Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis

Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang

***Abstract*—This paper presents an investigation into ways of integrating articulatory features into hidden Markov model (HMM)-based parametric speech synthesis. In broad terms, this may be achieved by estimating the joint distribution of acoustic and articulatory features during training. This may in turn be used in conjunction with a maximum-likelihood criterion to produce acoustic synthesis parameters for generating speech. Within this broad approach, we explore several variations that are possible in the construction of an HMM-based synthesis system which allow articulatory features to influence acoustic modeling: model clustering, state synchrony and cross-stream feature dependency. Performance is evaluated using the RMS error of generated acoustic parameters as well as formal listening tests. Our results show that the accuracy of acoustic parameter prediction and the naturalness of synthesized speech can be improved when shared clustering and asynchronous-state model structures are adopted for combined acoustic and articulatory features. Most significantly, however, our experiments demonstrate that modeling the dependency between these two feature streams can make speech synthesis systems more flexible. The characteristics of synthetic speech can be easily controlled by modifying generated articulatory features as part of the process of producing acoustic synthesis parameters.**

***Index Terms*—Articulatory features, hidden Markov model (HMM), speech production, speech synthesis.**

I. Introduction

THE hidden Markov model (HMM) has been used for automatic speech recognition (ASR) since the mid-1970s, and has since come to dominate that field. Recently, the HMM has also made significant progress as a method for speech synthesis, particularly within the last decade [1]–[3].

In this method, the spectrum, F0 and segment durations are modeled simultaneously within a unified HMM framework [1]. To synthesize speech, these features are directly predicted from the trained HMMs by means of the Maximum-Likelihood Parameter Generation Algorithm [2] which incorporates dynamic features. The predicted parameter trajectories are then sent to a parametric synthesizer to generate the speech waveform. This method is able to synthesize highly intelligible and smooth speech [4], [5]. In addition, several adaptation and interpolation methods can be applied to control the HMM's parameters and so diversify the characteristics of the generated speech [6]–[10]. In this way, HMM-based speech synthesis offers a far higher degree of flexibility compared to that afforded by the unit selection waveform concatenation method, which has been the leading method throughout the past decade.

Mainstream speech technology based on the HMM, including ASR and speech synthesis, has largely used features derived directly from the acoustic signal as the observation sequence to be modeled. However, an acoustic parameterization is not the only possible representation for speech; articulatory features also offer an effective description of a speech utterance. Here, we use "articulatory features" to refer to the quantitative positions and continuous movements of a group of human articulators. These articulators include the tongue, jaw, lips, velum, and so on. Various techniques are available which enable us to record the movement of these articulators, such as X-ray microbeam cinematography [11], electromagnetic articulography (EMA) [12], magnetic resonance imaging (MRI) [13], ultrasound [14], and video motion capture of the external articulators [15]. The acoustic and articulatory features for an utterance are inherently related, because it is the manipulation of the articulators that generates the acoustic signal. However, the physical nature of human speech production means that an articulatory parameterization of speech has certain attractive properties:

1) Due to physical constraints, articulatory features evolve in a relatively slow and smooth way. Hence, they are well suited for modeling with an HMM, which assumes a quasi-stationary stochastic process.
2) Articulatory features can provide a straightforward and simple explanation for speech characteristics. For example, to express the movement of the F2 formant from high to low is easy in terms of articulatory features (for example the tongue moving from the front of the mouth to the back) but is more complicated in the domain of standard acoustic parameters, such as mel-cepstra or line spectral frequencies (LSFs).
3) Since articulatory features may be acquired by capturing the positions of articulators directly, they are not influenced in the same way by acoustic noise and other environmental conditions, such as the frequency response of

Manuscript received April 08, 2008; revised December 16, 2008. Current version published June 26, 2009. This work was supported by the Marie Curie Early Stage Training (EST) Network, "Edinburgh Speech Science and Technology (EdSST)." The work of K. Richmond was supported by the Engineering and Physical Sciences Research Council (EPSRC). The work of J. Yamagishi was supported by the EPSRC and an EC FP7 collaborative project called the *EMIME* Project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

Z. Ling and R. Wang are with the iFlytek Speech Lab, University of Science and Technology of China, Hefei, 230027, China (e-mail: zhling@ustc.edu; rhw@ustc.edu.cn).

K. Richmond and J. Yamagishi are with the Center for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh EH8 9LW, U.K. (e-mail: korin@cstr.ed.ac.uk; jyamagis@inf.ed.ac.uk).

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

Digital Object Identifier 10.1109/TASL.2009.2014796

acoustic recorders, or the distance between the speaker's mouth and the microphone.

With potentially beneficial properties such as these in mind, several researchers have applied articulatory features to HMM-based ASR, and have reported positive results in terms of reducing recognition error [16]–[18]. Research on combining articulatory features with HMM-based parameter generation methods has also been previously described [19], [20]. In [19], an HMM-based acoustic-to-articulatory mapping method was proposed. In [20], which focused on speech synthesis, both articulatory and excitation parameters were modeled and generated using the framework of HMM-based speech synthesis. The generated articulatory parameters were then mapped to spectral coefficients using a Gaussian mixture model (GMM). Finally, the acoustic speech signal was generated from the mapped spectral coefficients and excitation parameters. In this paper, in contrast to [20], we explore several ways to simultaneously model and generate spectral and articulatory features using HMMs.

The work described here has been undertaken with two aims in mind. The first is to improve the naturalness of synthesized speech. It has previously been demonstrated that objective distance metrics calculated in terms of the acoustic parameterization of real and synthesized speech (e.g., mel-cepstral distortion or root mean square (RMS) error of line spectral frequencies (LSF)) correlate with human subjective perception of speech quality [21]. We therefore aim to reduce the distance between the generated and natural acoustic parameters and thus improve the naturalness of synthesized speech. The validity of this objective evaluation is also supported by previous work on an alternative optimization criterion for training HMM-based synthesis systems [22]. This work has likewise shown that the naturalness of synthesized speech can be improved by reducing the distance between the generated and natural acoustic parameters.

The second significant aim of this work is to broaden the flexibility of HMM-based speech synthesis. By flexibility, we refer to the capability, for example, to readily generate voices of different genders and ages, to simulate different accents of a language, and to approximate foreign loan words. A speech synthesis system can be applied more widely if it has greater flexibility.

As mentioned above, a major advantage of model-based parametric synthesis over unit selection is its flexibility. However, this flexibility comes from the application of data-driven learning and adaptation methods. As such, we are unfortunately still very much reliant upon, and constrained by, the availability of suitable data for model training and adaptation. For example, should we want to build a speech synthesizer with a child's voice, a certain amount of child speech data must be available, which can prove problematic. As another example, we might want to take a synthesizer trained on a specific English speaker's voice and extend it to enable synthesis of a foreign language such as Spanish. This would be useful for applications such as speech-to-speech translation, where a user would ideally be able to communicate in a foreign language with a voice resembling their own. However, this poses the problem of how to deal with a lack of Spanish speech data from the user; for example, Spanish has nasalized vowels which are not present in English. Unfortunately, while we might have relevant phonetic knowledge concerning the properties of speech (such as the differences between an adult's speech and that of a child, or the differences in phone inventories between two languages), it is very difficult to integrate such knowledge into current systems directly.

Articulatory features offer a useful approach to overcoming this limitation. Because articulatory features explicitly represent the speech production mechanism and have physiological meaning, it is far more convenient to modify them according to phonetic rules and linguistic knowledge than to modify acoustic features. For example, the articulatory features of an adult speaker could easily be scaled to simulate the shorter and more narrow vocal tract of a child speaker, while vowel nasalization could easily be realized by explicitly controlling the velar port opening.

To take advantage of this, in addition to adequately modeling articulatory features themselves, we need to model the relationship between the articulatory and acoustic domains. Specifically, we require the capability to produce acoustic features which appropriately reflect the state of the articulatory system. If successful, we would then be in a position to manipulate the articulatory representation of synthetic speech directly in order to change the characteristics of the synthesized audio speech signal. In other words, we would obtain "articulatorily controllable" speech synthesis. It would be possible to synthesize speech approximating a child's voice or to synthesize phones from a foreign language by modifying the articulatory features in the appropriate way and then reconstructing the acoustic parameters on the basis of these modified articulatory features. In many cases it would be possible, and quite desirable, to perform articulatory modification explicitly, according to phonetic knowledge and without requiring novel speech data from the target speaker.

Finally, in addition to speech synthesis in isolation, a unified statistical model for acoustic and articulatory features could be exploited by several other speech-related systems. For example, in an animated talking-head system, the speech synthesis and facial animation could make use of different parts of the unified model. This would facilitate coordination of coarticulation and synchronization between the audio and video streams. In a language tutoring system, the user could be guided not only by the synthesized speech but also by the articulator movements predicted simultaneously from the input text. The model could even be applied, for example, to assisting communication by speech in noisy environments; a portable hardware device to acquire a user's articulatory movements in real-time could be used in conjunction with a synthesis system able to incorporate articulatory features. Similarly, communication by whispered or silent speech (e.g., in an environment which requires silence or for laryngectomy patients) might become possible using speech synthesis driven by a user's articulatory movements.

In the following sections of this paper, we detail our method. A unified statistical model for the joint distribution of acoustic and articulatory features is estimated from parallel acoustic and articulatory training data. During synthesis, acoustic features are generated from the unified model using a maximum-likelihood criterion. In order to explore the influence of articula-

tory features on acoustic models, several variations of model structure are investigated in this work. These include: experiments where the HMM state tying tree is built using articulatory and acoustic features jointly ("shared clustering"); experiments to investigate the effect of synchrony in the articulatory and acoustic state sequences ("synchronous-state modeling"); and experiments where we introduce an explicit function to model the dependence of acoustic features on articulatory features ("dependent-feature modeling"). These experiments are conducted using a corpus of parallel acoustic and EMA recordings, and we evaluate the performance of the proposed method at improving the naturalness and flexibility of our HMM-based speech synthesis system.

II. Method

A. HMM-Based Parametric Speech Synthesis System

Fig. 1 shows a diagram of standard HMM-based speech synthesis systems. During training, the F0 and spectral parameters of $D_{\boldsymbol{X}}$ dimensions are extracted from the waveforms contained in the training set. Then a set of context-dependent HMMs λ are estimated to maximize the likelihood function $P(\boldsymbol{X} \mid \lambda)$ for the training acoustic features. Here, $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_N^\top]^\top$ is the observation feature sequence, $(\cdot)^\top$ means the matrix transpose, and N is the length of the sequence. The observation feature vector $\boldsymbol{x}_t \in \mathcal{R}^{3D_{\boldsymbol{X}}}$ for each frame consists of static acoustic parameters $\boldsymbol{x}_{S_t} \in \mathcal{R}^{D_{\boldsymbol{X}}}$ and their velocity and acceleration components as

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

where

$$\Delta \boldsymbol{x}_{S_t} = 0.5\boldsymbol{x}_{S_{t+1}} - 0.5\boldsymbol{x}_{S_{t-1}} \quad \forall t \in [2, N-1] \tag{2}$$

$$\Delta \boldsymbol{x}_{S_1} = \Delta \boldsymbol{x}_{S_2}, \quad \Delta \boldsymbol{x}_{S_N} = \Delta \boldsymbol{x}_{S_{N-1}} \tag{3}$$

and

$$\Delta^2 \boldsymbol{x}_{S_t} = \boldsymbol{x}_{S_{t+1}} - 2\boldsymbol{x}_{S_t} + \boldsymbol{x}_{S_{t-1}} \quad \forall t \in [2, N-1] \tag{4}$$

$$\Delta^2 \boldsymbol{x}_{S_1} = \Delta^2 \boldsymbol{x}_{S_2}, \quad \Delta^2 \boldsymbol{x}_{S_N} = \Delta^2 \boldsymbol{x}_{S_{N-1}}. \tag{5}$$

Therefore, the complete feature sequence $\boldsymbol{X}$ can be considered as a linear transform of the static feature sequence $\boldsymbol{X}_S = [\boldsymbol{x}_{S_1}^\top, \boldsymbol{x}_{S_2}^\top, \ldots, \boldsymbol{x}_{S_N}^\top]^\top$ as

$$\boldsymbol{X} = \boldsymbol{W}_{\boldsymbol{X}} \boldsymbol{X}_S \tag{6}$$

where $\boldsymbol{W}_{\boldsymbol{X}} \in \mathcal{R}^{3ND_{\boldsymbol{X}} \times ND_{\boldsymbol{X}}}$ is determined by the velocity and acceleration calculation functions in (2)–(5), [2]. A multispace probability distribution (MSD) [23] is used to model the F0 features. This addresses the problem that F0 is only defined for regions of voiced speech, while it takes a value of "unvoiced" for voiceless regions. The MSD provides a principled way to incorporate a distribution for F0 into the probabilistic framework of the HMM.

An HMM-based synthesizer typically contains a large number of context-dependent HMMs, with context features that are far more extensive and express far more fine-grained distinctions than those used in ASR HMM systems. This leads to data-sparsity problems, such as over-fitting in context-dependent models that have few training examples available and the problem that many valid combinations of context features will be completely unrepresented in the training set. To deal with this, a decision-tree-based model clustering technique that uses a minimum description length (MDL) criterion [24] to guide tree construction is applied after initial training to cluster context-dependent HMMs. The MDL criterion minimizes the description length of the model with respect to the training data at each split during the building of the decision tree in the top-down direction. The description length is defined as [24]

$$\mathcal{D}(\lambda) \equiv -\log P(\boldsymbol{X} \mid \lambda) + \frac{1}{2} D(\lambda) \log G + C \tag{7}$$

where $\log P(\boldsymbol{X} \mid \lambda)$ is the log likelihood function of the model for the training set, $D(\lambda)$ is the dimensionality of the model parameters, G is the total number of observed frames in the training set, and C is a constant. This criterion has been proved to find a decision-tree size that is close to optimal for the purpose of HMM-based speech synthesis model training [25]. Next, we take the state alignment results using the trained HMMs and use them to train context-dependent state duration probabilities [1]. A single-mixture Gaussian distribution is used to model the log-duration probability for each state. A decision-tree-based model clustering technique is similarly applied to these duration distributions.

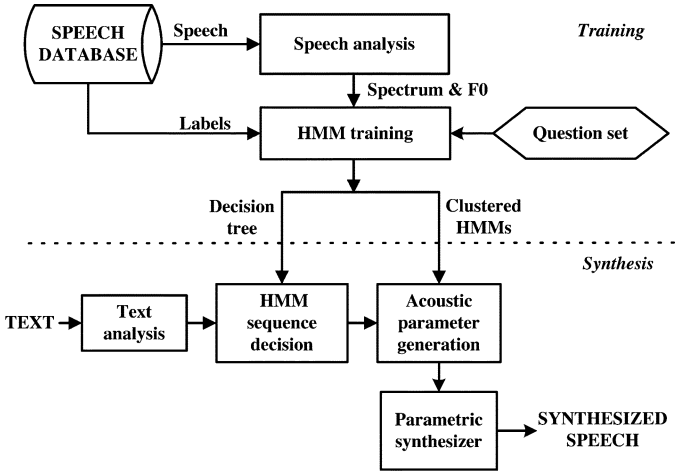


Fig. 1. Diagram of a typical HMM-based parametric speech synthesis system.

To perform synthesis, the result of front-end linguistic analysis on the input text sentence is used to determine the sentence HMM. This is done by consulting the decision-tree which was built to cluster HMM models during training. The Maximum-Likelihood Parameter Generation Algorithm [2] is then applied to generate the optimal static acoustic parameters, such that

$$\boldsymbol{X}_S^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{X} \mid \lambda) = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_{\boldsymbol{X}} \boldsymbol{X}_S \mid \lambda). \tag{8}$$

This equation can be solved by setting $\partial P(\boldsymbol{W}_{\boldsymbol{X}} \boldsymbol{X}_S \mid \lambda)/\partial \boldsymbol{X}_S = \boldsymbol{0}$. $\boldsymbol{X}_S^*$ can then be optimized directly once the state sequence is given [2]. Finally, these generated parameters are sent to a parametric synthesizer to reconstruct the speech waveform.

B. Integrating Articulatory Features

Our method of integrating articulatory features follows the same general framework of an acoustics-only HMM-based speech synthesis system. During training, with parallel acoustic and articulatory observation sequences of length N, a statistical model λ for the combined acoustic and articulatory features is estimated to maximize the likelihood function of their joint distribution $P(\boldsymbol{X},\boldsymbol{Y}\,|\,\lambda)$, where $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_N^\top]^\top$ denotes a given articulatory observation sequence. For each frame the articulatory feature vector $\boldsymbol{y}_t \in \mathcal{R}^{3D_{\boldsymbol{Y}}}$ is similarly composed of static features $\boldsymbol{y}_{S_t} \in \mathcal{R}^{D_{\boldsymbol{Y}}}$ and their velocity and acceleration components as

$$\boldsymbol{y}_t = [\boldsymbol{y}_{S_t}^\top, \Delta\boldsymbol{y}_{S_t}^\top, \Delta^2\boldsymbol{y}_{S_t}^\top]^\top \tag{9}$$

where $D_{\boldsymbol{Y}}$ is the dimensionality of the static articulatory features. At synthesis time, the acoustic features and articulatory features are simultaneously generated from the trained models based on a maximum-likelihood parameter generation method that considers explicit constraints of the dynamic features as

$$(\boldsymbol{X}_S^*, \boldsymbol{Y}_S^*) = \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} P(\boldsymbol{X},\boldsymbol{Y}\,|\,\lambda) \tag{10}$$

$$= \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} \sum_{\forall \boldsymbol{q}_{\boldsymbol{X}},\forall \boldsymbol{q}_{\boldsymbol{Y}}} P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda) \tag{11}$$

$$= \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} \sum_{\forall \boldsymbol{q}_{\boldsymbol{X}},\forall \boldsymbol{q}_{\boldsymbol{Y}}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S, \boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda) \tag{12}$$

where

$$\boldsymbol{Y}_S = [\boldsymbol{y}_{S_1}^\top, \boldsymbol{y}_{S_2}^\top, \ldots, \boldsymbol{y}_{S_N}^\top]^\top \tag{13}$$

$$\boldsymbol{Y} = \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S. \tag{14}$$

$\boldsymbol{W}_{\boldsymbol{Y}} \in \mathcal{R}^{3ND_{\boldsymbol{Y}} \times ND_{\boldsymbol{Y}}}$ is the matrix used to calculate a complete articulatory feature sequence based on static parameters. $\boldsymbol{q}_{\boldsymbol{X}} = \{q_{\boldsymbol{X}_1}, q_{\boldsymbol{X}_2}, \ldots, q_{\boldsymbol{X}_N}\}$ and $\boldsymbol{q}_{\boldsymbol{Y}} = \{q_{\boldsymbol{Y}_1}, q_{\boldsymbol{Y}_2}, \ldots, q_{\boldsymbol{Y}_N}\}$ denote the state sequence for acoustic and articulatory features, respectively. We solve (12) by keeping only the optimal state sequences in the accumulation and approximating it as a two-step optimization problem

$$(\boldsymbol{X}_S^*, \boldsymbol{Y}_S^*) \approx \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} \max_{\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S, \boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda) \tag{15}$$

$$= \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} \max_{\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S\,|\,\lambda, \boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}) \times P(\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda) \tag{16}$$

$$\approx \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S\,|\,\lambda, \boldsymbol{q}_{\boldsymbol{X}}^*,\boldsymbol{q}_{\boldsymbol{Y}}^*) \times P(\boldsymbol{q}_{\boldsymbol{X}}^*,\boldsymbol{q}_{\boldsymbol{Y}}^*\,|\,\lambda) \tag{17}$$

where

$$(\boldsymbol{q}_{\boldsymbol{X}}^*, \boldsymbol{q}_{\boldsymbol{Y}}^*) = \arg\max_{\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}} P(\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda) \tag{18}$$

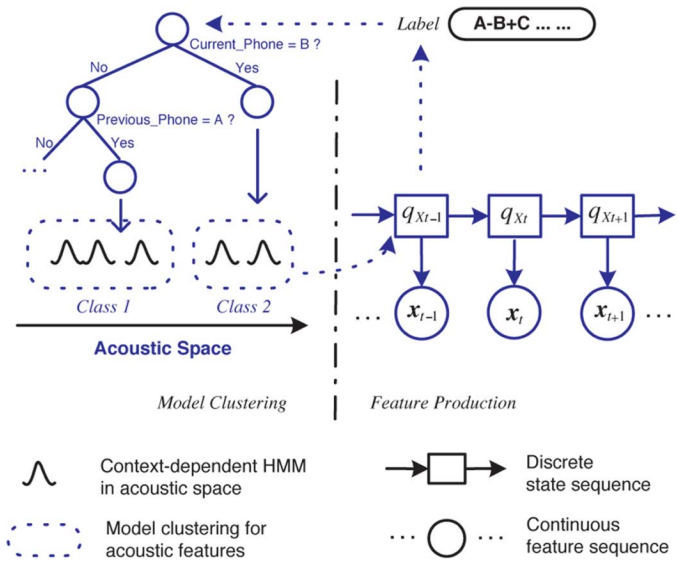


Fig. 2. Model structure of an HMM-based parametric speech synthesis system using only acoustic features.

is the set of optimal state sequences determined from the above duration probability $P(\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}}\,|\,\lambda)$, which is estimated based on the method proposed in [1].[1]

Before discussing how to train the joint distribution $P(\boldsymbol{X},\boldsymbol{Y}\,|\,\lambda)$ for the combined acoustic and articulatory features, let us look at the model structure of the acoustics-only HMM-based speech synthesis system, as shown in Fig. 2. For convenience, the acoustic space is illustrated as a single dimension in this figure. As indicated, the model structure can be considered as consisting of two parts. The first part is *model clustering*, through which parts of the acoustic space are populated with disjoint groups of clustered context-dependent HMMs. The second part is the *feature production* model, whereby an acoustic feature sequence is generated from the probability density functions (pdfs) of an HMM state sequence using a maximum-likelihood principle. Here, the set of context features associated with any given state in the sequence determines the class to which it belongs within the cluster tree. This class in turn determines the pdf parameters for the given state. For example, in Fig. 2, the context label "$A-B+C$" of state $q_{\boldsymbol{X}_{t-1}}$ indicates that the current phone is B, the previous phone is A and the next phone is C. We use context features such as these to "answer" the questions at each node in the decision tree and descend from the root node to the leaf cluster nodes. Hence, we determine that the state in this example belongs to the *Class 2* cluster in acoustic space. The model parameters of this class are then used to generate acoustic feature vector $\boldsymbol{x}_{t-1}$.

When acoustic and articulatory features are used in combination, we can thus investigate possibilities for model structure which consider these two aspects.

[1]For optimizing both the state sequences $(\boldsymbol{q}_{\boldsymbol{X}},\boldsymbol{q}_{\boldsymbol{Y}})$ and the feature vectors $(\boldsymbol{X}_S,\boldsymbol{Y}_S)$ simultaneously, an EM-based parameter generation algorithm [2] can be used instead of the above two-step optimization.

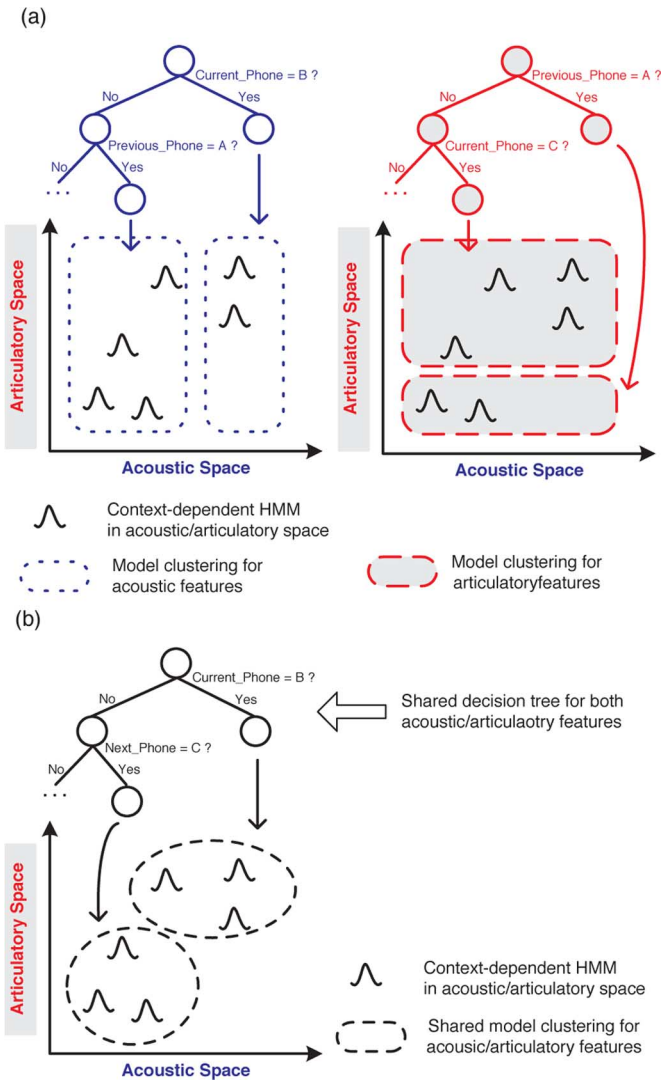


Fig. 3. Different model clustering approaches for combined acoustic and articulatory modeling. (a) *Separate Clustering*. (b) *Shared Clustering*.

1) *Model Clustering*. As Fig. 3 indicates, we can choose either to cluster the acoustic and articulatory model distribution parameters independently ["separate clustering," Fig. 3(a)], or to build a shared decision tree to cluster the distribution parameters for both feature types simultaneously ["shared clustering," Fig. 3(b)].
2) *Feature Production*. There are more variations available for feature production using combined acoustic and articulatory features. As shown in Fig. 4, we explore possibilities in terms of the synchrony between acoustic and articulatory state sequences on one hand, and the dependency between articulatory and acoustic features on the other. In the asynchronous-state model, the two feature sequences are assumed to be generated from different state sequences, whereas there is only one state sequence in the synchronous-state model. In the independent-feature model, the generation of acoustic features is assumed only to depend upon the current state, whereas it is also dependent upon the current articulatory features in the dependent-feature model.

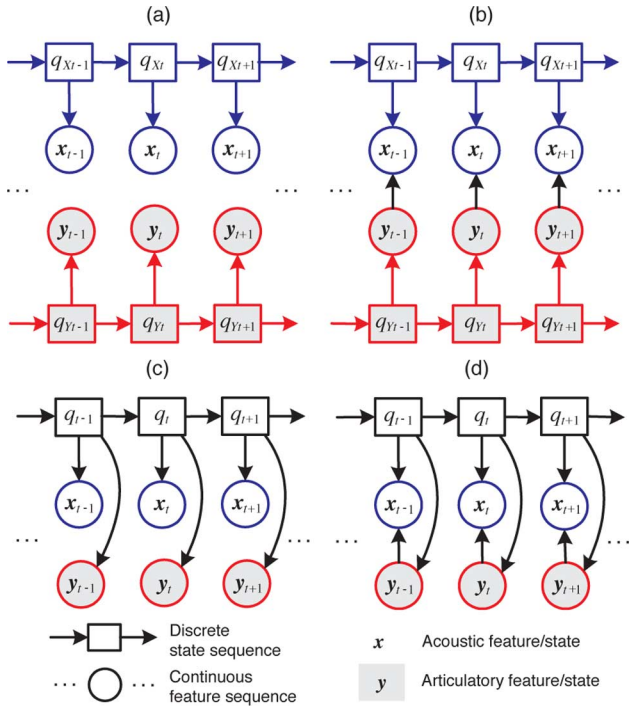


Fig. 4. Different feature production models for combined acoustic and articulatory modeling. (a) *Asynchronous & Independent*. (b) *Asynchronous & Dependent*. (c) *Synchronous & Independent*. (d) *Synchronous & Dependent*.

In total, we are presented with three variables to determine model structure: separate/shared clustering, asynchronous/synchronous-state, and independent/dependent-feature streams. Therefore, there is a total of eight model structures which are possible. In this paper, four of these are implemented and evaluated. This includes the *Baseline* system which is trained using acoustic features alone. For the purpose of our investigation here, we can consider the acoustic *Baseline* system as one of the possible eight systems since we compare systems only in terms of performance in the acoustic domain. Hence, for the sake of comparison with other systems, the *Baseline* system equates to the system with separate model clustering, asynchronous-state sequence and acoustic features independent of the articulatory stream.

The definition of the four systems and their corresponding subfigure indices in Figs. 3 and 4 are shown in Table I, where $\times$ means negative and $\surd$ means positive for the listed alternative for each factor. These four systems are sufficient to investigate the effect of the alternatives for all three factors. Having already described the *Baseline* system, we look at the other three systems in more detail next.

C. Shared Clustering System

Model clustering is an indispensable part of constructing an HMM-based speech synthesis system. Using decision-tree-based clustering, the robustness of model parameter estimation can be improved and the distribution parameters for context-dependent phones not present in the training set can be determined. In separate clustering, separate decision trees for the acoustic and articulatory feature streams are trained under the MDL criterion. Conversely, in shared clustering, a

TABLE I
DEFINITION OF DIFFERENT SYSTEMS

Name	Label	Model Structure			Fig.
		Shared Clustering	Synchronous-State	Dependent-Feature	
Baseline	BL	×	×	×	3(a)+4(a)
Shared Clustering	SC	√	×	×	3(b)+4(a)
State-Synchrony	SS	√	√	×	3(b)+4(c)
Feature-Dependency	FD	√	√	√	3(b)+4(d)

shared decision tree is built for both acoustic and articulatory models together. The same MDL criterion is followed and the tree building algorithm is similar to the shared tree clustering in [26]. The definition of description length is similar to (7) except that the log likelihood function $\log P(\boldsymbol{X}\,|\,\lambda)$ is replaced by $\log P(\boldsymbol{X},\boldsymbol{Y}\,|\,\lambda)$ and $D(\lambda)$ is set to the sum of the dimensionality of acoustic and articulatory models.

In the *Shared Clustering* system, the acoustic features are generated directly from the acoustic component of the models, as the two feature streams are assumed to be independent given their state sequences. Hence, (17) can be rewritten as

$$(\boldsymbol{X}_S^*,\boldsymbol{Y}_S^*) \approx \arg\max_{\boldsymbol{X}_S,\boldsymbol{Y}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S\,|\,\lambda,\boldsymbol{q}_{\boldsymbol{X}}^*) \times P(\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S\,|\,\lambda,\boldsymbol{q}_{\boldsymbol{Y}}^*) \tag{19}$$

$$\boldsymbol{X}_S^* \approx \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S\,|\,\lambda,\boldsymbol{q}_{\boldsymbol{X}}^*) \tag{20}$$

where the optimal state sequences $\boldsymbol{q}_{\boldsymbol{X}}^*$ and $\boldsymbol{q}_{\boldsymbol{Y}}^*$ are also predicted independently according to the duration probabilities for the acoustic and articulatory features, as there are no synchronicity constraints between them.

D. State-Synchrony System

In the *State-Synchrony* system, acoustic features and articulatory features are assumed to be generated from the same state sequence. This model structure can be approximated by two-stream HMM modeling. In the two-stream HMM, we have

$$P(\boldsymbol{X},\boldsymbol{Y}\,|\,\lambda) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{q}\,|\,\lambda) \tag{21}$$

$$= \sum_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{N} a_{q_{t-1}q_t} b(\boldsymbol{x}_t,\boldsymbol{y}_t) \tag{22}$$

$$b_j(\boldsymbol{x}_t,\boldsymbol{y}_t) = b_j(\boldsymbol{x}_t) b_j(\boldsymbol{y}_t) \tag{23}$$

$$b_j(\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_t;\boldsymbol{\mu}_{\boldsymbol{X}_j},\boldsymbol{\Sigma}_{\boldsymbol{X}_j}\right) \tag{24}$$

$$b_j(\boldsymbol{y}_t) = \mathcal{N}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_{\boldsymbol{Y}_j},\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\right) \tag{25}$$

where $\boldsymbol{q}_{\boldsymbol{X}} = \boldsymbol{q}_{\boldsymbol{Y}} = \boldsymbol{q} = \{q_1, q_2, \ldots, q_N\}$ denotes the state sequence shared by the two feature streams, π_j and a_{ij} represent initial state probability and state transition probability, respectively; $b_j(\cdot)$ means the state observation probability density function (pdf) for state j; and $\mathcal{N}(;\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The conventional Baum–Welch method [27] can be used to estimate the model parameters $\{a_{ij},\boldsymbol{\mu}_{\boldsymbol{X}_j},\boldsymbol{\Sigma}_{\boldsymbol{X}_j},\boldsymbol{\mu}_{\boldsymbol{Y}_j},\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\}$. The synchronous-state constraint not only influences the training of state duration probabilities, but also affects the calculation of the state occupancy probability for each frame in the Baum–Welch algorithm. As a result, the estimated acoustic and articulatory model parameters are different from those of the *Shared Clustering* system.

At synthesis time, the acoustic features can be generated in the same way as for the *Shared Clustering* system, with (19) and (20). Here, $\boldsymbol{q}_{\boldsymbol{X}}^* = \boldsymbol{q}_{\boldsymbol{Y}}^* = \boldsymbol{q}^*$ is decided by the duration probabilities that are trained using the single state alignment shared by the acoustic and articulatory features.

E. Feature-Dependency System

In the *Feature-Dependency* system, an explicit dependency between acoustic and articulatory features is considered. The generation of acoustic features is decided not only by the context-dependent acoustic model parameters but also by the simultaneous articulatory features. Accordingly, we modify (23) so that

$$b_j(\boldsymbol{x}_t,\boldsymbol{y}_t) = b_j(\boldsymbol{x}_t\,|\,\boldsymbol{y}_t) b_j(\boldsymbol{y}_t). \tag{26}$$

Several approaches have been proposed to model the dependency $b_j(\boldsymbol{x}_t\,|\,\boldsymbol{y}_t)$ between these two feature streams. In [16], articulatory features were discretized as $\boldsymbol{y}_t \in \{\boldsymbol{y}^{(1)},\boldsymbol{y}^{(2)},\ldots,\boldsymbol{y}^{(S)}\}$, where S denotes the size of the discrete space. Then, $b_j(\boldsymbol{x}_t\,|\,\boldsymbol{y}_t = \boldsymbol{y}^{(i)}), i = 1,2,\ldots,S$ were trained for each possible value of $\boldsymbol{y}_t$. In [19], a piecewise linear transform was used to model the dependency between these two feature streams for the acoustic-to-articulatory mapping. Similarly, a linear transform has been applied in multistream speech recognition [28] to model the dependency between different acoustic features.

In this paper, we too adopt the approach of using a linear transform to model the dependency of the acoustic features on the articulatory features. For a given state at a given time frame, we define the mean of the distribution for the acoustic features as the sum of two terms: a state-specific time-independent value (which is independent of the articulatory features) and a linear transform of the time-varying articulatory features (which introduces dependency). This is illustrated in Fig. 5. Note this linear transform matrix is also state-dependent. In this way, we introduce a globally *piecewise linear* mapping to model the relationship between the articulatory and acoustic features. Mathematically, such dependency can be expressed as

$$b_j(\boldsymbol{x}_t\,|\,\boldsymbol{y}_t) = \mathcal{N}\left(\boldsymbol{x}_t;\boldsymbol{A}_j\boldsymbol{y}_t + \boldsymbol{\mu}_{\boldsymbol{X}_j},\boldsymbol{\Sigma}_{\boldsymbol{X}_j}\right) \tag{27}$$

where $\boldsymbol{A}_j \in \mathcal{R}^{3D_{\boldsymbol{X}}\times 3D_{\boldsymbol{Y}}}$ is the linear transform matrix for state j. An expectation-maximization (EM) algorithm [29] can be used to estimate the model parameters. The re-estimation formulae can be derived as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{Y}_j} = \frac{\sum_{t=1}^{T}\gamma_j(t)\boldsymbol{y}_t}{\sum_{t=1}^{T}\gamma_j(t)} \tag{28}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t) \left(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}_j}\right)\left(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}_j}\right)^{\mathsf{T}}}{\sum_{t=1}^{T} \gamma_j(t)} \tag{29}$$

$$\hat{\boldsymbol{A}}_j = \left[\sum_{t=1}^{T} \gamma_j(t) \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{\boldsymbol{X}_j}\right)\boldsymbol{y}_t^{\mathsf{T}}\right] \cdot \left[\sum_{t=1}^{T} \gamma_j(t) \boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{T}}\right]^{-1} \tag{30}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{X}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t)(\boldsymbol{x}_t - \hat{\boldsymbol{A}}_j \boldsymbol{y}_t)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{31}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t) \left(\boldsymbol{x}_t - \hat{\boldsymbol{A}}_j \boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}_j}\right)\left(\boldsymbol{x}_t - \hat{\boldsymbol{A}}_j \boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}_j}\right)^{\mathsf{T}}}{\sum_{t=1}^{T} \gamma_j(t)} \tag{32}$$

where the hat symbol denotes the re-estimated parameters at each iteration, and $\gamma_j(t)$ is the occupancy probability of state j at time t. Model parameters taken from the *State-Synchrony* system are used as initial parameters for $\boldsymbol{\mu}_{\boldsymbol{X}_j}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}_j}$, $\boldsymbol{\mu}_{\boldsymbol{Y}_j}$, and $\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}$. $\boldsymbol{A}_j$ is set to be the zero matrix for the first iteration. In previous work, the joint distribution of acoustic and articulatory features has variously been modeled either in a context-independent way ([30], [31]) or in a context-dependent way with a separate transform matrix estimated for each state pdf $b_j(\cdot)$ ([19], [20]). Here, in contrast, the state-dependent transform matrices $\boldsymbol{A}_j$ are *tied* to a given class using a decision tree. The aim is to achieve a good balance between accuracy of cross-stream dependency modeling on one hand and a reduction of the number of parameters to be estimated on the other. Using a smaller number of tied transform matrices can help avoid over-fitting and improve robustness, but using too few tied matrices reduces the modeling power of the piecewise nonlinear mapping. In the experiments we present, we explore the effect of varying the number of tied transform matrices. Finally, to implement the tying of the transform matrices, we make use of the shared decision tree for the state pdfs of acoustic and articulatory features for convenience.

For the *Feature-Dependency* system, we consider two methods for parameter generation. Under the first method, we generate acoustic and articulatory parameters simultaneously from the unified model following a maximum-likelihood criterion similar to (17), such that

$$(\boldsymbol{X}_S^*, \boldsymbol{Y}_S^*) \approx \arg \max_{\boldsymbol{X}_S, \boldsymbol{Y}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \,|\, \lambda, \boldsymbol{q}^*). \tag{33}$$

The introduction of the transform matrix $\boldsymbol{A}_j$ in the *Feature-Dependency* system influences the calculation of $\gamma_j(t)$ and the estimation of all model parameters according to (28)–(32) at each iteration. Thus, the acoustic and articulatory features generated by this system are theoretically different from those generated by the *State-Synchrony* system. The joint distribution in (33) can be expressed as

$$\begin{aligned}\log P(&\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \,|\, \lambda, \boldsymbol{q}^*) \\ &= \boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \\ &\quad - \frac{1}{2}\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S + \boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{M}_{\boldsymbol{X}}\end{aligned}$$

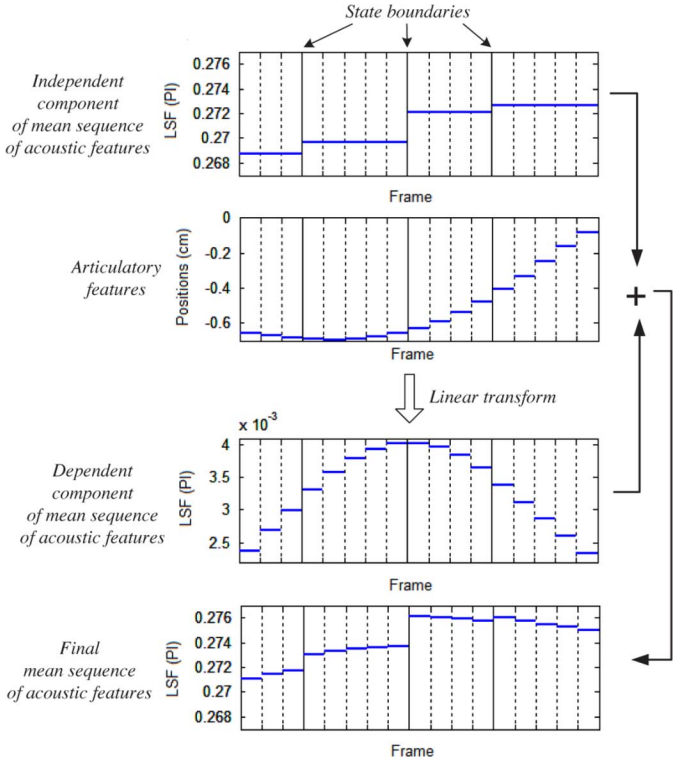


Fig. 5. Generation of HMM mean sequence of acoustic features in the *Feature-Dependency* system.

$$\begin{aligned}&- \frac{1}{2}\boldsymbol{Y}_S^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{Y}}^{\mathsf{T}}\left(\boldsymbol{U}_{\boldsymbol{Y}}^{-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{A}\right)\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \\ &+ \boldsymbol{Y}_S^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{Y}}^{\mathsf{T}}(\boldsymbol{U}_{\boldsymbol{Y}}^{-1}\boldsymbol{M}_{\boldsymbol{Y}} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{M}_{\boldsymbol{X}}) + K\end{aligned} \tag{34}$$

where

$$\boldsymbol{U}_{\boldsymbol{X}}^{-1} = \operatorname{diag}\left[\boldsymbol{\Sigma}_{\boldsymbol{X}_{q_1}}^{-1}, \boldsymbol{\Sigma}_{\boldsymbol{X}_{q_2}}^{-1}, \ldots, \boldsymbol{\Sigma}_{\boldsymbol{X}_{q_N}}^{-1}\right] \tag{35}$$

$$\boldsymbol{M}_{\boldsymbol{X}} = \left[\boldsymbol{\mu}_{\boldsymbol{X}_{q_1}}^{\mathsf{T}}, \boldsymbol{\mu}_{\boldsymbol{X}_{q_2}}^{\mathsf{T}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{X}_{q_N}}^{\mathsf{T}}\right]^{\mathsf{T}} \tag{36}$$

$$\boldsymbol{U}_{\boldsymbol{Y}}^{-1} = \operatorname{diag}\left[\boldsymbol{\Sigma}_{\boldsymbol{Y}_{q_1}}^{-1}, \boldsymbol{\Sigma}_{\boldsymbol{Y}_{q_2}}^{-1}, \ldots, \boldsymbol{\Sigma}_{\boldsymbol{Y}_{q_N}}^{-1}\right] \tag{37}$$

$$\boldsymbol{M}_{\boldsymbol{Y}} = \left[\boldsymbol{\mu}_{\boldsymbol{Y}_{q_1}}^{\mathsf{T}}, \boldsymbol{\mu}_{\boldsymbol{Y}_{q_2}}^{\mathsf{T}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{Y}_{q_N}}^{\mathsf{T}}\right]^{\mathsf{T}} \tag{38}$$

$$\boldsymbol{A} = \operatorname{diag}\left[\boldsymbol{A}_{q_1}, \boldsymbol{A}_{q_2}, \ldots, \boldsymbol{A}_{q_N}\right] \tag{39}$$

and K is a constant value. By setting

$$\frac{\partial \log P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \,|\, \lambda, \boldsymbol{q}^*)}{\partial \boldsymbol{X}_S} = \boldsymbol{0} \tag{40}$$

$$\frac{\partial \log P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S \,|\, \lambda, \boldsymbol{q}^*)}{\partial \boldsymbol{Y}_S} = \boldsymbol{0} \tag{41}$$

we can obtain the optimal trajectories for acoustic features $\boldsymbol{X}_S^*$ and articulatory features $\boldsymbol{Y}_S^*$ as follows:

$$\boldsymbol{X}_S^* = \left(\boldsymbol{W}_{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\right)^{-1}\boldsymbol{W}_{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}(\boldsymbol{M}_{\boldsymbol{X}} + \boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S^*) \tag{42}$$

$$\begin{aligned}\boldsymbol{Y}_S^* = &\left(\boldsymbol{W}_{\boldsymbol{Y}}^{\mathsf{T}}(\boldsymbol{U}_{\boldsymbol{Y}}^{-1} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{A} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{Z}^{-1}\boldsymbol{A})\boldsymbol{W}_{\boldsymbol{Y}}\right)^{-1} \\ &\cdot \boldsymbol{W}_{\boldsymbol{Y}}^{\mathsf{T}}\left(\boldsymbol{U}_{\boldsymbol{Y}}^{-1}\boldsymbol{M}_{\boldsymbol{Y}} + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{Z}^{-1}\boldsymbol{M}_{\boldsymbol{X}} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{M}_{\boldsymbol{X}}\right)\end{aligned} \tag{43}$$

where

$$Z^{-1} = \boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\left(\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\right)^{-1}\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}. \quad (44)$$

If we set $\boldsymbol{A} = \boldsymbol{0}$, (42) is equivalent to the standard parameter generation algorithm used with only acoustic features. Thus, the effect of dependent-feature modeling in parameter generation can be viewed as a modification to the mean sequence of acoustic features via $\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S^*$. In (44), $Z^{-1} \in \mathcal{R}^{3ND_{\boldsymbol{X}}\times 3ND_{\boldsymbol{Y}}}$ is a full matrix and N is the number of frames in a whole sentence. In order to alleviate the computational expense incurred by matrix inversion, Z^{-1} can be approximated by a band matrix with band width $3L \cdot D_{\boldsymbol{X}}$. The same method discussed in [32] is adopted here to achieve this approximation and so speed up the calculation.

The second method we consider here to generate acoustic parameters is to use natural articulatory features. This method would not generally apply under normal speech synthesis circumstances. However, for certain applications, such as speech enhancement in a noisy environment and speech reconstruction for laryngectomy patients based on articulatory movements, natural articulatory features could be available. Moreover, this method can be considered to be an upper bound on the performance of acoustic parameter generation in the *Feature-Dependency* system, which is helpful when evaluating the potential of the model structure. Once the natural articulatory parameters $\boldsymbol{Y}_S$ are given, the state observation pdf for the acoustic features can be determined using (27), which may in turn be used to generate acoustic parameters such that

$$\boldsymbol{X}_S^* \approx \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S \,|\, \lambda, \boldsymbol{q}^*, \boldsymbol{Y}_S) \quad (45)$$

where $\boldsymbol{q}^*$ is the state alignment determined for $\boldsymbol{Y}_S$ using the articulatory part of the trained model. The distribution can be found by simplifying (34) to

$$\log P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S \,|\, \lambda, \boldsymbol{q}, \boldsymbol{Y}_S) = -\frac{1}{2}\boldsymbol{X}_S^{\top}\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S + \boldsymbol{X}_S^{\top}\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}(\boldsymbol{M}_{\boldsymbol{X}} + \boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S) + K. \quad (46)$$

By setting

$$\frac{\partial \log P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S \,|\, \lambda, \boldsymbol{q}^*, \boldsymbol{Y}_S)}{\partial \boldsymbol{X}_S} = \boldsymbol{0} \quad (47)$$

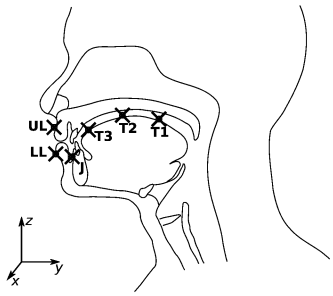
we can generate the optimal acoustic feature sequence $\boldsymbol{X}_S^*$ as

$$\boldsymbol{X}_S^* = \left(\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\right)^{-1}\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}(\boldsymbol{M}_{\boldsymbol{X}} + \boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}}\boldsymbol{Y}_S). \quad (48)$$

III. Experiments

A. Database

A multichannel articulatory database was used in our experiments. It contains the acoustic waveform recorded concurrently with EMA data. 1263 phonetically balanced sentences were read by a male British English speaker. The waveforms were available in 16-kHz PCM format with 16-bit precision. Six EMA receivers were used in our experiments. The positions of these receivers are shown in Fig. 6. For each receiver, coordinates in three dimensions were recorded at a sample rate of 200 Hz: the x- (left to right), y- (front to back) and z- (bottom to top) axes (relative to viewing the speaker's face from the front). All six receivers were placed in the midsagittal plane of the speaker's head, and their movements in the x-axis were very small. Therefore, only the y- and z-coordinates of the six receivers were used in our experiments, making a total of 12 static articulatory features.



Label	Location	Label	Location
T1	Tongue dorsum	J	Jaw
T2	Tongue body	LL	Lower lip
T3	Tongue tip	UL	Upper lip

Fig. 6. Placement of EMA receivers in the database used for the experiments.

B. System Construction

In order to build our HMM-based speech synthesis systems, we generated the context labels for the database using Unilex [33] and Festival [34] tools, and determined phone boundaries automatically using HTK [35]. 1200 sentences were selected for training and the remaining 63 sentences were used as a test set. The *Baseline* system was constructed using acoustic features alone. Fortieth-order frequency-warped LSFs [5] and an extra gain dimension were derived from the spectral envelope provided by STRAIGHT [36] analysis. The frame shift was set to 5 ms. A five-state, left-to-right HMM structure with no skips was adopted to train context-dependent phone models, whose covariance matrices were set to be diagonal. The HTS [37] toolkits were used to train the system.

Three systems integrating articulatory features were constructed, following the *Shared Clustering*, *State-Synchrony*, and *Feature-Dependency* modeling methods discussed above. In the *Feature-Dependency* system, $\boldsymbol{A}_j$ is defined as a three-block matrix corresponding to static, velocity and acceleration components of the feature vector in order to reduce the number of parameters that need to be estimated. As discussed in Section II-E, all state-dependent transform matrices $\boldsymbol{A}_j$ were tied to a given class. The optimal number of classes M to use was determined using the following two criteria.

1) *Maximum-likelihood criterion*. The optimal number of transforms is determined as that which maximizes the

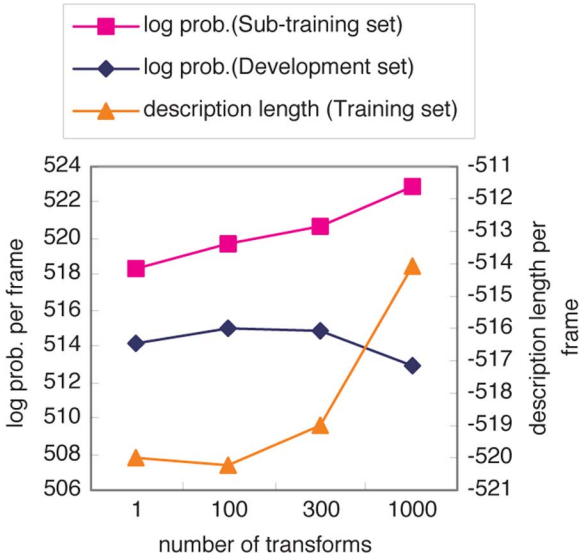


Fig. 7. Effect of varying the number of transforms in the *Feature-Dependency* system.

likelihood function $P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda)$ on a development set. We further subdivided the training set into what we will term a "sub-training set" and a development set that contained 63 sentences selected randomly. Four systems were trained on the subtraining set using $M = 1, 100, 300, 1000$, respectively. The average log probability per frame on the subtraining and development sets for different transform numbers was calculated, and these results are shown in Fig. 7.

2) *Minimum description length criterion*. The optimal transform number is determined so as to minimize the description length of the model with respect to the training set. The definition of description length here is similar to (5), except that $\log P(\boldsymbol{X} \mid \lambda)$ is replaced by $\log P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda)$ and $D(\lambda) = 3MD_{\boldsymbol{X}}D_{\boldsymbol{Y}} + C_D$, considering the three-block matrix structure of $\boldsymbol{A}_j$, where C_D is a constant that is independent from the number of transforms M. The description length per frame on the training set for $M = 1, 100, 300, 1000$ is also shown in Fig. 7.

In Fig. 7, we see that $M = 100$ leads to the best performance among the four configurations according to both criteria. That is, the *Feature-Dependency* system with 100 tied transform matrices results in the maximum probability for the development set and the minimum description length on the training set. Consequently, we used 100 transforms in the remainder of our experiments. The band width L for matrix $\boldsymbol{Z}^{-1}$ in (44) was set to 50.

C. Accuracy of Acoustic Parameter Prediction

As discussed above, various metrics for computing the distance between synthesized and natural acoustic features can be used as an objective measure to evaluate the naturalness of synthetic speech. Here, we use the root mean square error (RMSE) of the generated LSF feature sequences compared with the natural ones for the sentences in the test set to measure the accuracy of acoustic parameter prediction. The calculation for two LSF sequences $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_N]$ and $\tilde{\boldsymbol{L}} = [\tilde{\boldsymbol{l}}_1, \tilde{\boldsymbol{l}}_2, \ldots, \tilde{\boldsymbol{l}}_N]$ is defined as

$$\mathrm{RMSE}(\boldsymbol{L}, \tilde{\boldsymbol{L}}) = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \mathrm{Err}_{\mathrm{LSF}}^2(\boldsymbol{l}_t, \tilde{\boldsymbol{l}}_t)} \tag{49}$$

$$\mathrm{Err}_{\mathrm{LSF}}(\boldsymbol{l}_t, \tilde{\boldsymbol{l}}_t) = \sqrt{\sum_{d=1}^{D} w_{td}(l_{td} - \tilde{l}_{td})^2} \tag{50}$$

$$\boldsymbol{l}_t = [l_{t1}, l_{t2}, \ldots, l_{tD}]^{\mathsf{T}}, \tilde{\boldsymbol{l}}_t = [\tilde{l}_{t1}, \tilde{l}_{t2}, \ldots, \tilde{l}_{tD}]^{\mathsf{T}} \tag{51}$$

$$w_{td} = \begin{cases} 1/(l_{t2} - l_{t1}), & d = 1 \\ 1/\min(l_{td+1} - l_{td}, l_{td} - l_{td-1}), & d = 2 \ldots D-1 \\ 1/(l_{tD} - l_{tD-1}), & d = D. \end{cases} \tag{52}$$

where N is the sequence length, D is the dimensionality of the LSF vector for each time frame, and the function $\mathrm{Err}_{\mathrm{LSF}}(\cdot)$ defines the distance between two LSF vectors. Similar to the definition of quantization error in some speech coding algorithms [38], a Euclidean distance with perceptual weighting is used to emphasize the difference in frequency bands where two LSFs of adjacent order are close to each other, which corresponds to a peak in the spectral envelope. Finally, to simplify the calculation of RMSE in the following experiments, the LSFs were generated using state durations derived from state alignment performed on the natural speech.

Fig. 8 shows the objective evaluation results of predicted LSFs for the *Baseline*, *Shared Clustering*, and *State-Synchrony* systems. A t-test informs us that the differences between each two of these three systems are significant ($p < 0.05$). From this figure, we see that shared clustering improves the accuracy of LSF prediction. Table II lists the number of leaf nodes in the LSF decision tree in the three systems. We find that after integrating EMA features, shared model clustering generates a larger decision tree than the *Baseline* system under the same MDL criterion. This is an interesting result; as mentioned in Section II-C, the MDL criterion for the shared clustering has a larger dimensional penalty $D(\lambda)$ than for the separate model clustering. A larger penalty tends to reduce the number of leaf nodes in the decision tree. However, adding articulatory features has resulted in the opposite occurring. This implies the articulatory features discriminate more, in terms of variation of pronunciation, than the acoustic features. In other words, when building the decision tree, a given linguistic context feature may serve to split a cluster of models into distinct subgroups in terms of their articulatory parameterization, whereas in terms of their acoustic parameterization they might constitute only a single, homogeneous cluster. This may be explained by the nature of the EMA features, i.e., that they are more directly related to the speech production system than the corresponding acoustic features, and thus can provide supplementary information pertaining to context-dependence. Therefore, as our results show, shared clustering helps achieve a more reasonable model tying topology for the acoustic features compared with that of the *Baseline* system.

Meanwhile, comparing the *Shared Clustering* system with the *State-Synchrony* system in Fig. 8, we find that imposing the constraint of synchronous state alignment makes the prediction

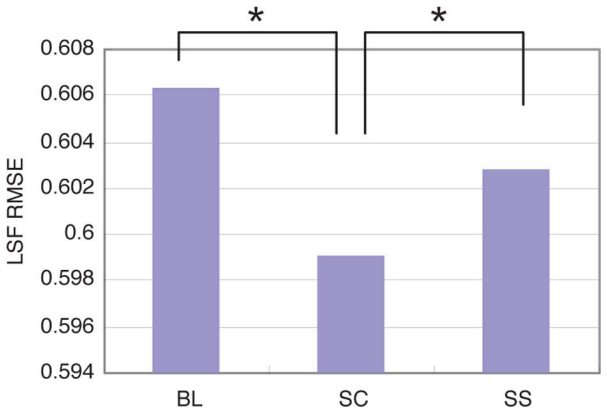


Fig. 8. Objective evaluation of LSF RMSE on *Baseline* ("BL"), *Shared Clustering* ("SC"), and *State-Synchrony* ("SS") systems. The definition of each system can be found in Table I. "$*$" indicates the difference between two systems is significant.

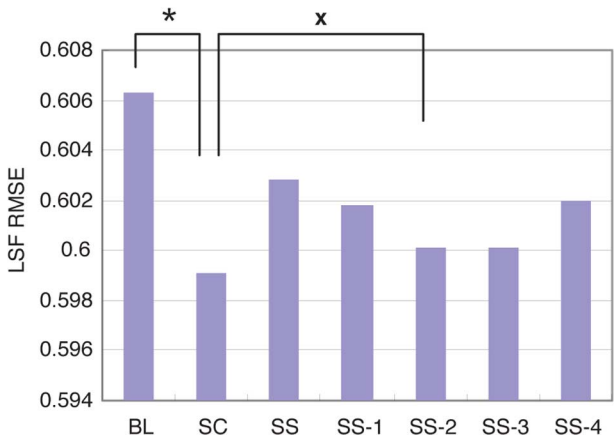


Fig. 9. Objective evaluation of LSF RMSE for *State-Synchrony* system with varying frame delay of articulatory features ("SS-1" to "SS-4"). The definition of each system can be found in Table I. "$*$" indicates the difference between two systems is significant and " x " indicates the difference is insignificant.

TABLE II
LSF DECISION TREE SIZE OF DIFFERENT SYSTEMS

System	Leaf Node Number
Baseline	2222
Shared Clustering	3481
State-Synchrony	3572

of LSF features worse. This is reasonable, since we expect a time delay between the movement of the articulators and the capturing of the corresponding generated speech waveform by the microphone. From this point of view, acoustic and articulatory features are asynchronous. An experiment was carried out to explore whether or not this asynchrony could be alleviated by a constant frame delay of EMA features in the *State-Synchrony* system. Fig. 9 shows the RMSE of predicted LSFs with a time delay of EMA features between one and four frames. As this figure shows, the optimal delay of EMA features is between two and three frames, which is consistent with the findings of previous related research [39], [40]. The best result of a *State-Synchrony* system with a constant EMA feature delay still cannot outperform the *Shared Clustering* system. This means the asynchrony between LSF and EMA features may not be entirely constant, but context-dependent. However, a t-test indicates that the difference between the *Shared Clustering* system and the *State-Synchrony* system with two-frame delay is not significant ($p = 0.36 > 0.05$). Therefore, the *State-Synchrony* system with two-frame delay is used as the initial model in the *Feature-Dependency* system.

Fig. 10 shows the evaluation results for the *Feature-Dependency* system. Two methods for acoustic parameter generation are tested. In this figure, we see that the accuracy of LSF prediction can be improved significantly by dependent-feature modeling when natural EMA features are provided ($p = 0.00 < 0.05$ between "SS-2" and "FD-N"). Unfortunately, dependent-feature modeling cannot improve the accuracy of LSF prediction if the natural EMA features are not given ($p = 0.92 > 0.05$ between "SS-2" and "FD"). This indicates the generated EMA features are not precise enough, compared with the natural ones. Thus, we make two observations on the basis of our results. On

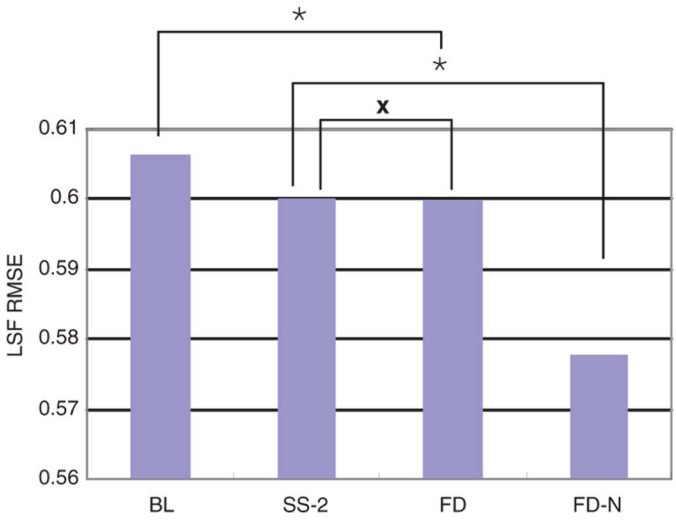


Fig. 10. Objective evaluation of LSF RMSE for *Feature-Dependency* system without natural EMA features in LSF generation ("FD") and with natural EMA features in LSF generation ("FD-N"). The definition of each system can be found in Table I. "$*$" indicates the difference between two systems is significant and "x" indicates the difference is insignificant.

one hand, dependent-feature modeling can describe the relationship between acoustic and articulatory features more reasonably and accurately. If one of them is given, we can generate the other feature more accurately. On the other hand, however, such a method does not help to predict both sets of features simultaneously.

D. Subjective Evaluation on Naturalness of Synthetic Speech

We conducted three groups of forced-choice listening tests to compare performance in terms of naturalness between 1) the *Baseline* and *Shared-Clustering* systems ("BL" versus "SC") 2) the *Baseline* and *Feature-Dependent* systems ("BL" versus "FD"), and 3) the *Feature-Dependent* systems with and without natural EMA features during LSF generation ("FD" versus "FD-N").

Twenty sentences were selected from the test set and synthesized by both systems in each test group. Each of these pairs of

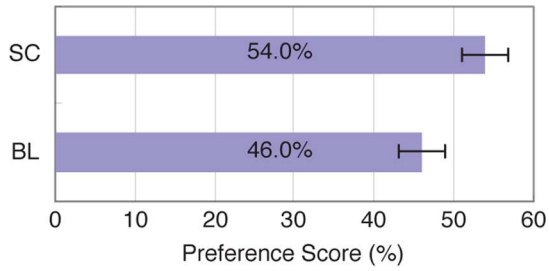


Fig. 11. Listener preference scores in forced choice between *Baseline* ("BL") and *Shared Clustering* ("SC") systems.

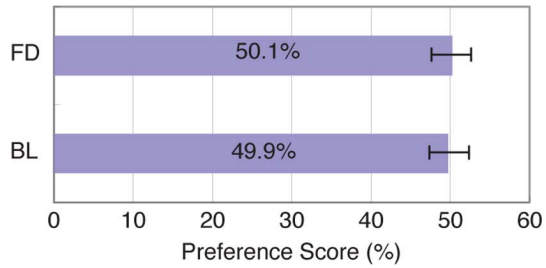


Fig. 12. Listener preference scores in forced choice between *Baseline* system ("BL") and *Feature-Dependency* system without natural EMA features in LSF generation ("FD").

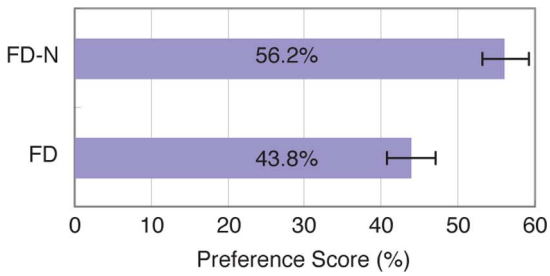


Fig. 13. Listener preference scores in forced choice between *Feature-Dependency* system without natural EMA features in LSF generation ("FD") and with natural EMA features in LSF generation ("FD-N").

synthetic sentences were evaluated by 40 listeners. Each pair of utterances was presented in both orders, making a total of 40 paired stimuli, and the overall order in which these pairs were presented to the subjects was randomized. The listeners were asked to identify which sentence in each pair sounded more natural. We then calculated the preference score of each listener for the two systems in each group. Figs. 11–13 show the average preference scores of all listeners with a 95% confidence interval for the three groups of tests.

In Fig. 11, we see a significant improvement when a shared decision tree is employed for model clustering after integrating articulatory features. This is consistent with the objective evaluation results for "BL" and "SC" in Fig. 8. Meanwhile, Fig. 12 shows that there is no significant difference in subjective preference between the *Baseline* system the *Feature-Dependency* system without natural EMA features in LSF generation. This means the "FD" system does not improve the naturalness of synthetic speech to the same extent as the "SC" system. Importantly, though, we equally find that synthetic speech quality is not degraded by the introduction of dependency of acoustic features on the articulatory features. In Figs. 9 and 10, we see that the

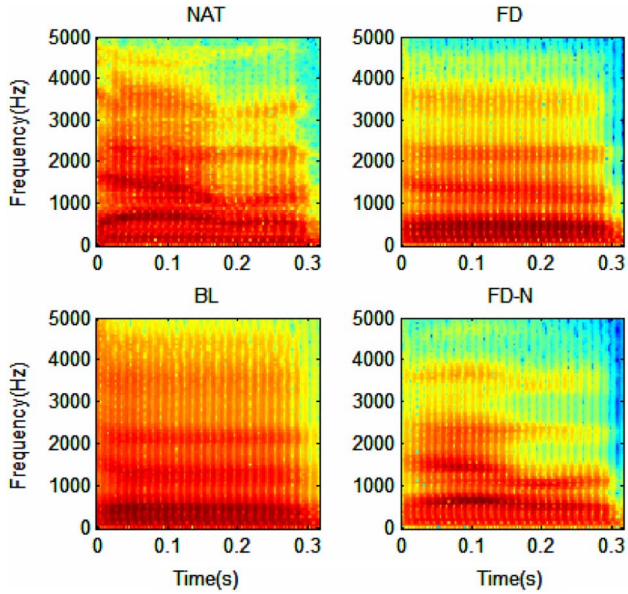


Fig. 14. Spectrograms for word "*dour*" from natural recording ("NAT") and speech synthesized by *Baseline* ("BL") and *Feature-Dependency* systems with and without natural EMA features during LSF generation ("FD-N" and "FD," respectively).

"FD" system cannot outperform the "SC" system in objective evaluation, and so the objective and subjective evaluation results are again consistent.

One inconsistency between the objective and subjective evaluation results is that the improvement of the "FD" system over the "BL" system is significant in terms of LSF RMSE but insignificant in the listening test. Note in Fig. 13, however, that once the natural EMA features are provided, the subjective evaluation results show the performance of the *Feature-Dependency* system can be improved significantly.

E. Articulatorily Controllable Acoustic Parameter Generation

In the *Feature-Dependency* system, the generation of acoustic features is determined not only by the acoustic models corresponding to the contextual information, but also by the concurrent articulatory features. This provides the possibility to control the generation of acoustic features by manipulating the articulatory features. Fig. 14 shows an example which demonstrates how articulatory features can affect the generation of acoustic features in addition to the effect of linguistic context information alone. This example shows the word "*dour*," which appears in the test set. This word is transcribed in the lexicon as /d uː r/.[2] However, during recording the speaker pronounced the word as /d aʊ ə/, resulting in a labelling mismatch. We can clearly see the effect exerted by the articulatory features by comparing the spectrograms of two variants of the *Feature-Dependency* system in Fig. 14. In one case, we have synthesized the word "dour" using our standard *Feature-Dependency* system ("FD"), whereas in the other case, we have applied the natural EMA features during parameter generation ("FD-N"). Notice that the spectrogram for the "FD" system is very similar to that produced by the *Baseline* system ("BL"). Subjectively, the pronunciation

[2]All phonetic symbols in this paper are in International Phonetic Alphabet (IPA) format.

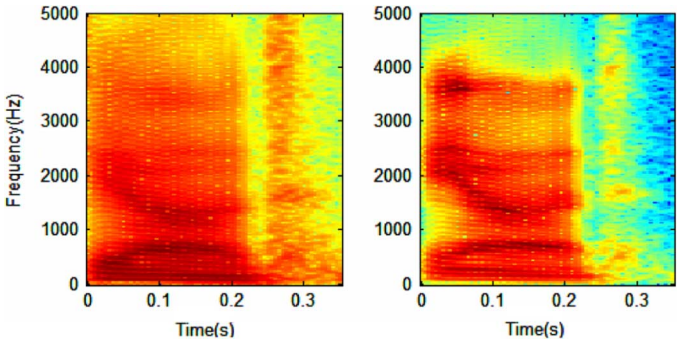


Fig. 15. Spectrograms for synthesized word "*yard*" using *Feature-Dependency* system without modification (left) and with a 1.5 scaling factor for the z-coordinates of all EMA receivers (right).

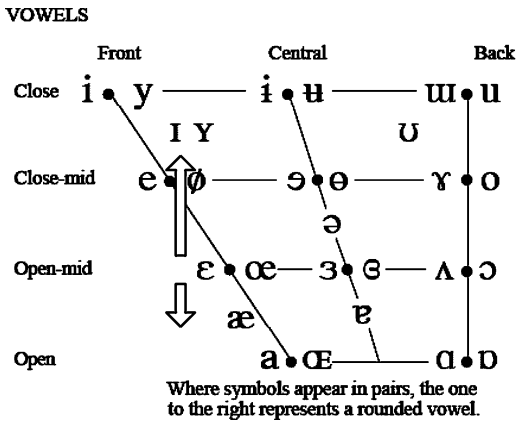


Fig. 16. IPA vowel chart. The arrows show the direction of vowel quality modification in our experiment.

for both these is the same as the lexicon entry: /d uː r/. However, the spectrogram for the "FD-N" system is far more similar to that of the natural recording ("NAT"), and the pronunciation for both of these is perceived as /d aʊ ə/. Since exactly the same context information and models are used for variants "FD" and "FD-N," it is clearly the use of the different EMA features in (42) and (48) that results in the differences we observe and hear. This effect is directly relevant to some of the potential applications we outlined in Section I, where natural articulatory features would be available at synthesis time, such as when using speech synthesis to assist speech communication in noisy or silent environments. More generally, however, this example demonstrates that the synthesized acoustic signal can be strongly affected by changing the underlying articulatory features. Consequently, we can achieve articulatory control over the synthesizer by modifying the generated articulatory features during acoustic parameter generation. Specifically, we can rewrite (42) as

$$\boldsymbol{X}_S^* = \left(\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}\boldsymbol{W}_{\boldsymbol{X}}\right)^{-1}\boldsymbol{W}_{\boldsymbol{X}}^{\top}\boldsymbol{U}_{\boldsymbol{X}}^{-1}(\boldsymbol{M}_{\boldsymbol{X}} + \boldsymbol{A}\boldsymbol{W}_{\boldsymbol{Y}} \cdot f(\boldsymbol{Y}_S^*)) \tag{53}$$

where $f(\cdot)$ is a modification function for the articulatory features $\boldsymbol{Y}_S^*$. Because articulatory parameters have a more straightforward, physiological meaning, it is much easier to control them than to directly control acoustic features in order to achieve desired modifications. Consequently, this makes the speech synthesis system more flexible. We should stress that in (53) the articulatory features $\boldsymbol{Y}_S^*$ are also generated. This means no input of natural articulatory features is required to carry out this modification, and so the modification can be performed for arbitrary novel synthetic utterances. In the following experiments, we will examine the effectiveness of this method in changing the overall character of synthesized speech and controlling the quality of a specific vowel.

Fig. 15 shows an example of globally modifying speech characteristics, where we increase the z-coordinates of EMA receivers to simulate a speaking style with a larger mouth opening and more effort. After modification, the formants become more pronounced and more easily distinguishable. We expect this modification could make the synthetic speech less muffled and more intelligible, especially in noisy conditions. A type-in listening test was carried out to investigate this. 100 semantically unpredictable sentences (SUS) were synthesized using the *Feature-Dependency* systems without modification and with a 1.2 scaling factor for the z-coordinates of all EMA receivers. To this, we then added babble noise, prerecorded in a dining hall, at 5-dB speech-to-noise ratio (SNR). Twenty-five native English listeners participated in the test. Each listener was presented with 12 sentences selected randomly and was asked to write down the words they heard. Finally, we calculated word error rate (WER) on all listeners for each system. The results show that the WER drops from 52% to 45% after this modification.

We carried out a further experiment in order to demonstrate the feasibility of controlling vowel quality by manipulating articulatory features in accordance with some phonetic motivation. We chose three front vowels /ɪ/, /ɛ/, and /æ/ in English for this experiment, as shown in Fig. 16.[3] The most significant difference in pronunciation between these three vowels is in tongue height. /ɪ/ has the highest position, /ɛ/ has the middle one, and /æ/ has the lowest position. In this experiment, $f(\cdot)$ is defined so as to modify the z-coordinates of EMA receivers T1, T2, and T3. Specifically, a positive (shift) modification means to raise the tongue and a negative value equates to lowering the tongue. Here, we neglect the naturally occurring differences of jaw position among these three vowels because a speaker can equally and easily pronounce them with a fixed jaw position. Five monosyllabic words ("*bet*," "*hem*," "*led*," "*peck*," and "*set*") with vowel /E/ were selected and embedded into the carrier sentence "Now we'll say ... again." In order to evaluate the effect of varying the extent of parameter tying for the transform matrix $\boldsymbol{A}_j$, three *Feature-Dependency* systems were built and tested. The first of these used a single global tied transform, the second used 100 transform classes, and the third used 3548 tied transform classes. We use the abbreviations"FD-1," "FD-100," and "FD-3548" to represent these three systems, respectively. The "FD-100" system was the same "FD" system used in previous experiments. For the "FD-3548" system, the number of transform matrices was set to the number of leaf nodes in the shared decision tree for the state pdf of acoustic and articulatory features. The modification distance was varied from -1.5 cm to 1.5

[3]The vowel chart of IPA is cited from "IPA Homepage" (http://www.arts.gla.ac.uk/IPA/index.html).

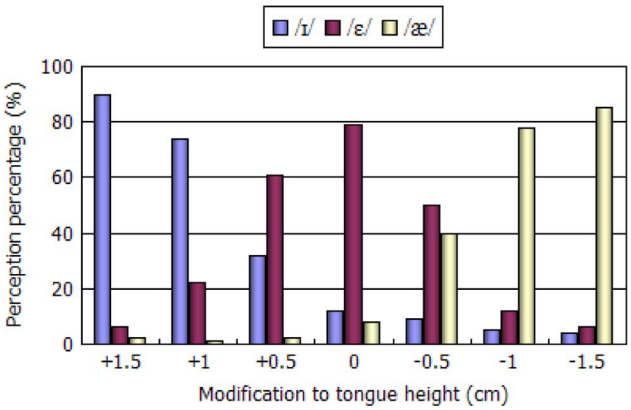


Fig. 17. Vowel quality perception after modifying the tongue height of EMA features when synthesising vowel /ɛ/ using the *Feature-Dependency* system with 100 tied transform classes.

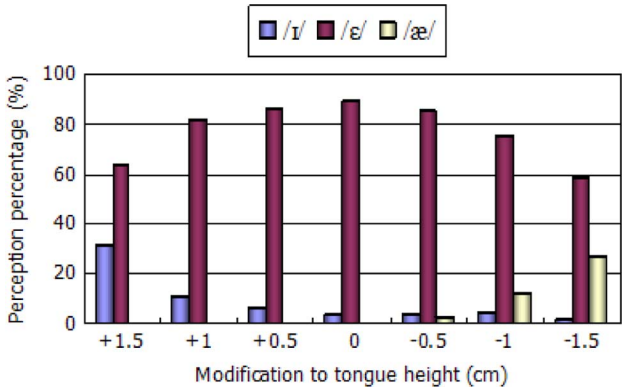


Fig. 18. Vowel quality perception after modifying the tongue height of EMA features when synthesizing vowel /ɛ/ using the *Feature-Dependency* system with a single global transform.

cm in 0.5-cm increments. Therefore, altogether we synthesized 35 samples using (53) for each system.[4]

When synthesizing using the "FD-3548" system, we found that the filters specified by the generated LSF parameters tended to be unstable, even after only a small modification of 0.5 cm for example. As a result, the quality of synthetic speech tended to be seriously degraded. This can be attributed to over-fitting in the models trained when a large number of transform matrices is used. Consequently, only the "FD-1" and "FD-100" systems were evaluated in the listening test, which we describe next.

Twenty listeners were asked to listen to the synthesized samples from each system and write down the key word in the carrier sentence they heard. Then, for each modification distance we calculated the percentages for how these three vowels were perceived.

The listening test results for the "FD-100" system in Fig. 17 clearly show the transition of vowel perception from /ɛ/ to /ɪ/ where we simulate raising the tongue by increasing the z-coordinates of EMA receivers T1, T2, and T3 in the modification function $f(\cdot)$. Conversely, we see a clear shift in vowel perception from /ɛ/ to /æ/ when simulating lowering the tongue. Meanwhile, the articulatory controllability of the "FD-1" system, shown in Fig. 18, is far more limited. There is no clear transition between vowels even after a modification of 1.5 cm. This experiment demonstrates that by using regression classes and selecting a suitable class number to model the articulatory-acoustic relationship throughout different linguistic contexts, we can achieve a balance between avoiding over-fitting to the training data and gaining effective articulatory control over the generated acoustic features.

Fig. 19 shows spectrograms for the synthesized variants of the word "*set*" which were generated by the "FD-100" system and used in the subjective evaluation. Spectrograms of the synthesized words "*sit*" and "*sat*" are also presented for comparison. Comparing these spectrograms, we notice that increasing the EMA features corresponding to the height of the tongue decreases the first formant and increases the second formant of the

[4]The speech samples used in this experiment can be found at http://www.cstr.ed.ac.uk/research/projects/artsyn/art_hmm/.

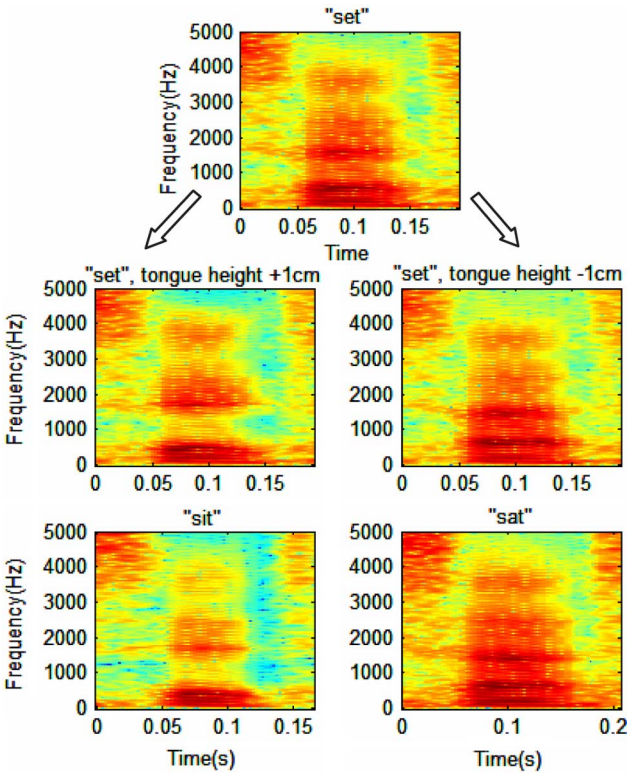


Fig. 19. Spectrograms of synthesized speech using the *Feature-Dependency* system with 100 transform classes (Top: word "*set*"; middle left: word "*set*," with the z-coordinates of T1, T2, and T3 increased by 1 cm; middle right: word "*set*," with the z-coordinates of T1, T2, and T3 decreased by 1 cm; bottom left: word "*sit*"; bottom right: word "*sat*").

/ɛ/ vowel, thus making it similar to /ɪ/. Conversely, lowering the tongue increases the first formant and decreases the second formant of the /ɛ/ vowel, which makes it similar to /æ/. This potential for modification can be employed to synthesize speech with different accents by using one unified model and specific phonetic rules which prescribe articulator movements. It is worth stressing again that this ability does not require any speech data for the target variation, in contrast to model adaptation and interpolation techniques.

IV. Conclusion

We have proposed a method for integrating articulatory features into an HMM-based parametric speech synthesis system. Three factors that influence the model structure have been explored in this paper: model clustering, synchronous-state modeling, and dependent-feature modeling. Our evaluation results have shown that the accuracy of acoustic parameter prediction, and the naturalness of synthesized speech which is correlated with this, can be improved significantly by modeling acoustic and articulatory features together in a shared-clustering and asynchronous-state system. Although dependent-feature modeling does not improve the accuracy of acoustic parameter generation unless the natural articulatory features are used, it in no way degrades speech quality in the absence of natural EMA features either. Moreover, we have clearly demonstrated that the parameter generation process becomes more flexible through the introduction of articulatory control. This offers the potential to manipulate both the global characteristics of the synthetic speech as well as the quality of specific phones, such as vowels. Importantly, this requires no additional natural articulatory data, and thus the technique can be employed to synthesize arbitrary novel utterances.

Finally, the experiments reported in this paper have shown that the naturalness of the *Shared Clustering* system is better than that of the *Feature-Dependency* system, but that the *Feature-Dependency* system can provide better flexibility for acoustic parameter generation. It is conceivable that a system using shared clustering, an asynchronous state sequence and a dependent-feature model structure [as shown in Fig. 4(b)], may combine all advantages. Our future work will include the implementation and evaluation of such a model structure.

Acknowledgment

The authors would like to thank Prof. P. Hoole of Ludwig-Maximilian University, Munich, for his great effort in helping record the EMA data and O. Watts of CSTR, University of Edinburgh, for proofreading the manuscript and making numerous helpful suggestions. The authors would also like to thank the associate editor and the anonymous reviewers for their insightful and helpful comments.

References

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

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, vol. 3, pp. 1315–1318.

[3] K. Tokuda, H. Zen, and A. W. Black, "HMM-based approach to multilingual speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004.

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

[5] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for Blizzard Challenge 2006: An improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

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

[7] 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.

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

[9] 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.

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

[11] S. Kiritani, "X-ray microbeam method for the measurement of articulatory dynamics: Technique and results," *Speech Commun.*, vol. 45, pp. 119–140, 1986.

[12] 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.

[13] T. Baer, J. C. Gore, S. Boyce, and P. W. Nye, "Application of MRI to the analysis of speech production," *Magn. Resonance Imag.*, vol. 5, pp. 1–7, 1987.

[14] Y. Akgul, C. Kambhamettu, and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," *Proc. IEEE Comput. Vis. Pattern Recog.*, vol. 124, pp. 298–303, 1998.

[15] Q. Summerfield, "Some preliminaries to a comprehensive account of audio visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds. Mahwah, NJ: Lawrence Erlbaum, 1987, pp. 3–51.

[16] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Commun.*, vol. 48, no. 2, pp. 161–175, 2006.

[17] K. Kirchhoff, G. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," in *Proc. ICASSP*, 2000, pp. 1435–1438.

[18] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 723–742, 2007.

[19] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, Mar. 2004.

[20] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda, "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *Proc. ICASSP*, 2006, pp. 93–96.

[21] T. P. Barnwell III, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. ICASSP*, 1980, pp. 706–709.

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

[23] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, 1999, pp. 229–232.

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

[25] 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. Audio Speech Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[26] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. Syst.*, vol. E86-D, no. 3, pp. 534–542, 2003.

[27] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164–171, 1970.

[28] Q. Cetin and M. Ostendorf, "Cross-stream observation dependencies for multi-stream speech recognition," in *Proc. Eurospeech*, 2003, pp. 2517–2520.

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

[30] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Proc. NOLISP*, 2007, pp. 263–272.

[31] T. Toda, W. A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, pp. 215–227, 2008.
[32] Y. Wu, "Research on HMM-Based Speech Synthesis," Ph.D. dissertation, Univ. of Sci. and Tech. of China, Hefei, 2006.
[33] S. Fitt and S. Isard, "Synthesis of regional english using a keyword lexicon," in *Proc. Eurospeech*, 1999, vol. 2, pp. 823–826.
[34] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system," in *Proc. 3rd ESCA Workshop Speech Synth.*, 1998, pp. 147–151.
[35] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*. Cambridge, U.K.: Cambridge University Engineering Department, 2002.
[36] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using 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.
[37] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop Speech Synth.*, 2007, pp. 294–299.
[38] *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, ITU-T Rec. G.723.1, 1996.
[39] C. Qin and M. A. Carreira-Perpiñán, "A comparison of acoustic features for articulatory inversion," in *Proc. Interspeech*, 2007, pp. 2469–2472.
[40] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. E. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Amer.*, vol. 100, no. 3, pp. 1819–1834, 1996.

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

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

Korin Richmond received the undergraduate M.A. degree in linguistics and Russian from the University of Edinburgh, Edinburgh, U.K., in 1995, the M.Sc. degree in cognitive science and natural language processing from the University of Edinburgh in 1997, and the Ph.D. degree from the Center for Speech Technology Research (CSTR), the University of Edinburgh in 2002 for a thesis titled "Estimating articulatory parameters from the acoustic speech signal," which clearly showed the advantage of utilizing a flexible probabilistic machine-learning framework in conjunction with corpora of acoustic–articulatory data for performing the inversion mapping.

He has been involved with human language and speech technology since 1991. He has worked as a Research Fellow at CSTR since 2000. Among other things, this work has included implementing a state-of-the-art unit selection synthesis module for CSTR's Festival speech synthesis system, called MultiSyn, which was included in the latest release of Festival. In addition to MultiSyn, he has also contributed as a core developer to the maintenance and further development of Festival and CSTR's Edinburgh Speech Tools C/C++ library since 2002. His research interests include speech synthesis and data-driven acoustic-articulatory modeling.

Dr. Richmond is a member of ISCA.

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

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

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

Ren-Hua Wang was born in Shanghai in August 1943.

Now he is a Professor and Ph.D. Supervisor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. His research interests include speech coding, speech synthesis and recognition, and multimedia communication. During the past 20 years, he was in charge of more than ten national key research projects in the information field.

Prof. Wang received the 2002 Second Class National Award for Science and Technology Progress, China, and 2005 Information Industries Significant Technology Award for Invention, China.