Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis

Junichi Yamagishi, *Member, IEEE*, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, *Member, IEEE*, Keiichi Tokuda, *Member, IEEE*, Simon King, *Senior Member, IEEE*, and Steve Renals, *Member, IEEE*

***Abstract*—This paper describes a speaker-adaptive HMM-based speech synthesis system. The new system, called "HTS-2007," employs speaker adaptation (CSMAPLR+MAP), feature-space adaptive training, mixed-gender modeling, and full-covariance modeling using CSMAPLR transforms, in addition to several other techniques that have proved effective in our previous systems. Subjective evaluation results show that the new system generates significantly better quality synthetic speech than speaker-dependent approaches with realistic amounts of speech data, and that it bears comparison with speaker-dependent approaches even when large amounts of speech data are available. In addition, a comparison study with several speech synthesis techniques shows the new system is very robust: It is able to build voices from less-than-ideal speech data and synthesize good-quality speech even for out-of-domain sentences.**

***Index Terms*—Average voice, HMM-based speech synthesis, HMM Speech Synthesis System, HTS, speaker adaptation, speech synthesis, voice conversion.**

I. Introduction

STATISTICAL parametric speech synthesis based on hidden Markov models (HMMs) [1], [2] is now well-established and can generate natural-sounding synthetic speech [3]. In this framework, we have pioneered the development of the HMM Speech Synthesis System, HTS (H Triple S) [4]. This research started by developing algorithms for generating a smooth parameter trajectory from HMMs [5]–[9]. Next, to simultaneously model the excitation parameters of speech as well as the spectral parameters, the multispace probability distribution (MSD) HMM [10] was developed. Using the logarithm of the fundamental frequency $(\log F_0)$ and its dynamic and acceleration features as the excitation parameters, the MSD-HMM enabled us to treat the $\log F_0$ sequence, which is a mixture of one-dimensional real numbers for voiced regions and symbol strings for unvoiced regions, in a probabilistic framework. To simultaneously model the duration parameters for the spectral and excitation components of the model, the MSD hidden semi-Markov model (MSD-HSMM) [11] was developed. The HSMM [12]–[14] is an HMM having explicit state duration distributions instead of transition probabilities, to directly model duration; it can generate more appropriate temporal structures for speech. These basic systems [1], [4], [11] employed a mel-cepstral vocoder with simple pulse or noise excitation, resulting in synthetic speech with a "buzzy" quality. To reduce buzziness, mixed or multi-band excitation techniques [15]–[17] have been integrated into the basic systems to replace the simple pulse or noise excitation and have been evaluated [18]–[21]. These basic systems also had another significant problem: the trajectories generated from the HMMs were excessively smooth due to statistical processing, resulting in synthetic speech with a "muffled" quality. To alleviate this problem, a parameter generation algorithm that considers the global variance (GV) of a trajectory to be generated was developed [22].

From the accumulation of these incremental improvements, several high-quality text-to-speech synthesis systems have been developed [20], [23]–[25]. They have demonstrated good performance in the Blizzard Challenges, which are open evaluations of corpus-based text-to-speech (TTS) synthesis systems [26]–[28]. In the Nitech-HTS system [20] used for the 2005 Blizzard Challenge, a high-quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [29] was used, in conjunction with MSD-HSMMs, mixed excitation, and the GV parameter generation algorithm. STRAIGHT explicitly uses F_0 information for removing the periodic components from the estimated spectrum: it interpolates missing frequency components considering neighboring harmonic components based on an F_0 adaptive smoothing process on a time-frequency region. This enables the generation of better spectral parameters and consequently more natural synthetic speech [20]. In the Nitech-NAIST-HTS system [23] for the Blizzard Challenge 2006, semi-tied covariance (STC) modeling [30], [31] was employed to enable the use of full-covariance Gaussians in the HSMMs, and the structure of the covariance matrices for

Manuscript received January 23, 2008; revised January 20, 2009. Current version published July 06, 2009. This work was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant 213845 (the EMIME project). The work of J. Yamagishi was supported by the Engineering and Physical Sciences Research Council (EPSRC) and EMIME. The work of Z. Ling was supported by the Marie Curie Early Stage Training (EST) Network, Edinburgh Speech Science and Technology (EdSST). S. King holds an EPSRC Advanced Research Fellowship. The associate editor coordinating the review of this manuscript for publication was Dr. Gaël Richard.

J. Yamagishi, S. King, and S. Renals are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9AB, U.K. (e-mail: jyamagis@inf.ed.ac.uk; simon.king@ed.ac.uk; s.renals@ed.ac.uk).

T. Nose is with Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502, Japan (e-mail: takashi.nose@ip.titech.ac.jp).

H. Zen and K. Tokuda are with Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, 466-8555 Japan (e-mail: zen@sp.nitech.ac.jp; tokuda@nitech.ac.jp).

Z.-H. Ling is with iFlytek Speech Lab, University of Science and Technology of China, Hefei, Anhui 230027, China (e-mail: zhling@ustc.edu).

T. Toda is with Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: tomoki@is.naist.jp).

Digital Object Identifier 10.1109/TASL.2009.2016394

the GV probability density functions (pdfs) was changed from diagonal to full. Although the use of GV parameter generation drastically reduces the muffled quality of synthetic speech, it was sometimes perceived as more artificial. One reason for this was that each acoustic feature dimension was optimized independently. This limitation was addressed by the use of full-covariance modeling in the HSMMs.

The above systems were *speaker-dependent*. In parallel, we have also been developing a *speaker-adaptive* approach in which "average voice models" are created using data from several speakers. The average voice models may then be adapted using a small amount of speech from a target speaker (e.g., [32] and [33]). This research started by transforming only the spectral parameters [34] using several speaker adaptation techniques developed for automatic speech recognition, such as maximum-likelihood linear regression (MLLR) [35]. To adapt spectral, excitation, and duration parameters within the same framework, extended MLLR adaptation algorithms for the MSD-HSMM have been proposed [32], [36], [37]. A more robust and advanced adaptation scheme, constrained structural maximum *a posteriori* linear regression (CSMAPLR), has been proposed and its effectiveness in HMM-based speech synthesis has been demonstrated [33].

We have also developed several techniques for training the average voice model. The average voice model is constructed using training data from several speakers. Because these data include many speaker-dependent characteristics that affect the adapted models and the quality of synthetic speech generated from them, we have employed a model-space speaker-adaptive training (SAT) algorithm [38] in order to reduce the negative influence of speaker differences [39]. In the SAT algorithm, the model parameters for the average voice model were obtained using a blind estimation procedure assuming that the speaker difference was expressed by linear transformations of the mean vectors of Gaussian pdfs in the average voice model. A similar model-space SAT algorithm for the MSD-HSMM was also derived [32]. Furthermore, applications to style adaptation (conversion of speaking styles and emotional expressions) and to multilingual/polyglot text-to-speech systems have also been reported [40]–[42]. By using the speaker-adaptive approach, we can obtain natural-sounding synthetic speech for a target speaker from as little as a hundred adaptation utterances, corresponding to about six minutes of speech data. In our experiments, we have shown that the synthetic speech generated using this approach is perceived as being more natural sounding, by many listeners, than that of a speaker-dependent (SD) system trained using thirty minutes of speech from the target speaker [32], [33]. The data-rich average voice model provides a strong prior for speech generation, with the target adaptation data being used to estimate speaker-specific characteristics.

In this paper, we outline a high quality speaker-adaptive HMM-based speech synthesis system. We then propose two new algorithms for acoustic modeling. This system was first evaluated in the 2007 Blizzard Challenge [28] and several issues were analyzed from additional evaluation tests. We then compare the system with several major competing TTS methods used in the 2007 Blizzard Challenge and assess its performance and potential.

We have combined several advances in the speaker adaptive approach with our existing speaker-dependent system that employs STRAIGHT, mixed excitation, HSMMs, GV, and full-covariance modeling. 1) First we propose a *feature-space speaker adaptive training (SAT) algorithm for HSMMs* to replace the standard embedded training used in the speaker-dependent system or the model-space SAT algorithm used in conventional speaker-adaptive systems. The feature-space SAT algorithm addresses two limitations of the model-space SAT algorithm mentioned in the next section and hence yields better speaker normalization of the average voice model. 2) Second, we propose a modeling technique for the average voice model called *mixed-gender modeling* to efficiently construct an average voice model from a limited amount of training data. 3) To adapt the average voice model, we utilize an algorithm combining CSMAPLR and maximum *a posteriori* (MAP) adaptation [43] for HSMMs. 4) We investigate a full-covariance modeling technique using the CSMAPLR transforms and adopt it instead of the STC transform. Although CSMAPLR is a speaker adaptation method rather than a full-covariance modeling method, it has the same transforms for the covariance matrices as STC and the additional MAP adaptation estimates the diagonal elements of the covariance matrix in a similar way to updating processes for STC. For CSMAPLR, multiple transforms are estimated using the robust SMAP criterion [44], which is expected to alleviate the artificiality and to improve the quality of synthetic speech. We describe the details of the resulting system, which we call "HTS-2007," assess its performance and discuss a number of outstanding issues.

II. HTS-2007 System

The HTS-2007 system, outlined in Fig. 1, consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

A. Speech Analysis

We use three kinds of parameters for the STRAIGHT mel-cepstral vocoder with mixed excitation: the STRAIGHT mel-cepstrum [20], $\log F_0$ and aperiodicity measures. These are the same as those of the Nitech-HTS 2005 speaker-dependent system. The mel-cepstral coefficients are obtained from a STRAIGHT spectral analysis in which F_0-adaptive spectral smoothing is carried out in the time–frequency domain to remove signal periodicity. The F_0 values are estimated using a three-stage extraction to reduce errors such as F_0 halving and doubling and to suppress voiced/unvoiced errors. First, using the instantaneous-frequency-amplitude-spectrum-based algorithm (IFAS) [45], the system extracts F_0 values for all speech data of each speaker within a common search range. Second, the F_0 range of each speaker is roughly determined based on a histogram of the extracted F_0 values. Third, F_0 values are re-extracted in the speaker-specific range using three methods: IFAS, a fixed-point analysis called TEMPO [46] and the ESPS get-F_0 tool [47], [48]. The final estimated value for F_0 at each frame is the median of the three extracted values.

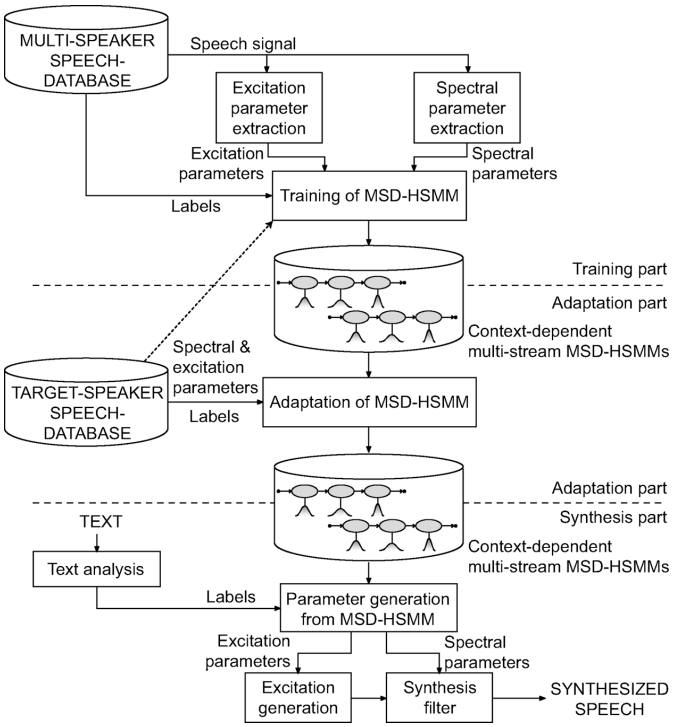


Fig. 1. Overview of the HTS-2007 speech synthesis system which consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

The aperiodicity measures for mixed excitation are based on a ratio between the lower and upper smoothed spectral envelopes, and averaged across five frequency sub-bands (0–1, 1–2, 2–4, 4–6, and 6–8 kHz).

In addition to these static features (STRAIGHT mel-cepstrum, $\log F_0$ and 5 aperiodicity measures), dynamic and acceleration features are also used, which are referred to as the first and second delta parameter vectors, corresponding to the first and second time derivative estimates of the static feature vector. Let $\boldsymbol{x}_t \in \mathcal{R}^L$ be the static vector at frame t. For a given static vector sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ with a length of T frames, the kth delta parameter vector for $\boldsymbol{x}_t$, $\Delta^k \boldsymbol{x}_t$, is defined by

$$\Delta^k \boldsymbol{x}_t = \sum_{i=-D_k}^{D_k} w^{(k)}(i) \boldsymbol{x}_{t+i}, \quad 0 \le k \le 2 \tag{1}$$

where $w^{(k)}(i)$ are coefficients used to obtain delta parameters and $D_0 = 0, w^{(0)}(0) = 1$. For example, if we set $D_1 = D_2 = 1$, then the $w^{(k)}(i)$ derived from numerical differentiation are

$$w^{(1)}(i) = i/2, \quad \text{for } i = -1, 0, 1 \tag{2}$$

$$w^{(2)}(i) = 3i^2 - 2, \quad \text{for } i = -1, 0, 1. \tag{3}$$

The static and delta feature vectors are combined and the observation vector at frame t, denoted by $\boldsymbol{o}_t \in \mathcal{R}^{3L}$, is

$$\boldsymbol{o}_t = \left[\boldsymbol{x}_t^\top, \Delta^1 \boldsymbol{x}_t^\top, \Delta^2 \boldsymbol{x}_t^\top\right]^\top \tag{4}$$

where $\cdot^\top$ denotes matrix transpose.

B. Acoustic Models and Labels

As in our previous systems, we utilize context-dependent multi-stream left-to-right MSD-HSMMs [11] in order to simultaneously model the above acoustic features and duration. The English phonetic and linguistic contexts that we employ contain phonetic, segment-level, syllable-level, word-level and utterance-level features [49]. Japanese phonetic and linguistic contexts used in the following experiments contain phonetic, mora-level [50], morpheme, accentual, breath-group-level, and utterance-level features [39]. In addition to this phonetic and linguistic information, we added speaker gender context labels when conducting the mixed-gender modeling described in Section II-D.

C. Speaker Adaptive Training

We estimated average voice models using the HSMMs described above, trained with the SAT algorithm from training data consisting of several speakers' speech. Previously, we had utilized a model-space SAT algorithm [38] using linear transformations of mean vectors of Gaussian pdfs in our average voice systems [32], [39]. Here, we employ a feature-space SAT algorithm [51] using linear transformations of feature vectors. There are two major reasons for the change from model-space to feature-space.

The first reason is computational feasibility. As reported in [51], in model-space SAT algorithms it is necessary to store a full matrix for each Gaussian pdf, or to store statistics for each Gaussian component for every speaker. In our *speaker-adaptive* HMM-based speech synthesis system, there are over 10 million Gaussians, which can make parameter estimation impractical. In particular, the embedded training procedures in which we could use the model-space SAT were restricted to only the training procedures in which the mean and covariance parameters were tied across several Gaussian pdfs [32], [39]. On the other hand, the feature-space SAT algorithm can be applied to all embedded training procedures.

The second reason is the additional use of Gaussian pdf covariance matrices for speaker normalization of the average voice model. A linear transformation of feature vectors can be viewed as a simultaneous linear transform of both mean vectors and covariance matrices using the same matrix [51], [52], and thus we may also regard the feature-space SAT algorithm as a constrained model-space algorithm.

We can derive feature-space SAT in the framework of the HSMM in a similar way to [32]. An N-state HSMM λ is specified by initial state probabilities $\{\pi_i\}_{i=1}^N$, state transition probabilities $\{a_{ij}\}_{i,j=1,i\neq j}^N$, state output probability distributions $\{b_i(\cdot)\}_{i=1}^N$, and state duration probability distributions $\{p_i(\cdot)\}_{i=1}^N$ (see Fig. 2). Let F be the total number of training speakers, $\boldsymbol{O} = \{\boldsymbol{O}_1, \cdots, \boldsymbol{O}_F\}$ be all the training data, and $\boldsymbol{O}_f = \{\boldsymbol{o}_{1_f}, \cdots, \boldsymbol{o}_{T_f}\}$ be training data of length T_f for speaker f. In the feature-space SAT algorithm, we assume that each state of the HSMM λ has an output pdf $b_i(\boldsymbol{o}_{t_f})$, characterized by a mean vector $\boldsymbol{\mu}_i \in \mathcal{R}^{3L}$ and a diagonal covariance matrix

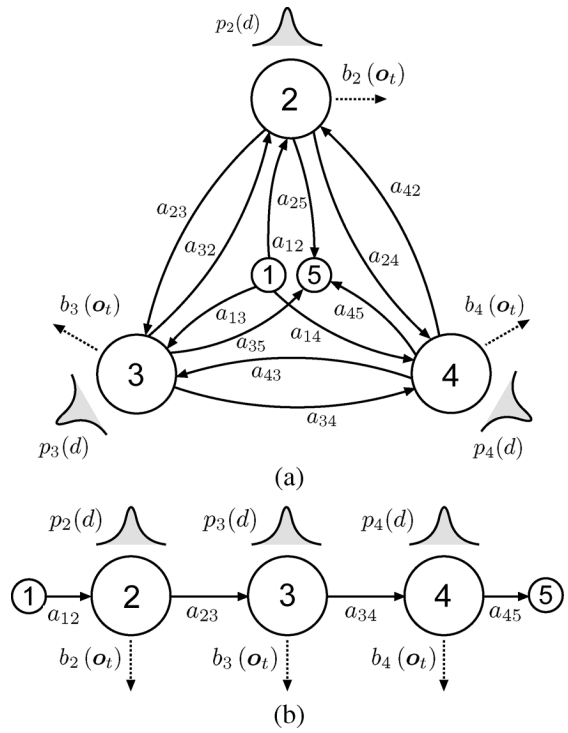


Fig. 2. Ergodic and left-to-right hidden semi-Markov models. Each state has a state output distribution $b_i(\boldsymbol{o}_t)$ and a state duration distribution $p_i(d)$. The state duration distributions directly model and control within-state duration instead of the self-transition Markov probabilities. For further explanation of the training, estimation and implementation and issues for HSMMs, see [11], [40], and [32]. (a) A five-state HSMM (one beginning null state, three emitting states, and one ending null state). (b) A five -state left-to-right HSMM (one beginning null state, three emitting states, and one ending null state).

$\boldsymbol{\Sigma}_i \in \mathcal{R}^{3L\times 3L}$, and a duration pdf $p_i(d)$ characterized by a scalar mean m_i and variance σ_i^2

$$b_i\left(\boldsymbol{o}_{t_f}\right) = |\boldsymbol{\zeta}_f|\mathcal{N}\left(\boldsymbol{\zeta}_f\boldsymbol{o}_{t_f}+\boldsymbol{\epsilon}_f;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right) \tag{5}$$

$$p_i(d) = |\chi_f|\mathcal{N}\left(\chi_f d+\nu_f;m_i,\sigma_i^2\right) \tag{6}$$

where $\boldsymbol{o}_{t_f}$ and d are the observation vector and duration, respectively, at state i, and $\boldsymbol{\zeta}_f \in \mathcal{R}^{3L\times 3L}$, $\boldsymbol{\epsilon}_f \in \mathcal{R}^{3L}$, χ_f, and ν_f are speaker-dependent linear transforms which normalize the observation vector and its duration for speaker f. These linear transforms can be estimated using the HSMM-based constrained maximum-likelihood linear regression (CMLLR) algorithm [33].

Re-estimation formulas based on the EM algorithm [53] for the Gaussian pdfs are given by

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\left(\boldsymbol{\zeta}_f\boldsymbol{o}_{s_f}+\boldsymbol{\epsilon}_f\right)}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t} d\cdot\gamma_t^d(i)} \tag{7}$$

$$\overline{\boldsymbol{\Sigma}}_i = \sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i) \frac{\sum_{s=t-d+1}^{t}\left(\boldsymbol{\zeta}_f\boldsymbol{o}_{s_f}+\boldsymbol{\epsilon}_f-\overline{\boldsymbol{\mu}}_i\right)\left(\boldsymbol{\zeta}_f\boldsymbol{o}_{s_f}+\boldsymbol{\epsilon}_f-\overline{\boldsymbol{\mu}}_i\right)^{\top}}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t} d\cdot\gamma_t^d(i)} \tag{8}$$

$$\overline{m}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\cdot(\chi_f d+\nu_f)}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{9}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\cdot(\chi_f d+\nu_f-\overline{m}_i)^2}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{10}$$

where $\gamma_t^d(i)$ is the state occupancy probability of being in state i of the HSMM for the period of time from $t-d+1$ to t given $\boldsymbol{O}_f$ and is defined as

$$\gamma_t^d(i) = \frac{1}{P(\boldsymbol{O}_f|\lambda)}\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_{t-d}(j)a_{ji}p_i(d)\prod_{s=t-d+1}^{t}b_i\left(\boldsymbol{o}_{s_f}\right)\beta_t(i). \tag{11}$$

Here, the observation probability of the training data $\boldsymbol{O}_f$ given the model λ, $P(\boldsymbol{O}_f|\lambda)$ and the forward and backward probabilities, $\alpha_t(i)$ and $\beta_t(i)$, can be written as

$$P(\boldsymbol{O}_f|\lambda) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{d=1}^{t}\alpha_{t-d}(j)a_{ji}p_i(d) \times\prod_{s=t-d+1}^{t}b_i\left(\boldsymbol{o}_{s_f}\right)\beta_t(i) \tag{12}$$

$$\alpha_t(i) = \sum_{d=1}^{t}\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_{t-d}(j)a_{ji}p_i(d) \times\prod_{s=t-d+1}^{t}b_i\left(\boldsymbol{o}_{s_f}\right) \tag{13}$$

$$\beta_t(i) = \sum_{d=1}^{T_f-t}\sum_{\substack{j=1\\ j\neq i}}^{N}a_{ij}p_j(d) \times\prod_{s=t+1}^{t+d}b_j\left(\boldsymbol{o}_{s_f}\right)\beta_{t+d}(j) \tag{14}$$

where $\alpha_0(i)=\pi_i$ and $\beta_T(i)=1$, and π_i is the initial state probability of being state i at time $t=1$. For further explanation of the training, estimation and implementation issues for HSMMs, see [11], [40], and [32].

D. Mixed-Gender Modeling and Training Procedures

In addition to phonetic and prosodic features, the variability of speech may be accounted for by speaker-dependent characteristics, some of which may be shared amongst all speakers of the same gender. If a large amount of training data for male and female speakers is available, then it is efficient to use gender-dependent average voice models as an initial model before speaker adaptation [33]. In practice, however, the available training data from one or both genders may be limited. For example, the CMU-ARCTIC speech database[1] includes four male and two female speakers. In such cases, it would not be the best choice to use gender-dependent average voice models.

A gender-independent average voice model may be used, but our previous work has shown that this results in a degradation in the naturalness and similarity of the resultant synthetic speech,

[1]A free database for speech synthesis, http://www.festvox.org/cmu_arctic/.

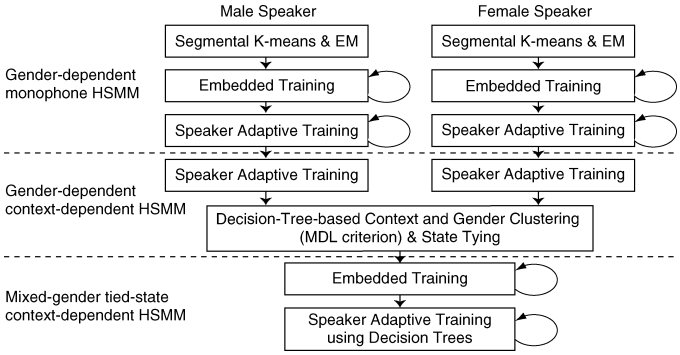


Fig. 3. Details of mixed-gender modeling. This modeling technique consists of speaker adaptive training and decision-tree-based context and gender clustering.

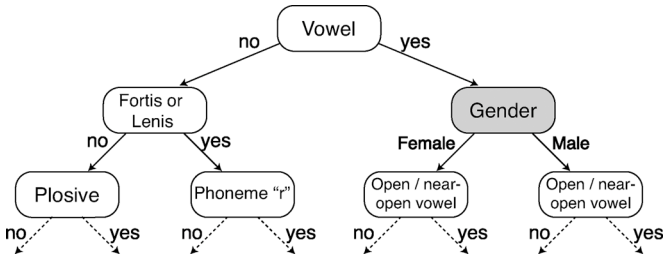


Fig. 4. Part of a constructed decision tree in the mixed-gender modeling. Genders of training speakers are split by using gender-related questions as well as other contexts.

after adaptation, compared with a gender-dependent average voice model. Alternatively, it is possible to use the gender-dependent average voice models simultaneously to enable them to complement one another and to perform soft decisions during the speaker adaptation [33]. However, we found no significant improvement between the results of the simultaneous use of the gender-dependent average voice models and those of the single gender-dependent average voice model. It required twice as many parameters for adaptation as the gender-dependent average voice model and seemed to suffer from the "curse of dimensionality." Therefore, we sought an approach which satisfies the following three conditions: 1) it reflects the gender-dependent characteristics as prior information; 2) it makes the best possible use of the training data from both genders, complementing one another if necessary; and 3) it does not increase the number of parameters required for speaker adaptation.

To achieve this, we propose a *mixed-gender modeling* technique, similar to *style-mixed modeling* [54]. Mixed-gender modeling includes speaker adaptive training and decision tree-based context and gender clustering, and is outlined in Fig. 3. In order both to normalize speaker-dependent characteristics and to conserve gender-dependent characteristics, we first train gender-dependent monophone HSMMs using the SAT algorithm with CMLLR global transforms. These are converted into gender-dependent context-dependent HSMMs, and the model parameters are re-estimated using the SAT algorithm again. Then, using the state occupancy probabilities obtained in the SAT framework, decision-tree-based context clustering (using a minimum description length (MDL) criterion [55]) is applied to the HSMMs, and the model parameters of the HSMMs at each leaf node of the decision trees are tied. We assume that the CMLLR transforms for the SAT algorithm remain unchanged during the clustering. The gender of each speaker is treated as a clustering context, and both the gender-dependent models undergo clustering at the same time. As a result, the gender information is included in the single resulting acoustic model. Note that a decision tree was constructed independently for each combination of state index and acoustic parameter (mel-cepstrum, $\log F_0$, aperiodicity) or duration. Hence, when the target feature is generally gender-specific, such as $\log F_0$, the gender will tend to be automatically split close to the root of the tree by using gender-related questions, and the pdfs of that feature retain their gender-dependent characteristics. For features which are less gender-dependent, gender will tend to be split deeper down the tree or not at all, thereby making efficient use of training data from both genders. Fig. 4 shows a part of the constructed decision tree for the mel-cepstral part of the fifth state of the HSMMs. In this part of the tree, we can see that vowels are gender-dependent, but consonants are not, which seems reasonable. We re-estimate the clustered HSMMs using SAT with piecewise linear regression functions. The decision trees constructed for the mixed-gender model are also used to determine the regression classes, since these automatically reflect both gender differences and phonetic and linguistic information.

E. Speaker Adaptation and Full-Covariance Modeling

In the speaker adaptation stage, we adapt the mixed-gender average voice model to that of the target speaker by using speech data plus gender information about the target speaker. We utilize a combined algorithm of HSMM-based CSMAPLR and MAP adaptation. The CSMAPLR adaptation simultaneously transforms the mean vectors and covariance matrices of state-output and state-duration distributions of the HSMMs as follows:

$$b_i(\boldsymbol{o}) = \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\zeta}_i'\boldsymbol{\mu}_i - \boldsymbol{\epsilon}_i', \boldsymbol{\zeta}_i'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}_i'^{\top}\right) \tag{15}$$

$$= |\boldsymbol{\zeta}_i|\mathcal{N}(\boldsymbol{\zeta}_i\boldsymbol{o} + \boldsymbol{\epsilon}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{16}$$

$$= |\boldsymbol{\zeta}_i|\mathcal{N}\left(\boldsymbol{H}_{b_i}\boldsymbol{\xi}_b; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{17}$$

$$p_i(d) = \mathcal{N}\left(d; \chi_i' m_i - \nu_i', \chi'\sigma_i^2\chi'\right) \tag{18}$$

$$= |\chi_i|\mathcal{N}(\chi_i d + \nu_i; m_i, \sigma_i) \tag{19}$$

$$= |\chi_i|\mathcal{N}\left(\boldsymbol{H}_{p_i}\boldsymbol{\xi}_p; m_i, \sigma_i\right) \tag{20}$$

where $\boldsymbol{\xi}_b = [\boldsymbol{o}^{\top}, 1]^{\top}$ and $\boldsymbol{\xi}_p = [d, 1]^{\top}$ are the extended observation and duration vectors. $\boldsymbol{H}_{b_i} = [\boldsymbol{\zeta}_i, \boldsymbol{\epsilon}_i] = [\boldsymbol{\zeta}_i'^{-1}, \boldsymbol{\zeta}_i'^{-1}\boldsymbol{\epsilon}_i'] \in \mathcal{R}^{3L\times(3L+1)}$ and $\boldsymbol{H}_{p_i} = [\chi_i, \nu_i] = [\chi_i'^{-1}, \chi_i'^{-1}\nu_i'] \in \mathcal{R}^{1\times 2}$ are, respectively, the linear transform matrices for the state output and duration pdfs.

To robustly estimate $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$, structural maximum *a posteriori* (SMAP) estimation [44] is used, in which tree structures group the distributions in the model and propagate priors for MAP estimation. Specifically, we first estimate a global transform at the root node of the tree structure using all adaptation data, and then propagate the transform to its child nodes as their priors. In the child nodes, transforms are estimated again using their adaptation data, based on MAP

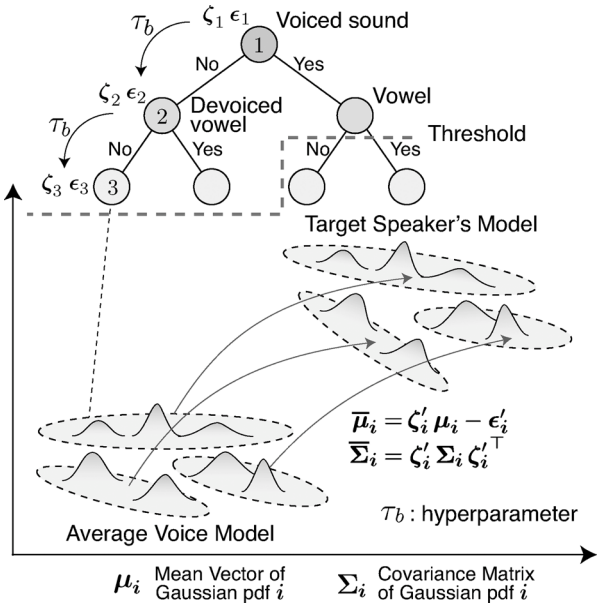


Fig. 5. Concepts of constrained structural maximum *a posteriori* linear regression. Transforms estimated at each node are propagated to its child nodes as their priors for MAP estimation. A recursive MAP-based estimation of the transforms from the root node to lower nodes is conducted.

estimation with the propagated priors $\boldsymbol{H}_{b_0} \in \mathcal{R}^{3L\times(3L+1)}$ and $\bar{\boldsymbol{H}}_{p_0} \in \mathcal{R}^{1\times 2}$. Then, the recursive MAP-based estimation of the transforms from the root node to lower nodes is conducted (see Fig. 5). For the tree structures of the distributions, the decision trees for the mixed-gender average voice model are used for the same reason as the above SAT algorithm with piecewise linear regression functions. Then, since the CSMAPLR adaptation algorithm estimates a piecewise linear regression, we update the linearly transformed model using MAP adaptation.

Another advantage of combining CSMAPLR and MAP adaptation is that we can efficiently construct full-covariance models. As we can see from (15), we may use the CSMAPLR transforms for full-covariance modeling, since $\boldsymbol{\Sigma}_i$ is a diagonal covariance matrix and $\boldsymbol{\zeta}_i'$ is a square matrix. In order to precisely model the full-covariance in the HSMMs, the following update procedures are used.

1) Train all parameters for the average voice model. Build the tree structures to group the distributions in the model.
2) Using the current transforms $\boldsymbol{H}_{b_i} = (\boldsymbol{\zeta}_i, \boldsymbol{\epsilon}_i)$, $\boldsymbol{H}_{p_i} = [\chi_i, \nu_i]$, and the average voice model, estimate the new transforms $\widehat{\boldsymbol{H}}_{b_i} = (\hat{\boldsymbol{\zeta}}_i, \hat{\boldsymbol{\epsilon}}_i)$ and $\widehat{\boldsymbol{H}}_{p_i} = [\hat{\chi}_i, \hat{\nu}_i]$ based on the SMAP criterion as follows.
   a) At the root node, estimate the initial transforms $\tilde{\boldsymbol{H}}_b$ and $\tilde{\bar{\boldsymbol{H}}}_p$ using the ML criterion (i.e. the CMLLR adaptation). Define the priors $\bar{\boldsymbol{H}}_{b_0}$ and $\bar{\boldsymbol{H}}_{p_0}$ for its child nodes as $\boldsymbol{H}_{b_0} = \tilde{\boldsymbol{H}}_b$ and $\boldsymbol{H}_{p_0} = \tilde{\bar{\boldsymbol{H}}}_p$.
   b) At each child node, estimate new transforms $\widetilde{\boldsymbol{H}}_b$ and $\widetilde{\bar{\boldsymbol{H}}}_p$ using the MAP criterion as follows:

$$\widetilde{\boldsymbol{h}}_{lb} = (\alpha \boldsymbol{p}_l + \boldsymbol{y}_l)\boldsymbol{G}_l^{-1} \tag{21}$$

$$\widetilde{\bar{\boldsymbol{H}}}_p = (\beta \boldsymbol{q} + \boldsymbol{z})\boldsymbol{K}^{-1} \tag{22}$$

where $\widetilde{\boldsymbol{h}}_{lb}$ is the lth row vector of $\widetilde{\boldsymbol{H}}_b$, $\boldsymbol{p}_l = [0, \boldsymbol{c}_l^\top]^\top$ and $\boldsymbol{q} = [0,1]^\top$. Note that $\boldsymbol{c}_l$ is the lth cofactor row vector of $\widetilde{\boldsymbol{H}}_b$. The terms $\boldsymbol{y}_l \in \mathcal{R}^{3L+1}$, $\boldsymbol{G}_l \in \mathcal{R}^{(3L+1)\times(3L+1)}$, $\boldsymbol{z} \in \mathcal{R}^2$, and $\boldsymbol{K} \in \mathcal{R}^{2\times 2}$ in these equations are given by

$$\boldsymbol{y}_l = \sum_{r\in R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\Sigma_r(l)}\mu_r(l)\sum_{s=t-d+1}^{t}\boldsymbol{\xi}_s^\top + \tau_b \boldsymbol{h}_{lb_0} \tag{23}$$

$$\boldsymbol{G}_l = \sum_{r\in R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\Sigma_r(l)}\sum_{s=t-d+1}^{t}\boldsymbol{\xi}_s\boldsymbol{\xi}_s^\top + \tau_b \boldsymbol{I}_{L+1} \tag{24}$$

$$\boldsymbol{z} = \sum_{r\in R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\sigma_r^2}m_r\boldsymbol{\phi}_r^\top + \tau_p \boldsymbol{H}_{p_0} \tag{25}$$

$$\boldsymbol{K} = \sum_{r\in R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\sigma_r^2}\boldsymbol{\phi}_r\boldsymbol{\phi}_r^\top + \tau_p \boldsymbol{I}_2 \tag{26}$$

where $\boldsymbol{h}_{lb_0}$ is the lth row vector of the prior $\boldsymbol{H}_{b_0}$, $\Sigma_r(l)$ is the lth diagonal element of diagonal covariance matrix $\boldsymbol{\Sigma}_r$, and $\mu_r(l)$ is the lth element of the mean vector $\boldsymbol{\mu}_r$. $\boldsymbol{I}_{(L+1)}$ and $\boldsymbol{I}_2$ are $(L+1)\times(L+1)$ and 2×2 identity matrices. τ_b and τ_p are positive hyperparameters of the prior distributions for the state output and duration distributions, respectively. R_b and R_p are, respectively, indices for the set of the distributions of the state output and duration distributions belonging to this node. Then α and β are scalar values that satisfy the following quadratic equations:

$$\alpha^2 \boldsymbol{p}_l\boldsymbol{G}_l^{-1}\boldsymbol{p}_l^\top + \alpha \boldsymbol{p}_l\boldsymbol{G}_l^{-1}\boldsymbol{y}_l^\top - \sum_{r\in R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)d = 0 \tag{27}$$

$$\beta^2 \boldsymbol{q}\boldsymbol{K}^{-1}\boldsymbol{q}^\top + \beta \boldsymbol{q}\boldsymbol{K}^{-1}\boldsymbol{z}^\top - \sum_{r\in R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r) = 0. \tag{28}$$

Since the cofactor $\boldsymbol{c}_l$ affects all row vectors of $\widetilde{\boldsymbol{H}}_b$, we update $\widetilde{\boldsymbol{H}}_b$ using an iterative method proposed in [51], whereas we can obtain a closed-form solution for the estimation of $\widetilde{\bar{\boldsymbol{H}}}_p$ in (22).
   c) Redefine the priors for its child nodes as $\boldsymbol{H}_{b_0} = \widetilde{\boldsymbol{H}}_b$ and $\boldsymbol{H}_{p_0} = \widetilde{\bar{\boldsymbol{H}}}_p$ and go to step b) until it reaches the leaf nodes or terminal nodes determined by thresholds.
   d) Assign transforms for the distribution i belonging to the leaf or terminal nodes to $\widehat{\boldsymbol{H}}_{b_i} = \widetilde{\boldsymbol{H}}_b$ and $\widehat{\bar{\boldsymbol{H}}}_{p_i} = \widetilde{\bar{\boldsymbol{H}}}_p$.
3) Using the estimated transforms $\widehat{\boldsymbol{H}}_{b_i} = [\hat{\boldsymbol{\zeta}}_i, \hat{\boldsymbol{\epsilon}}_i]$, $\widehat{\boldsymbol{H}}_{p_i} = [\hat{\chi}_i, \hat{\nu}_i]$, and the current average voice model, estimate $\widehat{\boldsymbol{\mu}}_i$, $\widehat{\boldsymbol{\Sigma}}_i$, $\widehat{m}_i$, and $\widehat{\boldsymbol{\Sigma}}_i$ for the average voice model based on the MAP criterion as follows:

$$\widehat{\boldsymbol{\mu}}_i = \frac{\upsilon_b \boldsymbol{\mu}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}(\hat{\boldsymbol{\zeta}}_i\boldsymbol{o}_s + \hat{\boldsymbol{\epsilon}}_i)}{\upsilon_b + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d} \tag{29}$$

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{\upsilon_b + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d} \times \Bigg[\boldsymbol{\Sigma}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i) \times \sum_{s=t-d+1}^{t}(\hat{\boldsymbol{\zeta}}_i\boldsymbol{o}_s + \hat{\boldsymbol{\epsilon}}_i - \widehat{\boldsymbol{\mu}}_i)(\hat{\boldsymbol{\zeta}}_i\boldsymbol{o}_s + \hat{\boldsymbol{\epsilon}}_i - \widehat{\boldsymbol{\mu}}_i)^\top + \upsilon_b(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i)^\top\Bigg] \tag{30}$$

$$\widehat{m}_i = \frac{\upsilon_p m_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot(\hat{\chi}_i d + \hat{\nu}_i)}{\upsilon_p + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{31}$$

$$\widehat{\sigma}_i^2 = \frac{1}{\upsilon_p + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)} \times \Bigg[\sigma_i^2 + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot(\hat{\chi}_i d + \hat{\nu}_i - \overline{m}_i)^2 + \upsilon_p(m_i - \widehat{m}_i)^2\Bigg] \tag{32}$$

where $\boldsymbol{\mu}_i$ and m_i are the current mean vectors of the state output and duration distributions of the average voice model for ith state. $\boldsymbol{\Sigma}_i$ and σ_i^2 are the current covariance matrices of the state output and duration distributions of the average voice model for ith state. υ_b and υ_p are positive hyperparameters of the prior distributions for the state output and duration distributions, respectively.

4) Go to step 2) until convergence, or an appropriate criterion is satisfied.
5) Transform the covariance matrices to full covariance using the updated parameters. Transform the mean vectors too.

F. Global Variance Parameter Generation Algorithm

Finally, we explain the GV parameter generation algorithm [22] for the CSMAPLR adapted model. The GV parameter generation algorithm is a penalized maximum-likelihood method. First, let us consider the problem of generating a parameter sequence from HSMM λ having N states, given the transforms $\Lambda = (\mathcal{H}_b, \mathcal{H}_p) = (\{\boldsymbol{H}_{b_j}\}_{j=1}^{N}, \{\boldsymbol{H}_{p_j}\}_{j=1}^{N})$ for CSMAPLR adaptation and frame length T in a maximum-likelihood sense [5]. In this approach, we obtain a suboptimal parameter sequence $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_T^*)$ without the dynamic and acceleration features as follows:

$$(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_T^*) = \underset{(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)}{\arg\max}\, P(\boldsymbol{O}|\lambda, \Lambda, T) \tag{33}$$

$$= \underset{(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)}{\arg\max} \sum_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda, \Lambda, T) \tag{34}$$

$$\simeq \underset{(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)}{\arg\max}\, \max_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda, \Lambda, T) \tag{35}$$

where $\boldsymbol{q} = (q_1, q_2, \ldots, q_T)$ is a hidden state sequence. Since this equation can be simply rewritten as

$$P(\boldsymbol{O}, \boldsymbol{q}|\lambda, \Lambda, T) = P(\boldsymbol{O}|\boldsymbol{q}, \lambda, \mathcal{H}_b, T)P(\boldsymbol{q}|\lambda, \mathcal{H}_p, T) \tag{36}$$

we first determine an optimal state sequence $\boldsymbol{q}^*$ by maximizing $P(\boldsymbol{q}|\lambda, \mathcal{H}_p, T)$, and then maximize $P(\boldsymbol{O}|\boldsymbol{q}, \lambda, \mathcal{H}_b, T)$ using $\boldsymbol{q}^*$. Here, we can obtain the optimal state sequence as

$$\boldsymbol{q}^* = \left(\underbrace{1, \ldots, 1}_{d_1}, \underbrace{2, \ldots, 2}_{d_2}, \ldots, \underbrace{N, \ldots, N}_{d_N}\right) \tag{37}$$

where state duration d_i is given by

$$d_i = \widetilde{m}_i + \rho\widetilde{\sigma}_i^2, \quad i = 1, 2, \ldots, N \tag{38}$$

$$\widetilde{m}_i = \chi_i' m_i - \nu_i', \quad i = 1, 2, \ldots, N \tag{39}$$

$$\widetilde{\sigma}_i^2 = \chi_i' \sigma_i^2 \chi_i', \quad i = 1, 2, \ldots, N \tag{40}$$

$$\rho = \left(T - \sum_{i=1}^{N}\widetilde{m}_i\right) \Big/ \sum_{i=1}^{N}\widetilde{\sigma}_i^2. \tag{41}$$

It is noted that the value of d_i is rounded to the nearest positive integer.

Given the optimal state sequence $\boldsymbol{q}^*$, we calculate a suboptimal parameter vector sequence $\boldsymbol{x}^*$. $P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, \mathcal{H}_b, T)$ is given by

$$\log P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, \mathcal{H}_b, T) = \sum_{t=1}^{T}\log\mathcal{N}\left(\boldsymbol{o}_t; \widetilde{\boldsymbol{\mu}}_{q_t^*}, \widetilde{\boldsymbol{\Sigma}}_{q_t^*}\right) \tag{42}$$

where

$$\widetilde{\boldsymbol{\mu}}_{q_t^*} = \boldsymbol{\zeta}'_{q_t^*}\boldsymbol{\mu}_{q_t^*} - \boldsymbol{\epsilon}'_{q_t^*} \quad t = 1, 2, \ldots, T \tag{43}$$

$$\widetilde{\boldsymbol{\Sigma}}_{q_t^*} = \boldsymbol{\zeta}'_{q_t^*}\boldsymbol{\Sigma}_{q_t^*}{\boldsymbol{\zeta}'_{q_t^*}}^\top \quad t = 1, 2, \ldots, T. \tag{44}$$

Although Kalman smoothing or regularization theory commonly uses $\Delta^1\boldsymbol{x}_t \sim \mathcal{N}(0, \sigma^2)$ or $\Delta^2\boldsymbol{x}_t \sim \mathcal{N}(0, \sigma^2)$ as continuity constraints, (42) constrains the static features obtained from (1)–(3) in the following way:

$$\Delta^1\boldsymbol{x}_t = -0.5\boldsymbol{x}_{t-1} + 0.5\boldsymbol{x}_{t+1} \tag{45}$$

$$\Delta^2\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t+1}. \tag{46}$$

We can obtain a smoothed parameter sequence $\boldsymbol{x}^*$ which maximizes $P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, \mathcal{H}_b, T)$ from these constraints [5].

In the GV parameter generation algorithm [22], we manipulate the objective function for $\boldsymbol{x}$ by adding a penalty term as follows:

$$\log P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, \mathcal{H}_b, T) + \omega\log\mathcal{N}(\boldsymbol{v}; \boldsymbol{\theta}, \boldsymbol{\kappa}) \tag{47}$$

where $\boldsymbol{v} \in \mathcal{R}^L$ is a GV vector having variance of each dimension of the parameter sequence $\boldsymbol{x}$ as shown in Fig. 6. L is the dimension of the static feature. Then, $\boldsymbol{\theta} \in \mathcal{R}^L$ and $\boldsymbol{\kappa} \in \mathcal{R}^{L\times L}$ are the mean vector and full-covariance matrix of the GV vectors estimated from the training data. ω is the weight for controlling the balance between these terms, and we set ω to $3T$, based on the number of Gaussian distributions included in the first term. The penalty term for the GV vector is intended to keep the variance of the generated trajectory as wide as that

TABLE I
DEFINITION OF HTS-2007 SYSTEM AND RELATIONSHIP TO PREVIOUS SYSTEMS

	System Structure				
	Speaker-dependent systems			*Speaker-adaptive systems*	
Module	Basic [1]	2005 [20]	2006 [23]	Basic [33], [59]	2007
Statistical					
Models	HMMs	HSMMs [11]	HSMMs	HSMMs [32]	HSMMs
Covariance	Diagonal	Diagonal	Semi-tied [30], [31]	Diagonal	CSMAPLR [33]
Training	ML (SD)	ML (SD)	ML (SD)	ML (SI)	ML (SI)
Clustering	MDL (SD)	MDL (SD)	MDL (SD)	MDL (SI)	MDL (SI)
Average voice	–	–	–	Gender-dependent	Mixed-gender
SAT	–	–	–	Model space [32]	Feature space
Adaptation	–	–	–	CSMAPLR+MAP [33]	CSMAPLR+MAP
Generation	ML [5]	GV [22]	GV	ML	GV
GV cov.	–	Diagonal	Full [23]	–	Full
Signal					
Spectrum	FFT	STRAIGHT smoothed	STRAIGHT smoothed	FFT	STRAIGHT smoothed
Source	Pulse/noise	Mixed [16]	Mixed	Pulse/noise	Mixed

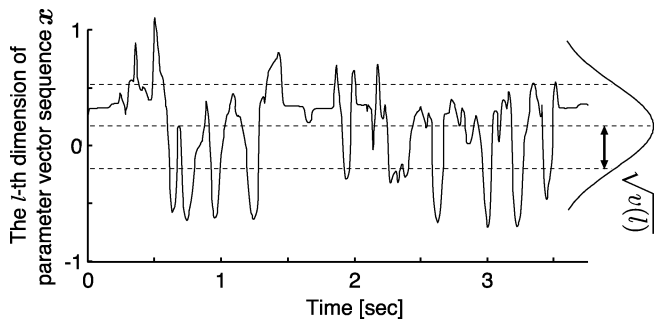


Fig. 6. A GV vector has variance of each dimension of the parameter sequence $\boldsymbol{x}$. $v(l)$ is the lth element of the GV vector $\boldsymbol{v}$.

of the target speaker, while maintaining an appropriate parameter sequence in the sense of maximum likelihood [22]. We use a Newton–Raphson maximization, and employ a sequence obtained from the maximization of $P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, \mathcal{H}_b, T)$, with a trajectory variance which is manipulated to $\boldsymbol{\theta}$, as the initial sequence for the numerical optimization. Here, it is possible to adapt the GV pdf using MAP adaptation. However, the number of parameters of the GV pdf is very small. Specifically, it is equal to the dimensionality of the static features. Hence, we directly estimate the GV pdf from the adaptation data in the following experiments.

Finally, an excitation signal is generated using mixed excitation (pulse plus a noise component band-filtered according to the five aperiodicity parameters) and pitch-synchronous overlap add (PSOLA) [56]. This signal is used excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients and thus generate the speech waveform. These vocoder modules are the same as those of the above Nitech-HTS 2005 speaker-dependent systems [20].[2]

[2]Comparison of these vocoder modules, our conventional vocoder with simple pulse or noise excitation and natural speech in analysis-synthesized speech was reported in [57]. Comparison of them in HMM-based speech synthesis was reported in [20]. Comparison of natural speech, vocoded speech, and HMM-based synthetic speech was reported in [58].

G. Relationship to Previous Systems

Table I shows definition of the proposed system and its relationship to previous systems. As can be seen from the table, two kinds of previous systems can be compared with the HTS-2007 system: a conventional speaker-adaptive system [33], [59] and our speaker-dependent systems for the 2005 and 2006 Blizzard Challenges [20], [23]. Comparing the conventional speaker-adaptive and the HTS-2007 systems, we can assess the effect of the use of STRAIGHT, mixed excitation, and GV. We have previously analyzed the relation between speaker-dependent and speaker-adaptive approaches without STRAIGHT, mixed excitation, and GV [33]. Considering our Blizzard Challenge 2005 and 2006 systems alongside the HTS-2007 system, we can compare speaker-dependent and speaker-adaptive approaches.

The offline procedures such as training, clustering, and adaptation for the HTS-2007 system require more computational costs than those for previous speaker-dependent systems since the system simply has to handle more data from several training speakers.[3] However, since we can concurrently conduct all the procedures per state, per stream, per speaker, and/or per subset of training data, grid computing clusters can straightforwardly deal with the procedures. Computational costs for the online procedures such as parameter generation and vocoding are the same as those for our 2006 systems.

III. EXPERIMENTS

A. Blizzard Challenge 2007

The Blizzard Challenge is an annual evaluation of corpus-based speech synthesis systems, in which participating teams build a synthetic voice from common training data, then synthesize a set of test sentences. Listening tests are used to evaluate the systems in term of naturalness, similarity to original speaker and intelligibility. The Blizzard Challenge 2005 used

[3]Computational costs for each frame are the same as those for our 2006 systems.

the CMU-ARCTIC speech database; in 2006, a database consisting of five hours of speech uttered by a male speaker was released by ATR from their ATRECSS corpus [60]. In the Blizzard Challenge 2007, an extended version of the 2006 corpus was released by ATR, containing eight hours of speech data uttered by the same male speaker [60].

B. Experimental Conditions

We carried out a number of subjective and objective evaluation tests to assess the performance of the new system and to evaluate the HSMM-based feature-space SAT algorithm and the mixed-gender modeling technique. In this section, we report on results using the CMU-ARCTIC and ATRECSS speech databases, employing systems that use a diagonal covariance model. The accuracy of the full-covariance modeling techniques depends strongly on the amount of speech data available; this evaluated in the next section.

The CMU-ARCTIC speech database contains a set of approximately one thousand phonetically balanced sentences uttered by four male speakers (AWB, BDL, JMK, and RMS) and two female speakers (CLB and SLT), with a total duration of about six hours. The ATRECSS speech database was released from ATR to be used in the 2007 Blizzard Challenge and contains the same sentences as CMU-ARCTIC, together with additional sentences, all uttered by a male speaker (EM001), with a total duration of about eight hours. It contains speech from three genres: conversation (3617 utterances), news (1930 utterances), and ARCTIC (1032 utterances). We used the U.S. English phone set "radio" of the Festival speech synthesis system [61], and obtained the phonetic and linguistic contexts from Festival utterance files (as distributed with these corpora) without any modifications.

Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5-ms shift. The feature vectors consisted of 24 STRAIGHT mel-cepstral coefficients (plus the zeroth coefficient), $\log F_0$, aperiodicity measures, and their dynamic and acceleration coefficients. We used five-state left-to-right context-dependent multistream MSD-HSMMs without skip transitions. Each state had a single Gaussian pdf with a diagonal covariance matrix. For speaker adaptation, the transformation matrices were triblock diagonal corresponding to the static, dynamic, and acceleration coefficients. We set the hyperparameters as $\tau_b = \tau_p = 1$ and $\upsilon_b = \upsilon_p = 50$. We set the number of frames T of speech data to be generated to $T = \sum_{i=1}^{N} \widetilde{m}_i$, that is, $\rho = 0$.

C. Implementation Issues

Since the HSMM-based feature-space SAT algorithm mentioned in Section II-C requires substantial computation [62], [63] and it was required to build systems within only one month in the Blizzard Challenge 2007, we had to simplify the training procedures for the average voice model used in our Blizzard Challenge 2007 entry. We first trained the acoustic models using the HMM-based feature-space SAT algorithm. We then roughly estimated initial duration pdfs from HMM trellises [64], and conducted the decision tree-based context and gender clustering for the duration pdfs. Using the tied duration pdfs, we applied the HSMM-based SAT algorithm with piecewise linear regression functions in order to normalize speaker characteristics included in the duration pdfs as well as other acoustic features.

Subsequent to the Blizzard Challenge 2007, we employed an efficient forward–backward algorithm for the HSMMs proposed by Yu and Kobayashi [62], [65], which makes training time for the HSMMs a factor of $ND/(N + D)$ times shorter, where N is the number of states used in an utterance. D is the maximum state duration.[4] Therefore, we were able to use the HSMM-based feature-space SAT algorithm in all the training procedures in additional experiments reported in Sections IV-C and V. The new efficient algorithm for HSMMs has been implemented and released in HTS version 2.1 [4].

D. Evaluation of the Proposed System

We first compared the system proposed in this paper with the conventional speaker-adaptive system [33], [59] in terms of the naturalness and similarity of the synthetic speech. Both systems were constructed using the same training data for the speaker-independent average voice model, and the same adaptation data for the target speaker. We chose male speaker AWB as the target speaker, using three male speakers (BDL, JMK, and RMS) and two female speakers (CLB and SLT) from the CMU-ARCTIC database as training speakers for the average voice model. The average voice model was trained using about 1000 sentences from each speaker, and the system was adapted to the target speaker using 100 sentences selected from the corpus randomly. Finally, a set of ten test sentences—which were not included in either the training or the adaptation data—were used for the subjective evaluations.

We carried out paired comparison tests via the internet, in which 28 subjects were presented with a pair of synthetic speech utterances generated from the adapted models in random order, and asked to indicate which sounded more natural. At the same time, we conducted an "ABX" comparison test to assess the adaptation performance of the average voice models of both systems. In this test, the subjects were presented with a reference utterance from the target speaker, in addition to the above pair of synthesized utterances, and asked which synthetic utterance was most similar to the reference. The same test sentences were used in both tests.

Fig. 7 shows the average preference scores (with 95% confidence interval) of the paired comparison and ABX tests. From this figure, we can see that the naturalness and similarity of the synthetic speech generated from the adapted model using the new system are both greatly improved compared with our previous system. In order to analyze which technique is responsible for this positive result, we separately investigated the effects of STRAIGHT, mixed excitation, feature-space SAT, mixed-gender modeling and GV parameter generation in some preliminary experiments. The results of these preliminary tests indicated that each of the methods had an effect, with the GV parameter generation making the largest single contribution. The

[4]The computational complexity of the new efficient algorithm is $\mathcal{O}(N(D + N)T)$, where N is the number of states used, D is the maximum state duration, and T is the number of total frames of the observations, whereas the conventional forward–backward algorithm requires $\mathcal{O}(N^2DT)$ computations [11], [14].

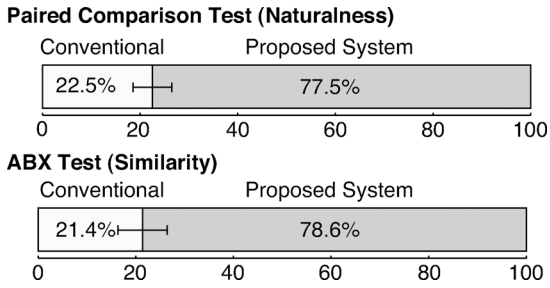


Fig. 7. Average preference scores of the paired comparison test and the ABX test using our conventional system [33], [59] and the proposed system. Target speaker is the English male speaker AWB.

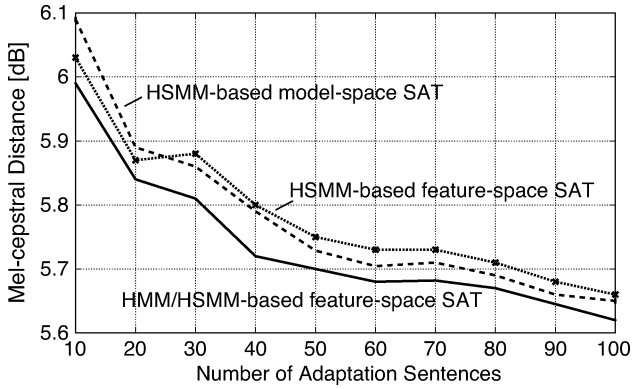


Fig. 8. Objective evaluation of the SAT algorithms: Average mel-cepstral distance (dB). Target speaker is the English male speaker EM001.

amount of adaptation data in these experiments was very limited. The introduction of the new techniques results in an increase in the number of parameters to be estimated. However, it proved possible to robustly apply the GV parameter generation algorithm using the adaptation data.

E. Evaluation of Feature-Space SAT

We evaluated the feature-space SAT algorithm using two types of objective evaluation: the average mel-cepstral distance for the spectral parameters and the RMSE of $\log F_0$. In these evaluations, we chose the male speaker EM001 as the target speaker and used six speakers—four male (AWB, BDL, JMK, and RMS) and two female (CLB and SLT)—from CMU-ARCTIC to train the average voice model. We constructed three kinds of gender-independent average voice model: one using model-space SAT in HSMM embedded training; a second using feature-space SAT in embedded HSMM training; and a third using feature-space SAT for both HMM and HSMM embedded training. Each average voice model was constructed using about 1100 training sentences from each speaker, and the amount of adaptation data ranged from 10–100 sentences. The test set consisted of a further 1000 test sentences from the target speaker. For simplification of the calculation of the average mel-cepstral distance and the RMSE

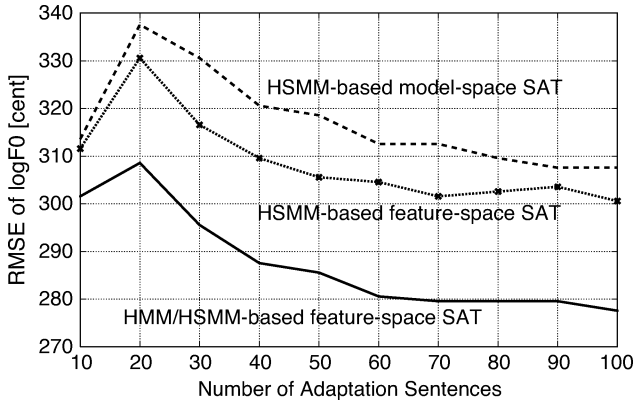


Fig. 9. Objective evaluation of the SAT algorithms: RMSE of $\log F_0$ (cent). Target speaker is the English male speaker EM001.

of $\log F_0$, the state duration of each HSMM was adjusted after Viterbi alignment with the corresponding natural utterance.[5]

The experimental results are shown in Figs. 8 and 9. Fig. 8 shows the average mel-cepstral distance between spectra generated from the adapted model and spectra obtained by analyzing the target speakers' natural utterances. Fig. 9 shows the RMSE of $\log F_0$ between F_0 patterns of synthetic and real speech. Silence, pause, and consonant regions were eliminated from the mel-cepstral distance calculation. The RMSE of $\log F_0$ was calculated in those regions where both the generated and the real F_0 were voiced, since F_0 is not defined in unvoiced regions. Comparing HSMM-based model-space and feature-space SAT only, one sees that the feature-space SAT gives slightly better results in the adaptation of the F_0 parameter, whereas the error of the feature-space SAT is slightly worse for adaptation of the spectral parameters. However, we can also see that when we consistently apply the feature-space SAT to all the embedded training procedures for HMMs and HSMMs, both the mel-cepstral distance and RMSE of $\log F_0$ decrease substantially.

F. Evaluation of the Mixed-Gender Modeling

We evaluated mixed-gender modeling using the same experimental conditions and evaluation measures as for SAT. We constructed gender-independent, gender-dependent, and mixed-gender average voice models, and adapted them to the target speaker using the same adaptation data. Figs. 10 and 11 show the average mel-cepstral distance and RMSE of $\log F_0$ between the synthetic and natural speech. As before, silence, pause, and consonant regions were eliminated from the mel-cepstral distance calculation, and the RMSE of $\log F_0$ was calculated in voiced regions only. Comparing the gender-dependent and mixed-gender average voice models, in the case of 10–50 adaptation sentences, we can see that the gender-dependent modeling has a lower error than the mixed-gender modeling, and is thus the most suitable average voice model to employ in the case of very small amounts of adaptation data. However, as the number of adaptation sentences increases, more of the decision tree nodes containing a question about gender can be used for determining the shared transforms. If

[5]In all the subjective evaluation tests, the state duration of each HSMM was automatically determined using (37)–(41).

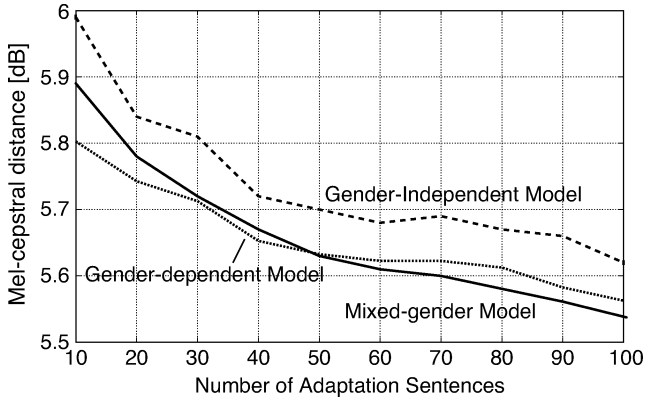


Fig. 10. Objective evaluation of the mixed-gender modeling: Average mel-cepstral distance (dB). Target speaker is the English male speaker EM001.

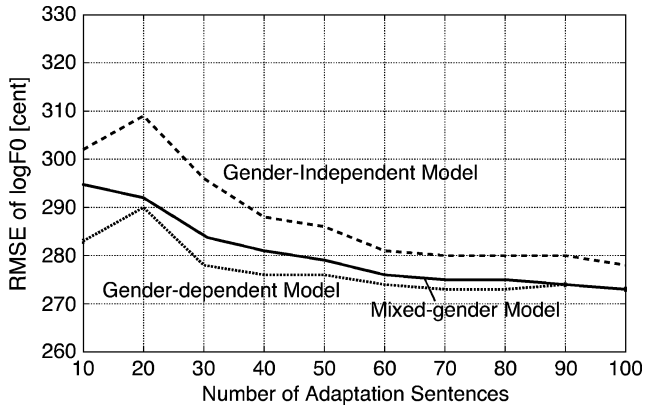


Fig. 11. Objective evaluation of the mixed-gender modeling: RMSE of $\log F_0$ (cent). Target speaker is the English male speaker EM001.

we compare the mel-cepstral distance of the gender-dependent and mixed-gender average voice models in the case of 50–100 adaptation sentences, we can see that mixed-gender modeling gradually becomes better. Mixed-gender modeling makes use of training data from both genders and can create leaf nodes common to both genders, as well as creating gender-dependent ones where necessary. On the other hand, we can see that mixed-gender modeling does not surpass gender-dependent modeling in terms of F_0 error. This is because, in the decision trees for F_0, gender was always split at the root node; hence, there are no mixed-gender leaf nodes.

G. Comparison With Nitech-HTS 2005

Finally, we conducted a comparison category rating (CCR) test to compare the performance of the new system with the speaker-dependent Nitech-HTS 2005 system. The only difference between this Nitech-HTS 2005 system and the system detailed by Zen *et al.* [20] is the dimension of the mel-cepstral coefficients. In [20], 39 mel-cepstral coefficients were used. However, this increases the number of parameters of the matrix for linear transformation. Hence, we used 24 mel-cepstral coefficients for both systems. The experimental condition on the training data in this subsection is the same as for the previous experiments.

We constructed the new system using the training data and adapted the resulting average voice model to the target speaker using 100 sentences of the target speaker EM001. The speaker-dependent system Nitech-HTS 2005 was built using 1000 sentences of the target speaker EM001. For reference, we also compared synthesized speech generated from an adapted model using the same 1000 sentences of the target speaker EM001 as adaptation data. Twenty-five experimental subjects were first presented with synthetic speech from Nitech-HTS 2005 as a reference, then with speech synthesized from the adapted models either using 100 sentences or 1000 sentences (in random order). The subjects were asked to compare the synthetic speech generated from the adapted models with the reference speech using a five-point scale: 2 for better, 1 for slightly better, 0 for almost the same, -1 for slightly worse, and -2 for worse than the reference speech.

The average values and their 95% confidence interval of each adapted model in the CCR tests were 0.140 ± 0.145 for 100 sentences and 0.424 ± 0.08 for 1000 sentences, respectively. The values indicate that the new system can synthesize speech of about the same quality as the Nitech-HTS 2005 system from only 100 adaptation sentences—that is, 10% of the training data for the speaker-dependent systems. This is a significant result, since the Nitech-HTS 2005 system performed very well in the Blizzard Challenge 2005. Furthermore, we can see that the synthetic speech generated from the new system using 1000 sentences is judged to be slightly better than that using 100 sentences and Nitech-HTS 2005 system. This result implies that the speaker-adaptive approach has the potential to surpass the usual speaker-dependent approach. We therefore decided to use the speaker-adaptive approach, even given the large amount of speech data provided in the Blizzard Challenge 2007.

H. Experimental Conditions for The Blizzard Challenge 2007

We used both the CMU-ARCTIC speech database and the ATRECSS speech database for the Blizzard Challenge 2007 as the training data for the average voice model, since the amount of speech data for the target speaker EM001 exceeded that of CMU-ARCTIC, and the purpose of the experiment was not rapid adaptation to a given target speaker, but rather improved quality. To investigate the effect of the corpus size, three systems could be submitted by all participants: one trained using all the speech data included in the released database (Voice A), a second trained using only the ARCTIC subset (Voice B), and a third system trained using a freely selected subset having the same duration of speech as that of the ARCTIC subset (Voice C). Because of the time-consuming training procedures of the HTS-2007 system, we constructed the HTS-2007 systems that use full-covariance models for Voices A and B only.

I. Results of the Blizzard Challenge 2007

Synthetic speech was generated for a set of 400 test sentences, including sentences from conversational, news and ARCTIC genres (used to evaluate naturalness and similarity) and semantically unpredictable sentences (used to evaluate intelligibility) [28]. To evaluate naturalness and similarity, five-point mean opinion score (MOS) and CCR tests were conducted. The scale

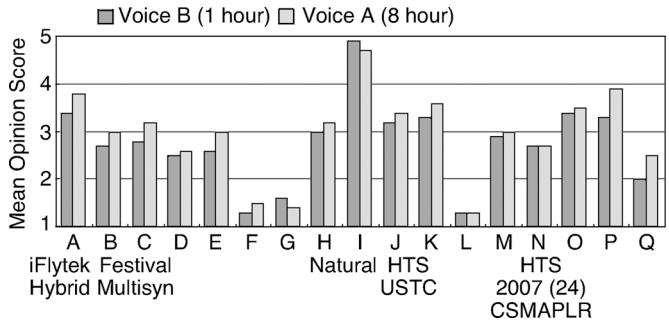


Fig. 12. Mean opinion scores of all systems in the Blizzard Challenge 2007. Target speaker is the English male speaker EM001.

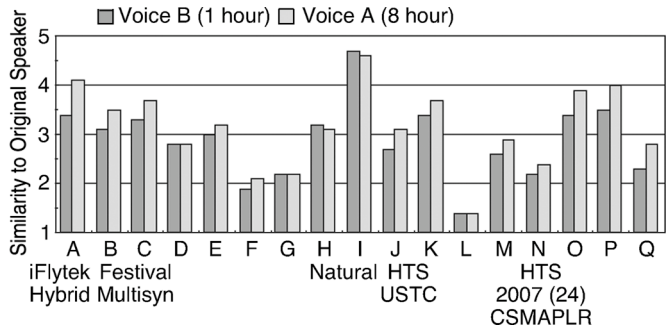


Fig. 13. Average similarity scores to original speaker of all systems in the Blizzard Challenge 2007. Target speaker is the English male speaker EM001.

for the MOS test was 5 for "completely natural" and 1 for "completely unnatural." The scale for the CCR tests was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to natural example sentences from the reference speaker (EM001). To evaluate intelligibility, the subjects were asked to transcribe semantically unpredictable sentences; average word error rates (WER) were calculated from these transcripts. The evaluations were conducted over a six week period via the internet, and a total of 402 listeners participated. For further details of these evaluations, see [28]. For overall analysis of these evaluations, see [66].

Figs. 12–14 show the evaluation results for Voice A (eight hours) and Voice B (one hour) of all 16 participating systems. In these figures, systems "N" corresponds to the HTS-2007 system. In addition "A", "B" and "J" correspond to the HTS system developed by USTC (HTS-USTC) [67], iFlytek hybrid system [67] and the Festival "Multisyn" speech synthesis system [68], respectively, with "I" corresponding to real speech.

These four systems represent the three current major competing TTS methods well: One method is the dominant, established and well-studied technique, "unit-selection." This method concatenates units of speech, selected from a corpus of the target speaker's speech, to create new utterances [69]; The next method is often termed "statistical parametric synthesis," in which a statistical model (usually a HMM) is trained on, or adapted to, the target speaker's speech. Our approach belongs to this category; The final method is a hybrid of the statistical parametric and unit-selection techniques [70], [71], which has been shown to generate very natural-sounding synthetic speech when clean speech data are available for the target speaker [70].

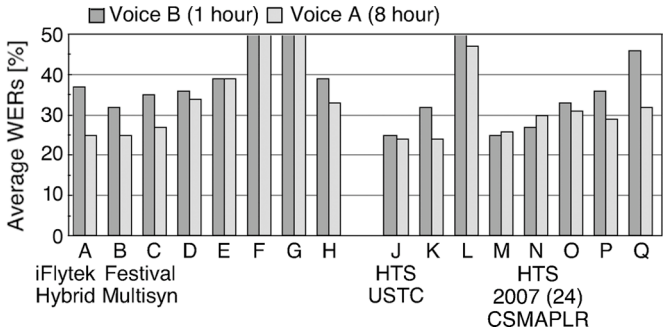


Fig. 14. Average word error rate (%) of all systems in the Blizzard Challenge 2007. Target speaker is the English male speaker EM001.

We give a brief overview of these systems and their relationship to one another, since we will focus on these systems in the following experiments.

Festival Multisyn

Festival [61] is a popular unit-selection speech synthesis system. In the 2007 Blizzard Challenge, Festival's new "Multisyn" module [72], which provides a flexible, general implementation of unit selection and a set of associated voice building tools, was used. HTS-2007 used the existing modules from Festival, resulting in different phonesets and front-end text-processing outputs.

HTS-USTC

The HTS-USTC speech synthesis system [67] is also HMM-based, with context-dependent HMMs for the STRAIGHT spectrum, $\log F_0$ and phone duration being estimated from a single speaker database. There are three principal differences between HTS-USTC and HTS-2007: 1) HTS-USTC used a minimum generation error (MGE) criterion [73], whereas HTS-2007 used the ML/MAP criterion; 2) HTS-USTC used line spectral pair (LSP) features whereas HTS-2007 used mel-cepstrum features to represent the spectrum. The order of those spectral coefficients was also different; 3) HTS-USTC only used data from the target speaker, whereas HTS-2007 was speaker-adaptive.

iFlytek Hybrid

The HTS-USTC and iFlytek systems [67] used the same underlying HMMs but different waveform generation methods. In the HTS-USTC system, speech parameters were generated directly from the statistical models using a parametric synthesiser to reconstruct the speech waveform. On the other hand, the iFlytek system adopted a waveform concatenation method, in which a maximum-likelihood criterion of the statistical models guided the selection of phone-sized candidate units from the single-speaker database [70], [71]. Both systems only used data from the target speaker.

From the results for MOS and CCR tests, we can see that the hybrid system was generally rated higher than other systems. In addition to this, we can see several interesting and important points regarding the HTS-2007 system: 1) the naturalness (Fig. 12) of Voice A for HTS-2007 was evaluated as worse than that of Festival ($p < 0.01$), whereas the naturalness of Voices A and B of another parametric system (HTS-

USTC) were significantly better than those of Festival ($p < 0.01$); 2) compared with the similarity scores for the Festival and HTS-USTC systems (Fig. 13), we observe that HTS-2007 has lower similarity scores in both Voices A and B ($p < 0.01$); 3) compared with the WER results for all the other systems (Fig. 14), we can see that systems which obtained WERs of less than 30% in both Voices A and B are HTS-USTC ("J"), "M", and HTS-2007 ("N") only. Although it is pleasing that the speaker-adaptive HTS-2007 system provides good intelligibility without requiring either manual adjustment of tuning parameters or manual modifications to the database, including speech and label files, the lower naturalness and similarity of the HTS-2007 voices need to be further explored. We analyze this next.

IV. Analysis and Improvement of the HTS-2007 System

Ideally, we would analyze the reasons for the lower naturalness and similarity of the HTS-2007 voices using the same speech database used for the 2007 Blizzard Challenge. However, the license agreement, concluded with ATR for the speech database, forced us to delete both the speech database and constructed systems immediately after the 2007 Blizzard Challenge. Moreover we identified a number of issues that made our analysis either from the above results or from perfectly simulated conditions difficult. Hence, in this section, we have utilized different speech databases for our analysis. In particular we have addressed five main aspects.

Amount of Available Speech Data (Section IV-A): Evaluations for Voice A and Voice B were separately performed in both the above MOS and CCR tests. Thus, because the listeners differed, strictly speaking, we cannot discuss the differences between Voice A and Voice B (that is, the effect of the amount of speech data available).

However, the speaker-adaptive HTS-2007 system works well on the limited amount of speech data available compared to the speaker-dependent HMM-based speech synthesis systems or unit-selection systems trained on enough amount of speech data. Hence, we have simultaneously evaluated the systems built on different amount of speech data and assess the effect of the amount of speech data available.

Configurations for HMMs (Section IV-A): In the previous comparison of HTS-2007 and HTS-USTC, the different criteria used for training/adaptation of HMMs, the spectral representation, and the order of spectral parameters appear to have had a decisive influence on the results.

The benefits of the MGE criterion and LSP features over the ML criterion and mel-cepstrum features were reported in [73]. However, the effect due to the order of spectral parameters is not clear. In particular full-covariance modeling, where the number of parameters to be estimated depends on the order of spectral parameters, should be dealt with both from the point of view of the order of spectral parameters and the amount of speech data available.

Number of target speakers (Section IV-B)

A single target speaker was used in the previous evaluation, rather than evaluating the systems using multiple speakers.

Text processing and contextual features for acoustic units (Section IV-C)

The results in Figs. 12–14 were influenced by the different phonesets and front-end text-processing used in each system. Since the front-end text-processing includes at least lexicon/dictionary, letter-to-sound rules/predictors for out-of-vocabulary words, part-of-speech tagging, pause/phrase break predictors, and accent/stress predictors, the accuracy of each module can affect the quality of synthetic speech. Moreover, the different front-end text-processing always results in different contextual features for acoustic units in HMM-based speech synthesis.

Open and new domain (Section IV-C)

All the test sentences used in the above MOS and CCR tests for naturalness and similarity were *closed/in-domain* sentences. The three genres used in the test sentences—conversation, news and ARCTIC—were the same as those predefined in the training corpus. Although some unit-selection methods have been developed for closed domain applications (and perform very well in such cases), it would be more desirable to be domain-independent and not require information about the domains of either training or test sentences. It would be better to evaluate the systems using new- and open-domain sentences.

Based on these points, we designed the following analysis. In Section IV-A, we first analyze the effects of the amount of speech data available and the order of mel-cepstral analysis in the HTS-2007 system. At the same time, we compare the speaker-adaptive approach of the HTS-2007 system with the previous speaker-dependent approaches, and compare the system using full-covariance modeling using CSMAPLR transforms with those using diagonal covariance and semi-tied covariances (STC) [30], since the relative performance of these methods depends on the amount of data available. In Section IV-B, we evaluate full-covariance modeling using multiple target speakers, since we found the effect of full-covariance modeling varies by speaker. In Section IV-C, we then reevaluate the selected four systems from above, using identical labels, in order to exclude any effect of differing phonesets and front-end text-processing.

A. Evaluation of Amount of Speech Data Available, Order of Mel-Cepstral Analysis and Full-Covariance Modeling

To investigate the effect of the amount of speech data available, we built two speaker-dependent systems (Nitech-HTS 2005, Nitech-NAIST-HTS 2006) and one speaker-adaptive system (HTS-2007) using several sets of sentences spoken by the target speaker EM001. These consisted of: 100 randomly chosen CMU-ARCTIC sentences (6-min duration); the 1032 CMU-ARCTIC sentences used for Voice B (1-h duration); all 6579 sentences used for Voice A (8-h duration). In all HTS-2007 systems, the speech data from the CMU-ARCTIC database was used as part of the training data for the average voice model. For reference, the Festival speech synthesis system using the same speech data of the speaker EM001 was also evaluated as a baseline unit-selection speech synthesis system.

We built the HTS-2007 systems using either 24 or 39 order STRAIGHT mel-cepstral coefficients for each voice, in order to investigate the effect of the model order of the STRAIGHT mel-cepstra. At the same time, systems using diagonal covariance and semi-tied covariance were also built, in order to evaluate full-covariance modeling techniques. In order to assess the effect on only the SMAP criterion and multiple transforms in CSMAPLR, systems with diagonal covariance or semi-tied covariance were built using the following procedures after step 5) for the CSMAPLR and MAP adaptation.

6) Diagonalize the covariance matrices of the transformed model from step 5).
7) Update the mean, diagonalized covariance, and weight of the transformed model based on the MAP criterion. Repeat the update.
8) Using the current semi-tied transform, estimate diagonal elements of the covariance matrices based on the MAP criterion.
9) Using the estimated diagonal elements of the covariance matrices, estimate the current semi-tied transform, which is equivalent to the transform of only the covariance matrices of (15), based on the ML criterion.
10) Go to step 8) unless convergence, or some other appropriate criterion is satisfied.
11) Transform the covariance matrices to full-covariance using the estimated semi-tied transform.

Models with diagonal covariance from step 7) and with semi-tied full-covariance from step 11) were compared to models with CSMAPLR-based full-covariance.

Tables II and III show the number of leaf nodes of the constructed decision trees and memory footprints corresponding to the acoustic models and linear transforms for each system. The number of leaf nodes for the Nitech-NAIST-HTS 2006 system is the same as for Nitech-HTS 2005. Since the number of leaf nodes corresponds to the number of parameter-tied Gaussian pdfs included in the model, we see that the HTS-2007 system can use many more Gaussians compared with speaker-dependent approaches. The memory footprints for the HTS-2007 systems depend on the condition of the speaker adaptation algorithms. For example, when we use a global transformation of the CSMAPLR adaptation only, the speaker-specific part of the memory footprint is 40–55 kB. The remainder of the memory usage is common to all speakers. However, since we focused not on memory requirements but on the quality of synthetic speech, we utilized combined piecewise CSMAPLR and MAP adaptation, which increased the memory footprint (Table III). If we diagonalize the covariance matrices of the adapted model in the parameter generation stage, it would be a better choice to transform the average voice model in advance. In this case, the footprint of the adapted model is identical to that of the average voice model and we can reduce the footprint for the transforms. With full-covariance matrices using the CSMAPLR transforms, the footprint for the transforms are also required.

TABLE II
NUMBER OF LEAF NODES OF CONSTRUCTED DECISION TREES FOR EACH SYSTEM OF EACH VOICE. (a) 6 min. (b) 1 h. (c) 8 h

(a)

System	Mel-cepstrum	$\log F_0$	Aperiodicity	Duration
2005 (24)	230	760	190	125
2005 (39)	170	777	168	138
2007 (24)	3,263	6,124	1,700	3,331
2007 (39)	2,311	11,613	1,504	3,204

(b)

System	Mel-cepstrum	$\log F_0$	Aperiodicity	Duration
2005 (24)	1,371	2,101	911	435
2005 (39)	961	2,096	850	459
2007 (24)	3,530	7,136	1,859	3,746
2007 (39)	2,508	13,034	1,735	3,557

(c)

System	Mel-cepstrum	$\log F_0$	Aperiodicity	Duration
2005 (24)	6,959	11,174	4,590	5,702
2005 (39)	4,598	21,189	3,994	5,110
2007 (24)	7,273	14,245	3,740	8,580
2007 (39)	5,285	31,411	3,747	8,438

Note that no compression techniques were applied to the piecewise CSMAPLR transforms.[6]

We evaluated naturalness and similarity. The reference speech included two recorded sentences spoken by target speaker EM001. In those tests, 33 subjects were presented with a set of synthetic speech utterances generated from the systems in random order.

In order to evaluate naturalness and similarity to the original speaker on out-of-domain sentences, 14 semantically unpredictable test sentences (as used in Blizzard 2007 [28]) were randomly chosen for each subject, from a set of 50 test sentences. Semantically unpredictable sentences were the only out-of-domain sentences in the 2007 Blizzard Challenge. Subjects were asked to rate them using a five-point scale, where 5 corresponded to natural (MOS test) or very similar (CCR test), and 1 corresponded to poor (MOS test) or very dissimilar (CCR test).

Fig. 15 shows the mean scores and 95% confidence intervals for the MOS and CCR tests. For both tests, there are significant differences between the HTS-2007 systems and the speaker-dependent systems when six minutes or one hour of target speech data is used. As the amount of training data available decreases, the differences become more significant. In order to make this speaker-adaptive approach beneficial even when large amounts of target speech data are available, we should train the average voice model from much larger amounts of speech data.

Further results from these experiments concern feature dimensionality and covariance modeling. In the CCR test, there are significant differences between the systems using 24th- or 39th-order STRAIGHT mel-cepstral coefficients when one or eight hours of target speech data are used. The higher feature dimensionality can improve the similarity of synthetic speech, when a large amount of speech data is available.

[6]Voice sizes for Festival above are about 233 MB and 2080 MB, respectively. Note that no compression techniques were applied to waveforms or utterance files.

TABLE III
MEMORY FOOTPRINT (MB) FOR EACH SYSTEM. (a) 6 min. (b) 1 h. (c) 8 h

(a)

		Mel-cepstral analysis			
		24th		39th	
System	Covariance	Acoustic models	Linear transforms	Acoustic models	Linear transforms
2005	Diagonal	0.35	–	0.42	–
2006	Semi-tied (Global)	0.35	0.04	0.42	0.06
2007	Diagonal	3.82	–	4.94	–
2007	CSMAPLR (Global)	3.82	0.04	4.94	0.06
2007	CSMAPLR (Piecewise)	3.82	5.50	4.94	7.29

(b)

		Mel-cepstral analysis			
		24th		39th	
System	Covariance	Acoustic models	Linear transforms	Acoustic models	Linear transforms
2005	Diagonal	1.64	–	1.70	–
2006	Semi-tied (Global)	1.64	0.04	1.70	0.06
2007	Diagonal	4.43	–	5.38	–
2007	CSMAPLR (Global)	4.43	0.04	5.38	0.06
2007	CSMAPLR (Piecewise)	4.43	10.63	5.38	17.37

(c)

		Mel-cepstral analysis			
		24th		39th	
System	Covariance	Acoustic models	Linear transforms	Acoustic models	Linear transforms
2005	Diagonal	8.12	–	9.29	–
2006	Semi-tied (Global)	8.12	0.04	9.29	0.06
2007	Diagonal	8.91	–	11.73	–
2007	CSMAPLR (Global)	8.91	0.04	11.73	0.06
2007	CSMAPLR (Piecewise)	8.91	30.33	11.73	37.37

Thus, we can conclude that one of the reasons the HTS-2007 system had poorer similarity scores in Fig. 13 is the use of 24th-order STRAIGHT mel-cepstral coefficients. The fact that the HTS-USTC system utilized 40th-order STRAIGHT LSP coefficients supports this finding. Contrary to this, in the MOS test there is a significant difference between systems using different order coefficients only in the case of six minutes of target speech data. The HTS-2007 system using 39-dimension mel-cepstra was found to be less natural than that using 24-dimension mel-cepstra only in the case of six minutes of target speech data, presumably due to the number of additional parameters that need to be estimated for the linear transform in the case of higher feature dimensionality. Although CSMAPLR-based full-covariance modeling had the highest scores in the CCR test, the differences were not significant. We discuss the effect of full-covariance modeling more fully in the next subsection.

We can see some important differences to the results reported earlier (Section III-I, Figs. 12–14). First, in Fig. 15(a), the naturalness scores of Voice A of HTS-2007 are now significantly better than those of Festival ($p < 0.01$), whereas before the naturalness (Fig. 12) of Voice A for HTS-2007 was evaluated as worse than that of Festival ($p < 0.01$). Moreover, the naturalness of synthetic speech generated from the Festival unit-selection speech synthesis system becomes much worse as the amount of target speech data becomes smaller ($p < 0.01$). It can be also seen that synthetic speech generated from the HTS-2007 system using six minutes of speech data was rated to be more natural than that of the unit-selection approach using one hour of speech data ($p < 0.01$). This is most likely due to differences in the type of test sentences used in these experiments. The test sentences used in the experiments reported in this subsection were semantically unpredictable sentences [74], with a simple grammatical structure *det-adj-noun-verb-det-adj-noun*, using words of between low and medium frequency. Table IV illustrates how the unit selection system makes more concatenations (as opposed to selecting contiguous units from the database) for the semantically unpredictable sentences. In Fig. 15(b), the similarity scores of the HTS-2007 system are comparable to those of Festival for Voice A and are better for Voice B, whereas we previously observed that HTS-2007 had lower scores in both Voices A and B ($p < 0.01$) than Festival (Fig. 13). In addition to the effect of the semantically unpredictable sentences described above, the differences in the order of the STRAIGHT mel-cepstral analysis also affected the results as shown in Fig. 15(b). These experiments indicate that unit-selection works well for in-domain sentences with eight hours of speech data. In particular, synthetic speech generated by unit-selection has good similarity. However, it loses similarity, and particularly naturalness, for out-of-domain sentences or when little speech data is available. On the other hand, the speaker-adaptive system proposed here is able to maintain naturalness and similarity even for out-of-domain sentences, or when little speech data is available.

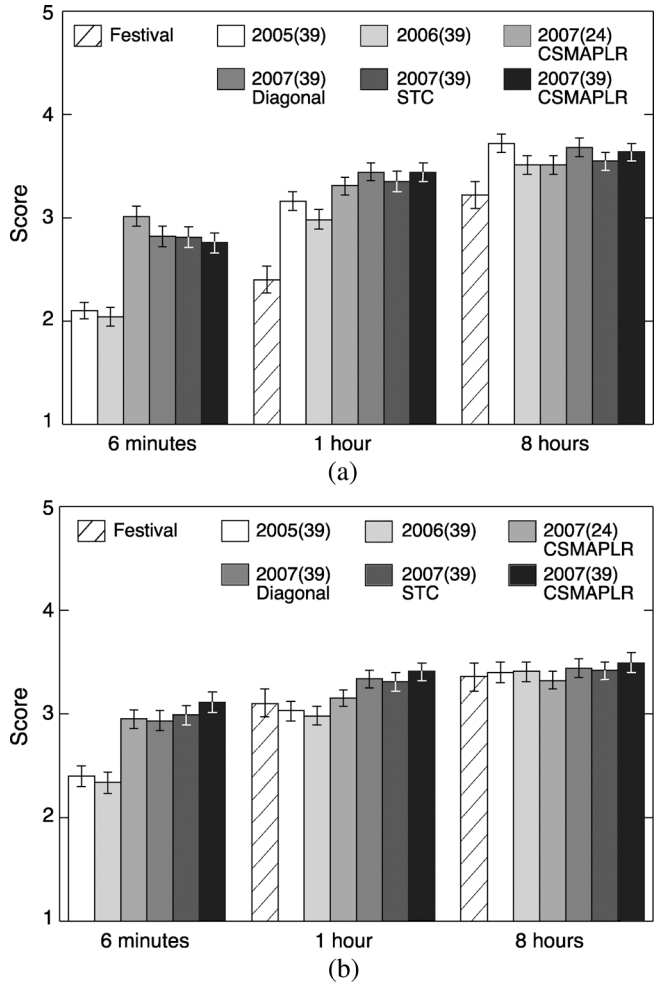


Fig. 15. Subjective evaluation of the English HTS-2007 and previous systems. Target speaker is the English male speaker EM001. (a) MOS test: naturalness. (b) CCR test: similarity.

TABLE IV
PERCENTAGES OF SELECTED DIPHONE UNITS WHICH WERE CONTIGUOUS IN THE CORPUS WITH THE PRECEDING SELECTED DIPHONE IN THE FESTIVAL MULTISYN SYSTEM. AVERAGE, MINIMUM, AND MAXIMUM PERCENTAGES PER UTTERANCE WERE CALCULATED FOR THE TEST SENTENCES FOR THE BLIZZARD CHALLENGE

Genre of test sentences	Average	Min	Max
In-domain	58.6%	36.4%	85.7%
Semantically unpredictable	54.1%	41.9%	68.0%

B. Evaluation of Full-Covariance Modeling With Multiple Target Speakers

The previous experiments involved the evaluation of a single target speaker in English. We also conducted experiments for Japanese speech synthesis using the Nitech-HTS 2005, Nitech-NAIST-HTS 2006, and HTS-2007 systems. To build the Japanese HTS-2007 systems, we used two data sets: first, the ATR Japanese speech database Set B,[7] containing a set of 503 phonetically balanced sentences each uttered by ten speakers (six male: MHO, MHT, MMY, MSH, MTK, and MYI; four female: FKN, FKS, FTK, and FYM), with a duration of about 30 minutes per speaker; Second, a database which contains the same sentences as those of the ATR Japanese speech database (Set B) uttered by a female speaker (FTY) and two male speakers (MJI and MMI), also with a duration of about 30 minutes per speaker. We utilized all the speakers for the training of a Japanese mixed-gender average voice model. Although the effect of full-covariance modeling in the English experiment above was not statistically significant, we found in preliminary experiments that the effect of full-covariance modeling varies by speaker. Thus, in this experiment, we used multiple target speakers for the adaptation of the average voice model. From the training corpus for the average voice model, we chose two female and two male speakers (FTY, FYM, MJI, and MYI) as target speakers. About 30 minutes of adaptation data for each target speaker was available. We also used one female and one male speaker (F109 and M001) as additional target speakers, not included in the training set. Speech data for the speaker F109 was obtained from the ATR Japanese speech database Set C,[8] containing a set of 100 phonetically balanced sentences of the ATR Japanese speech database (Set B), with a duration of about six minutes. Speech data for the speaker M001 was obtained from a Japanese database available from the HTS website,[9] which contains the same sentences as those of the ATR Japanese speech database (Set B), with a duration of about 30 minutes. About six minutes of speech data was used for F109. Two different amounts of data were used for M001: six minutes (the same set of sentences as for F109), and 30 minutes. The evaluation methods that we employed were the same MOS and CCR tests as in the above experiments on English. Ten Japanese male subjects were used, each listening to six test sentences randomly chosen from 50 test sentences from ATR Set B.

Fig. 16 shows the mean scores with 95% confidence interval for the MOS and CCR tests for the Japanese systems using the seven target speakers. Total scores and individual scores for each amount of speech data are shown. From the total scores, it can be seen that CSMAPLR-based full-covariance modeling slightly improves similarity of synthetic speech compared to that using diagonal covariance. Further, from the total scores we can also see that there are significant differences between the speaker-adaptive and speaker-dependent systems in both the MOS and CCR tests. The HTS-2007 system generates better quality synthetic speech than that of the speaker-dependent systems since the amount of speech data used for the target speakers is relatively small. The differences between the HTS-2007 and speaker-dependent systems become even clearer when only six minutes of speech data are used. These results are in agreement with our findings for English.

The effect of full-covariance modeling varied by speaker and did not have much effect for some speakers, while improving similarity others. Fig. 17 shows the scores for the CCR tests for the male speaker M001. Average scores for each amount of

[7]http://www.atr-p.com/sdb.html.

[8]http://www.atr-langue.com/product/index.html.

[9]http://hts.sp.nitech.ac.jp/?Download.

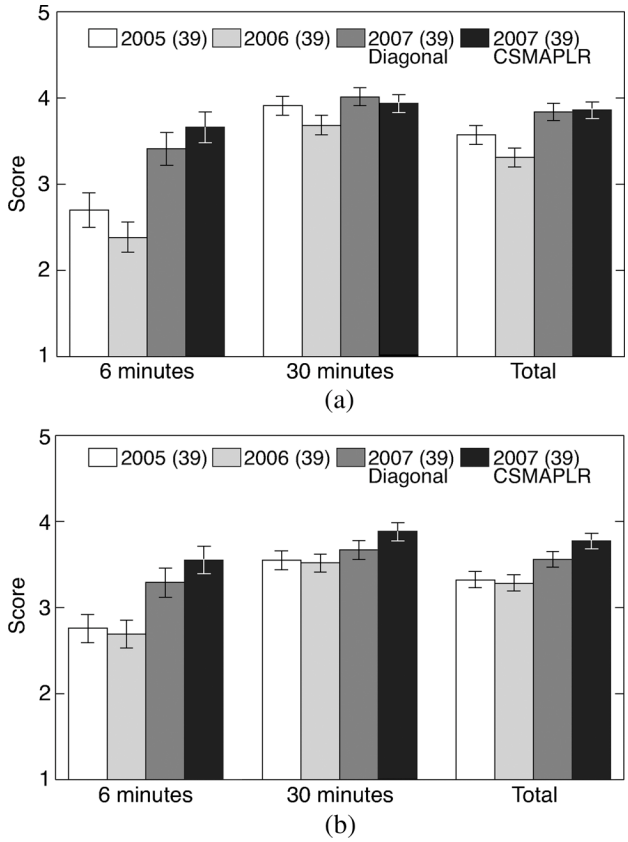


Fig. 16. Subjective evaluation of the Japanese HTS-2007 and past systems. Target speakers are six Japanese speakers. (a) MOS test: naturalness. (b) CCR test: similarity.

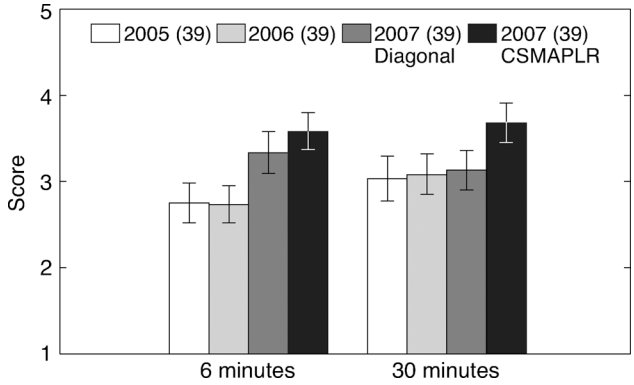


Fig. 17. Subjective evaluation using the CCR test of the Japanese HTS-2007 and past systems. Target speaker is M001.

speech data are shown. For this speaker, the CSMAPLR-based full-covariance modeling is highly effective.

C. Reevaluation With the Same Front-End Text-Processing

We now analyze the influence of the use of different front-end text processing. The easiest way to do this is to separate the influence of the front-end text processing and speech synthesis methods and compare the performance of several systems. In order to do that, we built voices for each of four synthesisers—Festival, HTS-2007, HTS-USTC, and iFlytek Hybrid systems—using the same front-end processing and the same corpus.

Since its use was limited to the 2007 Blizzard Challenge, we were not able to use the ATRECSS speech database, so we selected a different corpus for this evaluation. This corpus contains high quality clean speech data collected under controlled recording studio conditions by a male British English speaker with a received pronunciation (RP) accent. Subsets consisting of 768 randomly chosen sentences (about 1 h in duration), 3063 randomly chosen sentences (about 4 h in duration) and 6691 randomly chosen sentences (about 9.5 h in duration) were used. In all experiments, only target speaker data from the chosen subset was used to build the voice. For example, we did not utilize the full data set to train acoustic models used for segmentation, when building voices on the smaller sets. Note that the speaker-adaptive HTS-2007 system was trained on a substantial amount of clean speech data from other speakers, then adapted using the chosen subset of data from the target speaker. In all the procedures, MSD-HSMMs were used throughout.

Speech signals were sampled at 16 kHz. F_0 for use in all synthesis methods was estimated using the voting method described in Section II-A. The spectral analysis methods varied according to system: Festival uses 12 MFCC coefficients (in the join cost), HTS-2007 uses 39 mel-cepstral coefficients, HTS-USTC uses 40 LSP coefficients, and the iFlytek hybrid system uses 12 mel-cepstral coefficients. Each system may also have energy or the 0th coefficient.

In order to exclude differences in front-end text processing, we used the same labels and lexicon for the voice building and test sentence synthesis in all systems. The labels were generated using Unilex [75] and Festival's Multisyn module. Likewise, the same question set for the clustering of context-dependent HMMs was used in the HTS-2007, HTS-USTC, and iFlytek hybrid systems.

All the systems were used to synthesise the fairy tale "Goldilocks and the Three Bears" and the Festival, HTS-2007, and iFlytek hybrid systems were used to synthesise the story "The Little Girl and the Wolf" by James Thurber. Neither of these texts were in the training data. The reasons we used children's stories for the evaluation were 1) a new domain (this genre was not represented in the training data) and 2) increased naturalness compared with the semantically unpredictable sentences used in Section IV-A. The stories were split up into 12 and 22 utterances, respectively. In the "Little Girl" story, each utterance consisted of a single sentence, whereas each utterance consisted of two sentences in the "Goldilocks" story. 55 subjects (of whom 47 were native speakers) were presented with synthetic speech utterances from the various systems in a random order. They were then asked to score the naturalness of the utterance using MOS on a five point scale, where 5 corresponds to natural and 1 corresponds to unnatural. The listening tests were separately carried out for each story. For the "Goldilocks" story the systems using different amounts of speech data above were evaluated together.

Fig. 18 shows the mean opinion scores, with 95% confidence intervals for the "Little Girl" utterances. From this result, we can see that the HTS-2007 and hybrid system are rated as more natural than the Festival unit-selection system, even for out-of-domain children's story sentences. This confirms our hypothesis that the unit-selection system is less robust for out-of-domain

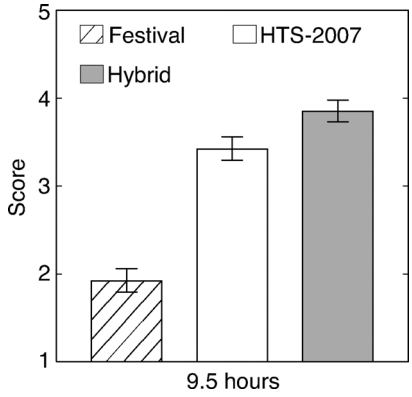


Fig. 18. Subjective evaluation using the "Little Girl" test utterances (one sentence per utterance) synthesized from voices built using the Festival, HTS-2007, HTS-USTC, and iFlytek Hybrid systems. The same front-end text-processing and the same corpus were used in all the systems.

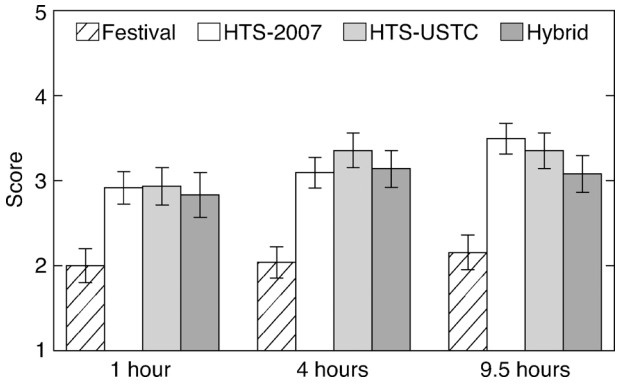


Fig. 19. Subjective evaluation using the "Goldilocks" test utterances (two sentences per utterance) synthesized from voices built using the Festival, HTS-2007, HTS-USTC, and iFlytek Hybrid systems. The same front-end text-processing and the same corpus were used in all the systems.

sentences. However, the hybrid system also uses a unit search and waveform concatenation method similar to that of the unit-selection system, but with a different unit selection criterion. Thus, we can conclude that it is the statistical models used in the HTS-2007 and hybrid systems that provide the robustness to the out-of-domain sentences. The models successfully guide unit selection in the hybrid system by using a maximum-likelihood criterion [70], [71]. In other words, the hybrid system finds better units to concatenate than the unit-selection system, given the same database. Fig. 19 shows the mean opinion scores, with 95% confidence intervals, for the "Goldilocks" utterances. From this figure we can also verify that 1) the unit-selection system is less robust for the out-of-domain sentences, 2) statistical parametric systems are robust by comparison, and 3) the hybrid system benefits from the robustness offered by the statistical parametric models. Comparing Fig. 19 and Fig. 18, we notice that subjects no longer rate the hybrid system as the most natural. Further work is needed to discover if this is because the test utterances consisted of two sentences, or whether there is some other reason. However, this is beyond the scope of this paper and thus we leave that analysis for the future. Compared to Fig. 12, it is surprising how strong the effects of test text and the front-end text processing are. This gives cause for concern and deserves further investigation, in order that we can better understand these various speech synthesis methods. The HTS-2007 system proposed here is comparable in quality to the HTS-USTC or iFlytek systems. The HTS-USTC system benefits from the use of the MGE criterion and LSP features. Integrating those advances into the HTS-2007 system should further improve the quality of synthetic speech.

D. Blizzard Challenge 2008

In the Blizzard Challenge 2008,[10] an English speech database consisting of 15 h of speech uttered by a British male speaker and a Mandarin speech database consisting of about 6 h of speech uttered by a Beijing female speaker were released by the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K., and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, respectively.

For the 2008 Blizzard Challenge, we used the same speaker-adaptive approach to HMM-based speech synthesis that was used for the 2007 challenge, but an improved system was built in which the multi-accented English average voice model was trained on 41 h of speech data with high-order mel-cepstral analysis using an efficient forward-backward algorithm for the HSMM, based on the analysis results above. The listener evaluation scores for the synthetic speech generated from this system was much better than in 2007: the system had the equal best naturalness on the small English data set and the equal best intelligibility on both small and large data sets for English, and had the equal best naturalness on the Mandarin data. In fact, the English system was found to be as intelligible as human speech [76]. These facts also underpin the importance of the above analysis results.

V. Robust Speech Synthesis

Our final experiment concerns what we consider to be a major advantage of the HTS-2007 system over other synthesis methods: it is speaker-adaptive. This system can create synthetic speech with diverse speaker characteristics by transforming the parameters of the average voice models using speaker adaptation techniques. Here, we report an experiment which tests this claim.

The ability to create diverse voices has many potential attractive commercial applications, such as virtual celebrity actors [77], as well as clinical applications such as synthetic replacement voices. The ability to create speech with the characteristics of a particular speaker could be combined with spoken language translation, to personalize speech-to-speech translation: a user's speech in one language can be used to produce corresponding speech in another language, while continuing to sound like the user's voice. This technology would also have applications in dubbing foreign-language television programmes or movies.

In many of these applications, the available speech for the target speaker will always suffer from noise or fluctuations in the recording conditions (changes in environment, microphone type and placement, etc.); this would be expected to significantly degrade the quality of the synthetic speech. Moreover, such "found" speech is unlikely to be phonetically balanced and

[10]http://www.synsig.org/index.php/Blizzard_Challenge_2008.

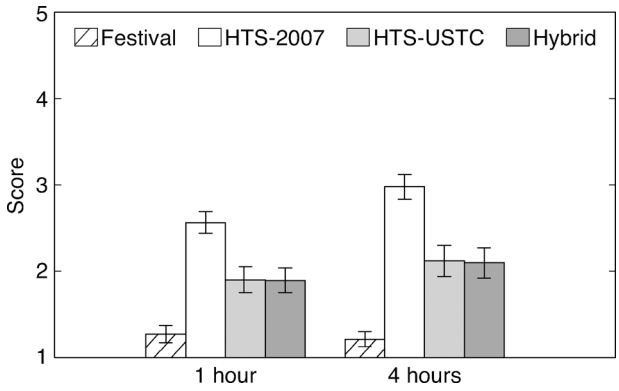


Fig. 20. Subjective evaluation using the "Goldilocks" test utterances (two sentences per utterance) synthesized from voices built using the Festival, HTS-2007, HTS-USTC, and iFlytek Hybrid systems with *noisy data*. The same front-end text-processing and the same corpus were used in all the systems.

will therefore lack some essential acoustic units. This causes severe problems in some systems: for example, concatenative systems must back off to some other unit, which may or may not sound acceptable.

It is not *impossible* to use unit-selection speech synthesis or other techniques in such applications. However, we would expect their performance to be severely impacted by the imperfect data quality. In this section, we therefore analyze how robust the Festival, HTS-2007, HTS-USTC, and iFlytek Hybrid systems are to such less favorable conditions. This is, as far as we know, a new research topic, which we have termed "Robust speech synthesis."

A. Experimental Conditions

The voices for each system were built in the same way as in Section IV-C except for the use of a different corpus. The corpus used here consists of noisy data and was constructed from audio freely available on the web, of a well-known American politician. These data were not recorded in a studio and have a small amount of background noise. The recording condition of the data is not consistent: the environment and microphone may vary. Subsets consisting of 978 randomly chosen sentences (about one hour in duration) and 3846 randomly chosen sentences (about 4 h in duration) were used. For details of this data, please see [77].

B. Evaluation of Speech Synthesis Systems Built From Imperfect and Noisy Data

The evaluation of the voices was also carried out in the same way as in Section IV-C. The same subjects evaluated the "Goldilocks" test utterances. Fig. 20 shows the mean opinion scores, with 95% confidence intervals, for the "Goldilocks" utterances. We can see completely different tendencies from this figure. Comparing Figs. 19 and 20, we notice first that the unit-selection method is very poor indeed on noisy data. This is because inconsistency in the recording conditions from session to session translates into inconsistency in the synthetic speech from unit to unit, which makes the resulting synthetic speech "patchy" and very unnatural sounding. The hybrid system is also vulnerable to the same problem to some extent, since it also concatenates waveforms to generate speech. The speaker-adaptive HTS-2007 system is clearly the most robust of the systems: its performance is least degraded by the use of noisy data. The naturalness of the HTS-2007 system increases as more data become available: the other systems are unable to improve naturalness by using more data. We believe that there are two principal reasons for the superior robustness of the speaker-adaptive HTS-2007 system. The first is that the average voice model is trained from a large amount of clean speech data. Therefore, the decision trees used for tying of HMM parameters are not affected by the noisy data at all. The second is that the speaker adaptation algorithms used in the system include feature transforms. These feature transforms are a generalization of several normalization techniques mentioned previously. They can normalize the fluctuations of the recording conditions, assuming that these can be approximated by linear or piecewise linear regression. The reasons the HTS-USTC system is worse on noisy data constitute a reversal from the advantages for the speaker-adaptive HTS-2007 system; both the estimation and tying of HMM parameters are affected by the noisy data. The MGE criterion used in the HTS-USTC system is especially sensitive to the noisy data.

Our results therefore demonstrate a newly discovered significant advantage of speaker-adaptive HMM-based speech synthesis: "robustness." This ability to generate a synthetic voice from noisy data further expands the potential applications of this technique, and of course dramatically increases the amount of existing data that can now be considered usable for speech synthesis.

VI. Conclusion

We have described the development and evaluation of a speaker-adaptive HMM-based speech synthesis system. The speaker-adaptive approach was further enhanced by two new algorithms: 1) feature-space adaptive training for HSMMs and 2) mixed-gender modeling, and two advanced techniques: 3) CSMAPLR+MAP speaker adaptation and 4) full-covariance modeling using the CSMAPLR transforms. These enhancements were successfully incorporated into our systems that employ STRAIGHT, mixed excitation, HSMMs, GV, and full-covariance modeling.

We demonstrated the effect of the new algorithms in the objective evaluations. In a subjective comparison with a conventional speaker-adaptive system, we showed that the GV algorithm results in synthetic speech of substantially higher quality. Furthermore, from several subjective comparisons with conventional speaker-dependent systems, we found that the speaker-adaptive approach is able to synthesize speech that is significantly better than that synthesized by speaker-dependent approaches in situations with realistic amounts of target speaker data, and bears comparison with those speaker-dependent approaches even when large amounts of speech data are available.

We also compared the performance of the proposed system with several other speech synthesis techniques, representative of the state of the art. From subjective evaluation results (including

the Blizzard Challenge 2007) we show that the new system generates high quality speech.[11] In particular, we have shown that the proposed system is robust, in several ways. It is able to synthesize speech well, even for out-of-domain sentences or when little speech data is available. It can also generate good-quality synthetic speech from less-than-ideal speech data where the data is not perfectly clean, recording conditions are not consistent, and/or the phonetic balance of the texts is not controlled. This robustness is unique to the proposed speaker-adaptive system and opens up possible novel applications for speech synthesis.

The current adaptation framework in HTS-2007 system is supervised: Although it does not require time-alignment information for the target speaker adaptation data, it does require complex context-dependent labels for that data. In order to build voices from only speech data, in a completely automatic fashion, we need to perform the speaker adaptation without such complex context-dependent labels. We now are developing methods to enable *unsupervised* speaker adaptation for speech synthesis, to enable adaptation either without labels or with only simple labels.

ACKNOWLEDGMENT

This research was conducted for the purpose of the Blizzard Challenge 2006 and 2007. The authors would like to thank the ATR Spoken Language Communication Research Laboratories and all the people who contributed to the Blizzard Challenge. The authors would also like to thank Dr. Y. Hifny Abdel-Haleem of the IBM T. J. Watson Research Center for his original idea on mixed-gender modeling and Professor H. Kawahara of Wakayama University for permission to use the STRAIGHT vocoding method. We also thank Dr. K. Richmond of the University of Edinburgh and Prof. T. Kobayashi of the Tokyo Institute of Technology for their valuable comments. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative. (http://www.edikt.org).

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-99*, Sep. 1999, pp. 2374–2350.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," (in Japanese) *IEICE Trans.*, vol. J83-D-II, no. 11, pp. 2099–2107, Nov. 2000.

[3] A. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP 2007*, Apr. 2007, pp. 1229–1232.

[4] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, The HMM-Based Speech Synthesis System (HTS) Version 2.0.1 [Online]. Available: http://www.hts.sp.nitech.ac.jp/

[5] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, May 1995, pp. 660–663.

[6] K. Tokuda, T. Masuko, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from HMM using dynamic features," (in Japanese) *J. Acoust. Soc. Japan*, vol. 53, no. 3, pp. 192–200, Mar. 1997.

[11]The Nitech-HTS 2005 and HTS-2007 (excluding full-covariance modeling) systems are available at the CSTR Festival online demonstration page: http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html.

[7] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP-96*, May 1996, pp. 389–392.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "HMM-based speech synthesis using dynamic features," (in Japanese) *IEICE Trans.*, vol. J79-D-II, no. 12, pp. 2184–2190, Dec. 1996.

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

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[11] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[12] J. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applicat. Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[13] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP-85*, Mar. 1985, pp. 5–8.

[14] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, vol. 1, no. 1, pp. 29–45, 1986.

[15] A. McCree and T. Barnwell, III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[16] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. 2nd MAVEBA*, Sep. 2001, pp. 13–15.

[17] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Audio Process.*, vol. 36, no. 8, pp. 1223–1235, Aug. 1988.

[18] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech'01*, Sep. 2001, 22632266.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," (in Japanese) *IEICE Trans.*, vol. J87-D-II, no. 8, pp. 1565–1571, Aug. 2004.

[20] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[21] A.-H. Ossama, A. S. Mahdy, and R. Mohsen, "Improving Arabic HMM based speech synthesis quality," in *Proc. Interspeech 2006*, Sep. 2006, pp. 1332–1335.

[22] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[23] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Trans. Inf. Syst.*, vol. E91-D, no. 6, pp. 1764–1773, Jun. 2008.

[24] 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*, Sep. 2006.

[25] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system—HTS-2007 system for the Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007 [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_008.html, paper 003.

[26] A. Black and K. Tokuda, "The Blizzard Challenge—2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Eurospeech 2005*, Sep. 2005, pp. 77–80.

[27] C. Bennett and A. Black, "The blizzard challenge 2006," in *Proc. Blizzard Challenge 2006*, Sep. 2006 [Online]. Available: http://festvox.org/blizzard/bc2006/eval_blizzard2006.pdf

[28] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007 [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_001.html, paper 001.

[29] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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 Commun.*, vol. 27, pp. 187–207, 1999.

[30] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, Mar. 1999.

[31] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classfication," in *Proc. ICASSP-98*, May 1998, pp. 661–664.

[32] 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, Feb. 2007.

[33] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009, 2007.

[34] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synth.*, Nov. 1998, pp. 273–276.

[35] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[36] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP-01*, May 2001, pp. 805–808.

[37] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation of pitch and spectrum for HMM-based speech synthesis," (in Japanese) *IEICE Trans.*, vol. J85-D-II, no. 4, pp. 545–553, Apr. 2002.

[38] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[39] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[40] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[41] L. Qin, Z. Ling, Y. Wu, B. Zhang, and R. Wang, "HMM-based emotional speech synthesis using average emotion model," in *Proc. ISCSLP-06 (Springer LNAI Book)*, Dec. 2006, pp. 233–240.

[42] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Commun.*, vol. 48, no. 10, pp. 1227–1242, 2006.

[43] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 294–300, Jul. 1996.

[44] K. Shinoda and C. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, Mar. 2001.

[45] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. Syst.*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.

[46] H. Kawahara, H. Katayose, A. Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech 1999*, Sep. 1999, pp. 2781–2784.

[47] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Elsevier, 1995, pp. 495–518.

[48] ESPS Programs Version 5.0. Entropic Research Laboratory Inc., 1993.

[49] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synth. Workshop*, Sep. 2002, pp. 227–230.

[50] H. Kubozono, "Mora and syllable," in *The handbook of Japanese Linguistics*, N. Tsujimura, Ed. Chichester, U.K.: Blackwell, 1995, pp. 31–61.

[51] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[52] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

[53] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[54] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 3, pp. 503–509, Mar. 2005.

[55] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[56] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–468, 1990.

[57] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. Interspeech 2006*, Sep. 2006, pp. 2266–2269.

[58] O. Watts, J. Yamagishi, K. Berkling, and S. King, "HMM-based synthesis of child speech," in *Proc. 1st Workshop Child, Comput., Interaction (ICMI'08 Post-Conf. Workshop)*, Oct. 2008.

[59] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP-07*, Apr. 2007, pp. 1233–1236.

[60] J. Ni, T. Hirai, H. Kawai, T. Toda, K. Tokuda, M. Tsuzaki, S. Sakai, R. Maia, and S. Nakamura, "ATRECSS—ATR English speech corpus for speech synthesis," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[61] A. Black, P. Taylor, and R. Caley, *The Festival Speech Synthesis System*. Edinburgh, U.K.: Univ. of Edinburgh, 1999.

[62] S.-Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Signal Process. Lett.*, vol. 10, no. 1, pp. 11–14, Jan. 2003.

[63] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMM's," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 213–217, May 1995.

[64] H. Zen, K. Tokuda, T. Masuko, T. Yoshimura, T. Kobayashi, and T. Kitamura, "State duration modeling for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 3, pp. 692–693, Mar. 2007.

[65] H. Zen, Implementing an HSMM-based speech synthesis system using an efficient forward-backward algorithm Nagoya Inst. of Technol., TR-SP-0001, Dec. 2007, Tech. Rep..

[66] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007 [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_003.html, paper 003.

[67] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Y. J. Chen, and G.-P. Hu, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007 [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_017.html, paper 017.

[68] K. Richmond, V. Strom, R. Clark, J. Yamagishi, and S. Fitt, "Festival Multisyn voices for the 2007 Blizzard Challenge," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007 [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_006.html, paper 006.

[69] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP-96*, May 1996, pp. 373–376.

[70] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *Proc. Interspeech 2006*, Sep. 2006, pp. 2034–2037.

[71] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback–Leibler divergence with likelihood criterion," in *Proc. ICASSP-07*, Apr. 2007, pp. 1245–1248.

[72] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Commun.*, vol. 49, no. 4, pp. 317–330, 2007.

[73] Y. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP-06*, May 2006, pp. 89–92 [Online]. Available: http://festvox.org/blizzard/bc2008/hts_Blizzard2008.pdf

[74] C. Benoit, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Commun.*, vol. 18, no. 4, pp. 381–392, 1996.

[75] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech 1999*, Sep. 1999, vol. 2, pp. 823–826.

[76] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Sep. 2008.

[77] M. Aylett and J. Yamagishi, "Combining statistical parametric speech synthesis and unit-selection for automatic voice cloning," in *Proc. LangTech 2008*, Feb. 2008 [Online]. Available: http://www.langtech.it/en/poster/03_AYLETT.pdf

Junichi Yamagishi (M'07) received the B.E. degree in computer science and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively.

He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) from 2004 to 2007. He was an Intern Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from 2003 to 2006. He was a Visiting Researcher at the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K. from 2006 to 2007. He is currently a Senior Research Fellow at the CSTR, University of Edinburgh, and continues the research on the speaker adaptation for HMM-based speech synthesis in an EC FP7 collaborative project called the *EMIME* project (www.emime.org). His research interests include speech synthesis, speech analysis, and speech recognition.

Dr. Yamagishi is a member of the ISCA, IEICE, and ASJ. He pioneered the use of speaker adaptation techniques in HMM-based speech synthesis in his doctoral dissertation "Average-voice-based speech synthesis," which won the Tejima Doctoral Dissertation Award in 2007.

Takashi Nose received the B.E. degree in electronic information processing from the Kyoto Institute of Technology, Kyoto, Japan, in 2001. He is currently pursuing the Ph.D. degree in the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. A major part of his Ph.D. research has focused on HMM-based expressive speech synthesis.

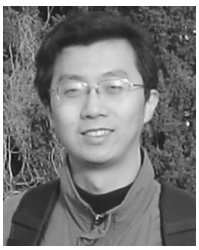
His research interests include speech synthesis, speech analysis, and speech recognition.

Mr. Nose is a member of the IEICE, ISCA, and ASJ.

Heiga Zen received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr.Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from the Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively.

During 2003, he was an Intern Researcher at the ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an Intern/Co-Op Researcher in the Human Language Technology Group, IBM T. J. Watson Research Center, Yorktown Heights, NY. He is currently a Postdoctoral Fellow at the Nagoya Institute of Technology. He is a main maintainer of the HMM-Based Speech Synthesis System (HTS). His research interests include statistical speech recognition and synthesis.

Dr. Zen received the Awaya Award and the Itakura Award from the Acoustical Society of Japan (ASJ) in 2006 and 2008, respectively, and he is a corecipient of the TELECOM System Technology Prize from the Telecommunications Advancement Foundation (TAF) in 2008 and the Information and Systems Society Best Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2008. He is a member of the ASJ, IPSJ, and ISCA.



Zhen-Hua Ling received the B.E. degree in electronic information engineering and the M.S. and Ph.D. degrees in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008, respectively.

From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K. He is currently a Joint Postdoctoral Researcher at University of Science and Technology of China and iFlytek Co., Ltd., Hefei, China. His research interests include speech synthesis, voice conversion, speech analysis, and speech coding.

Tomoki Toda (M'05) received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively.

From 2001 to 2003, he was an Intern Researcher at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He was a Research Fellow of the Japan Society for the Promotion of Science in the Graduate School of Engineering, Nagoya Institute of Technology, from 2003 to 2005. He was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, from October 2003 to September 2004. He is currently an Assistant Professor in the Graduate School of Information Science, NAIST, and a Visiting Researcher at the ATR Spoken Language Communication Research Laboratories. He is also a Visiting Researcher at the Department of Engineering, University of Cambridge, Cambridge, U.K., from March 2008 to August 2008. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, speech production, speech analysis, and speech recognition.

Dr. Toda received the TELECOM System Technology Award for Students and the TELECOM System Technology Award from the Telecommunications Advancement Foundation, in 2003 and 2008, respectively. He also received the Information and Systems Society Best Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE), Japan, in 2008. He has been a member of the Speech and Language Technical Committee of the IEEE Signal Processing Society since January 2007. He is a member of the ISCA, IEICE, and ASJ.

Keiichi Tokuda (M'89) received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively.

From 1989 to 1996, he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology, where he is now a Professor. He is also an Invited Researcher at the ATR Spoken Language Translation Research Laboratories, Japan, and was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.

He is a corecipient of the Paper Award and the Inose Award both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member of the ISCA, IPSJ, IEICE, ASJ, and JSAI.

Simon King (M'95–SM'08) received the M.A. (Cantab) degree in engineering and the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K., in 1992 and 1993, respectively, and the Ph.D. degree in speech recognition from the University of Edinburgh, Edinburgh, U.K., in 1998.

He has been involved in speech technology since 1992, and has been with the Centre for Speech Technology Research, University of Edinburgh, since 1993. He is a Reader in Linguistics and English Language and an EPSRC Advanced Research Fellow. His interests include concatenative and HMM-based speech synthesis, speech recognition, and signal processing, with a focus on using speech production knowledge to solve speech processing problems.

He is a member of the ISCA, serves on the steering committee for SynSIG (the special interest group on speech synthesis) and co-organizes the Blizzard Challenge. He is an Associate Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Steve Renals (M'91) received the B.Sc. degree in chemistry from the University of Sheffield, Sheffield, U.K., in 1986, the M.Sc. degree in artificial intelligence in 1987, and the Ph.D. degree in speech recognition and neural networks in 1990, both from the University of Edinburgh, Edinburgh, U.K.

He is a Professor in the School of Informatics, University of Edinburgh, where he is the Director of the Centre for Speech Technology Research. From 1991 to 1992, he was a Postdoctoral Fellow at the International Computer Science Institute, Berkeley, CA, and was then an EPSRC Postdoctoral Fellow in Information Engineering at the University of Cambridge, Cambridge, U.K. (1992–1994). From 1994 to 2003, he was a Lecturer and Reader at the University of Sheffield, moving to the University of Edinburgh in 2003. His research interests are in the area of signal-based approaches to human communication, in particular spoken language processing and machine learning approaches to modeling multimodal data. He has over 100 publications in these areas.

Dr. Renals is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and has been a member of the Technical Committee on Machine Learning and Signal Processing.