Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis

Zhen-Hua Ling, *Member, IEEE*, Li Deng, *Fellow, IEEE*, and Dong Yu, *Senior Member, IEEE*

***Abstract*—This paper presents a new spectral modeling method for statistical parametric speech synthesis. In the conventional methods, high-level spectral parameters, such as mel-cepstra or line spectral pairs, are adopted as the features for hidden Markov model (HMM)-based parametric speech synthesis. Our proposed method described in this paper improves the conventional method in two ways. First, distributions of low-level, un-transformed spectral envelopes (extracted by the STRAIGHT vocoder) are used as the parameters for synthesis. Second, instead of using single Gaussian distribution, we adopt the graphical models with multiple hidden variables, including restricted Boltzmann machines (RBM) and deep belief networks (DBN), to represent the distribution of the low-level spectral envelopes at each HMM state. At the synthesis time, the spectral envelopes are predicted from the RBM-HMMs or the DBN-HMMs of the input sentence following the maximum output probability parameter generation criterion with the constraints of the dynamic features. A Gaussian approximation is applied to the marginal distribution of the visible stochastic variables in the RBM or DBN at each HMM state in order to achieve a closed-form solution to the parameter generation problem. Our experimental results show that both RBM-HMM and DBN-HMM are able to generate spectral envelope parameter sequences better than the conventional Gaussian-HMM with superior generalization capabilities and that DBN-HMM and RBM-HMM perform similarly due possibly to the use of Gaussian approximation. As a result, our proposed method can significantly alleviate the over-smoothing effect and improve the naturalness of the conventional HMM-based speech synthesis system using mel-cepstra.**

***Index Terms*—Deep belief network, hidden Markov model, restricted Boltzmann machine, spectral envelope, speech synthesis.**

I. Introduction

THE hidden Markov model (HMM)-based parametric speech synthesis method has become a mainstream speech synthesis method in recent years [2], [3]. In this method, the spectrum, F0 and segment durations are modeled simultaneously within a unified HMM framework [2]. At synthesis time, these parameters are predicted so as to maximize their output probabilities from the HMM of the input sentence. The constraints of the dynamic features are considered during parameter generation in order to guarantee the smoothness of the generated spectral and F0 trajectories [4]. Finally, the predicted parameters are sent to a speech synthesizer to reconstruct the speech waveforms. This method is able to synthesize highly intelligible and smooth speech sounds [5], [6]. However, the quality of the synthetic speech is degraded due to three main factors: limitations of the parametric synthesizer itself, inadequacy of acoustic modeling used in the synthesizer, and the over-smoothing effect of parameter generation [7].

Many improved approaches have been proposed to overcome the disadvantages of these three factors. In terms of the speech synthesizer, STRAIGHT [8], as a high-performance speech vocoder, has been widely used in current HMM-based speech synthesis systems. It follows the source-filter model of speech production. In order to represent the excitation and vocal tract characteristics separately, F0 and a smooth spectral envelope without periodicity interference are extracted at each frame. Then, mel-cepstra [5] or line spectral pairs [6] can be derived from the spectral envelopes of training data for the following HMM modeling. During synthesis, the generated spectral parameters are used either to reconstruct speech waveforms directly or to recover the spectral envelopes for further speech reconstruction by STRAIGHT.

Acoustic modeling is another key component of the HMM-based parametric speech synthesis. In the common spectral modeling methods, the probability density functions (PDF) of each HMM state is represented by a single Gaussian distribution with diagonal covariance matrix and the distribution parameters are estimated under the maximum likelihood (ML) criterion [2]. Because the single Gaussian distributions are used as the state PDFs, the outputs of maximum output probability parameter generation tend to distribute near the modes (also the means) of the Gaussians, which are estimated by averaging observations with similar context descriptions in the ML training. Although this averaging process improves the robustness of parameter generation, the detailed characteristics of the spectral parameters are lost. Therefore, the reconstructed spectral envelopes are over-smoothed, which leads to a muffled voice quality in the synthetic speech. The existing refinements on acoustic modeling include increasing the number of Gaussians for each HMM state [4], reformulating HMM as a trajectory model [9], improving the model training criterion by minimizing the generation error [10].

Manuscript received January 25, 2013; revised April 18, 2013; accepted June 13, 2013. Date of publication June 18, 2013; date of current version July 22, 2013. This work was supported in part by the National Nature Science Foundation of China under Grant No.61273032 and in part by the China Scholarship Council Young Teacher Study Abroad Project. This paper is the expanded version of the conference paper published in ICASSP-2013 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

Z.-H. Ling is with the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: zhling@ustc.edu.cn).

L. Deng and D. Yu are with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; dongyu@microsoft.com).

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.2013.2269291

In order to alleviate the over-smoothing effect, many improved parameter generation methods have also been proposed, such as modifying the parameter generation criterion by integrating a global variance model [11] or minimizing model divergences [12], post-filtering after parameter generation [6], [13], using real speech parameters or segments to generate the speech waveform [14], [15], or sampling trajectories from the predictive distribution [16], [17], and so on. In this paper, we propose a new spectral modeling method which copes with the first two factors mentioned above. First, the raw spectral envelopes extracted by the STRAIGHT vocoder are utilized directly without further deriving spectral parameters from them during feature extraction. Comparing with the high-level[1] spectral parameters, such as mel-cepstra or line spectral pairs, the low-level spectral envelopes are more physically meaningful and more directly related with the subjective perception on the speech quality. Thus, the influence of spectral parameter extraction on the spectral modeling can be avoided. Similar approach can be found in [18], where the spectral envelopes derived from the harmonic amplitudes are adopted to replace the mel-cepstra for HMM-based Arabic speech synthesis and the naturalness improvement can be achieved. Second, the graphical models with multiple hidden layers, such as restricted Boltzmann machines (RBM) [19] and deep belief networks (DBN) [20], are introduced to represent the distribution of the spectral envelopes at each HMM state instead of single Gaussian distribution. An RBM is a bipartite undirected graphical model with a two-layer architecture and a DBN contains more hidden layers, which can be estimated using a stack of RBMs. Both of these two models are better in describing the distribution of high-dimensional observations with cross-dimension correlations, i.e., the spectral envelopes, than the single Gaussian distribution and Gaussian mixture model (GMM). The acoustic modeling method which describes the production, perception and distribution of speech signals is always an important research topic in speech signal processing [21]. In recent years, RBMs and DBNs have been successfully applied to modeling speech signals, such as spectrogram coding [22], speech recognition [23], [24], and acoustic-articulatory inversion mapping [25], where they mainly act as the pre-training methods for a deep autoencoder or a deep neural network (DNN). The architectures used in deep learning as applied to speech processing have been motivated by the multi-layered structures in both speech production and perception involving phonological features, motor control, articulatory dynamics, and acoustic and auditory parameters [26], [27]. The approaches of applying RBMs, DBNs, and other deep learning methods to the statistical parametric speech synthesis have also been studied very recently [1], [28]–[30]. In [28], a DNN-based statistical parametric speech synthesis method is presented, which maps the input context information towards the acoustic features using a neural network with deep structures. In [29], a DNN which is pre-trained by the DBN learning is adopted as a feature extractor for the Gaussian process based F0 contour prediction. Furthermore, RBMs and DBNs can be used as density models instead of the DNN initialization methods for the speech synthesis application. In [30], a single DBN model is trained to represent the joint distribution between the tonal syllable ID and the acoustic features. In [1], a set of RBMs are estimated to describe the distributions of the spectral envelopes in the context-dependent HMM states. In this paper, we extend our previous work in [1] by incorporating the dynamic features of spectral envelopes into the RBM modeling and developing RBMs to DBNs which has more layers of hidden units.

This paper is organized as follows. In Section II, we will briefly review the basic techniques of RBMs and DBNs. In Section III, we will describe the details of our proposed method. Section IV reports our experimental results. Section V gives the conclusion and the discussion on our future work.

II. Restricted Boltzmann Machines and Deep Belief Networks

A. Restricted Boltzmann Machines

An RBM is a kind of bipartite undirected graphical model (i.e., Markov random field) which is used to describe the dependency among a set of random variables using a two-layer architecture [19]. In this model, the visible stochastic units $\mathbf{v} = [v_1, \ldots, v_V]^\top$ are connected to the hidden stochastic units $\mathbf{h} = [h_1, \ldots, h_H]^\top$ as shown in Fig. 1(a), where V and H are the numbers of units of the visible and hidden layers respectively, and $(\cdot)^\top$ means the matrix transpose. Assuming $\mathbf{v} \in \{0,1\}^V$ and $\mathbf{h} \in \{0,1\}^H$ are both binary stochastic variables, the energy function of the state $\{\mathbf{v}, \mathbf{h}\}$ is defined as

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{V} a_i v_i - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j, \quad (1)$$

where w_{ij} represents the symmetric interaction between v_i and h_j, a_i and b_j are bias terms. The model parameters are composed of $\mathbf{a} = [a_1, \ldots, a_V]^\top$, $\mathbf{b} = [b_1, \ldots, b_H]^\top$, and $\mathbf{W} = \{w_{ij}\}_{V \times H}$. The joint distribution over the visible and hidden units is defined as

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right), \quad (2)$$

where

$$\mathcal{Z} = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right) \quad (3)$$

is the partition function which can be estimated using the annealed importance sampling (AIS) method [31]. Therefore, the probability density function over the visible vector $\mathbf{v}$ can be calculated as

$$P(\mathbf{v}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}} \exp\left(-E(\mathbf{v}, \mathbf{h})\right). \quad (4)$$

Given a training set, the RBM model parameters $\{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ can be estimated by maximum likelihood learning using the contrastive divergence (CD) algorithm [32].

[1]Here, the "level" refers to the steps of signal processing procedures involved in the spectral feature extraction. The high-level spectral parameters are commonly derived from the low-level ones by functional representation and parameterization.

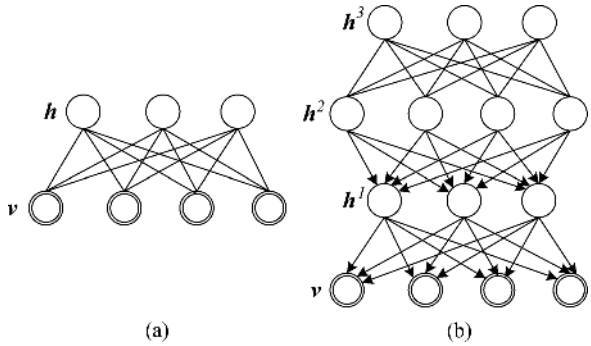

Fig. 1. The graphical model representations for (a) an RBM and (b) a three-hidden-layer DBN.

RBM can also be applied to model the distribution of real-valued data (e.g., the speech parameters) by adopting its Gaussian-Bernoulli form, which means $\mathbf{v} \in \mathcal{R}^V$ are real-valued and $\mathbf{h} \in \{0,1\}^H$ are binary. Thus, the energy function of the state $\{\mathbf{v}, \mathbf{h}\}$ is defined as

$$E(\mathbf{v},\mathbf{h}) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} h_j \frac{v_i}{\sigma_i}, \quad (5)$$

where the variance parameters σ_i^2 are commonly fixed to a predetermined value instead of learning from the training data [33].

B. Deep Belief Networks

A deep belief network (DBN) is a probabilistic generative model which is composed of many layers of hidden units [20]. The graphical model representation for a three-hidden-layer DBN is shown in Fig. 1(b). In this model, each layer captures the correlations among the activities of hidden features in the layer below. The top two layers of the DBN form an undirected bipartite graph. The lower layers form a directed graph with a top-down direction to generate the visible units. Mathematically, the joint distribution over the visible and all hidden units can be written as

$$\begin{aligned} &P(\mathbf{v},\mathbf{h}^1,\ldots,\mathbf{h}^L) \\ &\quad = P(\mathbf{v}|\mathbf{h}^1)P(\mathbf{h}^1|\mathbf{h}^2)\cdots P(\mathbf{h}^{L-2}|\mathbf{h}^{L-1})P(\mathbf{h}^{L-1},\mathbf{h}^L), \end{aligned} \quad (6)$$

where $\mathbf{h}^l = [h_1^l, \ldots, h_{H_l}^l]^\top$ is the hidden stochastic vector of the l-th hidden layer, H_l is the dimensionality of $\mathbf{h}^l$, and L is the number of hidden layers. The joint distribution $P(\mathbf{h}^{L-1}, \mathbf{h}^L)$ is represented by an RBM as (2) with the weight matrix $\mathbf{W}^L$ and the bias vectors $\mathbf{a}^L$ and $\mathbf{b}^L$. $P(\mathbf{v}|\mathbf{h}^1)$ and $P(\mathbf{h}^{l-1}|\mathbf{h}^l), l \in \{2,3,\ldots,L-1\}$ are represented by sigmoid belief networks [34]. Each sigmoid belief network is described by a weight matrix $\mathbf{W}^l$ and a bias vector $\mathbf{a}^l$. Assuming $\mathbf{v}$ are real-valued and $\mathbf{h}^l, l \in \{1,2,\ldots,L\}$ are binary, the dependency between $\mathbf{v}$ and $\mathbf{h}^1$ in the sigmoid belief network is described by

$$P(\mathbf{v}|\mathbf{h}^1) = \mathcal{N}(\mathbf{v}; \mathbf{W}^{1\top}\mathbf{h}^1 + \mathbf{a}^1, \mathbf{\Sigma}) \quad (7)$$

where $\mathcal{N}(\cdot)$ denotes a Gaussian distribution; $\mathbf{\Sigma} = diag\{\sigma_i^2\}$ and turns to an identity matrix when σ_i^2 are fixed to 1 during model training. For $l \in \{2,3,\ldots,L-1\}$, the dependency between two adjacent hidden layers is represented by

$$P(h_i^{l-1} = 1|\mathbf{h}^l) = g(a_i^l + \sum_j w_{ij}^l h_j^l) \quad (8)$$

where $g(x) = 1/(1+\exp(-x))$ is the sigmoid function. For an L-hidden-layer DBN, its model parameters are composed of $\{\mathbf{a}^1, \mathbf{W}^1, \ldots, \mathbf{a}^{L-1}, \mathbf{W}^{L-1}, \mathbf{a}^L, \mathbf{b}^L, \mathbf{W}^L\}$. Further, the marginal distribution of the visible variables for a DBN can be written as

$$P(\mathbf{v}) = \sum_{\mathbf{h}^1} \cdots \sum_{\mathbf{h}^L} P(\mathbf{v}, \mathbf{h}^1, \ldots, \mathbf{h}^L). \quad (9)$$

Given the training samples of the visible units, it is difficult to estimate the model parameters of a DBN directly under the maximum likelihood criterion due to the complex model structure with multiple hidden layers. Therefore, a greedy learning algorithm has been proposed and popularly applied to train the DBN in a layer-by-layer manner [20]. A stack of RBMs are used in this algorithm. Firstly, it estimates the parameters $\{\mathbf{a}^1, \mathbf{b}^1, \mathbf{W}^1\}$ of the first layer RBM to model the visible training data. Then, it freezes the parameters $\{\mathbf{a}^1, \mathbf{W}^1\}$ of the first layer and draws samples from $P(\mathbf{h}^1|\mathbf{v})$ to train the next layer RBM $\{\mathbf{a}^2, \mathbf{b}^2, \mathbf{W}^2\}$, where

$$P(h_j^1 = 1|\mathbf{v}) = g(b_j^1 + \sum_i w_{ij}^1 v_i). \quad (10)$$

This training procedure is conducted recursively until it reaches the top layer and gets $\{\mathbf{a}^L, \mathbf{b}^L, \mathbf{W}^L\}$. It has been proved that this greedy learning algorithm can improve the lower bound on the log-likelihood of the training samples by adding each new hidden layer [20], [31]. Once the model parameters are estimated, to calculate the log-probability that a DBN assigns to training or test data by (9) directly is also computationally intractable. A lower bound on the log-probability can be estimated by combining the AIS-based partition function estimation with approximate inference [31].

III. Spectral Envelope Modeling Using RBMs and DBNs

A. HMM-Based Parametric Speech Synthesis

At first, the conventional HMM-based parametric speech synthesis method is briefly reviewed. It consists of a training stage and a synthesis stage. During training, the F0 and spectral parameters are extracted from the waveforms contained in the training set. Then a set of context-dependent HMMs are estimated to maximize the likelihood function for the training acoustic features. Here $\mathbf{o} = [\mathbf{o}_1^\top, \mathbf{o}_2^\top, \ldots, \mathbf{o}_T^\top]^\top$ is the observation feature sequence and T is the length of the sequence. The observation feature vector $\mathbf{o}_t \in \mathcal{R}^{3D}$ for the t-th frame typically consists of static acoustic parameters $\mathbf{c}_t \in \mathcal{R}^D$ and their delta and acceleration components as

$$\mathbf{o}_t = \left[\mathbf{c}_t^\top, \triangle\mathbf{c}_t^\top, \triangle^2\mathbf{c}_t^\top\right]^\top, \quad (11)$$

where D is the dimension of the static component; the dynamic components are commonly calculated as

$$\Delta \mathbf{c}_t = 0.5\mathbf{c}_{t+1} - 0.5\mathbf{c}_{t-1} \qquad \forall t \in [2, T-1], \tag{12}$$

$$\Delta \mathbf{c}_1 = \Delta \mathbf{c}_2, \Delta \mathbf{c}_T = \Delta \mathbf{c}_{T-1} \tag{13}$$

and

$$\Delta^2 \mathbf{c}_t = \mathbf{c}_{t+1} - 2\mathbf{c}_t + \mathbf{c}_{t-1} \quad \forall t \in [2, T-1], \tag{14}$$

$$\Delta^2 \mathbf{c}_1 = \Delta^2 \mathbf{c}_2, \Delta^2 \mathbf{c}_T = \Delta^2 \mathbf{c}_{T-1}. \tag{15}$$

Therefore, the complete feature sequence $\mathbf{o}$ can be considered to be a linear transform of the static feature sequence $\mathbf{c} = [\mathbf{c}_1^\top, \mathbf{c}_2^\top, \ldots, \mathbf{c}_T^\top]^\top$ as

$$\mathbf{o} = \mathbf{M}\mathbf{c}, \tag{16}$$

where $\mathbf{M} \in \mathcal{R}^{3TD \times TD}$ is determined by the delta and acceleration calculation functions in (12)–(15)[4]. A multi-space probability distribution (MSD) [35] is applied to incorporate a distribution for F0 into the probabilistic framework of the HMM considering that F0 is only defined for voiced speech frames. In order to deal with the data-sparsity problem of the context-dependent model training with extensive context features, a decision-tree-based model clustering technique that uses a minimum description length (MDL) criterion [36] to guide the tree construction is adopted after initial training of the context-dependent HMMs. Next, a state alignment is conducted using the trained HMMs to train context-dependent state duration probabilities [2] for state duration prediction. A single-mixture Gaussian distribution is used to model the duration probability for each state. A decision-tree-based model clustering technique is similarly applied to these duration distributions.

At the synthesis stage, the maximum output probability parameter generation algorithm is used to generate acoustic parameters [4]. The result of front-end linguistic analysis on the input text is used to determine the sentence HMM λ. The state sequence $\mathbf{q} = \{q_1, q_2, \ldots, q_T\}$ is predicted using the trained state duration probabilities [2]. Then, the sequence of speech features are predicted by maximizing $P(\mathbf{o} \mid \lambda, \mathbf{q})$. Considering the constraints between static and dynamic features as in (16), the parameter generation criterion can be rewritten as

$$\mathbf{c}^* = \arg\max_{\mathbf{c}} P(\mathbf{M}\mathbf{c} \mid \lambda, \mathbf{q}), \tag{17}$$

where $\mathbf{c}^*$ is the generated static feature sequence. If the emission distribution of each HMM state is represented by a single Gaussian distribution, the closed-form solution to (17) can be derived. By setting

$$\frac{\partial P(\mathbf{M}\mathbf{c} \mid \mathbf{q}, \lambda)}{\partial \mathbf{c}} = \mathbf{0}, \tag{18}$$

we obtain

$$\mathbf{c}^* = \left(\mathbf{M}^\top \mathbf{U}_\mathbf{q}^{-1} \mathbf{M}\right)^{-1} \mathbf{M}^\top \mathbf{U}_\mathbf{q}^{-1} \mathbf{m}_\mathbf{q}, \tag{19}$$

where $\mathbf{m}_\mathbf{q} = [\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top]^\top$ and $\mathbf{U}_\mathbf{q} = \mathrm{diag}(\boldsymbol{\Sigma}_{q_1}, \ldots, \boldsymbol{\Sigma}_{q_T})$ are the mean vector and covariance matrix of the sentence as decided by the state sequence $\mathbf{q}$[4].

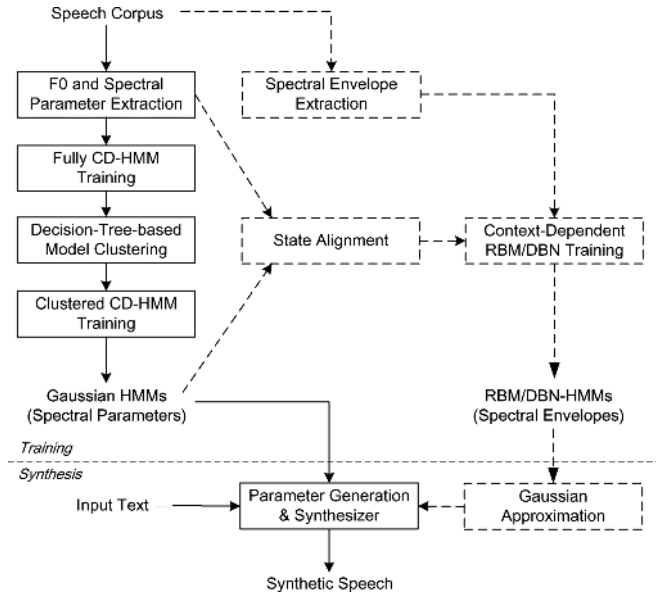

Fig. 2. Flowchart of our proposed method. The modules in solid lines represent the procedures of the conventional HMM-based speech synthesis using high-level spectral parameters, where "CD-HMM" stands for "Context-Dependent HMM." The modules in dash lines describe the add-on procedures of our proposed method for modeling the spectral envelopes using RBMs or DBNs.

B. Spectral Envelope Modeling and Generation Using RBM and DBN

In this paper, we improve the conventional spectral modeling method in the HMM-based parametric speech synthesis from two aspects. First, the raw spectral envelopes extracted by the STRAIGHT vocoder are modeled directly without further deriving high-level spectral parameters. Second, the RBM and DBN models are adopted to replace the single Gaussian distribution at each HMM state. In order to simplify the model training with high-dimensional spectral features, the decision trees for model clustering and the state alignment results are assumed to be given when the spectral envelopes are modeled. Thus, we can focus on comparing the performance of different models on the clustered model estimation. In current implementation, the conventional context-dependent model training using high-level spectral parameters and single Gaussian state PDFs is conducted at first to achieve the model clustering and state alignment results.

The flowchart of our proposed method is shown in Fig. 2. During the acoustic feature extraction using STRAIGHT vocoder, the original linear frequency spectral envelopes[2] are stored besides the spectral parameters. The context-dependent HMMs for conventional spectral parameters and F0 features are firstly estimated according to the method introduced in Section III-A. A single Gaussian distribution is used to model the spectral parameters at each HMM state. Next, a state alignment to the acoustic features is performed. The state boundaries are used to gather the spectral envelopes for each clustered context-dependent state. Similar to the high-level spectral parameters, the feature vector of the spectral envelope at each frame consists of static, velocity, and acceleration components

[2]The mel-frequency spectral envelopes can also be used here to represent the speech perception properties. In this paper, we adopt the linear frequency spectral envelope because it is the most original description of the vocal tract characters without any prior knowledge and assumption on the spectral parameterization and speech perception.

as (11)–(15). Then, an RBM or a DBN is estimated under the maximum likelihood criterion for each state according to the methods introduced in Section II. The model estimation of the RBMs or the DBNs is conducted only once using the fixed state boundaries. Finally, the context-dependent RBM-HMMs or DBN-HMMs can be constructed for modeling the spectral envelopes.

At synthesis time, the same criterion in (17) is followed to generate the spectral envelopes. The optimal sequence of spectral envelopes are estimated by maximizing the output probability from the RBM-HMM or the DBN-HMM of the input sentence. When single Gaussian distributions are adopted as the state PDFs, there is a closed-form solution as shown in (19) to this maximum output probability parameter generation with the constraints of dynamic features once the state sequence has been determined [4]. However, the marginal distribution defined in (4) for an RBM or in (9) for a DBN is much more complex than a single Gaussian, which makes the closed-form solution impractical. Therefore, a Gaussian approximation is applied before parameter generation to simply the problem. For each HMM state, a Gaussian distribution $\mathcal{N}(\mathbf{v};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is constructed, where $\mathbf{v}$ is the spectral envelope feature vector containing static, velocity, and acceleration components;

$$\boldsymbol{\mu} = \arg\max_{\mathbf{v}} \log P(\mathbf{v}) \tag{20}$$

is the mode estimated for each RBM or DBN and $P(\mathbf{v})$ is defined as (4) or (9); $\boldsymbol{\Sigma}$ is a diagonal covariance matrix estimated by calculating the sample covariances given the training samples of the state. These Gaussian distributions are used to replace the RBMs or the DBNs as the state PDFs at synthesis time. Therefore, the conventional parameter generation algorithm with the constraints of dynamic features can be followed to predict the spectral envelopes by solving a group of linear (19). By incorporating the dynamic features of the spectral envelopes during model training and parameter generation, temporally smooth spectral trajectories can be generated at synthesis time. The detailed algorithms of the mode estimation in (20) for an RBM and a DBN model will be introduced in the following subsections.

C. Estimating RBM Mode

Here, we consider the RBM of the Gaussian-Bernoulli form because the spectral envelope features are real-valued. Given the estimated model parameters $\{\mathbf{a},\mathbf{b},\mathbf{W}\}$ of an RBM, the probability density function (4) over the visible vector $\mathbf{v}$ can be further calculated as[3]

$$\begin{aligned} P(\mathbf{v}) &= \frac{1}{\mathcal{Z}}\sum_{\mathbf{h}} \exp\left(-E(\mathbf{v},\mathbf{h})\right) \\ &= \frac{1}{\mathcal{Z}}\sum_{\mathbf{h}} \exp\left(-\sum_{i=1}^{V}\frac{(v_i-a_i)^2}{2} + \mathbf{b}^\top\mathbf{h} + \mathbf{v}^\top\mathbf{W}\mathbf{h}\right) \\ &= \frac{1}{\mathcal{Z}}\exp\left(-\sum_{i=1}^{V}\frac{(v_i-a_i)^2}{2}\right) \\ &\quad \cdot \prod_{j=1}^{H}\sum_{h_j\in\{0,1\}} \exp(b_j h_j + \mathbf{v}^\top\mathbf{w}_j h_j) \\ &= \frac{1}{\mathcal{Z}}\exp\left(-\sum_{i=1}^{V}\frac{(v_i-a_i)^2}{2}\right)\prod_{j=1}^{H}\left(1+\exp(b_j+\mathbf{v}^\top\mathbf{w}_j)\right) \end{aligned} \tag{21}$$

where $\mathbf{w}_j$ denotes the j-th column of matrix $\mathbf{W}$. Because there is no closed-form solution to solve (20) for an RBM, the gradient descent algorithm is adopted here, i.e.,

$$\mathbf{v}^{(i+1)} = \mathbf{v}^{(i)} + \alpha\cdot\left.\frac{\partial \log P(\mathbf{v})}{\partial \mathbf{v}}\right|_{\mathbf{v}=\mathbf{v}^{(i)}}, \tag{22}$$

where i denotes the number of iteration; α is the step size;

$$\frac{\partial \log P(\mathbf{v})}{\partial \mathbf{v}} = -(\mathbf{v}-\mathbf{a}) + \sum_{j=1}^{H}\frac{\exp(b_j+\mathbf{v}^\top\mathbf{w}_j)}{1+\exp(b_j+\mathbf{v}^\top\mathbf{w}_j)}\mathbf{w}_j. \tag{23}$$

Thus, the estimated mode of the RBM model is determined by a non-linear transform of the model parameters $\{\mathbf{a},\mathbf{b},\mathbf{W}\}$. In contrast to the single Gaussian distribution, this mode is no longer the Gaussian mean which is estimated by averaging the corresponding training vectors under the maximum likelihood criterion.

Because the likelihood of an RBM is multimodal, the gradient descent optimization in (22) only leads to a local maximum and the result is sensitive to the initialization of $\mathbf{v}^{(0)}$. In order to find a representative $\mathbf{v}^{(0)}$, we firstly calculate the means of the conditional distributions $P(\mathbf{h}|\mathbf{v})$ for all training vectors $\mathbf{v}$. These means are averaged and made binary using a fixed threshold of 0.5 to get $\mathbf{h}^{(0)}$. Then, the initial $\mathbf{v}^{(0)}$ for the iteratively updating in (22) is set as the mean of $P(\mathbf{v}|\mathbf{h}^{(0)})$.

D. Estimating DBN Mode

Estimating the mode of a DBN model is more complex than dealing with an RBM. The marginal distribution of the visible variables in (9) can be rewritten as

$$P(\mathbf{v}) = \sum_{\mathbf{h}^1}\cdots\sum_{\mathbf{h}^L} P(\mathbf{v},\mathbf{h}^1,\ldots,\mathbf{h}^L) \tag{24}$$

$$= \sum_{\mathbf{h}^1} P(\mathbf{v}|\mathbf{h}^1)P(\mathbf{h}^1) \tag{25}$$

where $P(\mathbf{v}|\mathbf{h}^1)$ is described in (7) and $P(\mathbf{h}^1)$ can be calculated by applying

$$P(\mathbf{h}^{l-1}) = \sum_{\mathbf{h}^l} P(\mathbf{h}^{l-1}|\mathbf{h}^l)P(\mathbf{h}^l) \tag{26}$$

recursively for each l from $L-1$ to 2. The conditional distribution $P(\mathbf{h}^{l-1}|\mathbf{h}^l)$ is represented by (8) and $P(\mathbf{h}^{L-1})$ is the marginal distribution (4) of the RBM representing the top two hidden layers. Similar to the RBM mode estimation, the gradient descent algorithm can be applied here to optimize (25) once the values of $P(\mathbf{h}^1)$ are determined for all possible $\mathbf{h}^1$. However, this will lead to an exponential complexity with respect to the

[3]The variance parameters σ_i^2 in (5) are fixed to 1 to simplify the notation.

number of hidden units at each hidden layer due to the summation in (25) and (26). Thus, such optimization becomes impractical unless the number of hidden units is reasonably small.

In order to get a practical solution to the DBN mode estimation, an approximation is made to (24) in this paper. The summation over all possible values of the hidden units is simplified by considering only the optimal hidden vectors at each layer, i.e.,

$$P(\mathbf{v}) \simeq P(\mathbf{v}, \mathbf{h}^{1*}, \ldots, \mathbf{h}^{L*}), \tag{27}$$

where

$$\{\mathbf{h}^{L-1*}, \mathbf{h}^{L*}\} = \arg\max_{\{\mathbf{h}^{L-1}, \mathbf{h}^{L}\}} \log P(\mathbf{h}^{L-1}, \mathbf{h}^{L}) \tag{28}$$

and for each $l \in \{L-1, \ldots, 3, 2\}$

$$\mathbf{h}^{l-1*} = \arg\max_{\mathbf{h}^{l-1}} \log P(\mathbf{h}^{l-1}|\mathbf{h}^{l}). \tag{29}$$

The joint distribution $P(\mathbf{h}^{L-1}, \mathbf{h}^{L})$ in (28) is modeled by a Bernoulli-Bernoulli RBM according to the definition of DBN in Section II-B. Because $\mathbf{h}^{L-1}$ and $\mathbf{h}^{L}$ are both binary stochastic vectors, the iterated conditional modes (ICM) [37] algorithm is adopted to solve (28). This algorithm determines the configuration that maximizes the joint probability of a Markov random field by iteratively maximizing the probability of each variable conditioned on the rest. Applying the ICM algorithm here, we just update $\mathbf{h}^{L}$ by maximizing $P(\mathbf{h}^{L}|\mathbf{h}^{L-1})$ and update $\mathbf{h}^{L-1}$ by maximizing $P(\mathbf{h}^{L-1}|\mathbf{h}^{L})$ iteratively. Both of the two conditional distributions are multivariate Bernoulli distribution without cross-dimension correlation [20]. The optimal configuration at each step can be determined simply by applying a threshold of 0.5 for each binary unit. The initial $\mathbf{h}^{L-1}$ of the iteratively updating is set to be the $\mathbf{h}^{L-1*}$ which is obtained by solving (28) for the DBN with $L-1$ hidden layers.

For each l from $L-1$ to 2, (29) can be solved recursively according to the conditional distribution in (8). After $\{\mathbf{h}^{1*}, \ldots, \mathbf{h}^{L*}\}$ are determined, the mode of the DBN can be estimated by substituting (27) into (20). Considering $P(\mathbf{v}|\mathbf{h}^{1})$ is a Gaussian distribution as (7), we have

$$\begin{aligned} \boldsymbol{\mu} &\simeq \arg\max_{\mathbf{v}} \log P(\mathbf{v}, \mathbf{h}^{1*}, \ldots, \mathbf{h}^{L*}) \\ &= \arg\max_{\mathbf{v}} \log P(\mathbf{v}|\mathbf{h}^{1*}) \\ &= \mathbf{W}^{1\top}\mathbf{h}^{1*} + \mathbf{a}^{1}. \end{aligned} \tag{30}$$

IV. Experiments

A. Experimental Conditions

A 1-hour Chinese speech database produced by a professional female speaker was used in our experiments. It consisted of 1,000 sentences together with the segmental and prosodic labels. 800 sentences were selected randomly for training and the remaining 200 sentences were used as a test set. The waveforms were recorded in 16 kHz/16 bit format.

When constructing the baseline system, 41-order mel-cepstra (including 0-th coefficient for frame power) were derived from the spectral envelope by STRAIGHT analysis at 5 ms frame shift. The F0 and spectral features consisted of static, velocity, and acceleration components. A 5-state left-to-right HMM structure with no skips was adopted to train the context-dependent phone models. The covariance matrix of the single Gaussian distribution at each HMM state was set to be diagonal. After the decision-tree-based model clustering, we got 1,612 context-dependent states in total for the mel-cepstral stream. The model parameters of these states were estimated by maximum likelihood training.

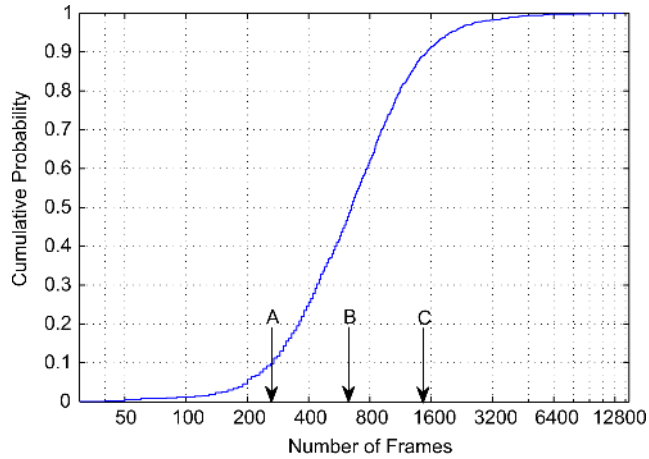

Fig. 3. The cumulative probability curve for the number of frames belonging to each context-dependent state. The arrows indicate the numbers of frames of the three example states used for the analysis in Section IV-B and Fig. 4.

In the spectral envelope modeling, the FFT length of the STRAIGHT analysis was set to 1024 which led to $513 \times 3 = 1539$ visible units in the RBMs and DBNs corresponding to the spectral amplitudes within the frequency range of $[0, \pi]$ together with their dynamic components. After the HMMs for the mel-cepstra and F0 features were trained, a state alignment was conducted on the training set and the test set to assign the frames to each state for the spectral envelope modeling and testing. The cumulative probability curve for the number of frames belonging to each context-dependent state is illustrated in Fig. 3. From this figure, we can see that the numbers of training samples vary a lot among different states. For each context-dependent state, the logarithmized spectral amplitudes at each frequency point were normalized to zero mean and unit variance. CD learning with 1-step Gibbs sampling (CD1) was adopted for the RBM training and the learning rate was 0.0001. The batch size was set to 10 and 200 epochs were executed for estimating each RBM. The DBNs were estimated following the greedy layer-by-layer training algorithm introduced in Section II-B.

B. Comparison Between GMM and RBM as State PDFs

At first we compared the performance of the GMM and the RBM in modeling the distribution of mel-cepstra and spectral envelopes for an HMM state. Three representative states were selected for this experiment, which have 270, 650, 1530 training frames and 60, 130, 410 test frames respectively. As shown in Fig. 3, the numbers of training frames of these three states correspond to the 0.1, 0.5, and 0.9 cumulative probabilities which are calculated over the numbers of the training frames of all the 1,612 context-dependent states. GMMs and RBMs were trained under the maximum likelihood criterion to model these three states. The covariance matrices in the GMMs were set to be

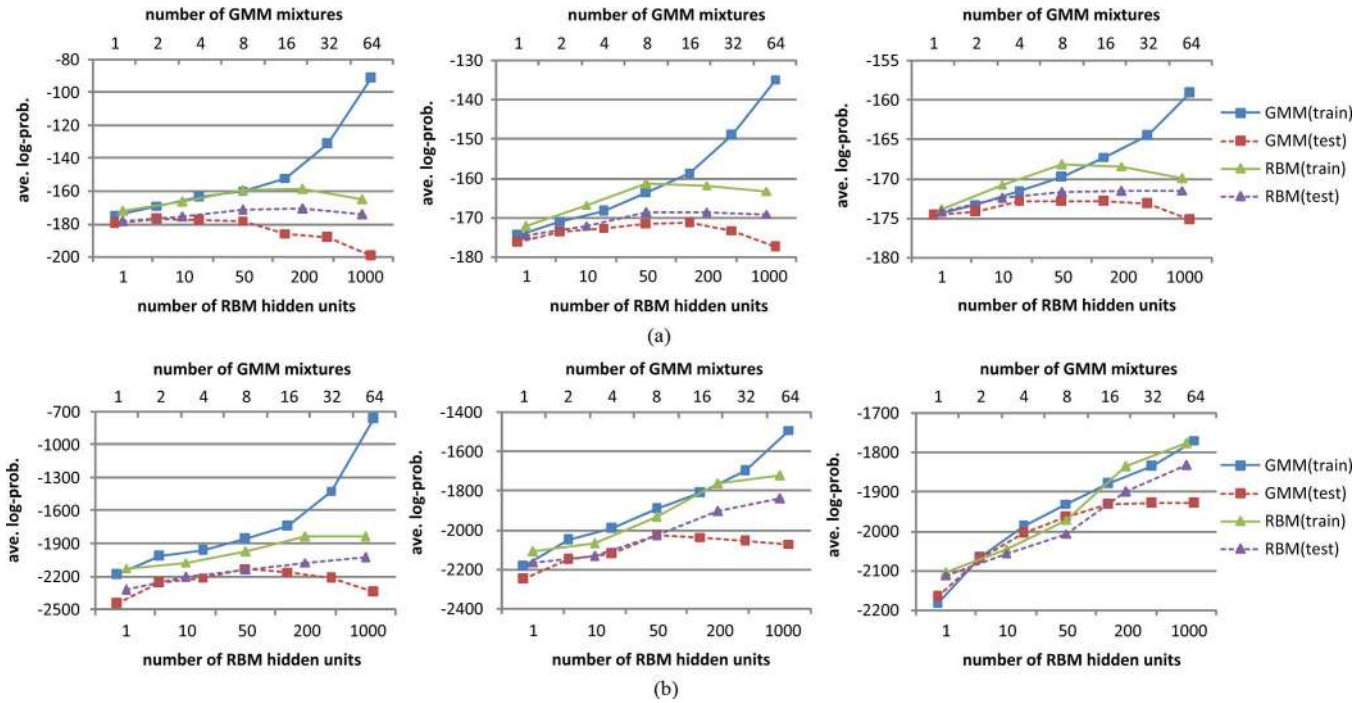

Fig. 4. The average log-probabilities on the training and test sets when modeling (a) the mel-cepstra and (b) the spectral envelopes of *state A* (left column), *state B* (middle column), and *state C* (right column) using different models. The number of training samples belonging to these three selected states are indicated in Fig. 3.

diagonal and the number of Gaussian mixtures varied from 1 to 64. The number of hidden units in the RBMs varied from 1 to 1,000. The average log-probabilities on the training and test sets for different models and states are shown in Fig. 4 for the mel-cepstra and the spectral envelopes respectively. Examining the difference between the training and test log-probabilities for both the mel-cepstra and the spectral envelopes, we see that the GMMs have a clear tendency of over-fitting with the increasing of model complexity. This over-fitting effect becomes less significant when a larger training set is available. On the other hand, the RBM shows consistently good generalization ability with the increasing of the number of hidden units. This can be attribute to utilizing the binary hidden units which create a information bottleneck and act as an effective regularizer during model training.

The differences between the test log-probabilities of the best GMM or RBM models and the single Gaussian distributions for the three states are listed in Table I. From Fig. 4 and Table I, we can see that the model accuracy improvements obtained by using the density models that are more complex than a single Gaussian distribution are relatively small when the mel-cepstra are used for spectral modeling. Once the spectral envelopes are used, such improvements become much more significant for both the GMM and RBM models. Besides, the RBM also gives much higher log-probability to the test data than the GMM when modeling the spectral envelopes. These results can be attributed to that the mel-cepstral analysis is a kind of decorrelation processing to the spectrums. A GMM with multiple components is able to describe the inter-dimensional correlations of a multivariate distribution to some extend even if the diagonal covariance matrices are used. An RBM with H hidden units can be considered as a GMM with 2^H structured mixture components according to (21). Therefore, it is good at analyzing the latent patterns embedded in the high-dimensional raw data

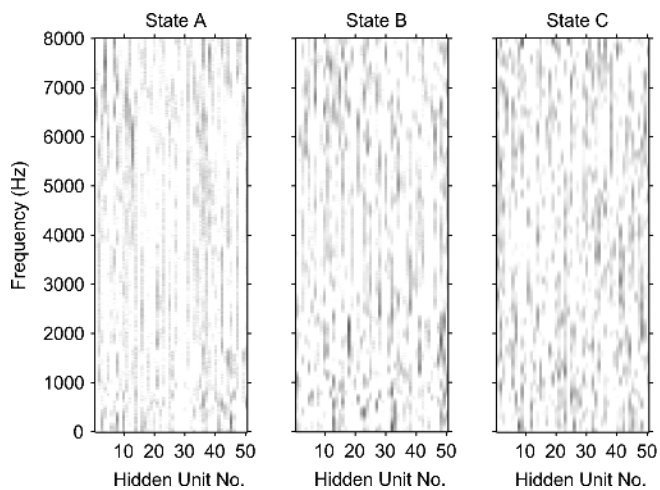

Fig. 5. Visualization of the estimated weight matrices $\mathbf{W}$ in the RBMs when modeling the spectral envelopes for the three states. The number of hidden units is 50 and only the first 513 rows of the weight matrices are drawn. Each column in the gray-scale figures corresponds to the weights connecting one hidden unit with the 513 visible units which compose the static component of the spectral envelope feature vector.

with inter-dimensional correlations. Fig. 5 shows the estimated weight matrices $\mathbf{W}$ in the RBMs when modeling the spectral envelopes for the three states. We can see that the weight matrices are somewhat sparse, indicating each hidden unit tries to capture the characteristics of the spectral envelope in some specific frequency bands. This is similar to a frequency analysis for spectral envelopes which makes use of the amplitudes of the critical frequency bands context-dependently.

For the spectral envelope modeling, we can further improve the model accuracy by training RBMs layer-by-layer and constructing a DBN. The average log-probabilities on the training and test sets when modeling the spectral envelopes using an

TABLE I
THE DIFFERENCES BETWEEN THE TEST LOG-PROBABILITIES OF THE BEST GMM OR RBM MODELS AND THE SINGLE GAUSSIAN DISTRIBUTIONS FOR THE THREE SELECTED STATES. THE NUMBERS IN THE BRACKETS INDICATE THE NUMBERS OF GAUSSIAN MIXTURES FOR THE GMMS AND THE NUMBERS OF HIDDEN UNITS THE RBMS WHICH LEAD TO THE HIGHEST LOG-PROBABILITIES ON THE TEST SET

State	mel-cepstra		spectral envelope	
	GMM	RBM	GMM	RBM
A	2.40(4)	9.32(200)	321.84(8)	421.67(1000)
B	4.80(8)	7.73(200)	221.66(8)	410.06(1000)
C	1.80(8)	3.14(1000)	237.07(32)	332.19(1000)

TABLE II
THE AVERAGE LOG-PROBABILITIES ON THE TRAINING AND TEST SETS WHEN MODELING THE SPECTRAL ENVELOPES USING AN RBM OF 50 HIDDEN UNITS AND A TWO-HIDDEN-LAYER DBN OF 50 HIDDEN UNITS AT EACH LAYER

State	RBM(50)		DBN(50-50)	
	train	test	train	test
A	-1968.267	-2133.704	-1862.665	-2033.919
B	-1930.347	-2025.088	-1852.420	-1943.159
C	-1970.269	-2006.260	-1837.336	-1875.578

TABLE III
SUMMARY OF DIFFERENT SYSTEMS CONSTRUCTED IN THE EXPERIMENTS

System	Spectral Features	State PDF
Baseline	mel-cepstra	single Gaussian
GMM(1)	spectral envelope	single Gaussian
GMM(8)	spectral envelope	GMM, 8 mixtures
RBM(10)	spectral envelope	RBM, 10 hidden units
RBM(50)	spectral envelope	RBM, 50 hidden units
DBN(50-50)	spectral envelope	2-hidden-layer DBN, 50 hidden units each layer
DBN(50-50-50)	spectral envelope	3-hidden-layer DBN, 50 hidden units each layer

RBM and a two-hidden-layer DBN are compared in Table II. Here, the lower bound estimation [31] to the log-probability of a DBN is adopted. From this table, we can observe a monotonic increase of test log-probabilities by using more hidden layers.

C. System Construction

Seven systems were constructed whose performance we compared in our experiments. The definitions of these systems are explained in Table III. As shown in Table I, the model accuracy improvement achieved by adopting the distributions more complicated than the single Gaussian is not significant when the mel-cepstra are used as spectral features. Therefore, we focus on the performance of spectral envelope modeling using different forms of state PDFs in our experiments. Considering the computational complexity of training state PDFs for all context-dependent states, the maximum number of hidden units in the RBM and DBN models were set to 50. All these systems shared the same decision trees for model clustering and the same state boundaries which were derived from the *Baseline* system. The F0 and duration models of the seven systems were identical.

D. Mode Estimation For the RBMS and DBNS

When constructing the *RBM(10)*, *RBM(50)*, *DBN(50–50)*, and *DBN(50–50-50)* systems, the mode of each RBM or DBN trained for a context-dependent state was estimated for

TABLE IV
AVERAGE LOG-PROBABILITIES OF THE SAMPLE MEANS AND THE ESTIMATED MODES FOR THE FOUR RBM OR DBN BASED SYSTEMS

System	Sample Means	PDF Modes
RBM(10)	-1652.1	-1488.0
RBM(50)	-1847.2	-1534.2
DBN(50-50)	-1604.5	-1430.2
DBN(50-50-50)	-1648.5	-1432.3

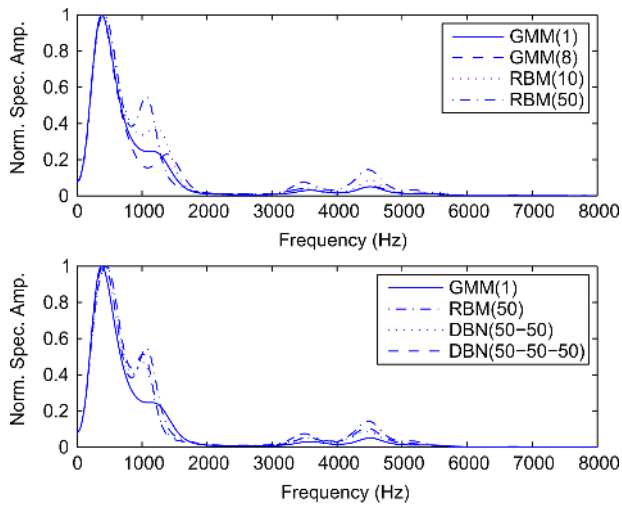

Fig. 6. The spectral envelopes recovered from the modes of different systems for one HMM state.

Gaussian approximation following the methods introduced in Section III-C and Section III-D. For each system, the average Log-Probabilities of the estimated modes and the sample means were calculated. The results are listed in Table IV. From this table, we see that the estimated modes have much higher log-probabilities than the sample means known to have the highest probability for a single Gaussian distribution. This means that when the RBMs or the DBNs are adopted to represent the state PDFs, the feature vector with the highest output probability is not the sample means anymore. This implies the superiority of RBM and DBN over single Gaussian distribution in alleviating the over-smoothing problem during parameter generation under the maximum output probability criterion.

The spectral envelopes recovered from the modes of different systems for one HMM state[4] are illustrated in Fig. 6. Here, only the static components of the spectral envelope feature vectors are drawn. The mode of the *GMM(1)* system is just the Gaussian mean vector. The mode of the *GMM(8)* system is approximated as the Gaussian mean of the mixture with the highest mixture weight. Comparing *GMM(8)* with *GMM*, we can see that using more Gaussian mixtures can help alleviate the over-smoothing effect on the spectral envelope. Besides, the estimated state mode of the RBM and DBN based systems have sharper formant structures than the GMM-based ones. Comparing *RBM(50)* with *RBM(10)*, we can see the advantages of using more hidden units in an RBM. While the differences among the estimated modes of the *RBM(50)*, *DBN(50–50)*, and *DBN(50–50-50)* systems are less significant. We will investigate the performance of these systems further by the following subjective evaluation.

[4]This state is not one of the three states used in Section IV-B.

TABLE V
SUBJECTIVE PREFERENCE SCORES (%) AMONG SPEECH SYNTHESIZED USING THE *BASELINE*, *GMM(8)*, *RBM(10)*, AND *RBM(50)* SYSTEMS, WHERE N/P DENOTES "NO PREFERENCE" AND p MEANS THE p-VALUE OF t-TEST BETWEEN THESE TWO SYSTEMS

Baseline	*GMM(8)*	*RBM(10)*	*RBM(50)*	N/P	p
18.67	48.00	–	–	33.33	0.0014
12.00	–	50.67	–	37.33	0.00
5.33	–	–	70.67	24.00	0.00
–	16.00	–	69.33	14.67	0.00
–	–	9.33	37.33	53.33	0.00

E. Subjective Evaluation

Because the mel-cepstrum extraction can be considered as a kind of linear transform to the logarithmized spectral envelope, the spectral envelope recovered from the mean of the mel-cepstra in a state is very close to the one recovered from the mean of the corresponding logarithmized spectral envelopes. Therefore, the *Baseline* and the *GMM(1)* systems had very similar synthetic results and the *Baseline* system was adopted as a representative for these two systems in the subjective evaluation to simplify the test design. For the *GMM(8)* system, the EM-based parameter generation algorithm [4] could be applied to predict the spectral envelope trajectories by iteratively updating. In order to get a closed-form solution, we made a single Gaussian approximation to the GMMs at synthesis time by only using the Gaussian mixture with the highest mixture weight at each HMM state.

The first subjective evaluation was to compare among the *Baseline*, *GMM(8)*, *RBM(10)*, and *RBM(50)* systems. Fifteen sentences out of the training database were selected and synthesized using these four systems respectively.[5] Five groups of preference tests were conducted and each one was to make comparison between two of the four systems as shown in each row of Table V. Each of the pairs of synthetic sentences were evaluated in random order by five Chinese-native listeners. The listeners were asked to identify which sentence in each pair sounded more natural. Table V summarizes the preference scores among these four systems and the p-values given by t-test. From this table, we can see that introducing the density models that are more complex than single Gaussian, such as GMM and RBM, to model the spectral envelopes at each HMM state can achieve significantly better naturalness than the single Gaussian distribution based methods. Compared with the *GMM(8)* system, the *RBM(50)* system has much better preference in naturalness. This demonstrates the superiority of RBM over GMM in modeling the spectral envelope features. A comparison between the spectral envelopes generated by the *Baseline* system and the *RBM(50)* system is shown in Fig. 7. From this figure, we can observe the enhanced formant structures after modeling the spectral envelopes using RBMs. Besides, we can also find in Table V that the performance of the RBM-based systems is influenced by the number of hidden units used in the model definition when comparing *RBM(10)* with *RBM(50)*. These results are consistent with the formant sharpness of the estimated modes for different systems shown in Fig. 6.

In order to investigate the effect of extending RBM to DBN with more hidden layers, another subjective evaluation was conducted among the *RBM(50)*, *DBN(50-50)*, and *DBN(50-50-50)* systems. Another fifteen sentences out of the training database were used and two groups of preference tests were conducted by five Chinese-native listeners. The results are shown in Table VI. We can see that there is no significant differences among these three systems at 0.05 significance level. Although we can improve the model accuracy by introducing more hidden layers as shown in Table II, the naturalness of synthetic speech can not be improved correspondingly. One possible reason is the approximation we make in (27) when estimating the DBN mode.

[5]Some examples of the synthetic speech using the seven systems listed in Table III can be found at http://staff.ustc.edu.cn/∼ zhling/DBNSyn/demo.html.

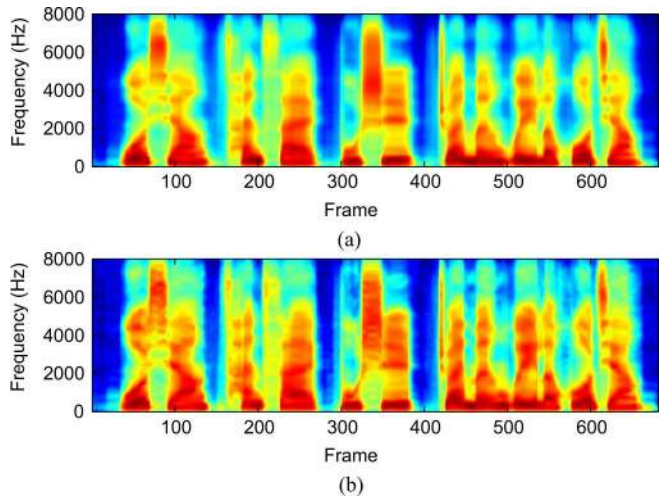

Fig. 7. The spectrograms of a segment of synthetic speech using (a) the *Baseline* system and (b) the *RBM(50)* system. These spectrograms are not calculated by STFT analysis on the synthetic waveform. For the *Baseline* system, the spectrogram is drawn based on the spectral envelopes recovered from the generated mel-cepstra. For the *RBM(50)* system, the spectrogram is drawn based on the generated spectral envelopes directly.

TABLE VI
SUBJECTIVE PREFERENCE SCORES (%) AMONG THE *RBM(50)*, *DBN(50-50)*, AND *DBN(50-50-50)* SYSTEMS

RBM(50)	*DBN(50-50)*	*DBN(50-50-50)*	N/P	p
25.33	17.33	–	57.33	0.2919
38.67	–	21.33	40.00	0.0520

F. Objective Evaluation

Besides the subjective evaluation, we also calculated the spectral distortions on the test set between the spectral envelopes generated by the systems listed in Table III and the ones extracted from the natural recordings. The synthetic spectral envelopes used the state boundaries of the natural recordings to simplify the frame alignment. Then, the natural and synthetic spectral envelopes at each frame were normalized to the same power and the calculation of the spectral distortion between them followed the method introduced in [38]. For the *Baseline* system, the generated mel-cepstra were converted to spectral envelopes before the calculation. The average spectral distortions of all the systems are listed in Table VII. We can see that the objective evaluation results are inconsistent with the subjective preference scores shown in Table V. For example, the *RBM(50)* system has significant better naturalness than the *Baseline* system in the subjective evaluation, while its average spectral distortion is the highest. The reason is that the spectral distortion in [38] is a Euclidean distance between two logarithmized spectral envelopes, which treats each dimension of the spectral envelopes independently and equally. However, the superiority of our proposed method is to provide

TABLE VII
AVERAGE SPECTRAL DISTORTIONS (SD) ON TEST SET BETWEEN THE SPECTRAL ENVELOPES GENERATED BY THE SYSTEMS LISTED IN AND THE ONES EXTRACTED FROM THE NATURAL RECORDINGS

system	ave. SD (dB)
Baseline	3.85
GMM(1)	3.77
GMM(8)	3.86
RBM(10)	3.89
RBM(50)	4.11
DBN(50-50)	4.10
DBN(50-50-50)	4.10

better representation of the cross-dimension correlations for the spectral envelope modeling, which can not be reflected by this spectral distortion measurement. Similar inconsistency between subjective evaluation results and objective acoustic distortions for speech synthesis has been observed in [12], [39].

V. CONCLUSION AND FUTURE WORK

We have proposed an RBM and DBN based spectral envelope modeling method for statistical parametric speech synthesis in this paper. The spectral envelopes extracted by STRAIGHT vocoder are modeled by an RBM or a DBN for each HMM state. At the synthesis time, the mode vectors of the trained RBMs and DBNs are estimated and used in place of the Gaussian means for parameter generation. Our experimental results show the superiority of RBM and DBN over Gaussian mixture model in describing the distribution of spectral envelopes as density models and in mitigating the over-smoothing effect of the synthetic speech.

As we discussed in Section I, there are also some other approaches that can significantly reduce the over-smoothing and improve the quality of the synthetic speech, such as the GV-based parameter generation [11] and the post-filtering techniques [6], [13]. In this paper, we focus on the acoustic modeling to tackle the over-smoothing problem. It worth to investigate alternative parameter generation and post-filtering algorithms that are appropriate for our proposed spectral envelope modeling method in the future.

This paper only makes some preliminary exploration on applying the ideas of deep learning into statistical parametric speech synthesis. There are still several issues in the current implementation that require further investigation. First, it is worth examining the system performance when the number of hidden units in the RBMs keeps increasing. As shown in Fig. 3, the training samples are distributed among many context-dependent HMM states in a highly unbalanced manner. Thus, it may be difficult to optimize the model complexity for all states simultaneously. An alternative solution is to train the joint distribution between the observations and the context labels using a single network [20] which is estimated using all training samples. Similar approach for the statistical parametric speech synthesis has been studied in [30], where the joint distribution between the tonal syllable ID and the spectral and excitation features is modeled using a multi-distribution DBN. Second, increasing the number of hidden layers in the DBNs didn't achieve improvement in our subjective evaluation. A better algorithm to estimate the mode of a DBN with less approximation is necessary. We plan as our future work to implement the Gaussian approximation according to (25) when the number of hidden units is reasonably small and compare its performance with our current implementation. Another strategy is to adopt the sampling outputs rather than the model modes during parameter generation. As better density models, the RBM and DBN are more appropriate than the GMM for generating acoustic features by sampling, which may help make the synthetic speech less monotonic and boring. Third, in the work presented in this paper, the decision tress for model clustering are still constructed using mel-cepstra and single Gaussian state PDF. To extend the RBM and DBN modeling from PDF estimation for the clustered states to model clustering for the fully context-dependent states will also be a task of our future work. Besides the spectral envelopes used in this paper, it is also straightforward to apply our proposed method to the modeling and generation of other forms of speech parameters, such as the articulatory features recorded by electromagnetic articulography (EMA) for articulatory movement prediction [40], the joint distribution between the acoustic features and the articulatory features for articulatory control of HMM-based speech synthesis [41], and the joint spectral distribution between the source and target speakers for voice conversion [42].

REFERENCES

[1] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis," in *Proc. ICASSP*, 2013, pp. 7825–7829.

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

[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, USA: Prentice-Hall, 2004.

[4] 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.

[5] 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.

[6] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for blizzard challenge 2006: an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge Workshop*, 2006.

[7] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, pp. 1039–1064, 2009.

[8] 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.

[9] 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.

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

[11] 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.

[12] Z.-H. Ling and L.-R. Dai, "Minimum kullback-leibler divergence parameter generation for HMM-based speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1492–1502, Jul. 2012.

[13] 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.

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

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

[16] K. Tokuda, H. Zen, and T. Kitamura, Reformulating the HMM as a trajectory model, Tech. Rep. of IEICE 2004.

[17] M. Shannon, H. Zen, and W. Byrnez, "The effect of using normalized models in statistical speech synthesis," in *Proc. Interspeech*, 2011, pp. 121–124.
[18] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," in *Proc. Interspeech*, 2006, pp. 1332–1335.
[19] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McClell, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 194–281.
[20] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[21] L. Deng and D. O'Shaughnessy, *Speech processing: A dynamic and optimization-oriented approach*. Boca Raton, FL, USA: CRC, 2003.
[22] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech*, 2010, pp. 1692–1695.
[23] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[24] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[25] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *Proc. NIPS 2011 Workshop Deep Learn. Unsupervised Feature Learn.*, 2011.
[26] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2702–2722, 1994.
[27] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *J. Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 3036–3050, 2000.
[28] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.
[29] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-Gaussian process hybrid model," in *Proc. ICASSP*, 2013, pp. 6885–6889.
[30] S.-Y. Kang, X.-J. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.
[31] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, Univ. of Toronto, Toronto, ON, Canada, 2009.
[32] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1711–1800, 2002.
[33] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Sci.*, vol. 313, no. 5786, pp. 504–507, 2006.
[34] R. M. Neal, "Connectionist learning of belief networks," *Artif. Intell.*, vol. 56, no. 1, pp. 71–113, 1992.
[35] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM (invited paper)," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.
[36] 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.
[37] J. E. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc., Ser. B*, vol. 48, no. 3, pp. 259–302, 1986.
[38] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 3–14, Jan. 1993.
[39] T. Toda, "Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis," in *Hidden Markov Models, Theory and Applications*, P. Dymarski, Ed. New York, NY, USA: InTech, 2011.
[40] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Commun.*, vol. 52, no. 10, pp. 834–846, 2010.
[41] 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.
[42] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2013.

Zhen-Hua Ling (M'10) received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008 respectively. From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. From July 2008 to February 2011, he was a joint postdoctoral researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an associate professor at University of Science and Technology of China. He also worked at the University of Washington, USA as a visiting scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Li Deng (SM'92–F'04) received the Ph.D. degree from the University of Wisconsin-Madison. He joined the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada in 1989 as an assistant professor, where he became a tenured full professor in 1996. In 1999, he joined Microsoft Research, Redmond, WA as a Senior Researcher, where he is currently a Principal Researcher. Since 2000, he has also been an Affiliate Full Professor and graduate committee member in the Department of Electrical Engineering at University of Washington, Seattle. Prior to MSR, he also worked or taught at the Massachusetts Institute of Technology, ATR Interpreting Telecom. Research Lab. (Kyoto, Japan), and HKUST. In the general areas of speech/language technology, machine learning, and signal processing, he has published over 300 refereed papers in leading journals and conferences and 4 books. He is a Fellow of the Acoustical Society of America, a Fellow of the IEEE, and a Fellow of ISCA. He served on the Board of Governors of the IEEE Signal Processing Society (2008–2010). More recently, he served as Editor-in-Chief for the *IEEE Signal Processing Magazine* (2009–2011), which earned the highest impact factor among all IEEE publications and for which he received the 2011 IEEE SPS Meritorious Service Award. He currently serves as Editor-in-Chief for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and General Chair of ICASSP-2013. His recent technical work since 2009 on and leadership in industry-scale deep learning with colleagues and academic collaborators has created significant impact in speech recognition, signal processing, and related applications.

Dong Yu (M'97–SM'06) joined Microsoft Corporation in 1998 and the Microsoft Speech Research Group in 2002, where he currently is a senior researcher. He holds a Ph.D. degree in computer science from the University of Idaho, an MS degree in computer science from Indiana University at Bloomington, an MS degree in electrical engineering from the Chinese Academy of Sciences, and a BS degree (with honor) in electrical engineering from Zhejiang University (China). His current research interests include speech processing, robust speech recognition, discriminative training, and machine learning. He has published over 120 papers in these areas and is the inventor/coinventor of more than 50 granted/pending patents.

His most recent work focuses on deep learning and its application in large vocabulary speech recognition. The context-dependent deep neural network hidden Markov model (CD-DNN-HMM) he co-proposed and developed has been seriously challenging the dominant position of the conventional GMM based system for large vocabulary speech recognition.

Dr. Dong Yu is a senior member of IEEE, a member of ACM, and a member of ISCA. He is currently serving as a member of the IEEE Speech and Language Processing Technical Committee (2013-) and an associate editor of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2011-). He has served as an associate editor of *IEEE Signal Processing Magazine* (2008–2011) and the lead guest editor of IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING–Special Issue on Deep Learning for Speech and Language Processing (2010–2011).