DNN-based stochastic postfilter for HMM-based speech synthesis

Ling-Hui Chen[1,2]*, Tuomo Raitio*[3]*, Cassia Valentini-Botinhao*[4]*, Junichi Yamagishi*[4,5]*, Zhen-Hua Ling*[1]

[1]University of Science and Technology of China, P.R. China
[2]iFLYTEK Research, USTC iFLYTEK Co., Ltd., Anhui, P.R. China
[3]Department of Signal Processing and Acoustics, Aalto University, Finland
[4]The Centre for Speech Technology Research, University of Edinburgh, UK
[5]National Institute of Informatics, Japan

{chenlh,zhling}@ustc.edu.cn, tuomo.raitio@aalto.fi, {cvbotinh,jyamagis}@inf.ed.ac.uk

Abstract

In this paper we propose a deep neural network to model the conditional probability of the spectral differences between natural and synthetic speech. This allows us to reconstruct the spectral fine structures in speech generated by HMMs. We compared the new stochastic data-driven postfilter with global variance based parameter generation and modulation spectrum enhancement. Our results confirm that the proposed method significantly improves the segmental quality of synthetic speech compared to the conventional methods.

Index Terms: HMM, speech synthesis, DNN, modulation spectrum, postfilter, segmental quality

1. Introduction

Statistical parametric speech synthesis is one of the most popular speech synthesis methods due to its flexibility and compact footprint [1]. It is known, however, that synthesised speech generated from statistical models still sounds "muffled". This is often attributed to the fact that fine spectral structures of natural speech are partly lost due to statistical averaging, and thus there is room for the improving the segmental quality.

There have been several successful attempts to improve the segmental quality of synthesised speech, including postfiltering to enhance spectral peaks [2] and the global variance (GV) parameter generation algorithm that enhances the dynamics within a speech utterance [3]. Recently an interesting approach based on the enhancement of the modulation spectrum (MS) was proposed in [4]. The aim of this method is to enhance the natural frequency modulation in the spectral parameter trajectories. These methods have been shown to improve the quality of synthetic speech, these approaches are based on empirical findings of acoustic differences between natural and synthetic speech, which tend to occur for most speakers.

Another possible way to reduce the gap between the segmental quality of natural and synthetic speech is to learn the acoustic differences directly from the data. If we have a parallel set of natural and synthetic speech, we can estimate the conditional probability of the acoustic differences, that is, the probability of natural speech given the muffled synthetic speech. One could then model and reconstruct the spectral fine structures through a data-driven statistical method. Conceptually this is similar to voice conversion techniques that consider the conditional probability of the parallel data [5].

In this paper we introduce a deep neural network (DNN) [6] to model the conditional probability of the acoustic differences. In voice conversion [7] this is typically done with a Gaussian mixture model (GMM) but a DNN was chosen here instead due to its abilities to model highly correlated and high dimensional data, allowing us to conduct spectral shaping directly in the spectral domain. We compared the proposed method with the GV and the recently proposed MS enhancement.

This paper is organised as follows: in Section 2, we explain the DNN-based approach and in Section 3 the MS enhancement. The experimental conditions and evaluation results are shown in Section 4. Analysis and discussions on what the DNN learns as well as the summary of our findings are given in Section 5.

2. DNN-based stochastic postfilter

This section introduces the DNN for stochastic modelling of the differences between spectra of synthesised and natural speech, similar to the approach used in voice conversion [6].

2.1. Model training

Fig. 1 shows the structure of a four-layer feedforward DNN. The DNN models the conditional probability $P(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ is the input synthesised spectral envelope and $\mathbf{y}$ is the corresponding natural spectral envelope. The chosen architecture is generatively trained layer-by-layer with a cascade of two restricted Boltzmann machines (RBMs) [8] and a Bernoulli bidirectional associative memory (BBAM) [9, 10], as illustrated in Fig. 1.

Both the RBM and BBAM are two-layer stochastic neural networks. They are generative models whose probabilistic distributions are defined by energy functions. The energy function of an RBM is defined as

$$E(\mathbf{v},\mathbf{h}) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^{V} \frac{v_i}{\sigma_i} \mathbf{w}_{i*}\mathbf{h} - \mathbf{b}^\top \mathbf{h}, \quad (1)$$

where v_i and a_i are the i-th elements in the visible variable vector $\mathbf{v}$ and bias vector $\mathbf{a}$, $\mathbf{h}$ and $\mathbf{b}$ are the hidden variable and bias vectors, $\mathbf{w}_{i*}$ is the i-th row vector of the weight matrix $\mathbf{W}$, and V is the number of units in the visible layer. $\mathbf{\Sigma} = \mathrm{diag}\{\sigma_1^2, \cdots \sigma_V^2\}$ is usually fixed to the diagonal covariance matrix of the training data [11]. The distribution described by an RBM can be written as

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

where $\mathcal{Z} = \sum_{\mathbf{h}} \int_{\mathbf{v}} \exp\{-E(\mathbf{v},\mathbf{h})\} d\mathbf{v}$ is the partition function. Unlike an RBM, there are no hidden layers in a BBAM.

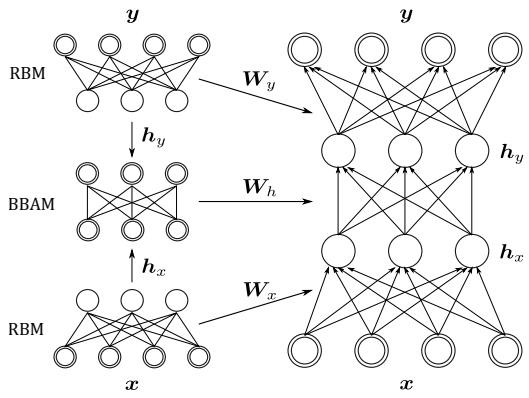

Figure 1: *Illustration of the structure and training procedure of proposed DNN.*

Similarly to RBMs, the distribution of a BBAM can be defined by an energy function:

$$E(\mathbf{v}_1, \mathbf{v}_2) = -\mathbf{b}_1^\top \mathbf{v}_1 - \mathbf{b}_2^\top \mathbf{v}_2 - \mathbf{v}_1^\top \mathbf{W} \mathbf{v}_2, \tag{3}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the binary variable vectors in the two visible layers, respectively, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the corresponding bias vectors.

During training, the two RBMs, θ_x and θ_y, are employed to model the distributions of synthesised and natural spectra, respectively. Then a BBAM θ_h is adopted to model the joint distribution of the hidden variables extracted from the two RBMs. The RBMs can be interpreted as feature extractors that extract high-order binary representations of spectral envelopes. The difference between synthesised and natural spectra is then modeled by the BBAM in the high-order space. Note that the RBMs can be replaced by deeper generative networks, e.g., deep belief networks (DBNs) [11, 12] or deep Boltzmann machines (DBMs) [13], to construct deeper mapping relationships.

2.2. Spectral enhancement

At the enhancement stage, the conditional distribution of $\mathbf{y}$ given a synthesised spectrum $\mathbf{x}$ can be derived from the DNN as:

$$P(\mathbf{y}|\mathbf{x}) \simeq \mathcal{N}(\mathbf{y}; \mathbf{\Sigma}_y^{\frac{1}{2}} (\mathbf{W}_y \mathbf{h}_y^* + \mathbf{a}_y), \mathbf{\Sigma}_y), \tag{4}$$

where $\mathbf{W}_y$, $\mathbf{a}_y$ and $\mathbf{\Sigma}_y$ are the parameters of the RBM θ_y and:

$$\mathbf{h}_y^* \sim P(\mathbf{h}_2|\mathbf{h}_x^*, \theta_h), \tag{5}$$

$$\mathbf{h}_x^* \sim P(\mathbf{h}_1|\mathbf{x}, \theta_x), \tag{6}$$

are samples of the hidden variables drawn from their conditional distributions. The conditional distribution is a Gaussian distribution with a diagonal covariance matrix. In this paper, we are using three consecutive frames of spectral envelopes as the input and output of the DNN. Therefore, the parameter generation algorithm in the HMM-based parametric speech synthesis method is adopted to generate enhanced spectral envelopes.

The proposed DNN-based enhancement can be time consuming since the model is applied directly to the high-dimensional spectra. To enhance a sentence of T frames, the computational complexity of this method is $O(NHT)$, where N is the dimensionality of the spectral envelope and H the number of units in each hidden layer. In this paper $N = 6147$ and $H = 2048$. The computational complexity of the GV method is $O(MKT)$, where M is the dimensionality of the spectral feature (e.g., mel-cepstrum) and K is the number of iterations for applying GV (note that $M \ll N$). In this paper $M = 60$ and $K = 1000$. If K and H are similar, the computational complexity of DNN is still about 100 times higher than that of GV which could be a limitation for using this method in real time.

3. Modulation spectrum (MS)

Short-term spectral analysis is one of the most predominant methods used in speech processing. Parameters that characterise the spectral envelopes can be derived in a number of ways (e.g., fast Fourier transform, linear prediction, cepstral analysis), and the changes in the glottal excitation and vocal tract shape are reflected in the temporal patterns of such parameters.

In the analysis of natural speech, the parameter trajectories of spectral coefficients exhibit rich modulation characteristics, whereas in statistical speech synthesis, the generated speech parameter trajectories are temporally over-smoothed due to the state-based statistical modelling and averaging thereof. The over-smoothing can be partly alleviated by speech parameter generation considering global variance (GV) [3]. This makes the overall scale of the generated trajectories more appropriate, but not their modulation characteristics (i.e. their spectral content). On the contrary, processing in the modulation spectrum (MS) domain, the frequency-dependent temporal modulations of the parameter trajectories can be enhanced [4].

The MS enhancement is studied in this paper for two reasons. First, it is a relatively new method [4], and evaluations comparing this method with other methods such as GV may still bring new information or confirm previous studies. Second, including the MS enhancement method in the comparison may yield information about the contributions of two separate aspects of speech quality: the spectral fine structure of speech and the spectral modulation in time. The proposed DNN-based method uses three consecutive frames as an input for modelling the output spectrum and should therefore be able to model modulation characteristics as the MS does.

3.1. Enhancement in the modulation spectrum domain

In this work, the spectrum of a speech frame is parametrised by the mel-cepstrum [14], resulting in a vector $\mathbf{c} = [c_1, c_2, \cdots, c_M]$ of length M, which is the order of the cepstral analysis. Short-term spectral analysis of a speech utterance yields a matrix $\mathbf{R} = [\mathbf{c}_1^\top, \mathbf{c}_2^\top, \cdots, \mathbf{c}_T^\top]$ of size $M \times T$, where T is the number of frames. The time trajectory of cepstral coefficient m is defined as $\mathbf{r}_m = [c_{m,1}, c_{m,2}, \cdots, c_{m,T}]$. The MS of trajectory $\mathbf{r}_m$ is defined as:

$$s_{m,f} = \log\left(|\mathcal{F}\{\mathbf{r}_m\}|\right), \tag{7}$$

where f is the modulation frequency bin, defined by the number of points in the Fourier analysis. The number of points in the FFT must be greater than the maximum number of frames T of an utterance. In order to evaluate the MS over a database, the MS of each utterance is evaluated for each coefficient. The MS statistics are assumed to be normally distributed:

$$s_{m,f} \sim \mathcal{N}\left(\mu_{m,f}, \sigma_{m,f}\right). \tag{8}$$

Fig. 2 illustrates the MS statistics $s_{m,f}$ of natural and synthetic speech over the whole speech database. We can see that synthetic speech has less modulated trajectories than natural speech. By modifying the MS of synthetic speech trajectories to be closer to the modulation characteristics of natural speech, the speech quality can be improved [4]. This can be done by the

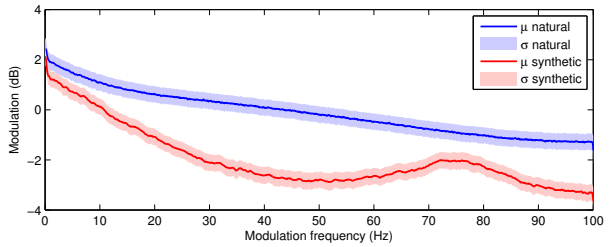

Figure 2: *Modulation spectra of natural and synthetic speech for the 16th mel-cepstral coefficient.*

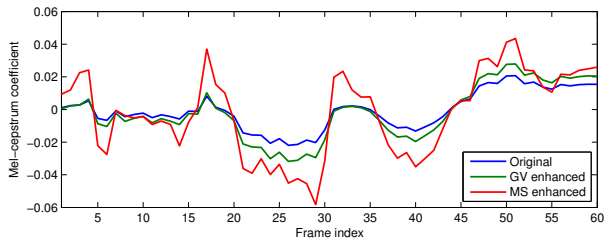

Figure 3: *Enhancing the 36th mel-cepstrum coefficient trajectory by global variance and modulation spectrum scaling.*

formula [4]:

$$s'_{m,f} = (1-\alpha)s_{m,f} + \alpha\left[\frac{\sigma^{(N)}_{m,f}}{\sigma^{(S)}_{m,f}}(s_{m,f} - \mu^{(S)}_{m,f}) + \mu^{(N)}_{m,f}\right], \tag{9}$$

where indices (N) and (S) indicate the parameters evaluated from natural and synthetic speech, respectively, and α defines the amount of shift from synthetic to natural MS. The enhanced trajectory is recovered by the inverse operation of Eq. 7 and preserving the original phase:

$$\mathbf{r}'_m = \mathcal{F}^{-1}\{e^{\mathbf{s}'_m + i\phi}\}, \tag{10}$$

where ϕ is the phase of the original parameter trajectory. Fig. 3 illustrates MS enhancement of a mel-cepstrum trajectory.

4. Evaluation

In this section we present the subjective quality evaluation[1]. First we describe the text-to-speech voice used in the experiments and the quality enhancement methods evaluated. We then present the design of the listening test and finally the results.

4.1. Voice and methods

The synthetic voice used in this evaluation was created from a high quality average voice model adapted to 2803 sentences recorded by a British male speaker, consisting of approximately 3 hours of material. All data was sampled at 48 kHz. We extracted the following acoustic features at a 5 ms shift: 59 mel-cepstral coefficients, mel scale F0 and 25 aperiodicity band energies extracted using STRAIGHT [15]. We used a hidden semi-Markov model as the acoustic model and the observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values, with one stream for the spectrum, three streams for F0 and one for the band-limited aperiodicity. Speech is synthesised in the frequency domain.

[1]Speech samples used in the evaluation can be found at: http://wiki.inf.ed.ac.uk/CSTR/Postfilter

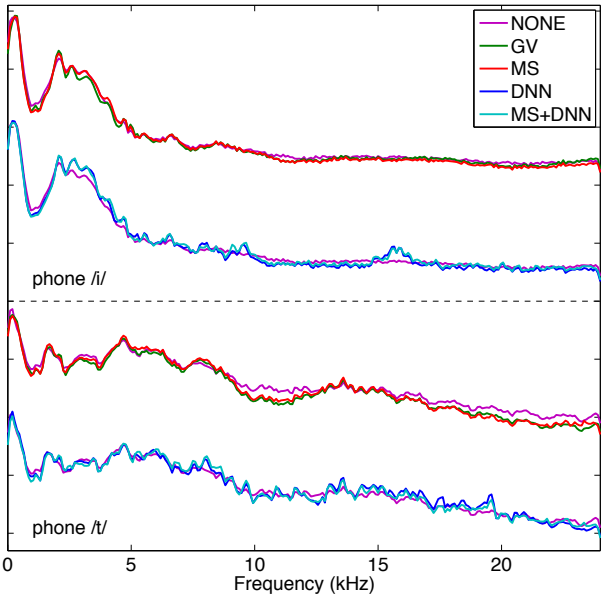

Figure 4: *Long term average spectrum of the synthetic instances of vowel /i/ (top) and consonant /t/ (bottom).*

Table 1 presents the four methods we evaluate. For the GV entry, we have applied the global variance method [3] to the mel-cepstral stream. We have also included a combination of the MS postfilter and DNN-based enhancement in order to assess the interaction between the two.

We trained a 4-layer DNN for the spectral enhancement. Three consecutive spectral envelopes were used as the input and output of the DNN. The FFT length for calculating spectral envelopes was set to 4096, which leads to $3 \times 2049 = 6147$ units in both input and output layers. The number of units in each of the two hidden layers was set to 2048. RBMs and BBAMs were trained using the contrastive divergence (CD) algorithm [16, 17]. The DNN was trained using paired synthetic and natural spectra aligned using dynamic time warping (DTW).

For the MS enhancement, the MS of the natural and the synthetic utterances were evaluated using Eqs. 7 and 8, i.e., for each file and each mel-cepstral coefficient trajectory, the MS was evaluated and statistics (mean μ and standard deviation σ) were estimated. In the Fourier analysis, 4096 points were used in order to exceed the maximum number of frames in an utterance in the database. After the evaluation the MS statistic, the synthetic trajectories were enhanced using Eq. 9. The value of α was set to 0.85 based on the findings in [4]. The MS enhanced mel-cepstra were then used for synthesising speech.

For the combination of MS and DNN enhancement, MS enhancement was first performed in the mel-cepstral domain, after which the mel-cepstrum was converted to spectrum. Then, a DNN was used to learn a mapping between the MS enhanced synthetic spectra and the natural spectra. Finally, DNN enhancement was applied to the MS enhanced mel-cepstra converted to spectra, and speech was synthesised from the enhanced spectra.

Table 1: *Methods evaluated.*

NONE	No enhancement
GV	Global variance [3]
DNN	Deep neural network
MS	Modulation spectrum [4]
MS+DNN	Combination of MS and DNN

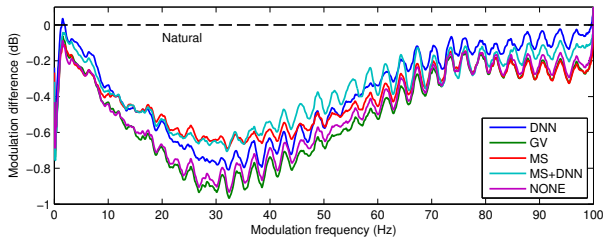

Figure 5: *Average difference in the modulation spectrum of different systems compared to natural speech.*

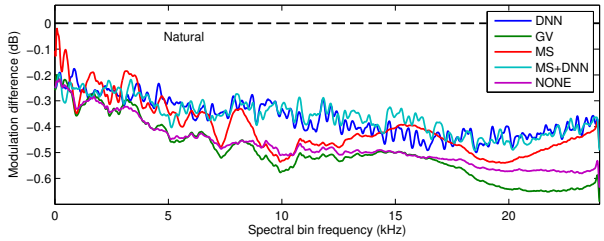

Figure 6: *Average difference in modulation per spectral bin of different systems compared to natural speech.*

4.2. Acoustic analysis

Fig. 4 shows the long term average spectrum calculated from instances of vowel /i/ and consonant /t/ across ten sentences generated by the five methods. We can clearly see that the proposed DNN-based method is more successful in reconstructing the spectral fine structure of speech compared to other methods.

Fig. 5 shows the difference in modulation spectrum with respect to natural speech for each method averaged across sentences. The figure shows that the baseline synthetic speech and GV have the least amount of modulation between 10 and 70 Hz, whereas the MS enhancement boosts modulations in that frequency range. The proposed DNN-based enhancement method increases modulation most at frequencies below 10 Hz and at high frequencies (60–100 Hz), and the combination of MS+DNN boosts most the mid-frequencies (40–60 Hz). However, there is still a large gap in modulation, most severe at 30 Hz, between all systems and natural speech.

Fig. 6 shows the difference in modulation relative to natural speech for each spectral bin. The modulation characteristics of unmodified synthesis (NONE) and GV are, as expected, mostly similar, but we can see that the MS enhancement boosts modulations in several frequency regions, most strongly at low frequencies. DNN-based methods, however, create a larger and more complex boost over all frequencies since they perform processing in the spectral domain instead of the mel-cepstral.

4.3. Listening experiment

We performed a listening experiment with 20 native English speakers. Participants were asked to rate the quality of synthesised sentences on a scale from 1 to 5, where a score of 5 is the best. Each participant rated 120 different sentences, which were divided equally across the five methods listed in Table 1. The experiment was designed so that across all participants, every sentence was rated for each method. The sentences were chosen from the first 12 sets of the Harvard dataset [18].

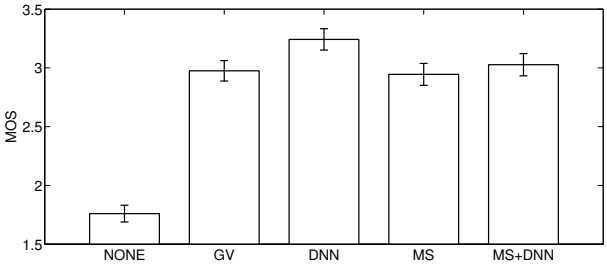

Figure 7: *Mean opinion scores obtained across 20 listeners.*

4.4. Results

We show the results in terms of mean opinion scores (MOS) calculated by averaging across all sentences and listeners. Fig. 7 presents the MOS obtained by each method and the 95% confidence intervals. We can see that all enhancement methods are rated significantly higher in quality than the unmodified baseline. The MS method and the MS+DNN combination reached quality scores as high as the scores obtained using GV. The DNN method scored significantly better than any other method.

5. Discussion and conclusions

Our results show that the DNN-based postfiltering method produced the highest quality synthetic speech. Possible reasons for this include:

- The DNN was trained in the spectral domain directly rather than in the mel-cepstrum domain, and was therefore able to learn the detailed spectral fine structures. The DNN was also able to learn the gap in speech dynamics between synthetic and natural speech in the spectral domain in a similar way to GV in the spectral domain [19].
- The DNN spectra are generated from an RBM trained on natural speech, which is equivalent to training a structured GMM that has a huge number of mixture components [12] (2^{2048} in this work). It is thus capable of capturing detailed spectral information. The patterns in spectra are analysed by RBMs and are represented by the binary hidden variables. The mapping between synthetic and natural spectra is learned by a BBAM in a binary hidden space since it is easier to analyse the differences in a binary space than a continuous one.
- The DNN can also learn modulation characteristics since it uses three consecutive frames for the mapping and because of a close relationship between DNN and MS. The FFT convolution is equivalent to the weighted sum in a network unit of the convolutional DNN [20], and the next deep layer of a DNN trained in the spectrum domain may therefore contain an MS-related representation. This also explains why the MS+DNN system results in the strongest modulation.

There is still a large difference in modulation spectrum between natural and synthetic speech centred around 30 Hz. Thus, our future work includes using deeper structures and more consecutive frames for DNN to better learn the modulation in a longer time period. As this paper studied the DNN as a speaker dependent postfilter, another future topic could be to study the DNN in a speaker independent fashion.

6. Acknowledgements

This work was supported by the EU FP7 (FP7/2007-13) project under grant agreement No. 287678 (Simple4All).

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge Workshop*, Pittsburgh, USA, September 2006.

[3] 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.

[4] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, May. 2014.

[5] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222 –2235, nov. 2007.

[6] L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Voice conversion using deep neural networks with multiple frame spectral envelopes," *Submitted to Interspeech*, 2014.

[7] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[8] P. Smolensky, "Information processing in dynamical systems: foundations of harmony theory," in *Parallel distributed processing: explorations in the microstructure of cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 194–281.

[9] B. Kosko, "Bidirectional associative memories," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 49–60, 1988.

[10] L.-J. Liu, L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Using bidirectional associative memories for joint spectral envelope modeling in voice conversion," in *Proc. ICASSP*, May. 2014.

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

[12] 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 Trans. on Audio, Speech and Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.

[13] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, University of Toronto, 2009.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, San Francisco, USA, March 1992, pp. 137–140.

[15] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.

[16] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 12, no. 14, pp. 1711–1800, 2002.

[17] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 599–619.

[18] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[19] Z.-H. Ling, Y. Hu, and L.-R. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis." in *Proc. Interspeech*, 2010, pp. 825–828.

[20] J. Andén and S. Mallat, "Deep scattering spectrum," *Submitted to IEEE transactions of Signal Processing*, 2013.