.

Table 4. *The preference scores of the baseline and the improved prosody generation with the models of longer units*

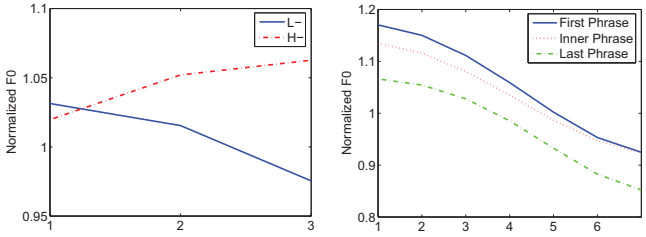
Baseline better	Improved better	No Preference
20%	39%	41%

To analyze the generated F0 contours on syllable and phrase levels, we cluster DCT coefficients in terms of TOBI labels for English, and tone types and the positions of current phrase in sentence for Mandarin. On syllable-level, Mandarin has four types of tones,indicated by numerical labels, English has three types of pitch accents: L*, L+H*, and H*, and two types of final boundary tones: L% and H%. On phrase-level, F0 contours are classified by the position in sentence: first, inner and last, for Mandarin since the majority of sentences are declarative, and the phrasal tones: L- and H- for English. The corresponding shapes of F0 contours on different levels are shown in Figure 1 and 2. They are consistent with TOBI labeling convention and phenomena observed by linguistics.



(a) English Syllable (b) Mandarin Syllable

Fig. 1. Shapes of syllable-level F0 contours



(a) English Phrase (b) Mandarin Phrase

Fig. 2. Shapes of phrase-level F0 contours.

6. CONCLUSION

We improve the prosody generation module in the conventional HMM-based TTS. Longer units of prosody are parameterized and modeled more properly. The prosody models of longer units are integrated into the baseline system to improve the prosody generation by maximizing the joint likelihood of state and longer units. The proposed prosody generation improves prosody prediction: the RMSE of phone durations are reduced by 2.1 and 3.3 ms and the RMSE of F0 trajectories are reduced by 0.87 and 0.67 Hz, in English and Mandarin synthesis. The synthesized speech by improved prosody generation also receives a higher preference score in a perceptional test, compared with that of baseline.

7. REFERENCES

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," in *Proc. ICASSP, 2000*.

[2] T. Toda and K. Tokuda, "Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," in *Proc. Eurospeech, 2005*.

[3] J. Latorre, K. Iwano, and S. Furui, "Combining Gaussian Mixture Model with Global Variance Term to Improve the Quality of an HMM-based Polyglot Speech Synthesizer," in *Proc. ICASSP, 2007*.

[4] 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," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[5] Y. Wu, R. Wang, and F. Soong, "Full HMM Training for Minimizing Generation Error in Synthesis," in *Proc. ICASSP, 2007*.

[6] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Investigation of State Duration Model based on Gamma Distribution for HMM-based Speech Synthesis." *IEICE Technical Report*, vol. 101, no. 352, pp. 57–62, 2001.

[7] M. Koo, H. Jeon, and S. Lee, "Context Dependent Phoneme Duration Modeling with Tree-Based State Tying," in *Proc. ICSLP, 2004*.

[8] B. Gao, Y. Qian, Z. Wu, and F. Soong, "Duration Refinement by Jointly Optimizing State and Longer Unit Likelihood," in *Proc. Interspeech, 2008*.

[9] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and Synthesising F0 Contours with the Discrete Cosine Transform," in *Proc. ICASSP, 2008*.

[10] Z. Wu, Y. Qian, F. Soong, and B. Zhang, "Modeling and Generating Tone Contour with Phrase Intonation for Mandarin Chinese Speech," *Accepted by ISCSLP, 2008*.

[11] J. Latorre and M. Akamine, "Multilevel Parametric-base F0 Model for Speech Synthesis," in *Proc. Interspeech, 2008*.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration Modeling For HMM-based Speech Synthesis," in *Proc. ICSLP, 1998*.

[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space Probability Distribution HMM," *IEICE Trans. Inf. & Syst., E85-D(3)*, pp. 455–464, 2002.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer, 2001.

[15] Y. Wu and R. Wang, "HMM-based Trainable Speech Synthesis for Chinese," *Journal of Chinese Information Processing*, vol. 20, no. 4, pp. 75–81, 2006.