Speech Synthesis Based on Hidden Markov Models

Keiichi Tokuda, *Member, IEEE,* Yoshihiko Nankaku, *Member, IEEE,* Tomoki Toda, *Member, IEEE,* Heiga Zen, *Member, IEEE,* Junichi Yamagishi, *Member, IEEE,* and Keiichiro Oura

***Abstract*—This paper gives a general overview of hidden Markov model (HMM)-based speech synthesis, which has recently been demonstrated to be very effective in synthesizing speech. The main advantage of this approach is its flexibility in changing speaker identities, emotions, and speaking styles. This paper also discusses the relation between the HMM-based approach and the more conventional unit-selection approach that has dominated over the last decades. Finally, advanced techniques for future developments are described.**

***Index Terms*—text-to-speech synthesis, hidden Markov model, HMM-based speech synthesis, statistical parametric speech synthesis, HTS**

I. Introduction

Text-to-speech (TTS) synthesis is a technique for generating intelligible, natural-sounding artificial speech for a given input text. It has been used widely in various applications including in-car navigation systems, e-book readers, voice-over functions for the visually impaired, and communication aids for the speech impaired. More recent applications include spoken dialogue systems, communicative robots, singing speech synthesizers, and speech-to-speech translation systems.

Typical TTS systems have two main components, text analysis and speech waveform generation, which are sometimes called *front-end* and *back-end*, respectively. In the text analysis component, given input text is converted into a linguistic specification consisting of elements such as phonemes. In the speech waveform generation component, speech waveforms are generated from the produced linguistic specification. The main focus of this paper is the speech waveform generation component and we omit details of the text analysis module [1].

Approaches for speech waveform generation from given text have progressed from knowledge- and rule-based ones to data-driven ones. In the early 1970s, the speech waveform generation component used very low dimensional acoustic parameters for each phoneme, such as formants, corresponding to vocal tract resonances [2]. In the 1980s, the speech waveform generation component used a small database of phoneme units called "diphones" (the second half of one phone plus the first half of the following) and concatenated them according to the given phoneme sequence by applying signal processing, such as linear predictive (LP) analysis, to the units [3].

In the 1990s, with the increase in the power and resources of computer technology and also the increase in speech and linguistics resources, larger speech databases were collected and used to select more appropriate speech units that match both phonemes and other linguistic contexts such as lexical stress, pitch accent, and part-of-speech information in order to generate high-quality natural sounding synthetic speech with appropriate prosody. This approach is generally called "unit selection," and various systems including commercial systems were developed resulting in a higher level of reading-style synthetic speech [4]–[8].

For applications, such as screen reader and newspaper read-out functions, reading-style synthetic speech may be sufficient. However, there are other potential applications where TTS systems are required to read out texts with expressivity. The unit selection method, however, restricts the output speech to the same style as that in the original recordings as no (or few) modifications to the selected pieces of recorded speech are normally done. If we need to generate synthetic speech with various speaking styles and emotions with this method, larger speech databases containing different speaking styles are always required. IBM's stylistic synthesis [9] is a good example; however, the size of the speech database becomes exponentially larger and further recording of a large quantity of speech with various speaking styles and emotions is obviously cost-inefficient and time consuming [10].

With such a need for more control over speech "variations," another data-driven approach called "statistical parametric speech synthesis" emerged in the late 1990s and has grown in popularity in recent years [11]–[14]. In this approach, several acoustic parameters are modeled using a time-series stochastic generative model. Statistical parametric speech synthesis which uses a hidden Markov model (HMM) as its generative model is typically called *HMM-based speech synthesis*. HMMs represent not only the phoneme sequences but also various contexts of the linguistic specification in a similar way to the unit selection approach, and acoustic parameters generated from HMMs selected according to the linguistic specification are used to drive a vocoder, which is a simplified speech production model, in which speech is represented by vocal tract parameters and excitation parameters, in order to generate a speech waveform.

Thanks to efficient and well-established machine learning algorithms, which mostly originated in the automatic speech

The research leading to these results was partly funded by JST CREST (uDialogue), EPSRC grants EP/I031022/1 (NST) and EP/J002526/1 (CAF).

K. Tokuda, Y. Nankaku and K. Oura are with the Department of Computer Science and Engineering, Nagoya Institute of Technology (NITech), Nagoya, 466-8555 Japan. E-mail: tokuda@nitech.ac.jp, nankaku@sp.nitech.ac.jp, uratec@nitech.ac.jp.

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

H. Zen was with Nagoya Institute of Technology, Nagoya, Japan and Toshiba Research Europe Ltd. Cambridge Research Lab., Cambridge, United Kingdom. He is now with Google, London, SW1W 9TQ United Kingdom. E-mail: heigazen@google.com

J. Yamagishi is with the Center for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9AB United Kingdom. E-mail: jyamagis@inf.ed.ac.uk.

Manuscript received April 19, 2012; revised November 21, 2012.

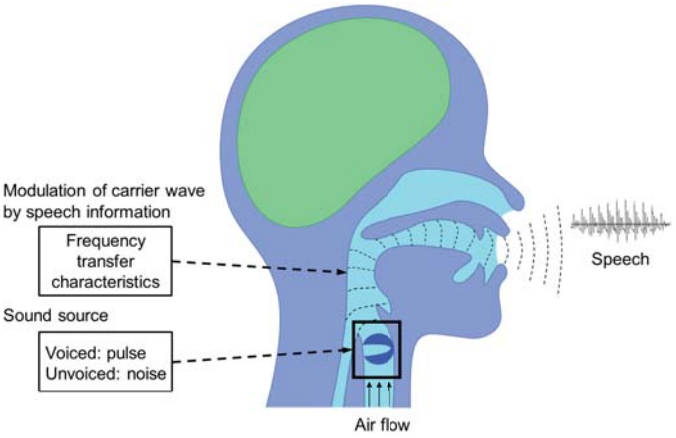

Fig. 1. Overview of human speech production.

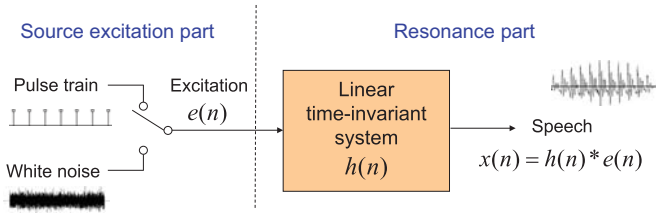

Fig. 2. Source-filter model that simulates human speech production shown in Fig. 1.

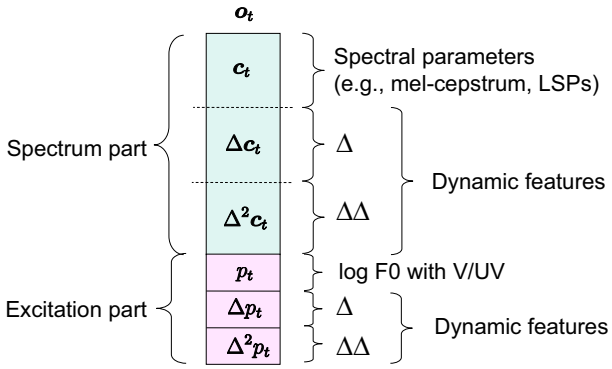

Fig. 3. Example of an observation vector at each frame.

recognition (ASR) field, such as Baum-Welch, Viterbi, and clustering algorithms [15] and various open-source toolkits that cover text analysis, signal processing, and HMMs [16]–[19], HMM-based speech synthesis has been a major topic in speech synthesis research and used worldwide by both academic and commercial organizations. About 76% of speech synthesis papers published in INTERSPEECH 2012, which is a major international conference on speech information processing, have used HMM-based approaches, and this trend strongly confirms the need for and potential of this new approach.

The quality of HMM-based synthetic speech has been improving, e.g., [20]–[23], and many techniques for controlling speech variations, e.g., [1], [24]–[35] have also been proposed. Commercial products based on the HMM-based speech synthesis approach, e.g., [36]–[39], have been available in the market.

The aim of this paper is to give a general overview of popular techniques used in HMM-based speech synthesis. Although many research groups have contributed to the recent progress in HMM-based speech synthesis, please note that the description given here is somewhat biased toward implementation of the HMM-based speech synthesis system called HTS [11], [40].

The rest of this paper is organized as follows. Section II introduces the fundamentals of the HMM-based speech synthesis system. Section III describes the flexibility of HMM-based speech synthesis, and open source software tools are introduced in Section IV. The relation between the HMM-based and unit selection speech synthesis approaches is discussed in Section VI, and the recent development is described in Section VII. Future directions are described in Section VIII. Concluding remarks are presented in the final section.

II. HMM-BASED SPEECH SYNTHESIS

A. Speech production and vocoder

It is well known that the speech production process (Fig. 1) may be approximated using a digital filter shown in Fig. 2. This implementation is based on the source filter theory of voice production [41] and is therefore called the source filter model. The most straightforward such model uses a white excitation (pulse train or noise) filtered with a single resonance filter to model the acoustic speech pressure wave, where spectral envelopes of the glottal flow, vocal tract resonance, and lip radiation effect are modeled all together by the single resonance filter. This model comprises: 1) voicing information, 2) fundamental frequency (F_0), and 3) spectral envelope represented by, e.g., mel-cepstral coefficients [42], and speech waveforms can be reasonably reconstructed from the sequence of these acoustic parameters. In HMM-based speech synthesis, HMMs predict these vocoder parameters from the given text. By concatenating spectral and excitation parameter vectors at each frame, we can form an observation vector at each frame. A typical form of the observation vector, which includes not only static but also dynamic features, will be mentioned in detail in Section II-C and is shown in Fig. 3. In addition to the mel-cepstral coefficients, various spectral representations, such as line spectral pairs (LSPs) [43], mel-generalized cepstral coefficients [44], and various excitation parameters (e.g., aperiodicities [45]) can also be used.

B. Hidden Markov model

Fig. 4 shows an example of a 3-state left-to-right HMM. An N-state HMM λ (e.g., corresponding to an utterance) is characterized by sets of initial-state probabilities $\{\pi_i\}_{i=1}^N$, state-transition probabilities $\{a_{ij}\}_{i,j=1}^N$, and state-output probability distributions $\{b_i(\cdot)\}_{i=1}^N$. The $\{b_i(\cdot)\}$ are typically assumed to

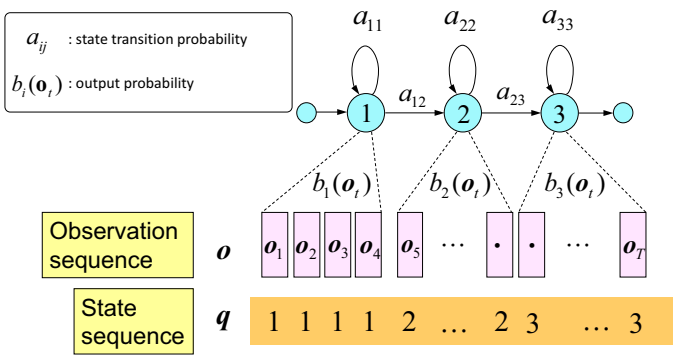

Fig. 4. Example of a 3-state, left-to-right hidden Markov model.

be single multivariate Gaussian distributions for simplicity;

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\right\} \tag{2}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are a d-by-1 mean vector and a d-by-d covariance matrix, respectively, d is the dimension of the acoustic parameters, and $\boldsymbol{o}_t$ is an observation vector, which consists of the vocoder parameters at frame t.

Since the HMM is a generative model, the basic concept of HMM-based speech synthesis is straightforward. Let $\boldsymbol{O} = [\boldsymbol{O}_1^\top, \boldsymbol{O}_2^\top, \ldots, \boldsymbol{O}_T^\top]^\top$, and $\mathcal{W}$ be a set of speech parameters and corresponding linguistic specifications (such as phoneme labels) to be used for the training of HMMs, respectively, and $\boldsymbol{o} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_{T'}^\top]^\top$ and w be speech parameters and corresponding linguistic specifications that we want to generate at synthesis time. The training of HMMs and synthesis from HMMs are simply written as follows:

$$\text{Training: } \lambda_{\max} = \arg\max_{\lambda} p(\boldsymbol{O} \mid \lambda, \mathcal{W}) \tag{3}$$

$$p(\boldsymbol{O} \mid \lambda, \mathcal{W}) = \sum_{\forall q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{O}_t) \tag{4}$$

$$\text{Synthesis: } \boldsymbol{o}_{\max} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \lambda_{\max}, w) \tag{5}$$

where $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence.

C. Speech Parameter Generation from HMM

Problem of Parameter Generation: The basic idea of the speech parameter generation algorithm is simple. The most probable speech parameter vector sequence given a set of HMMs and a text to be synthesized is determined as

$$\boldsymbol{o}_{\max} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \lambda_{\max}, w) \tag{6}$$

$$= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid \lambda_{\max}, w) \tag{7}$$

$$\approx \arg\max_{\boldsymbol{o}, \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q} \mid \lambda_{\max}, w) \tag{8}$$

$$= \arg\max_{\boldsymbol{o}, \boldsymbol{q}} p(\boldsymbol{o} \mid \boldsymbol{q}, \lambda_{\max}) P(\boldsymbol{q} \mid \lambda_{\max}, w) \tag{9}$$

$$\approx \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \boldsymbol{q}_{\max}, \lambda_{\max}) \tag{10}$$

$$= \arg\max_{\boldsymbol{o}} \prod_{t=1}^{T'} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{q_{\max,t}}, \boldsymbol{\Sigma}_{q_{\max,t}}) \tag{11}$$

where

$$\boldsymbol{q}_{\max} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \lambda_{\max}, w). \tag{12}$$

The maximization problem of Eq. (12) can easily be solved by state-duration probability distributions. The maximization problem of Eq. (10) is maximizing $p(\boldsymbol{o} \mid \boldsymbol{q}, \lambda)$ with respect to $\boldsymbol{o}$ given the pre-determined state sequence $\boldsymbol{q}_{\max}$.

Incorporating Dynamic Feature Constraints: From Eq. (2), $p(\boldsymbol{o} \mid \boldsymbol{q}_{\max}, \lambda)$ is maximized if $\boldsymbol{o}_t = \boldsymbol{\mu}_{q_t}$, $t = 1, 2, \ldots, T'$, that is, the speech parameter vector sequence becomes a sequence of the mean vectors. Because of the conditional independence of state-output probabilities assumed in the HMM, the mean vector sequence results in a step-wise sequence. This is unrealistic as speech parameters extracted from the natural speech vary smoothly. We can perceive discontinuities at state boundaries in a speech waveform that is resynthesized from the step-wise speech parameter sequence.

To avoid this problem, the speech parameter generation algorithm introduces the relationship between static and dynamic features as constraints of the maximization problem [46]. The use of the dynamic features (first and second-order time derivatives of speech parameters) as a part of the observation vector as shown in Fig. 3 is a simple but powerful mechanism for capturing time dependencies within the HMM framework. It greatly improves the performance of HMM-based automatic speech recognizers. It is assumed that the speech parameter vector $\boldsymbol{o}_t$ consists of the static feature[1] c_t and its dynamic feature[2] Δc_t as

$$\boldsymbol{o}_t = [c_t, \Delta c_t]^\top. \tag{13}$$

For simplicity, the dynamic feature $\Delta^2 c_t$ in Fig. 3 is omitted from this equation. The dynamic features are often calculated as regression coefficients from their neighboring static features, i.e.,

$$\Delta c_t = \sum_{\tau=-L}^{L} w(\tau) c_{t+\tau}, \tag{14}$$

where $\{w(\tau)\}_{\tau=-L}^{L}$ are window coefficients to calculate dynamic features. Usually, the maximum window length L is

[1]For notation simplicity, the static feature $\boldsymbol{c}_t$ is assumed to be a scalar value. Extension to vectors is straightforward. Usually the vector size of $\boldsymbol{c}_t$ is about $20 \sim 50$ depending on the sampling frequency.

[2] Using higher-order dynamic features is straightforward.

set to 1–4. To simplify the following discussion, the most straightforward case of Δc_t is considered;

$$\Delta c_t = c_t - c_{t-1}. \tag{15}$$

The relationship between the observation vector sequence $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \ldots, \boldsymbol{o}_{T'}^\top\right]^\top$ and static feature sequence $\boldsymbol{c} = [c_1, \ldots, c_{T'}]^\top$ can be arranged in a matrix form as

$$\begin{array}{ccc} \boldsymbol{o} & & \boldsymbol{W} \\ & & \begin{array}{cccccc} \cdots & t-2 & t-1 & t & t+1 & \cdots \end{array} \\ \left[\begin{array}{c} \vdots \\ \hline c_{t-1} \\ \Delta c_{t-1} \\ \hline c_t \\ \Delta c_t \\ \hline c_{t+1} \\ \Delta c_{t+1} \\ \hline \vdots \end{array}\right] & = & \left[\begin{array}{cccccc} \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \hline \cdots & 0 & 1 & 0 & & \cdots \\ \cdots & -1 & 1 & 0 & & \cdots \\ \hline \cdots & 0 & 0 & 1 & 0 & \cdots \\ \cdots & 0 & -1 & 1 & 0 & \cdots \\ \hline \cdots & & 0 & 0 & 1 & \cdots \\ \cdots & & 0 & -1 & 1 & \cdots \\ \hline \cdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{array}\right] \end{array} \left[\begin{array}{c} \vdots \\ \hline c_{t-2} \\ \hline c_{t-1} \\ \hline c_t \\ \hline c_{t+1} \\ \hline \vdots \end{array}\right]. \tag{16}$$

Under this *deterministic* relationship, maximizing the output probability with respect to $\boldsymbol{o}$ is equivalent to maximizing the output probability with respect to $\boldsymbol{c}$ as

$$\boldsymbol{c}_{\max} = \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{W}\boldsymbol{c}; \boldsymbol{\mu}_{\boldsymbol{q}_{\max}}, \boldsymbol{\Sigma}_{\boldsymbol{q}_{\max}}). \tag{17}$$

By equating the partial derivative of the logarithm of Eq. (17) with respect to $\boldsymbol{c}$ to $\boldsymbol{0}$, a set of linear equations to determine the most probable static feature vector sequence is derived as

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{W} \boldsymbol{c} = \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} \boldsymbol{\mu}_{\boldsymbol{q}} \tag{18}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_{T'}}^\top\right]^\top \tag{19}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{q}} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{q_1}, \ldots, \boldsymbol{\Sigma}_{q_{T'}}\right]. \tag{20}$$

The set of linear equations can be solved efficiently with the Cholesky decomposition with $\mathcal{O}(T')$ operations.

Example of Generated Parameters: Fig. 5 shows an example of statistics and generated speech parameters from a sentence-level HMM composed by concatenating two phoneme-level HMMs. The trajectory of the zero-th mel-cepstral coefficient $c(0)$ in the generated speech parameters and its dynamic features are shown. Each vertical dotted line represents a state output. Since the covariance matrix is assumed to be diagonal, each state has its mean and variance: each horizontal dashed line and the shaded area represent the state mean and the standard deviation of the state, respectively. The three trajectories, static, delta, and delta-delta, are constrained by Eq. (16),[3] and determined by maximizing their output probabilities. As a result, the trajectory is constrained to be realistic as determined from the statistics of both static and dynamic features.

[3] Please note that Eqs. (13)–(16) do not include delta-delta whereas Fig. 3 does.

D. Training part

Fig. 6 is a block diagram of a basic HMM-based speech synthesis system. It consists of training and synthesis parts as we mentioned earlier. The training part performs the maximum likelihood estimation of the HMM parameters by using the Baum-Welch algorithm. This process is similar to the one used for speech recognition: however, there are several differences that are worth mentioning.

Feature vectors and state output probabilities: Since we need to drive a source filter vocoder, HMMs need to model both spectral parameters, such as mel-cepstral coefficients, and excitation parameters, such as F_0, at the same time, whereas HMMs used in automatic speech recognition (ASR) typically use only spectral parameters, which are modeled by continuous distributions.

However, we cannot directly apply both the conventional discrete and continuous HMMs to F_0 pattern modeling since F_0 values are not defined in the unvoiced region, i.e., the observation sequence of an F_0 pattern is composed of one-dimensional continuous values and discrete symbols that represent "unvoiced" as shown in Fig. 7. Although several methods have been investigated for modeling F_0 sequences [47]–[49], the HMM-based speech synthesis system uses multi-space probability distributions [50] for modeling them. A typical multi-space probability distribution for F_0 modeling consists of a continuous distribution for voiced frames and a discrete distribution for unvoiced frames. By switching the continuous and discrete space according to the space label associated with each observation, it can model variable dimensional observation vector sequences, such as F_0 sequences, without heuristic assumptions. To keep synchronization between spectral parameters and F_0 parameters, they are modeled simultaneously by separate streams in a multi-stream HMM [51], which uses different state output probability distributions for modeling individual parts of the observation vector, i.e., the continuous distributions are used as stream-output probability distributions for modeling the spectral parameters and the multi-space probability distributions are used as those for modeling the F_0 parameters.

Explicit duration modeling: Each HMM also has its explicit state-duration probability distribution to model the temporal structure of speech [52] instead of transition probabilities. In the standard HMM case, the state duration probability exponentially decreases with increase of duration. However, it is too simple to control the temporal structure of the speech parameter sequence. Instead, HMM-based speech synthesis typically uses a semi-Markov structure in which the temporal structure is approximated by a Gaussian distribution [53].

Context dependency: Another difference is that linguistic specifications have been taken into account. In addition to phoneme information, HMM-based speech synthesis uses various linguistic contexts such as lexical stress, pitch accent, tone, and part-of-speech information for the context-dependent

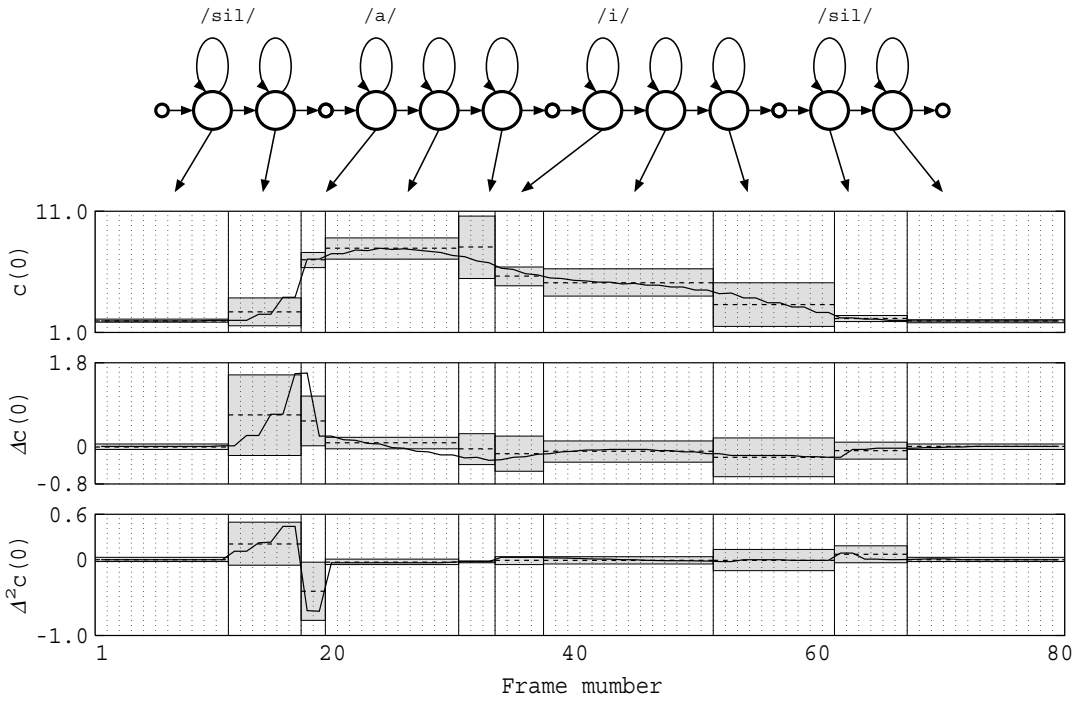

Fig. 5. Example of statistics and generated parameters from a sentence-level HMM composed of phoneme-level HMMs for /a/ and /i/. The dashed line and shading show the mean and standard deviation, respectively, of a Gaussian pdf at each state.

modeling of HMMs [54]. Although spectral parameters are mainly affected by phoneme information, prosodic and duration parameters may be affected by supra-segmental linguistic information. For example, the contexts used in the HTS English recipes [18] include the following contexts:

- Phoneme:
  - current phoneme
  - preceding and succeeding two phonemes
  - position of current phoneme within current syllable
- Syllable:
  - numbers of phonemes within preceding, current, and succeeding syllables
  - stress[4] and accent[5] of preceding, current, and succeeding syllables
  - positions of current syllable within current word and phrase
  - numbers of preceding and succeeding stressed syllables within current phrase
  - numbers of preceding and succeeding accented syllables within current phrase
  - number of syllables from previous stressed syllable
  - number of syllables to next stressed syllable
  - number of syllables from previous accented syllable
  - number of syllables to next accented syllable
  - vowel identity within current syllable
- Word:
  - estimate of the part of speech of preceding, current, and succeeding words
  - numbers of syllables within preceding, current, and succeeding words
  - position of current word within current phrase
  - numbers of preceding and succeeding content words within current phrase
  - number of words from previous content word
  - number of words to next content word
- Phrase:
  - numbers of syllables within preceding, current, and succeeding phrases
  - position of current phrase in major phrases
  - ToBI endtone of current phrase
- Utterance:
  - numbers of syllables, words, and phrases in utterance

Parameter tying: In practice, there are too many contextual factors in relation to the amount of speech data available. As the number of contextual factors we want to consider increases, their combinations also increase exponentially. Therefore, the context-dependent HMM parameters cannot be estimated accurately and robustly with a limited amount of training data. To overcome this problem, we always apply state tying techniques [56] to cluster similar states and to tie model parameters among several context-dependent HMMs so that we can estimate the model parameters more robustly. The state tying process is conducted in a hierarchical tree structure manner and the tree size is automatically determined based on an information criterion called minimum description length

[4] The lexical stress of the syllable as specified from the lexicon entry corresponding to the word related to this syllable.

[5] An intonational accent of the syllable predicted by a CART [55] (0 or 1).

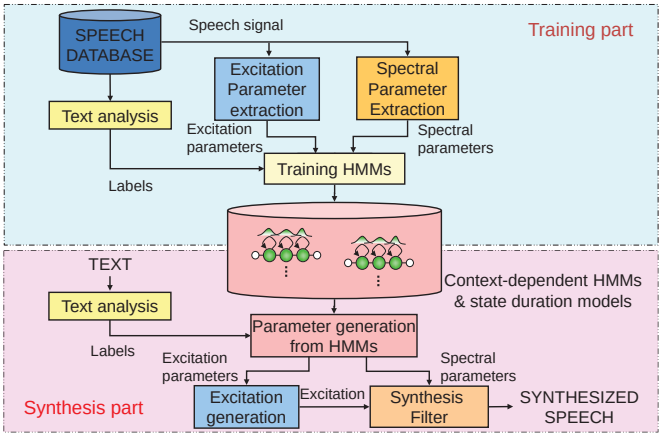

Fig. 6. Overview of the HMM-based speech synthesis system.

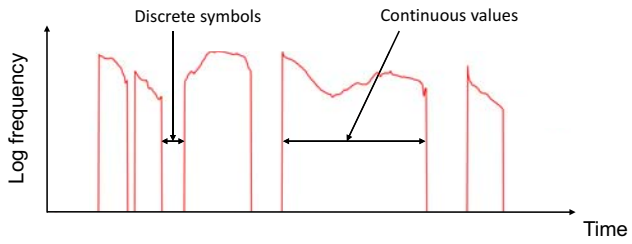

Fig. 7. Example of an observation sequence of an F_0 pattern.

(MDL) [57]. As the spectral, excitation, and duration parameters have different context dependency, they are clustered separately by using stream-dependent decision trees [11].

E. Synthesis part

The synthesis part of the system is shown in the lower part of Fig. 6. It first converts a given text to be synthesized into a sequence of context-dependent labels. According to the label sequence, a sentence-level HMM is constructed by concatenating context-dependent HMMs. The duration of each state is determined to maximize its probability based on its state duration probability distribution (Eq. (12)). Then a sequence of speech parameters including spectral and excitation parameters is determined so as to maximize its output probability using the speech parameter generation algorithm [46] (II-C). Finally, a speech waveform is re-synthesized directly from the generated spectral and excitation parameters by using a speech synthesis filter, such as the mel-log spectral approximation filter [42] for mel-cepstral coefficients and all-pole filter for linear prediction-based spectral parameter coefficients, as explained in II-A.

III. Flexibility of HMM-based speech synthesis

The main advantage of HMM-based speech synthesis over concatenative speech synthesis is its flexibility in changing voice characteristics, speaking styles, and emotions. Many techniques for controlling variation in speech have been proposed, and this section overviews major techniques to accomplish this, including adaptation, interpolation, eigenvoice, and multiple regression.

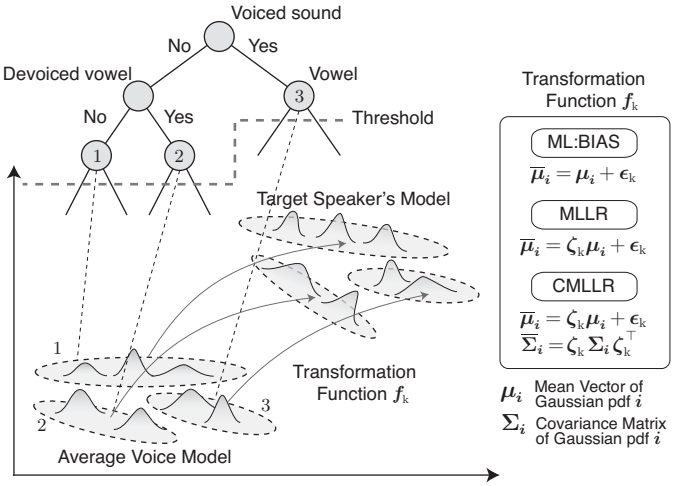

Fig. 8. Speaker adaptation techniques of an HMM-based speech synthesis system.

A. Speaker characteristics

Model adaptation (mimicking voices): Speaker adaptation is a technique for transforming existing speaker-independent acoustic models to match a target speaker using a very small amount of speech data [24]. This method starts with an "average voice model" and uses model adaptation techniques drawn from speech recognition such as maximum likelihood linear regression (MLLR) [58], [59], to adapt the speaker independent HMMs to a new speaker or to a new speaking style, as shown in Fig. 8.

The average voice model is a "canonical" speaker-independent HMM where inter-speaker acoustic variation is normalized using a technique called on speaker-adaptive training (SAT) [60], [61]. MLLR is one of the most important recent developments in speech recognition because this can effectively reduce acoustic mismatch between training data and test data. MLLR adaptation estimates a set of linear transforms to map Gaussian pdfs of the existing average voice model into a new adapted model so that the adapted model approximates given adaptation data better. Since the amount of adaptation data is limited, a regression class tree is normally used to cluster the Gaussian components based on acoustic similarity and to share the same MLLR transform [62]. In Fig. 8, there are three regression classes where the same transformation functions are shared.

Speaker adaptation is also a very exciting development in HMM-based speech synthesis. This adaptation allows text-to-speech synthesizers for a target voice to be built using much smaller amounts of training data than previously required. Prior to this, the development of a new voice required many hours of carefully annotated speech recordings from a single speaker. Speaker adaptive HMM-based synthesis requires as little as 5–7 minutes of recorded speech from a target speaker to generate a personalized synthetic voice [24]; hence, the average voice model can be easily transformed into a synthetic voice for any number of speakers [25]. The major adaptation techniques used for HMM-based speech synthesis are similar to those of ASR and include maximum a posteriori (MAP)

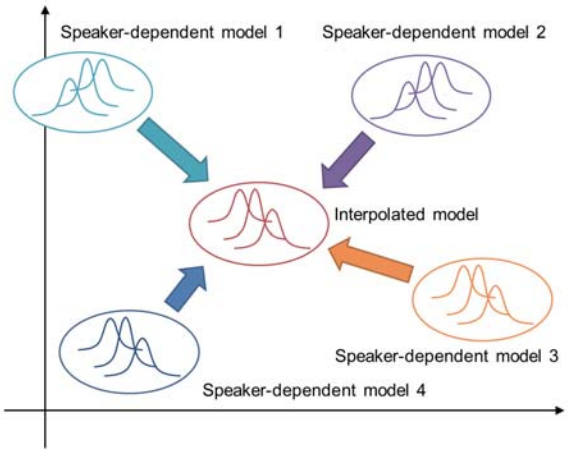

Fig. 9. Model interpolation techniques of an HMM-based speech synthesis system.

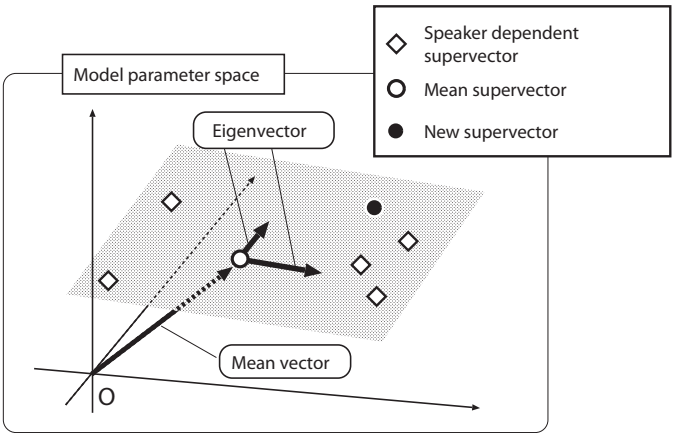

Fig. 10. Eigenvoice techniques of an HMM-based speech synthesis system. "Supervector" is a vector that consists of all HMM parameters, e.g., means of state Gaussians.

estimation [63], [64] and MLLR [26], [58].

MLLR performs linear transforms of mean vectors of the state output probability distributions of Eq (2)

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_k, \boldsymbol{\Sigma}_i) \tag{21}$$

where $\boldsymbol{\zeta}_k$ and $\boldsymbol{\epsilon}_k$ are a d-by-d matrix and a d-dimensional vector, respectively, and k denotes the k-th regression class. If we transform covariance matrices as well as the mean vectors of the state output probability distributions using the same matrices, this is called the constrained MLLR (CMLLR) [65] and the state output probability distribution is affine-transformed as follows:

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_k, \boldsymbol{\zeta}_k \boldsymbol{\Sigma}_i \boldsymbol{\zeta}_k^\top). \tag{22}$$

These transforms may be estimated using the standard maximum likelihood or MAP [24] criteria, and may be combined with other speaker adaptation techniques such as vocal tract length normalization [66].

Model interpolation (mixing voices): The model interpolation technique enables us to generate synthetic speech having intermediate voice characteristics among two or more than two representative pre-trained voice characteristics. The basic idea of interpolation was first proposed in the field of voice conversion, where pre-stored spectral patterns were interpolated among multiple speakers [67]. The same concept can also be applied to HMM-based speech synthesis, where HMM parameters are interpolated among some representative HMM sets [68] as shown in Fig. 9. The main difference between Iwahashi and Sagisaka's technique [67] and Yoshimura et al.'s one [68] was that as each speech unit was modeled by an HMM, mathematically well-defined statistical measures such as the Kullback-Leibler divergence could be used to interpolate the HMMs. Using the interpolation technique, we can synthesize speech with various voice characteristics [68], speaking styles [69], and dialects [70] that are not included in the training speech data.

Eigenvoices (producing voices): The use of the interpolation technique enables us to obtain various new voices by changing the interpolation ratio between representative HMM sets even if no data for the target voice are available. However, if we increase the number of representative HMM sets to enhance speech variations, it is not straightforward to determine the interpolation ratio to obtain the required voice. To address this problem, Shichiri et al. applied the "eigenvoice" technique based on principal component analysis (PCA) [71] to HMM-based speech synthesis [27]. The framework of probabilistic PCA can similarly be applied to HMM-based speech synthesis systems [28] to improve acoustic modeling.

The advantage of the eigenvoice approach is that it reduces the number of parameters to be controlled, which enables us to manually control the voice characteristics of synthesized speech by setting the weights (Fig. 10).

B. Expressive speech synthesis

Similar to the naturalness of synthesized speech, expression of emotion is one of the important issues which should be considered [72]. Various types of emotional/affective speech synthesis approaches have been proposed in the HMM-based speech synthesis framework.

For intuitively controlling the characteristics, Miyanaga et al. applied a multiple-regression approach to HMM-based speech synthesis to control voice characteristics intuitively [29], [30] in which mean vectors of state-output distributions were directly controlled with small-dimensional auxiliary features. The multiple-regression HMM was initially proposed to improve the accuracy of acoustic modeling for ASR by using auxiliary features that are correlated with the acoustic features [73]. Auxiliary features that have been used in ASR include fundamental frequency [73]; the auxiliary features that are often used in TTS, on the other hand, are more meta-level descriptions of speech such as specific voice characteristics, speaking styles, and emotions. This allows us to directly manipulate such expressivity, brightness, and emotions of specific words or phrases straightforwardly at the synthesis stage. In [29], [30], these auxiliary features for TTS are manually

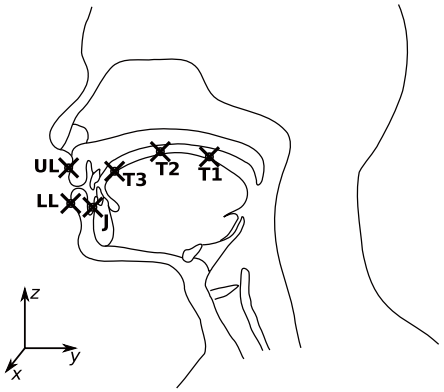

Fig. 11. Examples of articulatory features that can be integrated into HMM-based speech synthesis using regression approaches.

annotated through subjective listening tests prior to HMM training. A trial for estimating such voice characteristics, speaking styles, and emotions of speech data based on the same model has also been reported [31].

Another type of auxiliary feature that can be integrated into HMM-based speech synthesis is articulatory features, that is, the continuous movements of a group of speech articulators, for example the tongue, jaw, lips and velum shown in Fig. 11, recorded using human articulography techniques such as electromagnetic articulography [74]. An HMM-based speech synthesis system having such articulatory features as optional configurable inputs allows users to manipulate HMM-based synthetic speech via articulation [32], [33], which becomes a good link and contrast to conventional articulatory synthesis [1].

Furthermore, it is possible to combine the model adaptation approach mentioned earlier and the regression approach. By combining these techniques, we can synthesize speech with various voice characteristics, speaking styles, and emotions without having larger speech databases. For example, Tachibana et al. and Nose et al. proposed the combination of multiple-regression and adaptation techniques to achieve a multiple-regression technique with a small amount of speech data [34], [35].

For emphasis modeling, Badino et al. showed that HMM-based speech synthesis can produce recognizable variation when modeling emphasis of contrastive words [75]. On the other hand, Yu et al. proposed a two-pass decision tree and a factorized decision tree approach for word-level emphasis modeling instead of directly using emphasis context features [76].

C. Multilingual speech synthesis

Supporting multiple languages can easily be accomplished in HMM-based speech synthesis because the only language-dependent element is the set of contextual factors to be used. In other words, once we analyze and acquire the contextual factors required for the target languages, HMMs can be learned in a completely automatic manner, without the need for skilled human intervention because these technologies are primarily data-driven. This property is a key factor for efficiently developing speech synthesizers in new and/or multiple languages.

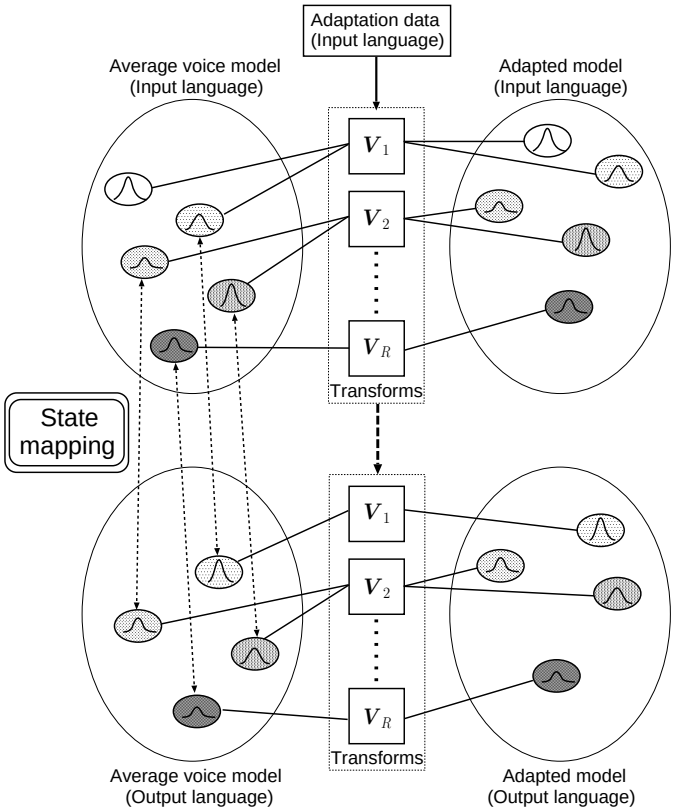

Fig. 12. Cross-lingual speaker adaptation techniques of HMM-based speech synthesis system using state mapping.

Identical model training methods/recipes can be used across languages and any improvements made to the training recipes will be automatically reflected in voices for all languages. To our knowledge, more than 40 different language systems have been or are being built by both various academic and commercial organizations.

As in the speech recognition field, multilingual/polyglot and cross-lingual acoustic modeling are also active research topics in the speech synthesis field, and we list a few interesting attempts in this subsection. Latorre et al. and Qian et al. proposed several techniques for building multilingual mixed-language speech synthesizers, where speech data in multiple languages are used simultaneously and HMM states are shared across these languages [77], [78]. Cross-lingual acoustic modeling for TTS in which a target speaker's voice in a new language is constructed from target speaker's speech data in a different language has also been proposed and developed. For example, Qian et al. and Wu et al. proposed cross-lingual speaker adaptation techniques using state mapping and tree structures learned from bilingual speakers and average voice models [78]–[80], respectively, shown in Fig. 12.

Zen et al. proposed a framework for estimating HMMs on data containing both multiple speakers and multiple languages, which attempts to factorize speaker-/language-specific characteristics in the data and then model them using separate transforms. Language-specific factors in the data are represented by transforms based on cluster mean interpolation with cluster-dependent decision trees, and acoustic variations caused by

speaker characteristics are handled by transforms based on constrained MLLR [81].

D. Singing voice

As we mentioned earlier in the multilingual section, HMM-based text-to-speech synthesis has only limited language dependency. Singing voice synthesis can be regarded as synthesis in a special language in which linguistic specifications and resulting contexts are derived from musical notes and lyrics [82], [83]. To construct a basic system, the speech database of the HMM-based text-to-speech synthesis system has to be replaced with a database of singing voices and the corresponding musical notes. The singing voice synthesis system can be constructed in almost the same manner as that of the HMM-based text-to-speech synthesis system. We just need to add several contextual factors specific for singing voices. For example, the contexts used for the singing voice synthesis system include the following:

- Phoneme:
  - current phoneme
  - preceding and succeeding two phonemes
  - position of current phoneme within current syllable
- Syllable:
  - numbers of phonemes within preceding, current, and succeeding syllables
  - positions of current syllable within current musical note and phrase
- Musical Note:
  - musical tone, key, beat, tempo, length, and dynamics of preceding, current, and succeeding musical notes
  - position of current musical note within current phrase
  - tied and slurred flags
  - distance between current musical note and preceding/succeeding accent and staccato
  - position of current musical note within current crescendo and decrescendo
- Phrase:
  - numbers of syllables within preceding, current, and succeeding phrases
- Song:
  - numbers of syllables, musical notes, and phrases in song

One of HMM-based singing speech synthesizers called "Sinsy" is available online [84]. In the online system, users can upload musical scores (MusicXML format) that specify measure, musical note, pitch, duration, lyrics, etc., and Sinsy automatically generates singing voices corresponding to the musical scores.

E. Small footprint

The footprints (required memory and disk storage size) of HMM-based speech synthesizers are usually significantly smaller than that of typical unit-selection synthesizers because only parameters of the HMMs are stored, instead of the speech waveforms. For example, the footprint of a standard system built using the publicly available toolkit HTS is normally less than 2 MBytes without the use of any compression techniques [14]. Thanks to this feature, statistical parametric speech synthesis systems are highly valued on embedded devices [85], and various commercial products have recently been released [36]–[39].

For some applications, such as server-client type speech services, the storage size of the TTS system is not usually an issue. For such cases, we can adaptively control the size of decision trees at synthesis time by storing larger decision trees and all corresponding HMM parameters [86].

For some applications, on the other hand, the footprint may also be further reduced without significant degradation in quality by eliminating redundant information. It was demonstrated that HMM-based speech synthesis systems whose footprints were about 100 KBytes could synthesize intelligible speech by using vector quantization, fixed-point numbers instead of floating-point numbers, and pruned decision trees. In addition, several techniques that suit embedded devices have been proposed, e.g., memory-efficient, low-delay speech parameter generation algorithms [87], [88] and tying model parameters [89].

IV. Open source software tools

Open source software tools have greatly facilitated research in the speech technology community as in many other fields. In particular, the HMM Toolkit (HTK) [16] and the Festival Speech Synthesis System [17] are the standard toolkits for speech recognition and synthesis research, respectively.

For HMM-based speech synthesis research, a toolkit called HTS [18] has been developed and is widely used. The number of downloads of HTS exceeds 10,000. This toolkit is released as a patch code to HTK under the 3-clause BSD license.[6] As its interface and functionality are very similar to those of HTK, researchers who are familiar with HTK can easily start using HTS. An application programming interface (API) for implementing run-time HMM-based speech synthesis called hts_engine API [90] has also been released. This toolkit is highly portable as it is written in C and uses only the standard C library. Based on this API, a Japanese TTS system called Open JTalk [91] and an English TTS system called Flite+hts_engine [90] have been released. The Festival Speech Synthesis System also uses this API to support HMM-based speech synthesis. A toolkit for performing reactive speech synthesis called pHTS is built upon hts_engine API [92].

The feature extraction part of HMM-based speech synthesis requires components for signal processing. The Speech Signal Processing Toolkit (SPTK) [19] provides most of the required signal processing functionality including linear predictive analysis, mel-cepstral analysis, and fundamental frequency extraction. The Edinburgh Speech Tools [93], the Snack Sound Tookit [94], and the ESPS Toolkit [95] are also often used for the feature-extraction part for HMM-based speech synthesis.

[6]Once the patch code is applied, users must obey the license of the HTK.

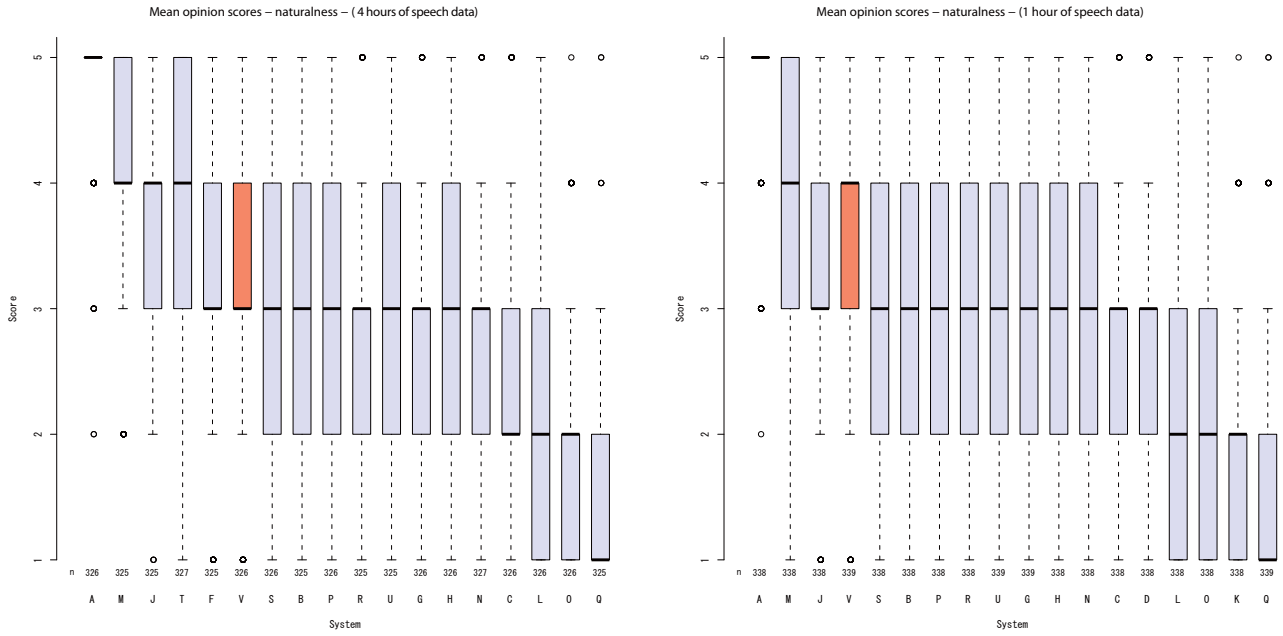

Fig. 13. Mean opinion scores on naturalness in the 2010 Blizzard Challenge EH1 task, where 4 hours of speech data were used (left figure) and EH2 task where 1 hour of speech data were used (right figure). The box-plots in orange represent HMM-based speech synthesis systems that can be constructed using the latest version of the HTS toolkit.

V. Performance and Evaluation

The Blizzard Challenge is an annual evaluation of corpus-based speech synthesis systems, in which participating teams build a synthetic voice from a common speech database then synthesize a set of test sentences. Listening tests are used to evaluate the systems in term of naturalness, similarity to original speakers, and intelligibility [96]–[100]. HMM-based speech synthesis systems have been evaluated since the 2005 challenge. We summarize the results of the challenges in the context of HMM-based approaches.

A. Intelligibility

It is often mentioned that text-to-speech has lower intelligibility than natural speech [101].

According to the Blizzard Challenge 2008 and the following challenges, however, synthetic speech generated from some HMM-based speech synthesis systems are occasionally found to *be as intelligible as natural human speech* in noiseless conditions [99].

In the latest 2011 Blizzard Challenge, even a benchmark HMM-based speech synthesis system that only uses the publicly released HTS toolkit was found to be as intelligible as natural human speech. To our knowledge, *this landmark achievement is a first for speech synthesis research* (see the 2008 and 2011 Blizzard Challenge for more details).

B. Naturalness and speaker similarity

According to the Blizzard Challenge evaluations, *no* text-to-speech synthesizer has yet been found to be as good as natural human speech in terms of both naturalness and speaker similarity.

The best quality of HMM-based speech synthesizers was found to be better than or comparable to the conventional unit selection systems on relatively smaller corpora consisting of one hour to a few hours of speech data. It is also found that combining unit-selection and HMM-based speech synthesis (which is mentioned in detail in the next section) resulted in better quality and speaker similarity on relatively larger corpora.

Fig. 13 shows some of the actual results of the 2010 Blizzard Challenge. In Blizzard Challenge 2010, two English speech databases consisting of four hours of speech uttered by a British male speaker and one hour of speech data uttered by a different British male speaker were released. To evaluate the naturalness of synthetic speech, a 5-point mean opinion score (MOS) was adopted, where 5 meant "completely natural" and 1 meant "completely unnatural." The evaluations were conducted over a six-week period via the Internet and in controlled sound booths. The figure shows the MOS results on naturalness corresponding to these provided corpora.

The system "V" marked in orange in the figure represents an HMM-based speech synthesis system constructed using the latest version of the HTS toolkit. On the four hours of speech corpus, "V" was not as good as the top hybrid systems such as "M." More importantly, there was no significant difference between the unit selection benchmark unit selection system "B" and the HMM-based system "V." On the one hour of speech corpus, "V" was evaluated as the second best.

VI. Relation to Unit Selection Approaches

A. Comparison with unit selection

In HMM-based speech synthesis systems, the distributions for individual speech components, such as the spectrum, exci-

tation (F_0), and duration, are clustered separately to effectively capture contextual dependencies specific to individual speech components. Some clustering-based systems for unit selection using HMM-based state clustering [102] have a similar structure to that of the HMM-based speech synthesis system. They often use regression trees (or CART [55]) for predicting prosodic parameters such as F_0 and duration. They are almost equivalent to the decision trees for F_0 and duration in the HMM-based speech synthesis system. On the other hand, if the waveform concatenation is used in the unit-selection systems, the leaves of one tree must have speech waveforms rather than spectral parameters; other trees are used to calculate target costs, to prune waveform candidates, or to give features to build the tree for speech waveforms.

The essential difference between the HMM-based speech synthesis system and unit-selection systems is that each cluster in the HMM-based speech synthesis system is represented by the probability distribution of the cluster instead of the multi-templates of speech units in the unit-selection systems. The HMM-based speech synthesis system generates speech parameter trajectories from the continuous probability distributions with the likelihoods of static and dynamic features, while the unit-selection systems select a sequence of speech units from multi-templates with target and concatenation costs. It should be noted that the likelihoods of static and dynamic features work as the target and concatenation costs, respectively. The use of continuous distributions enables us to achieve a continuous representation of speech parameter trajectories beyond a discrete representation by the multi-templates. This also causes a difference in the search space for the optimal parameter trajectories; whereas it is discrete for unit selection, it is continuous for the parameter generation from the continuous distributions. Thus, the parameter generation in the HMM-based speech synthesis system can be viewed as an analogue version of unit selection.

B. Hybrid approaches

There are hybrid approaches between unit-selection synthesis and HMM-based speech synthesis as a natural consequence of the viewpoints mentioned above. Some of these approaches use spectrum parameters, F_0 values, and durations (or some of them) generated from HMMs as "targets" for unit-selection synthesis [103]–[106]. Similarly, HMM likelihoods are used as "costs" for unit-selection synthesis [107]–[111]. Furthermore, some approaches use instances of frame samples in the state to approximate each state-output distribution, e.g., the HMM-based unit-selection system with frame-sized units [112] uses a frame-wise dynamic programming (DP) search calculating the dynamic feature as the difference between neighboring static features, which results from the ML-based parameter generation, the discrete HMM-based speech synthesis system [113] models each state-output distribution with discrete distributions using vector quantization based on a similar idea, and a frame-wise representation of state-output distribution was also investigated in an attempt at unifying unit-selection and HMM-based speech synthesis [114].

These hybrid approaches have several advantages. An over-smoothing problem, which is described in Section VII-B, is avoided by using natural acoustic instances as candidates of samples generated from the probability distributions. The quality degradation caused by vocoding is also avoided by using the waveform instances. Moreover, the HMM likelihoods help us design a complicated cost function sensitively capturing the context dependencies in each speech component. On the other hand, the hybrid approaches lose many of the advantages of the HMM-based speech synthesis system, such as a flexible control of voice quality and small footprint, as mentioned above. In the future, we may convert them into an optimal form of corpus-based speech synthesis by fusing HMM and unit-selection synthesis.

VII. The Recent Development

A. Excitation models

Speech samples synthesized using the basic HMM-based speech synthesis system sound somewhat buzzy since it uses a vocoder with a simple excitation model based on a periodic pulse-train and white-noise [11]. To mitigate this problem, high-quality vocoders such as mixed excitation linear prediction [115], [116], multi-band excitation [117], pitch synchronous residual codebook [118], the harmonic plus noise model (HNM) [119], [120], the flexible pitch-asynchronous harmonic/stochastic model [121], STRAIGHT [14], the glottal-flow derivative model [122], [123], and the glottal waveform [124], [125], have been implemented for the HMM-based speech synthesis system.

Most of these methods are based on the implementation of an excitation model through the use of additional parameters modeled by HMMs. However, they do not directly minimize the distortion between artificial excitation and speech residuals. Maia et al. have recently proposed a trainable technique of excitation modeling for HMM-based speech synthesis [126]. In this technique, mixed excitation is produced by inputting periodic pulse trains and white noise into two state-dependent filters, a voiced filter to model phase components depending on the glottal waveform and frequency-dependent periodic components and an unvoiced filter to model frequency-dependent aperiodic components. The filters are derived to maximize the likelihood of residual sequences over corresponding states through an iterative process. As a result, this technique directly minimizes the weighted distortion (i.e., Itakura-Saito distance [127]) between the generated excitation and speech residuals, which is equivalent to direct modeling of speech waveforms by using HMMs.

B. Avoiding over-smoothing

In HMM-based speech synthesis systems, the speech parameter generation algorithm is used to generate spectral and excitation parameters from HMMs to maximize their output probability density values under constraints between static and dynamic features. The statistical averaging in the modeling process improves robustness against data sparseness, and the use of dynamic-feature constraints in the synthesis process enables the generation of smooth trajectories. However, synthesized speech sounds are evidently muffled compared with

natural speech because the generated speech-parameter trajectories are often over-smoothed, i.e., detailed characteristics of speech parameters are removed in the modeling part and cannot be recovered in the synthesis part. Although using the advanced acoustic models may reduce this over-smoothing effect, this may still exist because the synthesis algorithm does not explicitly include a recovery mechanism.

The simplest way of compensating for over-smoothing is to emphasize the spectral structure by using a post-filter, which was originally developed for speech coding. The use of post-filtering techniques can reduce "buzziness" and muffled sounds [12], [115]. However, too much post-filtering often introduces artificial sounds and degrades the similarity of synthesized speech to the natural speech uttered by the original speaker.

Another way of compensating for over-smoothing is integrating multiple-level statistical models to generate speech-parameter trajectories. One of the most successful methods in this category is the speech parameter generation algorithm considering global variance (GV) [20]. A GV is defined as an intra-utterance variance of a speech-parameter trajectory, which is a second-order moment calculated over an utterance. We calculate GVs for all training utterances and approximately model their probability density by using a single multivariate Gaussian distribution. The speech parameter generation algorithm considering the GV maximizes not only the HMM likelihood function but also the objective function of the GV, which can be viewed as a penalty to prevent over-smoothing.

C. Trajectory HMMs

The speech parameter generation algorithm allows us to generate smoothly varying speech parameter trajectories from HMMs, while satisfying the statistics of both static and dynamic features. However, it also introduces an inconsistency between training and synthesis stages; dynamic feature constraints are ignored at the training stage but utilized explicitly at the synthesis stage, i.e.,

$$\lambda_{\max} = \arg\max_{\lambda} p(\boldsymbol{O} \mid \lambda, \mathcal{W}) \tag{23}$$

$$\boldsymbol{o}_{\max} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \lambda_{\max}, w) \mid_{\boldsymbol{o}=\boldsymbol{W}\boldsymbol{c}} \tag{24}$$

To avoid this inconsistency, Zen et al. explicitly introduced the dynamic feature constraints into the training stage and reformulated the HMM with dynamic features as a trajectory model [21]. This model, called a trajectory HMM, could overcome the assumption of conditional independence and constant statistics within an HMM state without the need for any additional parameters. The minimum generation error (MGE) training, which also uses the relationship between static and dynamic features at the training stage, can be viewed as estimating trajectory HMMs by a defined loss function, such as minimum mean squared error [22] or log spectral distortion [23]. Recent research showed that eliminating this inconsistency resulted in better predictive distribution of speech parameter trajectories [128]. The relationship between the trajectory HMM and Markov random field was also discussed [129].

VIII. Future directions

Thanks to the flexibility and adaptability of HMM-based speech synthesis, several new applications are emerging such as a) personalized speech-to-speech translation systems where a user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice [78]–[80] and b) voice banking and reconstruction; personalized speech synthesizers for individuals with vocal disabilities [130]. Further new TTS applications will be seen in the very near future.

There are many future directions that should be examined. For the quality of the synthesized speech, Kawahara et al. proposed pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region and an excitation method using instantaneous frequency calculation [45]. For F_0 modeling, Kameoka et al. proposed a statistical model of speech fundamental frequency contours [131] based on the formulation of the discrete-time stochastic process version of the Fujisaki model [132], which is known as a well-founded mathematical model representing the control mechanism of vocal fold vibration. On the other hand, Ni et al. focus on speech synthesis using "Big data" [133], including e-book, Internet radio, podcast, etc. In this approach, HMMs are trained with automatic speech transcription including errors. Black et al. proposed an approach to building TTS systems for low-resource languages [134]. This method allows building models without depending on the availability of part of speech taggers, or corpora with hand annotated breaks. Bellegarda proposed a framework for the analysis of emotion in texts for speech synthesis [135]. This approach translates plain texts into appropriate speaking styles automatically. Speech synthesis by physical simulation is also one of important directions. Kitamura et al. proposed an MRI-based articulatory speech synthesis system [136].

In the following, research topics that authors are interested in and working on are described.

A. Intelligibility of synthetic speech in noise

In a quiet listening environment, the intelligibility of state-of-the-art HMM-generated synthetic speech can be as good as that of natural speech [137]. However, in noisy environments, unmodified synthetic speech tends to reduce in intelligibility to a much greater extent than unmodified natural speech [138]. By modifying the synthetic speech via the statistical models, it is possible to control the characteristics of the generated speech and so generate synthetic speech that is more intelligible in noise than the natural speech used for training [138]–[140]. One way to do this is to use adaptation techniques based on natural speech produced in noise: so-called Lombard speech [141]–[143] or to enhance articulation degree of synthetic speech based on articulatory data [32].

B. Improvement of acoustic models

Additive models: One of the most important problems in HMM based speech synthesis is modeling the correlation between contextual factors and acoustic features, i.e., spectrum parameters, F_0 values, and durations. Typical context

dependent models, e.g., triphone HMMs, have direct dependencies of contexts, i.e., if a phonetic context is given, the Gaussian distribution is specified immediately. However, it is known that prosodic information, such as F_0, has an additive structure with multiple contextual factors [132]. Therefore, there probably exist more efficient model structures for representing the generation processes of observed data. The linear regression model is one approach for representing additive structures, and it is assumed that all the contextual factors independently affect acoustic features. However, it is difficult to find independent additive factors to obtain a good prediction of acoustic features. To overcome this problem, Nankaku et al. proposed an additive structure model that composes multiple decision trees [144]. This method can represent the intermediate structure of decision tree-based context clustering and linear regression models. The context clustering algorithm for the additive structure model that can automatically extract additive components by constructing multiple decision trees simultaneously has been proposed. Moreover, this method can automatically determine the appropriate number of additive components.

Bayesian approach: Bayesian learning is used to estimate the posterior distributions of model parameters from prior distributions and training data, whereas ML and MAP learning are used to estimate the parameter values (point estimates). This property enables us to incorporate prior knowledge into the estimation process and improve model generalization due to the marginalization effect of model parameters. It offers selection of a model's complexity in the sense of maximizing its posterior probability. Recently, Watanabe et al. applied the variational Bayesian-learning technique [145] to speech recognition [146], and Hashimoto et al. applied this idea to HMM-based speech synthesis [147]. Bayesian HMM-based speech synthesis determines $\boldsymbol{o}$ as

$$\boldsymbol{o}_{\max} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid w, \boldsymbol{O}, \mathcal{W}) \tag{25}$$

$$= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}, \boldsymbol{O} \mid w, \mathcal{W}) \tag{26}$$

$$= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o}, \boldsymbol{O}, \lambda \mid w, \mathcal{W}) d\lambda \tag{27}$$

$$= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o}, \boldsymbol{O} \mid w, \mathcal{W}, \lambda) p(\lambda) d\lambda \tag{28}$$

$$= \arg\max_{\boldsymbol{o}} \int p(\boldsymbol{o} \mid w, \lambda) p(\boldsymbol{O} \mid \mathcal{W}, \lambda) p(\lambda) d\lambda \tag{29}$$

Eq. (25) is the fundamental problem that needs to be solved in corpus-based speech synthesis, i.e., finding the most likely speech parameters $\boldsymbol{o}$ for a given word sequence w using the training data $\boldsymbol{O}$, and the corresponding word sequence $\mathcal{W}$. The equations above also indicate that $\boldsymbol{o}$ is generated from the predictive distribution, which is analytically derived from the marginalization of λ based on the posterior distribution estimated from $\boldsymbol{O}$. We can solve this maximization problem by using Bayesian speech parameter generation algorithms [147], which are similar to ML-based speech parameter generation algorithms [46]. One research topic in the Bayesian approach is how to set the hyperparameters of the prior distribution, because the quality of synthesized speech is sensitive to these. These hyperparameters have been set empirically in conventional approaches. Hashimoto et al. recently proposed a cross-validation (CV)-based technique of setting hyper-parameters [147] for Bayesian speech synthesis. It demonstrated that the CV-based Bayesian speech synthesizer achieved better quality synthesized speech than an ML-based one.

Unification with speech feature extraction: In typical parametric speech synthesis, feature extraction from speech signals and statistical modeling are separated and independently optimized. As a joint optimization method, Toda et al. proposed a statistical method for estimating the vocal tract transfer function from a speech signal based on the maximum a posteriori criterion [148]. This method effectively models harmonic components observed over an utterance by using a factor analyzed trajectory HMM, which is a unified model for spectral extraction and HMM-based spectral sequence modeling. By dealing with a mel-cepstrum sequence as a latent variable, the error in spectral extraction is effectively considered in the HMM training. On the other hand, Maia et al. proposed a method that combines the extraction of spectral parameters and excitation signal modeling in a fashion similar to the factor analyzed trajectory HMM [149]. The resulting joint estimation of acoustic and excitation model parameters can be interpreted as a waveform-level closed-loop training, where the distance between natural and synthesized speech is minimized. Similarly, Wu and Tokuda proposed a training algorithm that directly minimizes the generation error of harmonic components in the log spectral domain at LSP frequencies modeled using HMMs [23]. These methods can be regarded as a unification of feature extraction and statistical modeling. It is expected that a joint optimization of these two components can improve the performance of the whole system by using useful information of both components simultaneously.

Unification with text analysis: Standard TTS systems consist of two major modules: text analysis and speech synthesis. Conventionally, these two modules are constructed independently. The text analysis module including phrasing and prosodic models is trained using text corpora. On the other hand, the speech synthesis module including acoustic models (i.e., HMMs) is trained using a labeled speech database. If these two modules were combined and trained simultaneously as a unified model, we expect that the overall performance of a TTS system would be improved. Oura et al. defined a new integrated model for linguistic and acoustic modeling and proposed a joint optimization method of these two model parameter sets [150]. This method allows us to directly formulate the TTS problem of synthesizing a speech waveform from a word sequence. Another advantage of this method is to minimize the effort in hand-labeling of phrasing and prosodic events required in both linguistic and acoustic model training because these labels are regarded as latent variables in the model.

IX. Conclusions

This paper gave a general overview of HMM-based speech synthesis and its recent advances. HMM-based speech synthesis has started to be used in daily life, e.g., cellphones, smart phones, in-car navigation systems, and call centers. Although the quality of synthesized speech generated by HMM-based speech synthesis has been drastically improved recently, its naturalness is still far from that of actual human speech. In conversational speech, naturalness of prosody is still insufficient to properly convey nonverbal information, e.g., emotional expressions and emphasis. To fill the gap between natural and synthesized speech, the statistical approaches described in VIII will be more important in the future.

References

[1] P. Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[2] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, 1980.

[3] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, 1990.

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

[5] A. Breen and P. Jackson, "A phonologically motivated method of selecting nonuniform units," in *Proc. ICSLP*, 1998, pp. 2735–2738.

[6] R. E. Donovan and E. M. Eide, "The IBM trainable speech synthesis system," in *Proc. ICSLP*, 1998, pp. 1703–1706.

[7] B. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proc. Joint ASA, EAA and DAEA Meeting*, 1999, pp. 15–19.

[8] G. Coorman, J. Fackrell, P. Rutten, and B. Coile, "Segment selection in the L & H realspeak laboratory TTS system," in *Proc. ICSLP*, 2000, pp. 395–398.

[9] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <AHEM/> expressive speech synthesis," in *Proc. ISCA SSW5*, 2004, pp. 79–84.

[10] A. W. Black, "Unit selection and emotional speech," in *Proc. Eurospeech*, 2003, pp. 1649–1652.

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

[12] 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. the Blizzard Challenge Workshop*, 2006.

[13] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. Interspeech*, 2006, pp. 1762–1765.

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

[15] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Englewood Cliffs, N. J., 1993.

[16] "HTK," http://htk.eng.cam.ac.uk/.

[17] "Festival," http://www.festvox.org/festival/.

[18] "HTS," http://hts.sp.nitech.ac.jp/.

[19] "SPTK," http://sp-tk.sourceforge.net/.

[20] 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, 2007.

[21] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 153–173, 2006.

[22] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, 2006, pp. 89–92.

[23] Y.-J. Wu and K. Tokuda, "Minimum generation error training by using original spectrum as reference for log spectral distortion measure," in *Proc. ICASSP*, 2009, pp. 4013–4016.

[24] 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 & Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[25] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process.*, vol. 18, pp. 984–1004, Jul. 2010.

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

[27] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[28] K. Kazumi, Y. Nankaku, and K. Tokuda, "Factor analyzed voice models for HMM-based speech synthesis," in *ICASSP*, 2010, pp. 4234–4237.

[29] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. Interspeech*, 2004, pp. 1437–1439.

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

[31] T. Nose, Y. Kato, and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," in *Proc. ICASSP*, 2007, pp. 833–8346.

[32] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[33] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 1, pp. 207–219, 1 2013.

[34] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, "Speaker and style adaptation using average voice model for style control in HMM based speech synthesis," in *Proc. ICASSP*, 2008, pp. 4633–4636.

[35] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," in *IEICE Trans. Inf. Syst.*, 2009, vol. E92-D (3), pp. 489–497.

[36] SVOX AG, "SVOX Pico, a revolutionary new hidden markov model-based text-to-speech product for mobile phones," in *Press release*, 2007, http://www.svox.com/upload/pdf/PR_SVOX_Pico.pdf.

[37] Q. Bai, "The development of chinese TTS technology," in *Presentation given in SpeechTEK*, 2007.

[38] KDDI R&D Laboratories, "Development of downloadable speech synthesis software for mobile phones," in *Press release*, 2008, http://www.kddilabs.jp/press/detail_100.html.

[39] SVOX AG, "SVOX releases Pico: highest-quality sub-1MB TTS," in *Press release*, 2008, http://www.svox.com/upload/pdf/PR_SVOX_Pico_Release_Nov_08.pdf.

[40] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, 2007, pp. 294–299.

[41] G. Fant, "Acoustic theory of speech production," in *The Hague: Mouton*, 1970.

[42] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[43] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am*, vol. 57, pp. S35–S35, 1975.

[44] K. Tokuda, T. Masuko, T. Kobayashi, and S. Imai, "Mel-generalized cepstral analysis —a unified approach to speech spectral estimation," in *ICSLP-94*, 1994, pp. 1043–1046.

[45] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

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

[47] G. Freij and F. Fallside, "Lexical stress recognition using hidden Markov models," in *Proc. ICASSP*, 1988, pp. 135–138.

[48] U. Jensen, R. Moore, P. Dalsgaard, and B. Lindberg, "Modeling intonation contours at the phrase level using continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 8, no. 3, pp. 247–260, 1994.

[49] K. Ross and M. Ostendorf, "A dynamical system model for generating F0 for synthesis," in *Proc. ESCA/IEEE Workshop on Speech Synthesis*, 1994, pp. 131–134.

[50] 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, 2002.

[51] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The Hidden Markov Model Toolkit (HTK) version 3.4*, 2006, http://htk.eng.cam.ac.uk/.

[52] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, 1998, pp. 29–32.

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

[54] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002.

[55] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," *CRC Press*, 1984.

[56] J. J. Odell, *The use of context in large vocabulary speech recognition*, Ph.D. thesis, Cambridge Univ., Cambridge, U.K., 1995.

[57] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 21, no. 2, pp. 79–86, 2000.

[58] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[59] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition*, 2001, p. 119.

[60] 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.

[61] 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.

[62] M. Gales and S. Young, "The application of hidden markov models in speech recognition," 2008.

[63] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[64] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP*, 1997, pp. 1611–1614.

[65] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[66] L. Saheer, J. Yamagishi, P. N. Garner, and J. Dines, "Combining vocal tract length normalization with hierarchial linear transformations," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 4493–4496.

[67] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Commun.*, vol. 16, no. 2, pp. 139–151, 1995.

[68] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 2523–2526.

[69] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2484–2491, 2005.

[70] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of austrian german and viennese dialect in HMM-based speech synthesis," *Speech Commun.*, vol. 52, no. 2, pp. 164–179, 2010.

[71] R. Kuhn, J. C. Janqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, 2000.

[72] M. Schröder, "Emotional speech synthesis: A review," in *Proc. Eurospeech*, 2001, pp. 561–564.

[73] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *ICASSP*, 2001, pp. 513–516.

[74] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31, pp. 26–35, 1987.

[75] L. Badino, J. S. Andersson, J. Yamagishi, and R. A. J. Clark, "Identification of contrast and its emphatic realization in HMM-based speech synthesis.," in *Interspeech*, 2009, pp. 520–523.

[76] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. ICASSP*, 2010, vol. 1, pp. 4238–4241.

[77] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proc. ICASSP*, 2005, vol. 1, pp. 1–4.

[78] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin – English) TTS," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1231–1239, Aug. 2009.

[79] Y.-J. Wu, S. King, and K. Tokuda, "Cross-language speaker adaptation for HMM-based speech synthesis," in *Proc. ISCSLP*, 2008, pp. 9–12.

[80] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping," *Speech Commun.*, vol. 54, no. 6, pp. 703–714, 2012.

[81] H. Zen, N. Braunschweiler, S. Buchholz, M. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, pp. 1713–1724, 2012.

[82] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "HMM-based singing voice synthesis system," in *Proc. ICSLP*, 2006, pp. 2274–2277.

[83] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system — Sinsy," in *Proc. ISCA SSW7*, 2010, pp. 211–216.

[84] "Sinsy," http://www.sinsy.jp/.

[85] S.-J. Kim, J.-J. Kim, and M.-S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consumer Electronics*, vol. 52, no. 4, pp. 1384–1390, 2006.

[86] S. Kataoka, N. Mizutani, K. Tokuda, and T. Kitamura, "Decision-tree backing-off in HMM-based speech synthesis," in *Interspeech*, 2004, vol. 2, pp. 1205–1208.

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

[88] K. Koishida, K. Tokuda, T. Masuko, and T. Kobayashi, "Vector quantization of speech spectral parameters using statistics of static and dynamic features," *IEICE Trans. Inf. Syst.*, vol. E84-D, no. 10, pp. 1427–1434, 2001.

[89] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "A covariance-tying technique for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E93-D, no. 3, pp. 595–601, 2010.

[90] "hts_engine," http://hts-engine.sourceforge.net/.

[91] "Open JTalk," http://open-jtalk.sourceforge.net/.

[92] "pHTS," http://magephts.sourceforge.net/.

[93] "Speech Tools," http://www.cstr.ed.ac.uk/projects/speech_tools/.

[94] "Snack," http://www.speech.kth.se/snack/.

[95] "ESPS," http://www.speech.kth.se/software/#esps.

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

[97] C. L. Bennett and A. W. Black, "The Blizzard Challenge 2006," in *Proc. Blizzard Challenge Workshop*, Sep. 2006.

[98] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard Challenge Workshop (in Proc. SSW6)*, Aug. 2007.

[99] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, Sep. 2008.

[100] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, Edinburgh, UK, Sep. 2009.

[101] S. J. Winters and D. B. Pisoni, "Speech synthesis, perception and comprehension of," in *Encyclopedia of Language & Linguistics (Second Edition)*, Keith Brown, Ed., pp. 31–49. Elsevier, Oxford, second edition edition, 2006.

[102] R. Donovan and P. Woodland, "Improvements in an HMM-based speech synthesiser," in *Proc. Eurospeech*, 1995, pp. 573–576.

[103] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Proc. ISCA SSW5*, 2004, pp. 179–184.

[104] S. Rouibia and O. Rosec, "Unit selection for speech synthesis based on a new acoustic target cost," in *Proc. Interspeech*, 2005, pp. 2565–2568.

[105] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *Proc. ISCA SSW5*, 2004, pp. 37–42.

[106] J.-H. Yang, Z.-W. Zhao, Y. Jiang, G.-P. Hu, and X.-R. Wu, "Multi-tier non-uniform unit selection for corpus-based speech synthesis," in *Proc. the Blizzard Challenge Workshop*, 2006.

[107] X.-D. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, and J.-S. Liu, "Whistler: a trainable text-to-speech system," in *Proc. ICSLP*, 1996, pp. 2387–2390.

[108] H.-W. Hon, A. Acero, X.-D. Huang, J.-S. Liu, and M. Plumpe, "Automatic generation of synthesis units for trainable text-to-speech systems," in *Proc. ICASSP*, 1998, pp. 293–296.

[109] T. Okubo, R. Mochizuki, and T. Kobayashi, "Hybrid voice conversion of unit selection and generation using prosody dependent HMM," in *IEICE Trans. Inf. Syst.*, 2006, vol. E89-D (11), pp. 2775–2782.

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

[111] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jian, Z.-W. Zhao, J.-H. Yang, J. Chen, and G.-P. Hu, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Proc. Blizzard Challenge Workshop*, 2007.

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

[113] J. Yu, M. Zhang, J. Tao, and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," in *Proc. ICASSP*, 2007, pp. 709–712.

[114] P. Taylor, "Unifying unit selection and hidden markov model speech synthesis," in *Proc. Interspeech*, 2006, pp. 1758–1761.

[115] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2263–2266.

[116] X. Gonzalvo, C. Socoró, I. Iriondo, C. Monzo, and E. Martínez, "Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish," in *Proc. ISCA SSW6*, 2007, pp. 362–367.

[117] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," in *Proc. Interspeech*, 2006, pp. 1332–1335.

[118] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitchsynchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. ICASSP*, 2009, pp. 3793–3796.

[119] C. Hemptinne, *Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system*, Master thesis, IDIAP Research Institute, 2006.

[120] S.-J. Kim and M.-S. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 378–381, 2007.

[121] E. Banos, D. Erro, A. Bonafonte, and A. Moreno, "Flexible harmonic/stochastic modeling for HMM-based speech synthesis," *V Jornadasen Tecnologias del Habla*, pp. 145–148, 2008.

[122] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *Proc. ISCA SSW6*, 2007, pp. 113–118.

[123] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.

[124] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, 2008, pp. 1881–1884.

[125] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[126] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. ISCA SSW6*, 2007, pp. 131–136.

[127] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 23 (1), pp. 67–72, 1975.

[128] M. Shannon, H. Zen, and W. Byrne, "The effect of using normalized models in statistical speech synthesis," in *Proc. Interspeech*, 2011, pp. 121–124.

[129] H. Zen, M. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Signal Process.*, vol. 20, no. 3, pp. 794–365, 2012.

[130] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," in *Acoustical Science & Technology*, 2012, vol. 33, pp. 1–5.

[131] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F0 contours," in *Proc. SAPA*, 2010, pp. 43–48.

[132] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn. (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[133] J. Ni and H. Kawai, "On the effects of transcript errors across dataset sizes on hmm-based voices," in *Proc. the Autumn Meeting of ASJ*, 2011, pp. 339–342.

[134] A. Parlikar and A. W. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *Proc. ICASSP*, 2011, vol. 1, pp. 4013–4016.

[135] J. R. Bellegarda, "A data-driven affective analysis framework toward naturally expressive speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1113–1122, 2011.

[136] T. Kitamura, H. Takemoto, P. Mokhtari, and T. Hirai, "MRI-based time-domain speech synthesis system," in *Proc. ASA/ASJ joint meeting*, 2006.

[137] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, Sep. 2008, vol. 5.

[138] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kyoto, Japan, Sep. 2010.

[139] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kyoto, Japan, Sep. 2010.

[140] D. Bonardo and E. Zovato, "Speech synthesis enhancement in noisy environments," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2853–2856.

[141] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-based lombard speech synthesis," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 2781–2784.

[142] T. Dutoit B. Picart, T. Drugman, "Continuous control of the degree of articulation in hmm based speech synthesis," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1797–1800.

[143] R. K. Moore and M. Nicolao, "Reactive speech synthesis: Actively managing phonetic contrast along an h&h continuum," in *7th International Congress of Phonetics Sciences (ICPhS).*, 2011, pp. 1422–1425.

[144] Y. Nankaku, K. Nakakura, H. Zen, and K. Tokuda, "Acoustic modeling with contextual additive structure for HMM-based speech recognition," in *Proc. ICASSP*, 2008, pp. 4621–4624.

[145] M. Beal, "Variational algorithms for approximate bayesian inference," *Ph.D. Thesis, University of London*, 2003.

[146] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational bayesian estimation and clustering for speech recognition," in *IEEE Trans. Speech Audio Process*, 2004, vol. 12 (4), pp. 365–381.

[147] K. Hashimoto, H. Zen, Y. Nankaku, and K. Tokuda, "A bayesian approach to HMM-based speech synthesis," in *Proc. ICASSP*, 2009, pp. 4029–4032.

[148] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," in *Proc. ICASSP*, 2008, pp. 3925–3928.

[149] R. Maia, H. Zen, and M. J. F. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," in *Proc. ISCA SSW7*, 2010, pp. 88–93.

[150] K. Oura, Y. Nankaku, T. Toda, K. Tokuda, R. Maia, S. Sakai, and S. Nakamura, "Simultaneous acoustic, prosodic, and phrasing model training for TTS conversion systems," in *Proc. ISCSLP*, 2008, pp. 1–4.

Keiichi Tokuda received his B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, in1984 and his M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1986 and 1989. From 1989 to 1996, he was a Research Associate in the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor in the Department of Computer Science, Nagoya Institute of Technology, where he is currently a Professor. He is also an Invited Researcher at the National Institute of Information and Communications Technology (NICT), formally known as the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan from 2000, and was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. His research interests include speech synthesis and speech recognition. He published over 70 journal papers and over 160 conference papers in the research area, and received 5 paper awards. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is currently a member of ISCA Advisory Council and an associate editor of IEEE Transactions on Audio, Speech and Language Processing.

Yoshihiko Nankaku received his B.E. degree in Computer Science, and his M.E. and Ph.D. degrees in the Department of Electrical and Electronic Engineering from Nagoya Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004. After a year as a postdoctoral fellow at the Nagoya Institute of Technology, he became an Assistant Professor at the same Institute. He was a visiting researcher at the Department of Engineering, University of Cambridge, Cambridge, UK, from May to October 2011. His research interests include statistical machine learning, speech recognition, speech synthesis, image recognition, and multimodal interface. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visting Researcher at NICT, Kyoto, Japan since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan and was a Visiting Researcher at ATR until March 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech processing such as speech synthesis and speech analysis. He published over 30 journal papers and 100 conference papers in this research area, and received 8 paper awards including the 2009 Young Author Best Paper Award from the IEEE SPS. He was a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009.

Heiga Zen received his A.E. degree from the Suzuka National College of Technology, Suzuka, Japan, in 1999, and his B.E., M.E., and Ph.D. degrees from the Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006. From 2004 to 2005, 2006 to 2008, and 2008 to 2011, he worked as an intern/co-op researcher at the IBM T. J. Watson Research Center, Yorktown Heights, NY, U.S.A., Research Associate at Nagoya Institute of Technology, and Research Engineer at the Toshiba Research Europe Cambridge Research Laboratory, Cambridge, U.K.. Presently, he is a Research Scientist at Google, London, U.K. His research interests include speech recognition and synthesis. Dr. Zen was awarded the 2006 ASJ Awaya Award, 2008 ASJ Itakura Award, 2008 TAF TELECOM System Technology Award, 2008 IEICE Information and Systems Society Best Paper Award, and 2009 IPSJ Yamashita SIG Research Award. He is a member of ASJ and IPSJ, and has been a member of the Speech and Language Processing Technical Committee since 2012.

Junichi Yamagishi is a lecturer and holds an EPSRC Career Acceleration Fellowship in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. He is also an associate professor of National Institute of Informatics (NII) in Japan. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has been in CSTR and has authored and co-authored about 100 refereed papers in international journals and conferences. His work has led directly to three large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. A recent coauthored paper was awarded the 2010 IEEE Signal Processing Society Best Student Paper Award and cited as a "landmark achievement of speech synthesis." He was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustic Society of Japan for his achievements in adaptive speech synthesis. In 2012 he was an area chair for the Interspeech conference and elected to membership of the IEEE Signal Processing Society Speech & Language Technical Committee. He is an external member of the Euan MacDonald Centre for Motor Neurone Disease Research. He has been Principal Investigator at Edinburgh on EPSRC and JST projects totalling over £1.8m.

Keiichiro Oura received his Ph.D. degree in Computer Science and Engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2010. He was an intern/co-op resercher at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan, from September 2007 to December 2007. From March 2008 to November 2009, he was a postdoctral fellow of the EMIME project at the Nagoya Institute of Technology. From December 2009 to Mar. 2012, he was a postdoctral fellow of the SCOPE project at Nagoya Institute of Technology. He is currently a postdoctral fellow of the CREST project at the Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis. He received the ISCSLP Best Student Paper Award, the Yamashita SIG Research Award and the ASJ Itakura Award in 2008, 2010 and 2013, respectively. He is a member of the Acoustical Society of Japan and Information Processing Society of Japan.