Improved Prosody Generation by Maximizing Joint Probability of State and Longer Units

Yao Qian, *Member, IEEE*, Zhizheng Wu, Boyang Gao, and Frank K. Soong, *Fellow, IEEE*

***Abstract*—The current state-of-the-art hidden Markov model (HMM)-based text-to-speech (TTS) can produce highly intelligible, synthesized speech with decent segmental quality. However, its prosody, especially at phrase or sentence level, still tends to be bland. This blandness is partially due to the fact that the state-based HMM is inadequate in capturing global, hierarchical suprasegmental information in speech signals. In this paper, to improve the TTS prosody, longer units are first explicitly modeled with appropriate parametric distributions. The resultant models are then integrated with the state-based baseline models in generating better prosody by maximizing the joint probability. Experimental results in both Mandarin and English show consistent improvements over our baseline system with only state-based prosody model. The improvements are both objectively measurable and subjectively perceivable.**

***Index Terms*—Discrete cosine transforms (DCTs), speech synthesis, statistical distributions.**

I. Introduction

In recent years, corpus-driven speech synthesis system trained as hidden Markov models (HMMs) has steadily gained its popularity in text-to-speech (TTS) research and application. In this framework, spectral envelope, fundamental frequency (F0), and duration are modeled simultaneously by the corresponding HMMs [1]. In synthesis, for a given text, speech parameter trajectories are generated by the trained HMMs in the maximum probability sense with the dynamic ("delta") feature constraints [2]. Speech waveform is finally synthesized from the generated spectral and excitation parameters via source-filter based production model. Compared with the unit selection based speech synthesis [3], [4], HMM-based synthesis is trained in a more unified statistical criterion, i.e., maximum-likelihood (ML) principle in a parametric form. The speech generated by the HMMs is fairly smooth and rarely exhibits concatenation glitches which occur occasionally in conventional unit-selection synthesis. Characteristics of the synthesized speech also can be easily controlled by transforming the HMM parameters in a statistically tractable metric like likelihood function, e.g., segmental and supra-segmental parameters of the generated speech have been changed flexibly [5], [6]. However, overly smoothed parameter trajectories due to statistical averaging in HMM acoustic modeling still tend to make synthesized speech sound not as lively as desired.

Many research attempts have been tried in order to improve the performance of HMM-based speech synthesis [7]. To reduce the over-smoothing problem of trajectory and the resultant degraded synthesized speech quality, a parameter generation algorithm was proposed by considering the global variance (GV) of generated parameters in synthesis [8]. The probability of GV is used to boost (restore) the dynamic range of generated speech trajectory. An extension which adopts a Gaussian mixture model for characterizing the GV term was used to improve the quality of an HMM-based polyglot speech synthesizer [9]. To improve the acoustic modeling accuracy, a trajectory model by imposing the explicit relationship between the static and the dynamic features was proposed to improve the acoustic model accuracy [10]. It can overcome the assumption of conditional independence and piecewise statistics within a state without any additional parameters. In [11], minimum generation error (MGE) was proposed as an alternative criterion in HMM training. It adjusts ML-trained HMM parameters to minimize the generation errors between synthesized and original parameter trajectories in the training data.

With the above improvements, the segmental quality of synthesized speech is improved. However, synthesized speech prosody, particularly at the phrase and sentence levels, still tends to be somewhat bland. The relative ineffectiveness of GV or MGE in producing lively prosody is due to the fact that a state-based HMM is still inadequate in modeling a global, hierarchical prosodic structure and contextual effects in longer units like phrases or sentences. Furthermore, the decision tree-based state tying, which is commonly used in HMM-based TTS, is difficult to capture robustly the underlying additive structure of the features [12]. To model the hierarchical and additive structure of prosody, multi-layer models, in which each layer represents one component of prosody, have been tried in speech synthesis systems [13]–[19]. In some systems, different layers of prosody are modeled and generated simultaneously. Gradient tree boosting, which can iteratively build the regression trees from modeling (prediction) residuals in the training and output the sum of the regression trees in generation, was used to model prosody [16], [17]. In other systems [15], [18], [19], prosody models at different levels were first built separately and

Manuscript received August 31, 2009; revised July 02, 2010 and October 13, 2010; accepted November 18, 2010. Date of publication December 06, 2010; date of current version June 01, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Keiichi Tokuda.

Y. Qian and F. K. Soong are with Microsoft Research Asia, Beijing 100190, China (e-mail: yaoqian@microsoft.com; frankkps@microsoft.com).

Z. Wu was with Microsoft Research Asia, Beijing 100190, China. He is now with the School of Computer Engineering, Nanyang Technological University, 639798 Singapore (e-mail: wuzz@ntu.edu.sg).

B. Gao was with Microsoft Research Asia, Beijing 100190, China. He is now with the Department of Mathematics and Information, Ecole Centrale de Lyon, 69134 Ecully Cedex, France (e-mail: boyang.gao@ec-lyon.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2097248

1558-7916/$26.00 © 2010 IEEE

integrated together by maximizing their joint probabilities in parameter generation. In addition to using multi-layer models, the prosody models of long-term units can also be improved with more appropriate parametric distributions. Gamma distribution, which can model random variables with positive values, is more appropriate for modeling duration [20]–[22]. Discrete cosine transform (DCT), which can represent signal in terms of a linear combination of cosine functions at different frequencies, is a good parametric representation of smooth F0 trajectories [19], [23], [24].

In this paper, we investigate how to use gamma distribution for modeling durations and DCT for parameterizing F0 of longer units [18]. The longer-unit models are integrated with state-level models in generation and their joint probability is maximized [18]. The longer-unit model integration and the joint probability optimization is similar to GV constraint [8] and utterance length constraint [25] in parameter generation.

The rest of paper is organized as follows. A review on prosody modeling and generation in conventional HMM-based TTS system is given in Section II. In Section III, we investigate the statistical distributions of duration of longer units and corresponding parametric representations of F0 contours. The algorithm for prosody generation by maximizing the joint probability of different units and corresponding experimental results are presented in Sections IV and V, respectively. In Section VI, a conclusion is given.

II. Prosody Modeling and Generation in Conventional HMM-Based TTS System

In a conventional HMM-based TTS system, the state duration is explicitly modeled with a single Gaussian distribution where parameters are estimated by using the state occupancy counts in the Baum–Welch or Viterbi training procedure [25]. F0 features are modeled by multi-space probability distribution HMM (MSD-HMM) [26], which can characterize stochastically the piece-wise continuous F0 trajectory for both voiced and unvoiced frames. MSD models two, discrete and continuous probability spaces for unvoiced regions and voiced F0 contours, respectively. It models F0 in a stream separated from the spectral feature stream.

In synthesis, the parameter trajectories are generated based on the maximum-probability principle. The state duration is given as the mean of the corresponding Gaussian distribution. F0 trajectory is generated with the dynamic feature constraints. For a given HMM model λ, an F0 vector of T components, $\boldsymbol{F} = [f_0, f_1, \ldots, f_{T-1}]^\top$, is generated by maximizing $\log P(\boldsymbol{O}|\lambda)$ with respect to $\boldsymbol{O} = \boldsymbol{W}\boldsymbol{F}$, where $\boldsymbol{W} = [\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{T-1}]^\top$ is the weight matrix of static, delta, and delta-delta coefficients. In our implementation, $\boldsymbol{w}_t^{(0)} = [0, 1, 0]$, $\boldsymbol{w}_t^{(1)} = [-0.5, 0, 0.5]$ and $\boldsymbol{w}_t^{(2)} = [-1, 2, -1]$ are used. If the state sequence $\boldsymbol{Q} = [q_0, \ldots, q_{T-1}]$ is given according to the state duration statistics, we set

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

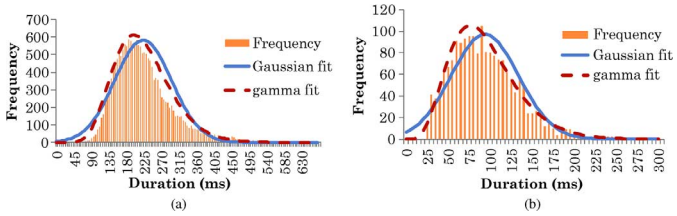


Fig. 1. Histograms for duration of different units. (a) 3-phone syllables; (b) phone of "ao."

and obtain the optimal solution $\boldsymbol{F}$ by solving the weighted least squares equations

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

where $\boldsymbol{U}^{-1} = diag[\boldsymbol{U}_{q_0}^{-1}, \boldsymbol{U}_{q_1}^{-1}, \ldots, \boldsymbol{U}_{q_{T-1}}^{-1}]$ and $\boldsymbol{M} = [\boldsymbol{\mu}_{q_0}^\top, \boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_{T-1}}^\top]^\top$ are covariance matrix and mean vector of F0, respectively.

III. Prosody Modeling for Longer Units

Richer prosodic contexts can be used to capture the co-articulation effects in longer speech units like words, phrases, or sentences. However, in practice, limited by insufficient training data, we usually have to cluster models of long and rich contexts into generalized ones in order to predict abundant, unseen contexts in test. State tying via a clustered classification and regression tree (CART) is therefore commonly used in conventional HMM-based TTS. CART is an effective and efficiently handling messy data, missing values, or predictor variables measured in different scales, but with its own limitations, e.g., difficulty in capturing underlying additive structure of prosody which is universally observed across different languages. Multi-layered models seem to be more capable of modeling the hierarchical structure of speech prosody. The decision tree-based state tying is inappropriate to model hierarchical prosodic structure at a sentence or phrase level. This is true even when the high-level prosody questions are included in the tree-building node splitting process. In this paper, we propose to model the prosody of longer units explicitly and integrate them with state-level model in parameter generation. Additionally, we also investigate the problem of how to parameterize the prosody of longer units with appropriate distributions.

A. Duration Modeling

In [22], we have compared gamma and Gaussian distributions in their model fitting behavior to the duration distributions of longer units, like phone and syllables. The data is obtained by forced aligning speech training data with corresponding acoustic HMMs trained for TTS. Syllables are clustered into groups according to their length in term of number of phones. The distributions (histograms) of syllable and phone durations resemble a gamma distribution more than a Gaussian counterpart as shown in Fig. 1(a) and (b), where the duration histograms of 3-phone syllables and a diphthong "ao" of English are depicted, respectively.

TABLE I
PERCENTAGE OF LEAF NODES WHICH GAMMA FITS BETTER

Gamma better (%)	state	phone	syllable
English	50.3	62.9	61.9
Mandarin	48.7	58.6	55.5

The distribution is further tested by Chi-Square statistics [27]. We test the distribution of durations in each leaf node of the decision trees at state, phone and syllable levels with the Chi-Square test of goodness-of-fit. Table I shows the percentage of leaf nodes where gamma distributions fit better than Gaussians, i.e., with a smaller value of χ^2 statistics. The difference between Gaussian and gamma is more distinctive in English than in Mandarin.

We use gamma distribution to model durations in the following form:

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

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$. We use the method of moments estimator[1] to estimate the parameters of the gamma distribution. The expected value and variance of the random variable x in a gamma distribution are $E(x) = ab$ and $Var(x) = ab^2$, 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 variables in a leaf node.

B. F0 Modeling

In [24], we investigated the parametric representations of F0 contours. Two parametric forms, natural cubic spline (NCS) and discrete cosine transform (DCT), are investigated for representing F0 trajectories. A natural cubic spline is a piece-wise cubic function defined in terms of the sample points called knots $x_1, \ldots, x_k$, satisfying $x_1 < x_2 < \ldots < x_{k-1} < x_k$ and its first and second derivatives are continuous at all intermediate knots, $x_2, \ldots, x_{k-1}$. The so-called a "natural" cubic spline has additional constraints that the second derivatives, at the first and last knots are zero. The discrete cosine transform is linear and invertible. It represents N discrete samples in terms of a weighted sum of cosine basis functions of different frequencies. The most commonly used DCT is

$$c_n = \frac{2}{T}\sum_{t=0}^{T-1} s_t \cos\left[\frac{\pi}{T} n \left(t + \frac{1}{2}\right)\right], \quad n = 0, \ldots, N-1 \tag{4}$$

[1]We use the method of moments for its simplicity.

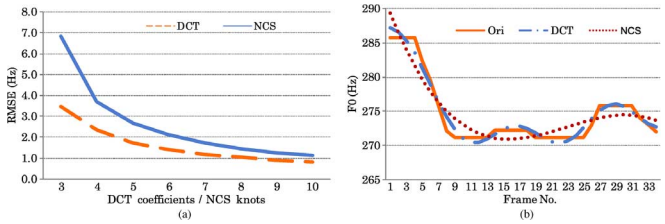


Fig. 2. Fitting performances of NCS and DCT. (a) Fitting errors by 18 000 syllables. (b) An example of F0 fitting.

where $s_0, \ldots, s_{T-1}$ are of length T samples and represented by N coefficients of DCT, $c_0, \ldots, c_{N-1}$. DCT can be wrote in matrix form $\boldsymbol{C}$, as shown in (5) at the bottom of the page.

Similarly, the inverse DCT is defined as

$$s_t = \frac{1}{2}c_0 + \sum_{n=1}^{N-1} c_n \cos\left[\frac{\pi}{T} n \left(t + \frac{1}{2}\right)\right], \quad t = 0, \ldots, T-1. \tag{6}$$

F0 curves extracted from 18 000 Mandarin syllables in a female continuous speech corpus are used to check the fitting performances of NCS and DCT. Fig. 2(a) shows the fitting errors in term of root mean square error (RMSE) with respect to different number of DCT coefficients and NCS knots. As shown in the figure, DCT outperforms NCS when they use the same number of coefficients/knots. As the number of DCT coefficients and NCS knots used increases, the contours represented by DCT and NCS approach the original F0 contours. The fitting errors in term of the number of coefficients show DCT is a more effective parametric representation of F0 curves. An example by using seven DCT coefficients or NCS knots to approximate original F0 contour is also shown in Fig. 2(b).

DCT uses a set of smooth orthogonal basis functions, where the first coefficient is the mean (average) of input signal samples and the rest coefficients are the weights for the corresponding cosine functions at different frequencies, which can represent the "shapes" of F0 contours. The orthogonality property of DCT basis function also makes DCT computation efficient. Therefore, we use DCT to parameterize the F0 contours at both syllable and phrase levels. Context-dependent Gaussian pdfs are used to model DCT coefficients at different levels.

IV. PROSODY GENERATION BY MAXIMIZING THE JOINT PROBABILITY OF DIFFERENT UNITS

A. Duration Generation

Speech durations can be predicted more precisely if the duration information of states and longer units like phone and syllables are jointly considered. This idea was originally proposed

$$\boldsymbol{C} = \frac{2}{T}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \cos\left[\frac{\pi}{T} * 1 * \left(0 + \frac{1}{2}\right)\right] & \cos\left[\frac{\pi}{T} * 1 * \left(1 + \frac{1}{2}\right)\right] & \cdots & \cos\left[\frac{\pi}{T} * 1 * \left(T - 1 + \frac{1}{2}\right)\right] \\ \vdots & \vdots & \ddots & \vdots \\ \cos\left[\frac{\pi}{T} * (N-1) * \left(0 + \frac{1}{2}\right)\right] & \cos\left[\frac{\pi}{T} * (N-1) * \left(1 + \frac{1}{2}\right)\right] & \cdots & \cos\left[\frac{\pi}{T} * (N-1) * \left(T - 1 + \frac{1}{2}\right)\right] \end{bmatrix} \tag{5}$$

by Wu [5], where the state duration is jointly estimated with the phone duration. In this paper, we extend it further to syllable duration. Also, we investigate the possibility of using gamma pdf as a possible distribution function in addition to Gaussian pdf. The probability of state durations is jointly maximized in conjunction with the probabilities of phone and syllable durations. For a given speech with J syllables, the objective function of duration sequence $\boldsymbol{D} = [d_0, \ldots, d_{J-1}]$, $L(\boldsymbol{D})$, is defined as

$$L(\boldsymbol{D}) = \sum_j \left[\sum_n \left[\sum_k \log p_{j,n,k}(d_{j,n,k}) + \alpha \log p_{j,n}(d_{j,n}) \right] + \beta \log p_j(d_j) \right] \quad (7)$$

subject to

$$\sum_k d_{j,n,k} = d_{j,n} \quad (8)$$

$$\sum_n d_{j,n} = d_j \quad (9)$$

where $d_{j,n,k}$ is the duration of state k, phone n, and syllable j, and $p_{j,n,k}(\cdot)$ is the corresponding probability density function. The phone and syllable pdfs, $p_{j,n}(\cdot)$ and $p_j(\cdot)$, are similarly defined. Parameters, α and β, are used to adjust the relative weights of phone and syllable durations probabilities, respectively.

We use Gaussian and gamma distributions for modeling phone and syllable durations as a refinement of the Gaussian modeled state duration. Despite the fact that the histograms of phone and syllable durations look more gamma like than Gaussian, Gaussian is still used here for two reasons: 1) Gaussian has been used for modeling duration information as a default benchmark distribution; 2) Gaussian is chosen for its mathematical tractability of its the first- and second-order moment sufficient statistics. To maximize $L(\boldsymbol{D})$ of Gaussian distributions, we construct the Lagrangian and find its gradient:

$$F(d_{j,n,k}, \lambda_1, \lambda_2) = L(\boldsymbol{D}) + \lambda_1 \left(\sum_k d_{j,n,k} - d_{j,n} \right) + \lambda_2 \left(\sum_n d_{j,n} - d_j \right) \quad (10)$$

$$\nabla F(d_{j,n,k}, \lambda_1, \lambda_2) = \begin{pmatrix} \frac{d_{j,n,k}-\mu_{j,n,k}}{\sigma^2_{j,n,k}} + \alpha \frac{d_{j,n}-\mu_{j,n}}{\sigma^2_{j,n}} + \beta \frac{d_j-\mu_j}{\sigma^2_j} \\ \sum_k d_{j,n,k} - d_{j,n} \\ \sum_n d_{j,n} - d_j \end{pmatrix} = 0. \quad (11)$$

It gives us

$$d_{j,n,k} = \mu_{j,n,k} + \left[-\alpha \frac{d_{j,n} - \mu_{j,n}}{\sigma^2_{j,n}} - \beta \frac{d_j - \mu_j}{\sigma^2_j} \right] \sigma^2_{j,n,k} \quad (12)$$

$$d_{j,n} = \frac{\sigma^2_{j,n} \left[M_{j,n} - \beta \frac{d_j - \mu_j}{\sigma^2_j} V_{j,n} \right] + \alpha \mu_{j,n} V_{j,n}}{\sigma^2_{j,n} + \alpha V_{j,n}} \quad (13)$$

$$d_j = \frac{\sigma^2_j \left[\sum_n \frac{\sigma^2_{j,n} M_{j,n}}{D_{j,n}} + \alpha \sigma^2_j \sum_n \frac{\mu_{j,n} V_{j,n}}{D_{j,n}} \right] + \beta \mu_j \sum_n \frac{\sigma^2_{j,n} M_{j,n}}{D_{j,n}}}{\sigma^2_j + \beta \sum_n \frac{\sigma^2_{j,n} V_{j,n}}{D_{j,n}}} \quad (14)$$

where

$$M_{j,n} = \sum_k \mu_{j,n,k} \quad (15)$$

$$V_{j,n} = \sum_k \sigma^2_{j,n,k} \quad (16)$$

$$D_{j,n} = \sigma^2_{j,n} + \alpha \sum_k \sigma^2_{j,n,k}. \quad (17)$$

When we compare gamma and Gaussian distributions for modeling state durations, we find the difference is rather small as shown in Table I. By further checking the duration histograms and the corresponding goodness of fit, we find the state duration in terms of the number of frames per state ranges from 1 to 5, for 90% of states (five-state HMM phone model used in our system) and the values of χ^2 test statistics in gamma fit are close to those in Gaussian fit in most cases. Therefore, we use gamma distributions to model phone and syllable durations only in (7). Similar to maximize $L(\boldsymbol{D})$ of Gaussian distributions by the method of Lagrange multipliers (10), we have

$$d_{j,n,k} - \mu_{j,n,k} - \rho_{j,n} \sigma^2_{j,n,k} = 0 \quad (18)$$

$$d^2_{j,n} + \left[\left(\frac{\alpha}{b_{j,n}} - \beta \frac{a_j - 1}{d_j} + \frac{1}{b_j} \right) V_{j,n} - M_{j,n} \right] d_{j,n} - \alpha (a_{j,n} - 1) V_{j,n} = 0 \quad (19)$$

$$2d^2_j - \sum_n \sqrt{\left(P^2_{j,n} - 4G_{j,n}\right) d^2_j + 2H_{j,n} P_{j,n} d_j + H^2_{j,n}} + d_j \sum_n P_{j,n} + \sum_n H_{j,n} = 0 \quad (20)$$

where

$$\rho_{j,n} = -\alpha \left(\frac{1}{b_{j,n}} + \frac{1 - a_{j,n}}{d_{j,n}} \right) - \beta \left(\frac{1}{b_j} + \frac{1 - a_j}{d_j} \right) \quad (21)$$

$$G_{j,n} = a(1 - a_{j,n}) V_{j,n} \quad (22)$$

$$P_{j,n} = -M_{j,n} + \frac{\alpha}{b_{j,n}} V_{j,n} + \frac{\beta}{b_j} V_{j,n} \quad (23)$$

$$H_{j,n} = \beta (1 - a_j) V_{j,n} \quad (24)$$

and $a_{j,n}$, $b_{j,n}$, a_j, b_j denote the parameters of the gamma distributions associated with $d_{j,n}$ and d_j, respectively. $M_{j,n}, V_{j,n}, D_{j,n}$ are defined the same as they are in the Gaussian case. The optimization is performed by solving two quadratic (19) and (20). Note that (20) has no closed-form solution; however, $\sum_n \sqrt{(P^2_{j,n} - 4G_{j,n}) d^2_j + 2H_{j,n} P_{j,n} d_j + H^2_{j,n}}$ can be well approximated by a linear function of d_j derived from Taylor series expansion, which leads to a solvable quadratic equation

$$2d^2_j + \left[\sum_n P_{j,n} - \sum_n f'_{j,n}(\hat{d}_j) \right] d_j + \sum_n \left[H_{j,n} - f_{j,n}(\hat{d}_j) \right] + \hat{d}_j \sum_n f'_{j,n}(\hat{d}_j) = 0 \quad (25)$$

where $f_{j,n}(x) = \sqrt{(P_{j,n}^2 - 4G_{j,n})x^2 + 2H_{j,n}P_{j,n}x + H_{j,n}^2}$ and $\hat{d}_j$ is the point where Taylor series expansion is performed and in practice we use μ_j. Solving (19) and (25), we obtain the solution of $d_{j,n,k}$.

B. F0 Generation

Similar to the duration generation, the probability of the F0 trajectory can be jointly maximized in conjunction with the probabilities of syllable and phrase contours. A similar approach was performed in [19], where DCT coefficients on phone and syllable levels are modeled and the resultant models are used together in parameter generation. However, the artificially interpolated F0 values [19] for unvoiced phones do not reflect the actual F0 contours. In our approach, F0 contours at higher levels, i.e., syllable and phrase, without interpolating the unvoiced segments are adopted to improve F0 generation. The log probability of F0 trajectory $\boldsymbol{F}$, $L(\boldsymbol{F})$, is defined as

$$L(\boldsymbol{F}) = \log P(\boldsymbol{O}_s|\boldsymbol{Q}_s, \lambda_s) + \alpha \log P(\boldsymbol{O}_y|\lambda_y) + \beta \log P(\boldsymbol{O}_h|\lambda_h) \tag{26}$$

with respect to

$$\boldsymbol{O}_s = \boldsymbol{W}_s\boldsymbol{F}, \quad \boldsymbol{O}_y = \boldsymbol{W}_y\boldsymbol{C}_y\boldsymbol{F} \quad \text{and} \quad \boldsymbol{O}_h = \boldsymbol{C}_h\boldsymbol{Z}\boldsymbol{F}$$

where λ_s is the state HMM, λ_y and λ_h are the HMMs on syllable and phrase levels, respectively; $\boldsymbol{Q}_s$ is the state sequence given by the duration model. Voiced/unvoiced decision for each frame is given by the state-level MSD. $\boldsymbol{C}_y$ is the DCT matrix for F0 contour on voiced part of syllable. At the phrase level, DCT matrix $\boldsymbol{C}_h$ is performed on the F0 mean of each consistent syllable. $\boldsymbol{Z}$ is the matrix to get the mean of F0s on each syllable; To make F0 trajectory locally continuous, $\boldsymbol{W}_s$, 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, $\boldsymbol{W}_y$ 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 probability of syllable-level and phrase-level DCT models. When we set $\alpha = 0$ and $\beta = 0$, only state-level is considered. The relationship between $\boldsymbol{O}_y$ and $\boldsymbol{F}$ in terms of three DCT coefficients can be arranged in matrix form, as shown in (27) at the bottom of the page, where $\boldsymbol{C}_{j,T_j}$, $j = 1, 2, \ldots, J$, is the DCT matrix, i.e., (5), for the jth syllable with T_j frames; $c_{j,n}, j = 1, 2, \ldots, J$; n =0,1,2, is the nth DCT coefficient for the jth syllable; and Δ, Δ^2 are the corresponding dynamic features of first DCT coefficient.

To maximize the probability $L(\boldsymbol{F})$, we set

$$\frac{\partial L(\boldsymbol{F})}{\partial \boldsymbol{F}} = 0 \tag{28}$$

and obtain the solution as

$$\bar{\boldsymbol{F}} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{29}$$

$$\boldsymbol{A} = \boldsymbol{W}_s^T\boldsymbol{U}_s^{-1}\boldsymbol{W}_s + \alpha\left[(\boldsymbol{W}_y\boldsymbol{C}_y)^T\boldsymbol{U}_y^{-1}\boldsymbol{W}_y\boldsymbol{C}_y\right] + \beta\left[(\boldsymbol{C}_h\boldsymbol{Z})^T\boldsymbol{U}_h^{-1}\boldsymbol{C}_h\boldsymbol{Z}\right] \tag{30}$$

$$\boldsymbol{b} = \boldsymbol{W}_s^T\boldsymbol{U}_s^{-1}\boldsymbol{M}_s + \alpha\left[(\boldsymbol{W}_y\boldsymbol{C}_y)^T\boldsymbol{U}_y^{-1}\boldsymbol{M}_y\right] + \beta\left[(\boldsymbol{C}_h\boldsymbol{Z})^T\boldsymbol{U}_h^{-1}\boldsymbol{M}_h\right] \tag{31}$$

where $\boldsymbol{U}_s$ and $\boldsymbol{M}_s$ are covariance matrix and mean vector of F0s at state-level while $\boldsymbol{U}_y, \boldsymbol{U}_h, \boldsymbol{M}_y$, and $\boldsymbol{M}_h$ are the corresponding covariance matrices and mean vectors of DCT coefficients at syllable and phrase levels.

$$5J\left\{\begin{array}{c}\boldsymbol{O_y}\\ \left[\begin{array}{c} c_{1,0}\\ \Delta c_{1,0}\\ \Delta^2 c_{1,0}\\ c_{1,1}\\ c_{1,2}\\ \hline c_{2,0}\\ \Delta c_{2,0}\\ \Delta^2 c_{2,0}\\ c_{2,1}\\ c_{2,2}\\ \hline \vdots\\ \vdots\\ \hline c_{J,0}\\ \Delta c_{J,0}\\ \Delta^2 c_{J,0}\\ c_{J,1}\\ c_{J,2}\end{array}\right]\end{array}\right. = \begin{array}{c}\boldsymbol{W_y}\\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \cdots & 0\\ 2 & 0 & 0 & -1 & 0 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0\\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0\\ -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & \cdots & 0\\ -1 & 0 & 0 & 2 & 0 & 0 & -1 & \cdots & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0\\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \hline 0 & \cdots & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & \cdots & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0\\ 0 & \cdots & 0 & -1 & 0 & 0 & 2 & 0 & 0\\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{array}\right]\end{array} \begin{array}{c}\boldsymbol{C_y}\\ \left[\begin{array}{cccc} \boldsymbol{C}_{1,T_1} & 0 & \cdots & 0\\ 0 & \boldsymbol{C}_{2,T_2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \boldsymbol{C}_{J,T_J}\end{array}\right]\end{array} \begin{array}{c}\boldsymbol{F}\\ \left.\left[\begin{array}{c} f_1\\ f_2\\ f_3\\ \vdots\\ f_{T-1}\\ f_T\end{array}\right]\right\}T\end{array} \tag{27}$$

TABLE II
TRAINING, DEVELOPING, AND TESTING SETS (#SENTENCES) OF THE TWO SPEECH CORPORA

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

TABLE III
NUMBER OF LEAF NODES IN THE DECISION TREES

	Duration		F0	
	English	Mandarin	English	Mandarin
state(baseline)	1,513	329	4,095	2,004
phone/syllable	592	129	1,589	773
syllable/phrase	324	113	50	56

V. EXPERIMENTS AND RESULTS

A. Experimental setup

Two phonetically and prosodically rich, speaker-dependent continuous speech corpora, one in American English and one in Mandarin Chinese, are used in our experiments. Each corpus was recorded by a female, professional native speaker in broadcast news style. Each database is divided into three parts: training, developing, and testing sets. The corresponding size of each set is given in number of sentences as shown in Table II. The training set is used for training the prosody model. Developing set is used to determine the appropriate weights (α, β) in (7) and (26). The testing set is used to measure the performance of the new prosody generation algorithm.

Speech signals are sampled at 16 kHz. The spectral analysis is performed by a 25-ms Hamming window, shifted every 5-ms. Spectral envelopes are estimated by STRAIGHT [28] and LPC modeled first and ultimately represented by 40th-order LSPs and their dynamic counterparts. F0 is extracted on a short-time basis by applying the robust algorithm for pitch tracking (RAPT) [29] without manual corrections, then pass them through a five-point medium filter to reduce spurious pitch extraction errors, finally normalize the filtered pitch values with the mean of sentence F0 contours. Five-state, left-to-right HMM phone models are adopted in our baseline system.

State durations of the training data are automatically obtained by forced alignments with the baseline models. Phone and syllable durations are obtained by accumulating the durations of the constituent states. Both gamma and Gaussian distributions are used to model durations at phone and syllable levels. State durations are still modeled by Gaussian distribution. F0 contours from voiced parts of syllable are used to F0 modeling at 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 50 ms, seven DCT coefficients, delta and delta-delta features of the first DCT coefficient of preceding, current and succeeding syllables are used to represent syllable-level F0 contour. Our previous analysis also show the DCT with seven coefficients can achieve the balance between the fitting error and the parsimonious number of coefficients [24]. At the phrase-level, three DCT coefficients are used to represent a contour that passes through the F0 mean of each constituent syllable. At 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 quin-phone; stress, TOBI labels, continuity of F0 contour on contextual quin-syllable; POS on contextual tri-word; the position of phone, syllable and word in phrase and sentence; and the length of syllable, word, and phrase in number of phone, syllable, and word for both Mandarin and English. The same question set is used for prosody modeling at different levels. The questions for splitting the nodes of tree are automatically selected in the ML sense. Minimum description length (MDL) criterion [30] for balancing model complexity and training data size is used as a stopping criterion for different level units clustering in decision tree growing. We set MDL factor equal to one in decision tree growing, the corresponding numbers of leaf nodes in the decision trees: state, phone, and syllable for duration models and state, syllable, and phrase for F0 models, are shown in Table III. Although the number of training sentences in Mandarin is larger than that in English, the length of each sentence in Mandarin is much shorter than that in English and as a result the number of leaf nodes in Mandarin is much smaller than that in English.

For a given text sequence to synthesize, duration and F0 are generated by our proposed method as mentioned in Section IV. F0 is further multiplied by the mean of F0s in training data for synthesis since it is normalized during models training. Spectral parameter (LSP) trajectories are first generated from trained HMMs by the conventional approach, then formant sharpened based on LSP frequencies [31] to reduce the over-smoothing problem of HMMs and the resultant degraded synthesized speech quality.

B. 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 the original speech, we first measure the root mean squared error (RMSE) of phone and syllable durations between original and synthesized speech. F0 distortions are then measured by RMSE and the 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 system. The duration and F0 in the baseline system are from the state-level model.

To find the optimal α and β values for maximizing the joint probabilities of state and longer unit prosody, we use the development set via a grid search in the two dimensional space of (α, β). The grid search for finding the best (α, β) values in gamma distribution of duration is shown in Fig. 3. The best (α, β) values for integrating gamma distributions of phone and syllable are (0.5, 0.3) and (0.5, 2.1) for English and Mandarin, respectively. While RMSE and correlation are two common metrics for evaluating F0 model performance objectively, we use correlation as the sole criterion in the grid search since correlation is more relevant to the subjective quality of generated

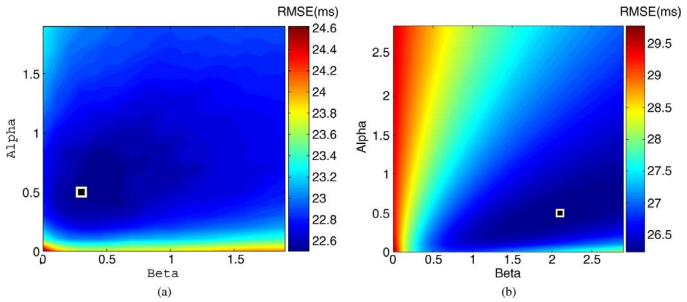


Fig. 3. Full grid search of optimal (α, β) values for integrating of the gamma distributions of phone and syllable durations. The best values are marked with squares. (a) English. (b) Mandarin.

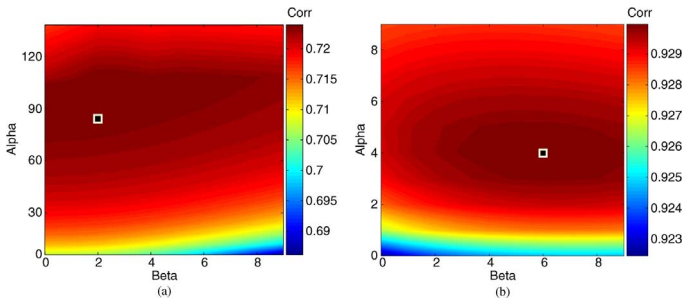


Fig. 4. Full grid search of optimal (α, β) values for integrating DCT F0 models of syllables and phrases. The best values are marked with squares. (a) English. (b) Mandarin.

F0. The optimal (α, β) values search for integrating DCT syllable and phrase models is shown in Fig. 4 and the resultant values are (80, 2) and (4, 6) for English and Mandarin, respectively. When we set $(\alpha = 0)$ or $(\beta = 0)$, only phone or syllable is integrated to the duration model, and only syllable or phrase is integrated to the F0 model. Figs. 3 and 4 also show that the α and β values are sensitive to the performance. The α and β distributions are fairly speaker and language-dependent since the dynamic ranges of duration and F0 vary from one speaker (language) to another.

The RMSE results of phone and syllable durations predicted by the duration models of state (baseline) and integrated with duration models of phone and syllable are shown in Table IV, where the RMSE results of syllable duration are listed in brackets. In our experiments, pause durations are not modified by integrating higher level model and excluded in the calculation of the RMSE results since their dynamic ranges are too large. Table IV shows that integrating phone and syllable duration models can reduce phone RMSE of 23.46 ms of English baseline to 21.35 ms, and 30.1 ms of Mandarin baseline to 26.78 ms, i.e., relative improvements of 9.9% and 11.1% are obtained for English and Mandarin corpora, respectively. Similar relative improvements are obtained on most RMSE results of syllable duration except Mandarin syllable. The syllable duration RMSE for the Mandarin corpus is reduced from 55.34 ms to 36.94 ms, i.e., a relative improvement of 33.2%, obtained by incorporating syllable gamma duration models into state+phone duration models. Mandarin is known as a syllabically paced tonal language. Compared with English, Mandarin has a simpler and more restricted syllabic structure and the syllable durations are rather stable, e.g., 4 or 5 syllables per second generally. Therefore, the syllable durations can be well

TABLE IV
RMSE FOR BASELINE AND IMPROVED DURATION GENERATION WITH DURATION MODELS OF PHONE AND SYLLABLE

RMSE(ms)	English		Mandarin	
phone (syllable)	Gaussian	Gamma	Gaussian	Gamma
state(baseline)	23.46 (38.29)		30.10 (55.94)	
state+phone	21.96 (32.82)	21.44 (34.01)	29.54 (54.34)	29.86 (55.34)
state+phone+syllable	21.52 (32.77)	21.35 (33.36)	27.91 (39.06)	26.78 (36.94)

TABLE V
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+syllable	12.60	0.75	20.88	0.92
state+syllable+phrase	12.59	0.75	20.72	0.92

regulated by the syllable level duration models. We cannot tell which distribution, gamma or Gaussian, performs better from the Table IV. By further checking the goodness of fit shown in Table I, we find that although more leaf nodes fit gamma better than Gaussian, some bad cases which are much farther away from gamma than Gaussian are observed, i.e., the values of χ^2 test statistics in bad cases of gamma fit are much larger than those in Gaussian fit. This might be the possible reason why gamma distribution underperforms Gaussian distribution in some RMSE results in Table IV.

Table V shows the RMSE and correlation coefficients between original and generated F0 trajectories for baseline and integrated with models at both syllable and phrase levels. RMSE improvements of 0.87 and 0.67 Hz 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 much further improved significantly. This may due to the fact that Mandarin is a tonal language and the contextual HMM with lexical tone label in training regulates well the F0 contour movement in F0 generation, even in the baseline Mandarin system.

The improved prosody generation on both duration and F0 is further evaluated by a perceptual test. 50 Mandarin and 50 English sentences, which are selected from the testing set sentences and synthesized by the baseline and the improved prosody generations, are evaluated in an AB preference test participated by six subjects. 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 can be perceived but difficult to choose which one is better). The preference scores between the baseline and the improved systems consisting of improved duration modeling, improved F0 modeling and improved both are shown in Fig. 5. It shows that the speech synthesized by the improved systems is preferred than the baseline subjectively. The system with both duration and F0 modeling improvement is the best. Its preference score (39%) is significantly higher than the baseline system (20%) (A significance test: $\alpha = 0.001$, CI $= [0.0964, 0.2836]$). While the perception difference between baseline and improved system with either improved duration or F0 is not distinctive.

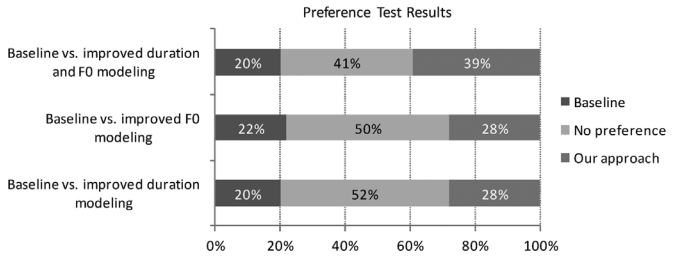


Fig. 5. Preference scores of the baseline and the improved systems with the models of longer units.

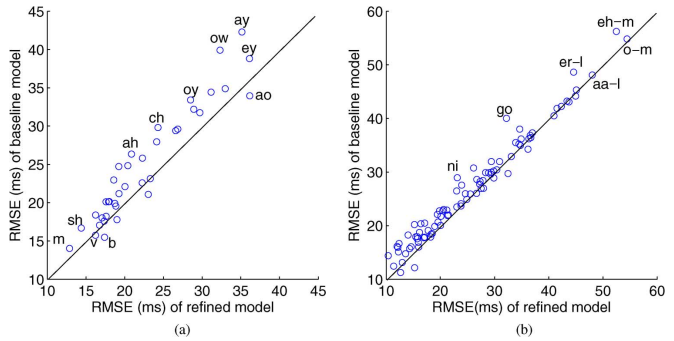


Fig. 6. Scatter diagram of phone RMSE pairs predicted by baseline and refined model by phone and syllable gamma distributions. (a) All English phones. (b) All Mandarin phones.

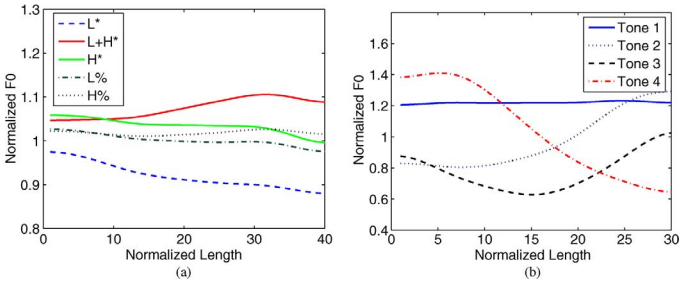


Fig. 7. Shapes of syllable-level F0 contours. (a) English syllable. (b) Mandarin Syllable.

Fig. 6 shows the scattered diagrams of RMSE pairs of phone durations predicted by the baseline and the refined model, plotted for each individual phone in English and Mandarin. Note that only for a few phones, predicted phone duration is not improved by the refined model.

To analyze the generated F0 contours at both 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. At syllable level, Mandarin has four types of tones, indicated by their numerical labels, English has three types of pitch accents: L^*, $L + H^*$, and H^*, and two types of final boundary tones: L% and H%. At 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 Figs. 7 and 8. They are consistent with TOBI labeling convention and the general contour shapes observed by linguistics.

VI. 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 probability of state and longer units. The proposed prosody generation improves prosody prediction: the RMSE of syllable durations are reduced by 5.5 and 19.0 ms and the RMSE of F0 trajectories are reduced by 0.87 and 0.67 Hz, in synthesized English and Mandarin, respectively. The synthesized speech generated by the improved prosody generation module also preferably perceived in subjective listening test, compared with that of baseline.

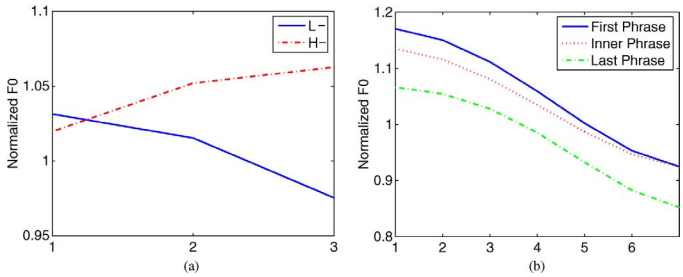


Fig. 8. Shapes of phrase-level F0 contours. (a) ENglish phrase. (b) Mandarin phrase.

References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[3] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR μ-talk speech synthesis system," in *Proc. ICSLP*, 1992, pp. 483–486.

[4] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[5] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[6] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[7] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[8] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-Based speech synthesis," in *Proc. Eurospeech*, 2005, pp. 373–376.

[9] 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, pp. 1241–1244.

[10] 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.*, vol. 21, no. 1, pp. 153–173, 2007.

[11] Y.-J. Wu, R.-H. Wang, and F. K. Soong, "Full HMM Training for Minimizing Generation Error in Synthesis," in *Proc. ICASSP*, 2007, pp. 517–520.

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

[13] X.-J. Sun, "F0 generation for speech synthesis using a multi-tier approach," in *Proc. ICSLP*, 2002, pp. 2077–2080.

[14] S. Sakai, "Additive modeling of english f0 contour for speech synthesis," in *Proc. ICASSP*, 2005, pp. 277–280.

[15] Y.-J. Wu and R.-H. Wang, "HMM-based trainable speech synthesis for Chinese," *J. Chinese Inf. Process.*, vol. 20, no. 4, pp. 75–81, 2006.

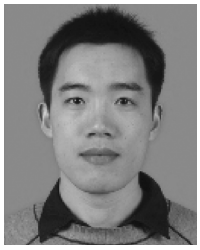
[16] Y. Qian, H. Liang, and F. K. Song, "Generating natural F0 trajectory with additive trees," in *Proc. Interspeech*, 2008, pp. 2126–2129.

[17] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Commun.*, vol. 50, no. 5, pp. 405–415, 2008.

[18] Y. Qian, Z.-Z. Wu, and F. K. Song, "Improved prosody generation by maximizing joint likelihood of state and longer units," in *Proc. ICASSP*, 2009, pp. 3781–3784.
[19] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Proc. Interspeech*, 2008, pp. 2274–2277.
[20] 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 Tech. Rep.*, vol. 101, no. 352, pp. 57–62, 2001.
[21] S. J. Park, M. W. Koo, and C. S. Jhon, "Context-Dependent phoneme duration modeling with Tree-Based state tying," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 3, pp. 662–666, 2005.
[22] B.-Y. Gao, Y. Qian, Z.-Z. Wu, and F. K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Proc. Interspeech*, 2008, pp. 2266–2269.
[23] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and synthesising F0 contours with the discrete cosine transform," in *Proc. ICASSP*, 2008, pp. 3973–3976.
[24] Z.-Z. Wu, Y. Qian, F. K. Soong, and B. Zhang , "Modeling and generating tone contour with phrase intonation for Mandarin Chinese speech," in *Proc. ISCSLP*, 2008, pp. 121–124.
[25] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, 1998, pp. 29–32.
[26] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D(3), pp. 455–464, 2002.
[27] G. W. Snedecor, *Statistical Methods*. Ames: Iowa State Univ. Press, 1989.
[28] H. Kawahara, I. Masuda Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.
[29] A. D. Talkin, "chapter A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elservier, 1995.
[30] K. Shinoda and T. Watanable, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn(E)*, vol. 21, no. 2, pp. 79–86, 2000.
[31] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for blizzard challenge 2006 an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge 2006 Workshop*, 2006.

Yao Qian (M'06) received the B.S. and M.S. degrees in computer science and linguistics from Shanghai Normal University, Shanghai, China, in 1995 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Chinese University of Hong Kong, Hong Kong, in 2005.

In 2005, she joined the Speech Group of Microsoft Research Asia (MSRA), Beijing, China. Her current research interests include trainable text-to-speech synthesis, singing voice synthesis, and prosody modeling for speech synthesis, recognition, and understanding.

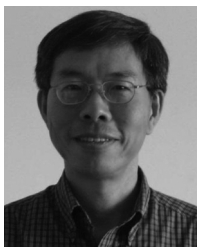


Zhizheng Wu received the B.E. degree in computer science from Hangzhou Dianzi University, Hangzhou, China, in 2006, and the M.E. degree from Nankai University, Tianjin, China, in 2009. He is currently pursuing the Ph.D. degree at the School of Computer Engineering, Nanyang Technological University, Singapore.

His current research interests include voice transformation, speaker recognition, and trainable text-to-speech synthesis.

Boyang Gao received the B.S. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2006 and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2009. He is currently pursuing the Ph.D. degree in the Department of Mathematics and Information, Ecole Centrale de Lyon, Ecully, France, with the Laboratory of LIRIS.

He is currently working on structured semantic description for music information retrieval (MIR) systems. His research interests include MIR, general-purpose computation on graphics hardware, acoustic signal analysis and classification, multimedia retrieval systems and machine learning.



Frank K. Soong (M'81–SM'91–F'10) received the B.S. degree from National Taiwan University, Taipei, the M.S. degree from the University of Rhode Island, Kingston, and Ph.D. degree from Stanford University, Stanford, CA, all in electrical engineering.

He joined Bell Labs Research, Murray Hill, NJ, in 1982, worked there for 20 years, and retired as a Distinguished Member of Technical Staff in 2001. In Bell Labs, he had worked on various aspects of acoustics and speech processing, including: speech coding, speech and speaker recognition, stochastic modeling of speech signals, efficient search algorithms, discriminative training, dereverberation of audio and speech signals, microphone array processing, acoustic echo cancellation, and hands-free noisy speech recognition. He was also responsible for transferring recognition technology from research to AT&T voice-activated cell phones which were rated by the Mobile Office Magazine as the best among competing products evaluated. He visited Japan twice as a visiting researcher: first from 1987 to 1988, at the NTT Electro-Communication Labs, Musashino, Tokyo; then from 2002 to 2004, at the Spoken Language Translation Labs, ATR, Kyoto. In 2004, he joined Microsoft Research Asia (MSRA), Beijing, China, to lead the Speech Research Group. He is a Visiting Professor of the Chinese University of Hong Kong (CUHK) and the codirector of the CUHK-MSRA Joint Research Lab, recently promoted to a National Key Lab of Ministry of Education, China.

Dr. Soong was the corecipient of the Bell Labs President Gold Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package. He was the cochair of the 1991 IEEE International Arden House Speech Recognition Workshop. He has served the IEEE Speech and Language Processing Technical Committee of the Signal Processing Society as a committee member and associate editor of the TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He published extensively and coauthored more than 200 technical papers in the speech and signal processing fields.