SPECTRAL MODELING USING NEURAL AUTOREGRESSIVE DISTRIBUTION ESTIMATORS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Xiang Yin, Zhen-Hua Ling, Li-Rong Dai

National Engineering laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
byx1030@mail.ustc.edu.cn, zhling@ustc.edu.cn, lrdai@ustc.edu.cn

ABSTRACT

This paper describes a new approach which utilizes neural autoregressive distribution estimators (NADE) for the spectral modeling in statistical parametric speech synthesis. In order to alleviate the over-smoothing effect on the generated spectral structures, a restricted Boltzmann machine (RBM) modeling method has been proposed in our previous work, where the RBM is adopted to represent the joint distribution of high-dimensional and physically meaningful spectral envelopes. However, the RBM can not provide a tractable partition function even in a moderate size. In this paper, we introduce NADE to model the distribution of mel-cepstra and spectral envelopes at each HMM state considering its simplicity in evaluating the probability of given observations. At the stage of synthesis, the spectral parameters derived from the mode of each context-dependent NADE are used to replace the Gaussian mean vector in the parameter generation process. Experimental results show that the NADE is able to model the distribution of the spectral features with better accuracy than the RBM model. Furthermore, our proposed method improves the naturalness of the conventional HMM-based speech synthesis system using mel-cepstra significantly and outperforms the RBM-based spectral modeling.

Index Terms— Speech synthesis, hidden Markov model, neural autoregressive distribution estimator, restricted Boltzmann machine

1. INTRODUCTION

The hidden Markov model (HMM)-based parametric speech synthesis method has been proposed in 1990's [1] and become a mainstream speech synthesis method in recent years. In this method, the spectrum, F0 and duration are modeled simultaneously within a unified framework of HMMs [2]. STRAIGHT [3] is a widely used speech vocoder which extracts a smooth spectral envelope at each frame. Then, mel-cepstra [4] or line spectral pairs [5] are calculated from the spectral envelopes for the HMM modeling. At the stage of synthesis, these features are predicted from the HMMs through the maximum likelihood parameter generation (MLPG) algorithm under the constraint between static and dynamic features [6]. Then the spectral envelopes are recovered from the predicted spectral parameters and are used to reconstruct the speech waveforms by STRAIGHT. This method can synthesize highly intelligible and smooth speech sounds [7].

However, the quality of its synthetic speech degrades due to the over-smoothing issue of the generated acoustic features. One reason is the inadequacy of the acoustic modeling. In order to address this problem, some methods have been proposed. An RBM-based spectral envelope modeling method was proposed in [8]. In this method, the spectral envelopes extracted by STRAIGHT vocoder were modeled by an RBM for each HMM state. At synthesis time, the estimated mode vectors of the trained RBMs were used to replace the Gaussian mean vectors for parameter generation. Additionally, dynamic features of spectral envelopes have also been incorporated into the RBM modeling and the deep belief networks (DBN) have also been adopted in [9].

An RBM is a kind of bipartite undirected graphical model which has been applied to speech synthesis [8] and voice conversion [15], [16]. However, it does not provide a tractable partition function for computing the probability of an observation. Not knowing the exact value of partition function makes it hard to evaluate how well the distribution estimated by the RBM fits the observations. The NADE proposed in [10] is inspired by the RBM and its model structure is similar to a fully visible sigmoid belief network (FVSBN) [11]. It can solve the difficulty of partition function calculation by decomposing the joint distribution of observations into tractable conditional distributions. Therefore, in this paper, we propose to adopt NADE as the form of the state PDFs instead of RBM [8].

This paper is organized as follows. Section 2 describes the details of our proposed method, including a brief review on the RBM and NADE. Section 3 gives the experimental results and section 4 concludes this paper.

2. METHODS

2.1. Restricted Boltzmann machines

An RBM is a Markov random field with a two-layer architecture which is used to describe the cross-dimension dependency among a set of random variables [12]. In this model, a set of weights $\boldsymbol{W}$ connect the visible stochastic units $\boldsymbol{v}=[v_1,...,v_V]^{\mathrm{T}}$ to the hidden stochastic units $\boldsymbol{h}=[h_1,...,h_H]^{\mathrm{T}}$ as shown in Fig. 1.a), where V and H refer to the unit numbers of the visible and hidden layers. More specifically, the energy function of the Gaussian-Bernoulli RBM is defined as

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

where $\boldsymbol{a}=[a_1,...,a_V]^{\mathrm{T}}$, $\boldsymbol{b}=[b_1,...,b_H]^{\mathrm{T}}$, and $\boldsymbol{W}=\{w_{ij}\}_{V\times H}$ are model parameters. The joint distribution over the real-valued

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

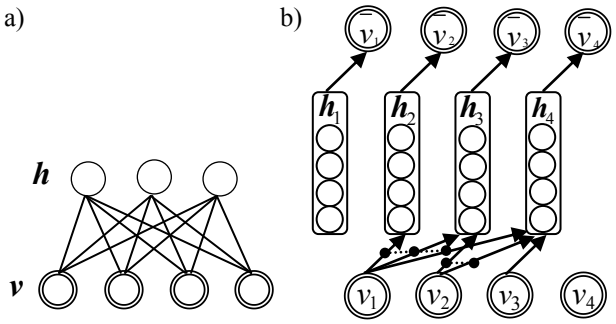

Fig. 1. The graphical model representations for a) an RBM and b) an NADE. $\bar{v}_i$ is short for $P(v_i = x \mid \boldsymbol{v}_{<i})$ and the j -th dimension of $\boldsymbol{h}_i$ denotes $P(h_j = 1 \mid \boldsymbol{v}_{<i})$. Arrows connected by a dash line relates to connections with shared or tied parameters.

visible units and binary hidden units is defined as

$$P(\boldsymbol{v}, \boldsymbol{h}) = \exp(-E(\boldsymbol{v}, \boldsymbol{h}))/Z, \tag{2}$$

where Z is known as the partition function and ensures that $P(\boldsymbol{v}, \boldsymbol{h})$ is a valid probability density function and sums to 1. The marginalized probability of $\boldsymbol{v}$ is related to the free-energy $F(\boldsymbol{v})$ by $P(\boldsymbol{v}) \equiv \exp^{-F(\boldsymbol{v})}/Z$ and

$$F(\boldsymbol{v}) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2} - \sum_{j=1}^{H} (1 + \exp(b_j + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{w}_j)), \tag{3}$$

Although the RBM has been shown to be a powerful model in representing the distribution of high-dimensional observations, computing its partition function becomes intractable with even just a moderate number of hidden variables, e.g. around 30. Therefore, some approximations to the gradients are necessary when estimating the model parameters using the contrastive divergence (CD) algorithm [13].

2.2. Neural autoregressive distribution estimator

The neural autoregressive distribution estimator (NADE) [10] is a tractable model inspired by the RBM and it is derived from the FVSBN. This model decomposes the joint distribution of observations into tractable conditional distributions to solve the difficulty of estimating partition functions. In this paper, NADE is applied to model the distribution of spectral features. Therefore, the Gaussian-Bernoulli NADE is adopted as shown in Fig. 1.b), which means $\boldsymbol{v} \in \Re^V$ are real-valued and $\boldsymbol{h} \in \{0,1\}^H$ are binary. The distribution of each visible unit v_i is expressed as a Gaussian function of the vector $\boldsymbol{v}_{<i} \equiv \{v_k, \forall k < i\}$, and the probability distribution of observation is defined by:

$$P(\boldsymbol{v}) = \prod_{i=1}^{V} P(v_i \mid \boldsymbol{v}_{<i}) \tag{4}$$

$$P(v_i = x \mid \boldsymbol{v}_{<i}) = = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - a_i - \boldsymbol{U}_{i,:} \boldsymbol{h}_i)^2}{2} \right), \tag{5}$$

$$\boldsymbol{h}_i = \sigma(\boldsymbol{b} + \boldsymbol{W}_{:,<i} \boldsymbol{v}_{<i}), \tag{6}$$

where $\boldsymbol{U}$ is a separate set of weights for the connections from the hidden units to the outputs. $\boldsymbol{U}_{i,:}$ denotes the i -th row of matrix $\boldsymbol{U}$, $\boldsymbol{W}_{:,<i}$ indicates the columns of matrix $\boldsymbol{W}$ whose subscripts are less than i . $\boldsymbol{h}_i$ is a vector of H dimensions whose elements are the posterior probability $P(h_j = 1 \mid \boldsymbol{v}_{<i})$. $\sigma(t) \equiv (1 + e^{-t})^{-1}$ is the logistic sigmoid function.

The parameter derivatives of negative log-likelihood function $C \equiv -\log P(\boldsymbol{v})$ are calculated as

$$\frac{\partial C}{\partial a_i} = a_i + \boldsymbol{U}_{i,:} \boldsymbol{h}_i - v_i, \tag{7}$$

$$\frac{\partial C}{\partial b} = \sum_{k=1}^{V} \frac{\partial C}{\partial a_k} \boldsymbol{U}_{k,:}^{\mathrm{T}} \boldsymbol{h}_k (1 - \boldsymbol{h}_k), \tag{8}$$

$$\frac{\partial C}{\partial \boldsymbol{W}_{:,i}} = v_i \sum_{k=i+1}^{V} \frac{\partial C}{\partial a_k} \boldsymbol{U}_{k,:}^{\mathrm{T}} \boldsymbol{h}_k (1 - \boldsymbol{h}_k), \tag{9}$$

$$\frac{\partial C}{\partial \boldsymbol{U}_{i,:}} = \frac{\partial C}{\partial a_i} \boldsymbol{h}_i^{\mathrm{T}}, \tag{10}$$

Given a training set, the NADE model parameters $\{\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{b}\}$ can be estimated using the algorithm of stochastic gradient descent directly without any approximations [10]. This is the superiority of the NADE over the RBM which needs some approximations to calculate the gradients in the CD algorithm [13].

2.3. NADE for spectral modeling

Recently, NADE has been proved to be an efficient multivariate binary distribution estimator and performs similarly to a large (but intractable) RBMs on several datasets [10]. It has also been applied to modeling multinomial distributions in [14]. In this paper, we adopt NADE as the form of the state PDFs and investigate its ability in modeling and generating spectral features for HMM-based speech synthesis which relates to the work of modeling the continuous stochastic distribution in [8] [9].

During the acoustic feature extraction using the STRAIGHT vocoder, the spectral envelopes are saved besides the mel-cepstra. The conventional method using mel-cepstra and single Gaussian state PDFs is conducted first for context-dependent HMM training to obtain the model clustering decision tress and the state alignment results. Then, at the stage of spectral modeling using NADEs, two approaches are considered here.

One is to model the mel-cepstra. NADEs are estimated for modeling the static features of mel-cepstra under the maximize log-likelihood criterion at each context-dependent HMM state. At synthesis time, the spectral parameters derived from the mode of each NADE are used to replace the Gaussian mean vector in the trained HMMs for the process of parameter generation.

Another is to model the spectral envelopes. This approach is similar to the method in [8]. NADEs are trained for modeling the static features of spectral envelopes at each context-dependent HMM state. At synthesis time, the mel-cepstra are derived from the estimated mode of each NADE and used to replace the Gaussian mean vector of the static spectral parameters in the trained context-dependent HMMs.

2.4. Estimating the mode of an NADE

Given the model parameters $\{\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{a}, \boldsymbol{b}\}$ of an NADE which are estimated by directly maximize the average log-likelihood of the parameters on the training set, the mode of the NADE is defined by

$$\boldsymbol{v}^* = \arg\max_{\boldsymbol{v}} \log P(\boldsymbol{v})$$
$$= \arg\max_{\boldsymbol{v}} \sum_{i=1}^{V} \log P(v_i \mid \boldsymbol{v}_{<\mathbf{i}}) \quad (11)$$

where V means the unit number of visible layer. Eq. (11) can be solved by sequentially determining each v_i according to $P(v_i \mid \boldsymbol{v}_{<\mathbf{i}})$ which is a Gaussian distribution as shown in Eq. (5). During the estimating of $\boldsymbol{v}^*$, two kinds of initialization are used. One is the normal initialization. $\boldsymbol{h}_1$ is directly calculated by Eq. (6) with initial $\boldsymbol{v}_{<1}$ set as zero. Another is the binary initialization. We firstly calculate the means of $\boldsymbol{h}_i$ $(i = 1...V)$ for all training vectors by Eq. (6), these means are averaged and made binary using a fixed threshold of 0.5 to get $\boldsymbol{h}_1$. Finally, we can use either of these two initialization methods to calculate $\boldsymbol{h}_1$ and further get the mode of NADE $\boldsymbol{v}^*$ by utilizing Eq. (5) and (6) iteratively.

This process is more efficient than the mode estimation for the RBMs [8]. In contrast to the single Gaussian distribution, this estimated mode avoids the averaging effect and maintains the detailed characteristics of the spectral parameters.

3. EXPERIMENTS

3.1. Experimental conditions

A Chinese speech database recorded by a professional female speaker was used in our experiments. It consists of 1,000 sentences together with the segmental and prosodic labels. 800 sentences were selected randomly for training and the remaining 200 sentences were taken as a test set. Speech waveforms were recorded in 16kHz/16bit format. The acoustic features, including the logarithmized F0, 41-order mel-cepstra (including 0-th order), were extracted from the spectral envelope with a 5ms frame shift by STRAIGHT analysis. The F0 and spectral features included static, velocity, and acceleration components. A 5-state left-to-right with no skip HMM structure was used to train context-dependent phone models. The covariance matrix of the single-mixture Gaussian distribution at each HMM state was set to be diagonal. Decision-tree-based model clustering was applied in the context-dependent model training and we got 1,612 context-dependent states in total for the mel-cepstral stream.

During the modeling of mel-cepstra, the visible units of NADE correspond to 40-order mel-cepstra (excluding 0-th order). During the modeling of spectral envelopes, the FFT length of the STRAIGHT analysis was set to 1024, so there are 513 visible units in the NADEs. For each context-dependent state, the spectral amplitudes at all frequency points were logarithmized. These two features were normalized to zero mean and unit variance. The learning rate was 0.001 and 200 epochs were executed for estimating each NADE. When it comes to the RBM training, we refer to the configuration in [8].

3.2. NADE training

At first we take the performance of the RBM and NADE in modeling the distribution of mel-cepstra and spectral envelopes on a specific state for comparison. A context-dependent state was chosen in our experiment. There are 520 samples in this state for training and 200 samples used for test. The number of hidden units in the RBMs and NADEs ranged from 1 to 1,000. The average log-likelihood (ALL) on the training and test sets for different models are shown in Table 1 for the mel-cepstra and the spectral envelopes respectively. With the increase of hidden unit numbers, the RBM and NADE both show good ability of generalization without the tendency of over-fitting. From Table 1, we can also see that, when the number of hidden units is more than one, NADEs have much better log-likelihood on the train data and test data than the RBMs either for the mel-cepstra or the spectral envelopes. The reason is that NADE can provide exact gradients of model parameters without any approximations to the partition function. This demonstrates the superiority of the NADE over the RBM which needs some approximations to calculate the gradients by the CD algorithm [13].

Table 1. The average log-likelihood (ALL) on the training and test sets when using RBM and NADE to model the (a) mel-cepstra and (b) the spectral envelopes of a specific state. The numbers in the brackets means the hidden unit numbers of the RBMs and NADEs.

(a)

Model	ALL train	ALL test	number of parameters
RBM(1)	-55.146	-55.296	81
RBM(10)	-51.293	-52.443	450
RBM(50)	-50.969	-52.233	2090
RBM(200)	-52.605	-53.535	8240
RBM(1000)	-55.055	-55.715	41040
NADE(1)	-56.761	-56.347	121
NADE(10)	-45.497	-48.109	850
NADE(50)	-43.950	-47.066	4090
NADE(200)	-44.329	-47.080	16240
NADE(1000)	-45.403	-47.960	81040

(b)

Model	ALL train	ALL test	number of parameters
RBM(1)	-642.577	-636.646	1027
RBM(10)	-628.847	-624.837	5653
RBM(50)	-573.244	-591.263	26213
RBM(200)	-551.845	-572.693	103313
RBM(1000)	-552.721	-560.905	514513
NADE(1)	-626.159	-635.720	1540
NADE(10)	-516.682	-530.044	10783
NADE(50)	-481.835	-490.604	51863
NADE(200)	-477.288	-483.079	205913
NADE(1000)	-477.829	-480.993	1027513

Considering the computational cost in the RBM and NADE training, the number of hidden units was set to 50 in the following experiments. At last, six systems were constructed for comparison as listed in Table 2. The ALL on the whole training set for these six systems are compared in Table 3.

3.3. Mode estimation for the NADEs

In the mode estimation of MCP-NAD, we used the binary initialization and obtained speech with better perception than that of using normal initialization. However, in the mode estimation of SPE-NAD systems, the binary initialization would bring some noises to the speech and reduce its quality. Therefore, we used normal initialization instead. The ALL over all states are listed in Table 4 for six systems. From this table, we see that the NADE m-

Table 2. Summary of different systems constructed in the experiments. The numbers in the brackets means the hidden unit numbers of the RBMs and NADEs.

System	Spectral Features	State PDF
MCP-GAU	mel-cepstra	single Gaussian
MCP-RBM	mel-cepstra	RBM(50)
MCP-NAD	mel-cepstra	NADE(50)
SPE-GAU	spectral envelope	single Gaussian
SPE-RBM	spectral envelope	RBM(50)
SPE-NAD	spectral envelope	NADE(50)

Table 3. Average log-likelihood on the whole training set for all six systems.

System	ALL
MCP-GAU	-56.758
MCP-RBM	-54.430
MCP-NAD	-47.797
SPE-GAU	-727.915
SPE-RBM	-613.469
SPE-NAD	-487.055

Table 4. Average log-likelihood of Gaussian means, the RBM and NADE modes for the NADEs trained in the MCP-NAD and SPE-NAD respectively.

System	ALL
MCP-GAU	-47.831
MCP-RBM	-46.469
MCP-NAD	-37.758
SPE-GAU	-666.176
SPE-RBM	-516.128
SPE-NAD	-471.415

odes have much higher log-likelihood than the Gaussian means known to have the highest probability for a single Gaussian distribution. This implies the superiority of NADE over Gaussian mixture model in avoiding the averaging effect during the process of maximum likelihood parameter generation.

The spectral envelopes which correspond to the Gaussian mean of the SPE-GAU system, the estimated mode of the SPE-RBM system and the SPE-NAD system for one state are illustrated in Fig. 2. We can see that the estimate state mode of the SPE-RBM and the SPE-NAD have much sharper formant structure and less over-smoothing than the envelopes of the SPE- GAU.

3.4. Subjective evaluation

The subjective evaluation was to compare among the MCP-GAU, MCP-RBM, MCP-NAD, SPE-RBM, SPE-NAD systems. In these systems, there were no differences in the modeling and generation of duration and pitch. Therefore, the evaluation focused on the difference of naturalness that caused by the different spectral modeling method. Fifteen sentences out of training set were randomly selected and synthesized using these five systems respectively. Six groups of preference tests were conducted and each one was to make comparison between two of the five systems as shown in Table 5. Each of the pairs of synthetic sentences were evaluated in random order by seven Chinese-native listeners. Table 5 shows the preference scores and the *p*-values given by *t*-test. From this table, we can see that introducing NADEs to be the density models can achieve significantly better naturalness than the single Gaussian distribution based system. Besides, adopting spectral envelopes as the features in NADE is better than that of using mel-cepstra for preserving the cross-dimensional correlations. Finally, the NADE-based systems outperform the RBM-based ones for both mel-cepstra and spectral envelopes.

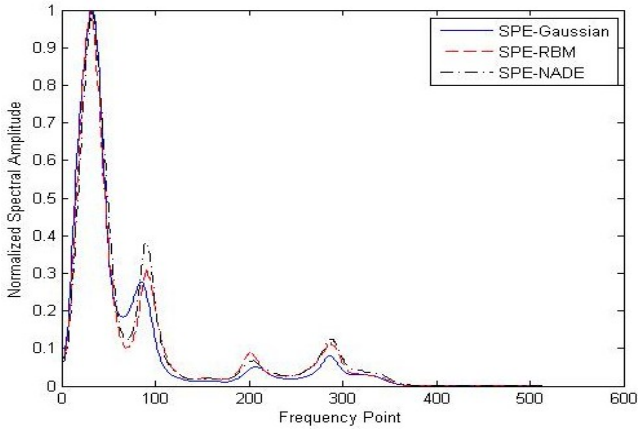

Fig. 2. The spectral envelopes which correspond to the Gaussian mean of the SPE-GAU system, the estimated mode of the SPE-RBM system and the SPE-NAD system for one state.

Table 5. Subjective preference scores (%) among speech synthesized by MCP-GAU, MCP-RBM, MCP-NAD, SPE-RBM and SPE-NAD systems, where N/P indicates "No Preference" and *p* means the *p*-value given by *t*-test between two compared systems.

MCP-GAU	MCP-RBM	MCP-NAD	SPE-RBM	SPE-NAD	N/P	p
11.43	40.00	--	--	--	48.57	0.00
11.43	--	78.09	--	--	10.48	0.00
	10.48	69.52	--	--	20.00	0.00
14.28	--	--	--	79.05	6.67	0.00
--	--	24.76	--	58.10	17.14	0.027
--	--	--	28.33	58.33	13.34	0.00

4. CONCLUSIONS

We have proposed an NADE-based spectral modeling method in this paper. The spectral envelopes extracted by STRAIGHT vocoder and the mel-cepstra derived from the envelopes are modeled by an NADE for each HMM state. At synthesis time, the mode vectors of the trained NADE are calculated and replace the Gaussian means for parameter generation. Our experimental results show the superiority of NADEs over Gaussian mixture models in describing the distribution of spectral envelopes as a density model and in alleviating the over-smoothing effect at the synthesis time. When comparing the ability of model generalization between RBMs and NADEs, the experimental results show that NADEs demonstrates better performance than RBMs due to the accurate calculation of gradients at training time. Incorporating the dynamic features of mel-cepstra and spectral envelopes into NADE modeling and extending the spectral features from the spectral envelopes to the FFT spectrum will be the tasks of our future work.

5. ACKNOWLDEGEMENT

This work was supported by the National Nature Science Foundation of China (Grant No.61273032) and by the Science and Technology Development of Anhui Province, China (Grants No. 13Z02008-5).

6. REFERENCES

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

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

[3] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instant-neous-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] K. Tokuda, T. kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *ICASSP*, 1995, pp. 660-663.

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

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

[9] Z.-H Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129-2139, 2013.

[10] Larochelle. H, and Murray. I, "The neural autoregressive distribution estimator," in *JMLR:W&CP*, 2011, pp. 15:29-37.

[11] Neal, R. M, "Connectionlist learning of belief networks," in *Artificial Intelligence*, 1992, pp. 56, 71-113.

[12] 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.

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

[14] Larochelle. H, and Lauly. S, "A neural autoregressive topic model," *advance in Neural information processing systems*, 25, 2012

[15] 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, pp. 3052–3056.

[16] Z.-Z Wu, E.S. Chng, and H.-Z. Li, "Conditional restricted Boltzmann machine for voice conversion," in *ChinaSIP*, 2013