MODELING SPECTRAL ENVELOPES USING RESTRICTED BOLTZMANN MACHINES FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Zhen-Hua Ling[1,2]*, Li Deng*[3]*, Dong Yu*[3]

[1]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]Department of Electrical Engineering, University of Washington, WA, USA
[3]Microsoft Research, Redmond, WA, USA

zhling@ustc.edu.cn, {deng,dongyu}@microsoft.com

ABSTRACT

This paper presents a new spectral modeling method for statistical parametric speech synthesis. In contrast to the conventional methods in which 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 new method directly models the distribution of the lower-level, un-transformed or raw spectral envelopes. Instead of using single Gaussian distributions, we adopt restricted Boltzmann machines (RBM) to represent the distribution of the spectral envelopes at each HMM state. We anticipate these will give superior performance in modeling the joint distribution of high-dimensional stochastic vectors. The spectral parameters are derived from the spectral envelope corresponding to the estimated mode of each context-dependent RBM and act as the Gaussian mean vector in the parameter generation procedure at synthesis time. Our experimental results show that the RBM is able to model the distribution of the spectral envelopes with better accuracy and generalization ability than the Gaussian mixture model. As a result, our proposed method can significantly improve the naturalness of the conventional HMM-based speech synthesis system using mel-cepstra.

Index Terms— Speech synthesis, hidden Markov model, restricted Boltzmann machine, spectral envelope

1. INTRODUCTION

The hidden Markov model (HMM)-based parametric speech synthesis method has become a mainstream speech synthesis method in recent years [1, 2]. In this method, the spectrum, F0 and segment durations are modeled simultaneously within a unified HMM framework [1]. STRAIGHT [3] as a high-performance speech vocoder is widely used in current HMM-based speech synthesis systems [4, 5]. It extracts a smooth spectral envelope without periodicity interference at each frame. Then, mel-cepstra [4] or line spectral pairs [5] can be derived from the spectral envelopes of training data for the subsequent HMM modeling. The probability density functions (PDF) of each HMM state is commonly represented by a single Gaussian distribution [1]. At synthesis time, the spectral parameters are predicted so as to maximize their output probabilities from the HMM of the input sentence [2]. Then the spectral envelopes are recovered from the generated spectral parameters and are used for waveform reconstruction using STRAIGHT. Because the single Gaussian distributions are used as the state PDFs, the parameter generation outputs tend to distribute near the modes (also the means) of the Gaussians, which are estimated by averaging observations with similar context descriptions in the training set. Although this averaging process improves the robustness of parameter generation, the detailed characteristics of the spectral parameters are lost. The reconstructed spectral envelopes are over-smoothed, which leads to a muffled voice quality in the synthetic speech.

In this paper, we aim to improve the conventional spectral modeling method in HMM-based speech synthesis in two aspects. First, the distributions of the spectral envelopes are modeled directly to avoid the influence of spectral parameter extraction on the process of spectral modeling. Second, a restricted Boltzmann machine (RBM), rather than the single Gaussian distribution, is adopted as the form of the state PDFs in order to better describe the distribution of high-dimensional spectral envelopes and alleviate the over-smoothing problem at synthesis time.

This paper is organized as follows. In Section 2, we will describe the details of our proposed method, including a brief review of the RBM. Section 3 reports our experimental results and Section 4 gives the conclusions.

2. METHODS

2.1. 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 [6]. In this model, the visible stochastic units $\boldsymbol{v} = [v_1, ..., v_V]^\top$ are connected to the hidden stochastic units $\boldsymbol{h} = [h_1, ..., h_H]^\top$, where V and H are the unit numbers of the visible and hidden layers. In this paper, we apply RBMs to model the distribution of spectral envelopes, and the visible units correspond to the spectral amplitudes at all frequency points. The Gaussian-Bernoulli RBM, in which $\boldsymbol{v} \in \mathcal{R}^V$ are real-valued and $\boldsymbol{h} \in \{0,1\}^H$ are binary, is suitable for this task, and so is adopted here. The energy function of the state $\{\boldsymbol{v}, \boldsymbol{h}\}$is defined as

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

This work is partially funded by the National Nature Science Foundation of China (Grant No.61273032) and the China Scholarship Council Young Teacher Study Abroad Project.

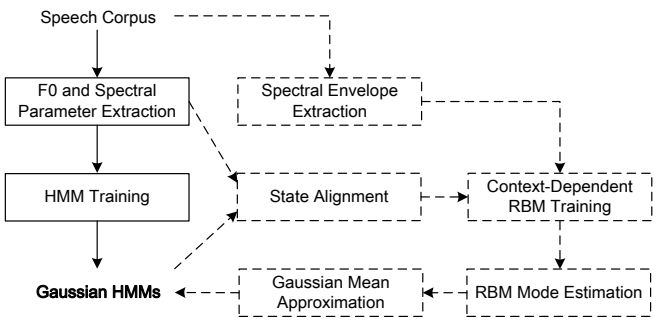

Fig. 1. Flowchart of our proposed method. The modules in solid lines represent the procedures of context-dependent model training in the conventional HMM-based speech synthesis. The modules in dash lines describe the add-on procedures of our proposed method.

where $\boldsymbol{a} = [a_1, ..., a_V]^\top$, $\boldsymbol{b} = [b_1, ..., b_H]^\top$, and $\boldsymbol{W} = \{w_{ij}\}_{V \times H}$ are model parameters. The joint distribution over the visible and hidden units is defined as

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

where

$$\mathcal{Z} = \int_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right) d\boldsymbol{v} \tag{3}$$

is the partition function which can be estimated using the annealed importance sampling (AIS) method [7]. Therefore, the probability density function over the visible vector $\boldsymbol{v}$ can be calculated as

$$\begin{aligned} P(\boldsymbol{v}) &= \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right) \\ &= \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{h}} \exp\left(-\sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2} + \boldsymbol{b}^\top \boldsymbol{h} + \boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{h}\right) \\ &= \frac{1}{\mathcal{Z}} \exp\left(-\sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2}\right) \\ &\qquad \cdot \prod_{j=1}^{H} \sum_{h_j \in \{0,1\}} \exp(b_j h_j + \boldsymbol{v}^\top \boldsymbol{w}_j h_j) \\ &= \frac{1}{\mathcal{Z}} \exp\left(-\sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2}\right) \prod_{j=1}^{H} \left(1 + \exp(b_j + \boldsymbol{v}^\top \boldsymbol{w}_j)\right), \end{aligned} \tag{4}$$

where $\boldsymbol{w}_j$ denotes the j-th column of matrix $\boldsymbol{w}$. Given a training set, the RBM model parameters $\{\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{b}\}$ can be estimated by maximum likelihood learning using the contrastive divergence (CD) algorithm [8, 9].

2.2. RBM for modeling and generating spectral envelopes

In recent years, RBMs have been applied to modeling speech signals, such as speech recognition [10, 11, 12], spectrogram coding [13], and acoustic-articulatory inversion mapping [14] where it mainly acts as a pre-training method for a deep autoencoder or a deep neural network. In this paper, we treat the RBM as a density model and investigate its ability in modeling and generating the spectral envelopes for HMM-based speech synthesis. The flowchart of our proposed method is shown in Fig.1.

In order to make minimum modification to the original model training and parameter generation procedures, the RBM-based spectral envelope modeling method is implemented as a post-processing step performed on the trained context-dependent Gaussian HMMs using the conventional spectral parameters such as mel-cepstra or line spectral pairs. During the acoustic feature extraction using the STRAIGHT vocoder, the original spectral envelopes are stored besides the spectral parameters. After training the context-dependent HMMs, a state alignment to the acoustic features is performed. The state boundaries are used to gather the spectral envelopes for each context-dependent state and an RBM is estimated under the maximum likelihood criterion for each state. Finally, the context-dependent RBM-HMMs can be constructed for modeling the spectral envelopes.

At synthesis time, the optimal sequence of spectral envelopes is estimated to maximize the output probability from the RBM-HMMs of the input sentence. Because the dynamic features of the spectral envelopes are not considered yet in this paper, the trained RBMs cannot generate the continuous sequence of spectral envelopes directly. Therefore, an approximate approach is applied by deriving the spectral parameters from the estimated mode of each RBM and using these parameters to replace the Gaussian mean vector of the static spectral parameters in the trained context-dependent HMMs. One benefit of this approximation is that it keeps the synthesis part of the conventional system intact.

2.3. Estimating RBM mode

Given the model parameters $\{\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{b}\}$ of an RBM which are estimated using the CD algorithm [8] on the training set, the mode of the RBM is defined by

$$\boldsymbol{v}^* = \arg\max_{\boldsymbol{v}} \log P(\boldsymbol{v}), \tag{5}$$

where

$$\begin{aligned} \log P(\boldsymbol{v}) = &-\frac{1}{2}(\boldsymbol{v} - \boldsymbol{a})^\top (\boldsymbol{v} - \boldsymbol{a}) \\ &+ \sum_{j=1}^{H} \log\left(1 + \exp(b_j + \boldsymbol{v}^\top \boldsymbol{w}_j)\right) - \log \mathcal{Z} \end{aligned} \tag{6}$$

according to (4). In contrast to the single Gaussian distribution, this mode is not the average of training vectors any more. The gradient descent algorithm is adopted to solve (5), i.e.,

$$\boldsymbol{v}^{(i+1)} = \boldsymbol{v}^{(i)} + \alpha \cdot \left.\frac{\partial \log P(\boldsymbol{v})}{\partial \boldsymbol{v}}\right|_{\boldsymbol{v}=\boldsymbol{v}^{(i)}}, \tag{7}$$

where i denotes the number of iteration, α is the step size and

$$\frac{\partial \log P(\boldsymbol{v})}{\partial \boldsymbol{v}} = -(\boldsymbol{v} - \boldsymbol{a}) + \sum_{j=1}^{H} \frac{\exp(b_j + \boldsymbol{v}^\top \boldsymbol{w}_j)}{1 + \exp(b_j + \boldsymbol{v}^\top \boldsymbol{w}_j)} \boldsymbol{w}_j. \tag{8}$$

Because the RBM is multimodal, the gradient descent optimization in (7) only leads to a local maximum and the result is sensitive to the initialization of $\boldsymbol{v}^{(0)}$. In order to find a representative $\boldsymbol{v}^{(0)}$, we firstly calculate the means of the conditional distributions $P(\boldsymbol{h}|\boldsymbol{v})$ for all training vectors $\boldsymbol{v}$. These means are averaged and made binary using a fixed threshold of 0.5 to get $\boldsymbol{h}^{(0)}$. Then, the initial $\boldsymbol{v}^{(0)}$ for the iterative updating in (7) is set as the mean of $P(\boldsymbol{v}|\boldsymbol{h}^{(0)})$.

Table 1. The average log-probabilities on the training and test sets when modeling (a) the mel-cepstra and (b) the spectral envelopes of a specific state using different models. The numbers in the brackets indicate the Gaussian mixtures numbers for the GMMs and the hidden unit numbers for the RBMs. "*diag*" and "*full*" denote using diagonal and full covariance matrices respectively.

(a)

	ave. log-prob.		number of
	train	test	parameters
GMM(1)-*diag*	-58.176	-56.380	82
GMM(4)-*diag*	-51.188	-53.097	328
GMM(16)-*diag*	-40.869	-59.492	1,312
GMM(32)-*diag*	-29.973	-72.056	2,624
GMM(1)-*full*	-30.883	-54.648	902
RBM(1)	-56.464	-55.244	83
RBM(10)	-52.416	-52.660	461
RBM(50)	-51.840	-53.636	2,141
RBM(200)	-53.554	-55.020	8,441
RBM(1000)	-55.797	-56.940	42,041

(b)

	ave. log-prob.		number of
	train	test	parameters
GMM(1)-*diag*	-727.915	-728.647	1,026
GMM(4)-*diag*	-599.642	-648.818	4,104
GMM(16)-*diag*	-485.072	-665.609	16,416
GMM(32)-*diag*	-379.980	-717.523	32,832
GMM(1)-*full*	2207.177	-89202.438	132,354
RBM(1)	-685.799	-700.938	1,027
RBM(10)	-629.906	-649.823	5,653
RBM(50)	-587.146	-628.222	30,317
RBM(200)	-576.461	-617.480	103,313
RBM(1000)	-562.439	-583.169	514,513

3. EXPERIMENTS

3.1. 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. The waveforms were recorded in 16kHz/16bit 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 5ms 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.

In the spectral envelop modeling, the FFT length of the STRAIGHT analysis was set to 1024 which meant 513 visible units were used in the RBMs. For each context-dependent state, the spectral amplitudes at each frequency point were logarithmized and 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.

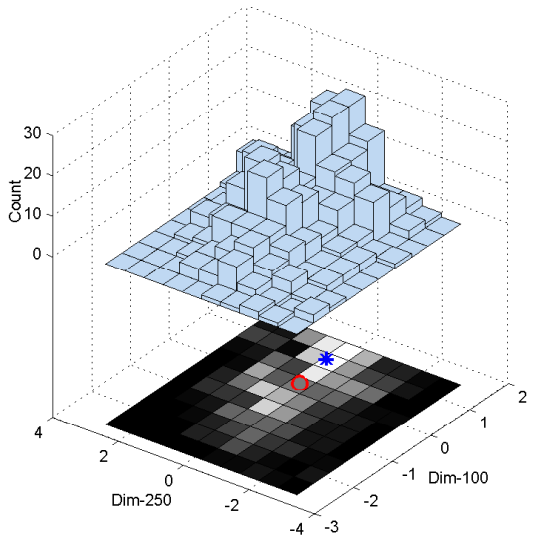

Fig. 2. The histogram and its gray-scale mapping for two dimensions of the spectral envelopes in the training set used in Table 1(b). The red circle and the blue star indicate the mean of the GMM(1)-*diag* model and the estimated mode of the RBM(50) model respectively.

3.2. RBM training

First, we compared the performance of the Gaussian mixture model (GMM) and the RBM in modeling the distribution of mel-cepstra and spectral envelopes for a specific state. A context-dependent state with 720 samples was used in this experiment. 520 samples were used for training the GMM and RBM models. The remaining 200 samples were used as a test set. The number of Gaussian mixtures varied from 1 to 32 and the number of hidden units in the RBMs varied from 1 to 1000. The average log-probabilities on the training and test sets for different models are shown in Table 1 for the mel-cepstra and the spectral envelopes respectively. Examining the difference between the training and test log-probabilities in both tables, we see that the GMMs have a clear tendency for over-fitting as model complexity increases. On the other hand, the RBM shows consistently good generalization ability as the number of hidden units increases. From Table 1(a), we can see that the best GMM and the best RBM have very close log-probabilities on the test set when modeling the mel-cepstra. Once the spectral envelopes are used, the RBMs give much higher log-probabilities to the test data than the GMMs as shown in Table 1(b). This can be attributed to the fact that the mel-cepstral analysis serves to decorrelate the spectral parameters, while the RBMs are able to analyze the latent patterns embedded in the high-dimensional raw data with inter-dimensional correlations. The histogram for two dimensions of the spectral envelopes in the training set is illustrated in Fig. 2. We can observe the multimodal distribution of the training samples and the space containing the estimated RBM mode has higher sample frequency than that of the Gaussian mean.

Considering the computational cost of RBM training, the number of hidden units were set to 50 in our following experiments. It took about 57 hours to train the RBMs for all the context-dependent

Table 2. Average log-probabilities on the training database for the SPE-Gaussian and SPE-RBM systems.

	ave. log prob.
SPE-Gaussian	-727.915
SPE-RBM	-614.123

Table 3. Average log-probabilities of the Gaussian means and the RBM modes for the RBMs trained in the SPE-RBM system.

	ave. log prob.
SPE-Gaussian means	-672.363
SPE-RBM modes (*initial*)	-556.077
SPE-RBM modes (*optimized*)	-518.800

HMM states using a server with a 2.5GHz Intel Xeon E5420 CPU. Finally, three systems were constructed for comparison.

- **MCEP-Gaussian** The baseline system, in which the mel-cesptrums were used for spectral modeling and the state PDFs were in the form of single Gaussian distributions.
- **SPE-RBM** Our proposed method in Section 2.2, in which RBMs were adopted to model the spectral envelopes of each state.
- **SPE-Gaussian** The only difference to our proposed method was that the single Gaussian distribution instead of RBM was used for spectral envelope modeling.

The average log-probabilities on the training database for the SPE-Gaussian and SPE-RBM systems are compared in Table 2.

3.3. Mode estimation for the RBMs

After the modes of all context-dependent RBMs in the SPE-RBM system were estimated following the method introduced in Section 2.3, the log-probabilities of the initial $\boldsymbol{\nu}^{(0)}$ and the optimized $\boldsymbol{\nu}^{*}$ were calculated for each RBM. The average log-probabilities over all states are listed in Table 3 together with the results calculated using the Gaussian mean vectors of the SPE-Gaussian system. From this table, we see that the initial RBM modes have much higher log-probability than the Gaussian means known to have the highest probability for a single Gaussian distribution. The log-probability of the RBM modes increases further after the iterative optimization. Comparing Table 2 and 3, we can find that the Gaussian means have much lower log-probabilities than the training samples once they are described using the RBMs. This implies the superiority of the RBMs over the GMMs in avoiding the use of the sample means during parameter generation under the maximum output probability criterion.

The spectral envelopes corresponding to the Gaussian mean of the MCEP-Gaussian system, the Gaussian mean of the SPE-Gaussian system, and the estimated mode of the SPE-RBM system for one state are illustrated in Fig. 3. We can see that the spectral envelope recovered from the state mean of the MCEP-Gaussian system is very close to that from the state mean of the SPE-Gaussian system. The estimated state mode of the SPE-RBM system has much sharper formant structure and less over-smoothing than the other two envelopes.

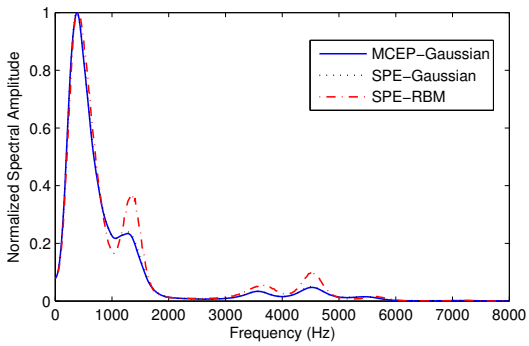

Fig. 3. The spectral envelopes corresponding to the Gaussian mean of the MCEP-Gaussian system, the Gaussian mean of the SPE-Gaussian system, and the estimated mode of the SPE-RBM system for one state.

Table 4. Subjective preference scores (%) between speech synthesized using the MCEP-Gaussian and SPE-RBM systems, where N/P denotes "No Preference" and p means the p-value of t-test between these two system.

MCEP-Gaussian	SPE-RBM	N/P	p
14.67	61.33	24.00	0.00

3.4. Subjective evaluation

Because the MCEP-Gaussian and SPE-Gaussian systems had very similar parameter generation results, only the MCEP-Gaussian and the SPE-RBM systems were compared in a preference test on the naturalness of synthetic speech. Fifteen sentences out of the training database were selected and synthesized using these two systems respectively. Five Chinese-native listeners with no hearing problems took part in the test. Table 4 shows the preference scores between these two systems and the p-values given by t-test. We see that the SPE-RBM system has significantly better naturalness than the MCEP-Gaussian system.[1]

4. CONCLUSIONS

We have proposed an RBM-based spectral envelope modeling method in this paper. The spectral envelopes extracted by STRAIGHT vocoder are modeled by an RBM for each HMM state. At synthesis time, the mode vectors of the trained RBMs are estimated and used in place of the Gaussian means for parameter generation. Our experimental results show the superiority of RBMs over Gaussian mixture models in describing the distribution of spectral envelopes as a density model and in alleviating the over-smoothing effect at synthesis time. Incorporating the dynamic features of spectral envelopes into RBM modeling and extending RBM to deep belief networks (DBN) [15] or deep Boltzmann machines (DBM) [7] will be the tasks of our future work.

[1] Some examples of the synthetic speech using these two methods can be found at http://staff.ustc.edu.cn/~zhling/SPERBM-ICASSP2013/demo.html.

5. 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 *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 *ICASSP*, 2000, vol. 3, pp. 1315–1318.

[3] 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 Communication*, vol. 27, pp. 187–207, 1999.

[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.-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 *Blizzard Challenge Workshop*, 2006.

[6] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing*, D.E. Rumelhart and McClelland J.L., Eds., vol. 1, chapter 6, pp. 194 – 281. MIT Press, 1986.

[7] R. Salakhutdinov, *Learning deep generative models*, Ph.D. thesis, University of Toronto, 2009.

[8] G.E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1711–1800, 2002.

[9] G.E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] A. Mohamed, G.E. Dahl, and G.E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[11] 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, 2012.

[12] G.E. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Andrew Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] 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 *Interspeech*, 2010, pp. 1692–1695.

[14] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," in *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] G.E Hinton, S. Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.