Statistical Parametric Speech Synthesis using Weighted Multi-distribution Deep Belief Network

Shiyin Kang and Helen Meng

Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
{sykang, hmmeng}@se.cuhk.edu.hk

Abstract

This paper presents a weighted multi-distribution deep belief network (wMD-DBN) for context-dependent statistical parametric speech synthesis. We have previously proposed the use of MD-DBN for speech synthesis, which models simultaneously both spectrum and fundamental frequency (F_0), and has demonstrated the potential to generate high-dimensional spectra with high quality and to produce natural synthesized speech. However, the model showed only mediocre performance on low-dimensional data, such as the F0 and voiced/unvoiced (V/UV) flag, resulting in a vibrating pitch contour in the synthesized voice. To address this problem, this paper investigates the use of an extra weighting vector on the acoustic output layer of the MD-DBN. It reduces the dimensional imbalance between spectrum and pitch parameters by giving different weighting coefficients to the spectrum, F0 and the V/UV flag in the training procedure. Experimental results show that wMD-DBN can generate smoother pitch contours and improve the naturalness of the synthesized speech.

Index Terms: speech synthesis, deep belief network, restricted Boltzmann machine

1. Introduction

Statistical parametric speech synthesis [1] has achieved great success in the past decade. The crux of this approach is to build an acoustic model that maps the textual content to the acoustic parameters. Previously, Hidden Markov Models (HMMs) was a dominant approach [2], where decision trees are used to cluster a large number of context-dependent HMM states into Gaussian Mixture Model (GMM) leaf nodes [3], and adopted as a regression model for acoustic parameter prediction. Recently, Multi-distribution Deep Belief Network (MD-DBN) has been investigated as an alternative acoustic model for speech synthesis [4]. In this approach, the MD-DBN models the joint distribution between the input symbolic representation based on text (e.g. syllables) and the output syllable-level acoustic parameters. These parameters consist of multiple frames of spectrum, F_0, and the voiced/unvoied (V/UV) flags. Experiments show that the spectrum generated by MD-DBN has less distortion, resulting in a clearer voice in the synthesized speech. At the same time, several teams also reported promising results on other variants of deep learning networks for speech synthesis, including Deep Neural Networks (DNNs) [5, 6], Restricted Boltzmann Machines (RBMs) [7, 8] and the DNN-Gaussian Process hybrid model [9]. These studies show that deep learning networks have certain advantages over the HMM-based approach [10]. First, high dimensional data with cross-dimension correlations can be well modeled by the deep learning networks, which bypass the independence assumption that often introduced by the use of GMM with diagonal covariance matrix in the HMM-based approach. Thus the deep learning networks perform well on spectrum modelling, and can also handle the correlation between the spectrum and F_0 when they are modeled simultaneously. Second, all training data are modelled in a centralized network so as to avoid data fragmentation [11]. Instead of using thousands of GMM leaf nodes to piece the acoustic space together as in the HMM-based approach, only one deep learning network is used to portray the whole acoustic space, which potentially reduces the training data requirements and increases the efficiency of model parameters.

However, these attempts have not yet achieved satisfactory performance on F_0 modelling. It can be seen that the one-dimensional F_0 does not contribute to the model as much as the high-dimensional spectrum data in the training procedure. Furthermore, the use of the weight decay regularizer during training [12] also tends to equalize the the learned weights for each dimension. As a result, the generated F_0 contour is more noisy and degrades the quality of the synthesized voices.

To address this problem, this paper presents a novel, weighted MD-DBN (wMD-DBN) for speech synthesis. The enhancement is achieved through the use of a weighting vector added to the visible layer of the MD-DBN, so that the weight of spectrum, F_0 and the V/UV flags can be controlled individually. The basic idea is simple – to reduce the dimensional imbalance between the spectrum and F_0 by duplicating the F_0 elements in the acoustic vector until the number of dimensions allocated for F_0 is equal to the number of dimensions of the spectrum. Duplicating the number of dimensions of F_0 is equivalent to giving F_0 a positive integer weight.

The rest of the paper is organized as follows. The mathematical details of wMD-DBN are given in Section 2. The wMD-DBN-based speech synthesis framework is described in Section 3. Section 4 reports the experimental results, and Section 5 gives the conclusions.

2. Weighted Multi-distribution Deep Belief Network

A wMD-DBN is a specialized Deep Belief Network (DBN) [12] designed for modeling text contextual factors and speech acoustic parameters in speech synthesis tasks. It differs with the standard DBN in two aspects: 1) a multi-distribution visible layer is used to model data with different distributions jointly, i.e., the spectrum and the log F_0 with assumed Gaussian distribution, the V/UV flags with binary distribution (i.e., the Bernoulli distribution) and the phoneme identities with *Categorical* distribution (i.e., the generalized Bernoulli distribution); 2) an extra weighting vector is attached to the visible layer to give individual control on each dimension. In acoustic modeling, the wMD-DBN allows balanced control between the spectrum and F_0, where the spectrum usually has many more dimensions and tends to take up more weight during the model learning process. There is a greedy layer-wised learning algorithm [12] for estimating the connection weights in a DBN, which is equivalent to training each adjacent layer pair as an Restricted Boltzmann Machine (RBM)[13] from bottom-up.

An RBM is an undirected graphical model with one layer of stochastic visible units and one layer of stochastic hidden units. There is no direct interaction between units in the same layer and is thus "restricted". Here we generalize the three types of RBMs described in [4] into an unified Weighted Multi-distribution RBM (wMD-RBM) as the building blocks of the wMD-DBN. In the wMD-RBM, the visible layer units $\boldsymbol{v} = [\boldsymbol{v}^{g\mathrm{T}}, \boldsymbol{v}^{b\mathrm{T}}, \boldsymbol{l}^{c\mathrm{T}}]^{\mathrm{T}}$ have three different distributions, including Gaussian units $\boldsymbol{v}^g$, binary units $\boldsymbol{v}^b$ and Categorical units $\boldsymbol{l}^c$, and the hidden layer units $\boldsymbol{h}$ are binary. Accordingly the weighting vector is defined as $\boldsymbol{q} = [\boldsymbol{q}^{g\mathrm{T}}, \boldsymbol{q}^{b\mathrm{T}}, \boldsymbol{q}^{c\mathrm{T}}]^{\mathrm{T}}$. Assuming the Gaussian distribution has zero mean and unit variance, the "energy" of a visible-hidden configuration $(\boldsymbol{v}, \boldsymbol{h})$ in a wMD-RBM is defined by

$$\begin{aligned} E(\boldsymbol{v}, \boldsymbol{h}; \Theta) = &- \boldsymbol{h}^{\mathrm{T}} \boldsymbol{W}^g \boldsymbol{Q}^g \boldsymbol{v}^g \\ &+ \frac{1}{2} (\boldsymbol{Q}^g \boldsymbol{v}^g)^{\mathrm{T}} \boldsymbol{Q}^g \boldsymbol{v}^g \\ &- \boldsymbol{h}^{\mathrm{T}} \boldsymbol{W}^b \boldsymbol{Q}^b \boldsymbol{v}^b - \boldsymbol{a}^{b\mathrm{T}} \boldsymbol{Q}^b \boldsymbol{v}^b \\ &- \boldsymbol{h}^{\mathrm{T}} \boldsymbol{W}^c \boldsymbol{Q}^c \boldsymbol{l}^c - \boldsymbol{a}^{c\mathrm{T}} \boldsymbol{Q}^c \boldsymbol{l}^c \\ &- \boldsymbol{b}^{\mathrm{T}} \boldsymbol{h}, \end{aligned} \tag{1}$$

where $\boldsymbol{Q} = \mathrm{diag}\left[\boldsymbol{Q}^g, \boldsymbol{Q}^b, \boldsymbol{Q}^c\right] = \mathrm{diag}\left[\boldsymbol{q}\right]$ is the weighting diagonal matrix defined by the weighting vector $\boldsymbol{q}$; $\Theta = \{\boldsymbol{W}^g, \boldsymbol{W}^b, \boldsymbol{W}^c, \boldsymbol{a}^b, \boldsymbol{a}^c, \boldsymbol{b}\}$ is the set of the model parameters; $\boldsymbol{W}^g, \boldsymbol{W}^b, \boldsymbol{W}^c$ are the weight matrices of the symmetric connection between the hidden layer units $\boldsymbol{h}$ and the Gaussian visible layer units $\boldsymbol{v}^g$, the Bernoulli visible layer units $\boldsymbol{v}^b$ and the *Categorical* visible layer units $\boldsymbol{l}^c$; $\boldsymbol{a}^b$, $\boldsymbol{a}^c$ and $\boldsymbol{b}$ are their bias terms. The joint distribution of the configuration $(\boldsymbol{v}, \boldsymbol{h})$ is given as

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

where

$$Z = \int_{\boldsymbol{v}^g} \sum_{\boldsymbol{v}^b} \sum_{\boldsymbol{l}^c} \sum_{\boldsymbol{h}} \exp\{-E(\boldsymbol{v}^g, \boldsymbol{v}^b, \boldsymbol{l}^c, \boldsymbol{h})\} d\boldsymbol{v}^g \tag{3}$$

is the partition function.

Given a training data set, the model parameters Θ of the wMD-RBM can be estimated based on the maximum likelihood criterion by stochastic gradient descent algorithm [14]. To reduce the computational cost, the contrastive divergence (CD) algorithm has been proposed by using Gibbs Sampling [15]. The conditional probabilities used by Gibbs sampling can then be derived as:

$$\begin{aligned} &P(h_j = 1 | \boldsymbol{v}) = \\ &\sigma(\sum_i w^g_{ij} q^g_i v^g_i + \sum_i w^b_{ij} q^b_i v^b_i + \sum_i w^c_{ij} q^c_i l^c_i + b_j), \end{aligned} \tag{4}$$

$$P(v^g_i | \boldsymbol{h}) = \mathcal{N}\left(v^g_i; \sum_i w^g_{ij} h_j, 1\right), \tag{5}$$

$$P(v^b_i = 1 | \boldsymbol{h}) = \sigma\left(\sum_j w_{ij} h_j + a^b_i\right), \tag{6}$$

$$P(l^c_i = 1 | \boldsymbol{h}) = \frac{\exp\left(\sum_j w^c_{ij} h_j + a^c_i\right)}{\sum_k \exp\left(\sum_j w^c_{kj} h_j + a^c_k\right)}, \tag{7}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

is a sigmoid function, $\mathcal{N}(\cdot; \mu, \sigma)$ denotes the Gaussian distribution with the mean μ and the variance σ. It can be seen that the weighting vector $\boldsymbol{q}$ only exists in the bottom-up inference direction of the Gibbs sampling process. In the top-down generation direction, the conditional probabilities are same as in MD-DBN training.

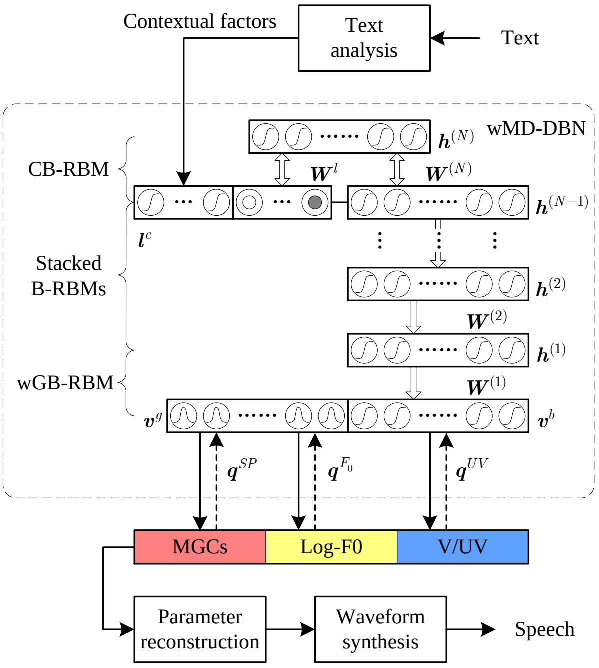

Figure 1: Architecture of the wMD-DBN for speech synthesis.

3. wMD-DBN-based Speech Synthesis

Figure 1 illustrates the speech synthesis framework using wMD-DBN. The text-to-acoustic mapping is achieved by modeling the joint distribution between the input contextual factors and the output acoustic parameters with wMD-DBN. For each Mandarin syllable, 39 commonly used contextual factors [16] are used to form the input contextual vector $\boldsymbol{l}^c$. 21 of them are symbolic identities (e.g. initial, final and tone of syllable).

They are modeled by the Categorical distribution units. 18 of them are numerical features (e.g. position of the syllable in current word, number of words in the previous phrase). They are normalized to the range of $(0.03, 0.97)$ and modeled by the Bernoulli distribution units. On the output side, the acoustic parameters of each syllable are represented by a 2700-dimensional super-vector $[\boldsymbol{v}^g, \boldsymbol{v}^b]$, which is concatenated by 100 uniformly-spaced frames of 24-order mel-cepstral coefficients (MCEP) [18] with log-energy, log-F_0, and the V/UV flags extracted from within the syllable. The log-F_0 values in the unvoiced region are interpolated. Both the MCEPs and the log-F_0 are normalized to have zero-mean and unit-variance before training the wMD-DBN.

The training procedure is similar to the training of MD-DBN [4], except for the bottom layer weighed Gaussian-Bernoulli RBM (wGB-RBM), where the acoustic super-vector is weighted by the weighting vector $[\boldsymbol{q}^{SP}, \boldsymbol{q}^{F_0}, \boldsymbol{q}^{UV}]$ in the bottom-up inference step as shown in the equation 4. After the wGB-RBM is trained, we can stack up as many layers of Bernoulli RBMs (B-RBMs) as we want to construct the deep architecture. At last, a Categorical-Bernoulli RBM (CB-RBM) is used to model the joint distribution between contextual factors $\boldsymbol{l}^c$ and the acoustic parameter transformations $\boldsymbol{h}^{(N-1)}$.

At synthesis time, the contextual factors $\boldsymbol{l}^c$ are first determined for each syllable by the text analysis module. Then alternative Gibbs sampling using $P(h_i^{(N)} = 1|\boldsymbol{l}^c, \boldsymbol{h}^{(N-1)})$ and $P(h_j^{(N-1)} = 1|\boldsymbol{h}^{(N)})$ are conducted to update $\boldsymbol{h}^{(N-1)}$ while keeping the input $\boldsymbol{l}^c$ clamped, until convergence or a maximum number of iterations is reached. Then, the acoustic super-vector $[\boldsymbol{v}^g, \boldsymbol{v}^b]$ are generated with a single down-pass from $\boldsymbol{h}^{(N-1)}$ to $\boldsymbol{h}^{(1)}$. The MCEPs and F_0 are recovered by the global mean and variance. Finally, the generated acoustic features are interpolated according to the predicted syllable durations and are sent into the Mel Log Spectral Approximation (MLSA) filter [18] to reconstruct the speech waveforms. Note that the weighting vector $[\boldsymbol{q}^{SP}, \boldsymbol{q}^{F_0}, \boldsymbol{q}^{UV}]$ is not used in the synthesis procedure.

4. Experiments

4.1. Experiment Setup

A manually transcribed Mandarin corpus recorded from a female speaker is used for the experiments. The training set contains 1,000 utterances with a total length of 80.9 minutes, including 23,727 syllable samples. Another test set with 100 utterances is used for model architecture optimization and the evaluations.

A HMM-based speech synthesis system is built as the evaluation baseline using a standard recipe [16]. 24-order MCEPs and log-F_0 together with their Δ and Δ^2 are modeled by the multi-stream MSD-HMMs. Each syllable HMM has a left-to-right topology with 10 states. Initially, 416 mono-syllable HMMs are estimated as the seed for training the context-dependent HMMs. The candidate question set for the decision tree-based context clustering contains 2596 effective questions (i.e., the questions which can divide the parameter space into 2 non-empty sub-spaces). In the synthesis stage, speech parameters including MCEPs and log-F0 are generated obtained by the maximum likelihood parameter generation algorithm [19], and are later used to synthesis the speech waveform using the MLSA vocoder. The duration predicted by the HMM baseline are used for all the synthesis systems.

Two DBN-based systems are built for the evaluation. One uses the proposed wMD-DBN model, the other one uses the MD-DBN without the extra weighting control. Both of these two systems use the same input contextual factors and the output acoustic parameters as described in the previous section. The RBMs are trained using CD algorithm with a mini-batch size of 200 training samples. The weighting vectors are set to 0.32, 4.0 and 4.0 for the spectrum, F_0 and the V/UV flags respectively. This setting is trying to balance the data dimensions between the spectrum $(800 = 0.32 \times 2500)$ and the excitation including F_0 and V/UV flags $(800 = 4.0 \times (100 + 100))$. For wGB-RBMs and GB-RBMs 400 epochs are executed with a learning rate of 0.01 while for B-RBMs and CB-RBM 200 epochs are executed with a learning rate of 0.05. During the weight updates, we apply a momentum of 0.9 and a weight decay of 0.001. The model parameters are obtained from the mean of the final 20 epochs. The training procedure is accelerated by the NVIDIA CUDA BLAS library on one Tesla K20 GPU. For an wMD-DBN with 5 hidden layers and 2000 units per layer, the training takes about one hour. Since it eliminates the serialized decision tree training procedure as in the HMM system, the wall time of training a wMD-DBN is actually shorter using the proper parallel computing hardware.

4.2. Objective Evaluation

In the objective evaluation, we investigates the relationship between the performance of the acoustic parameter prediction and the architecture of the wMD-DBN, including the number of the hidden layers and the number of the units in each layer. This investigation also helps us to find the optimized model architecture for the following subjective evaluation.

Figure 2 shows the MCEP distortion of the synthesized speech. Measuring the spectral distortion [20] is a commonly-used method for objective evaluation in voice conversion as well as in speech synthesis. Here we measure the Euclidean distance between the MCEPs of synthesized speech and that of original speech recording. The two speech samples are aligned using the dynamic time wrapping algorithm (DTW). The MCEP distortion of the HMM baseline in this evaluation is 9.0 dB, while the wMD-DBN approach achieves better result with a minimal distortion of 8.7 dB. It can be seen that the distortion tends to decrease with more hidden layers in the model.

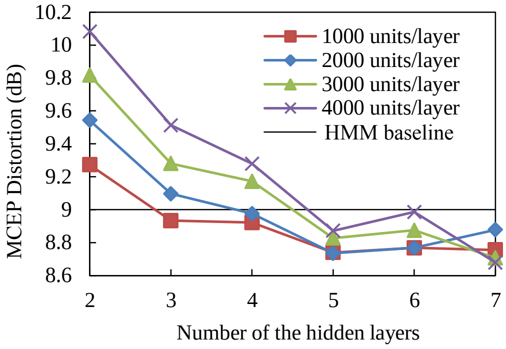

Figure 2: The distortions (dB) of the MCEPs predicted by the wMD-DBN-based systems. The MCEP distortion of the HMM baseline systems is 9.0 dB.

Figure 3 gives the F_0 contours predicted by the three systems. It can be seen that with the weighting control, wMD-DBN can generate a more smooth F_0 contour compared to the previous MD-DBN approach. Both the wMD-DBN and the

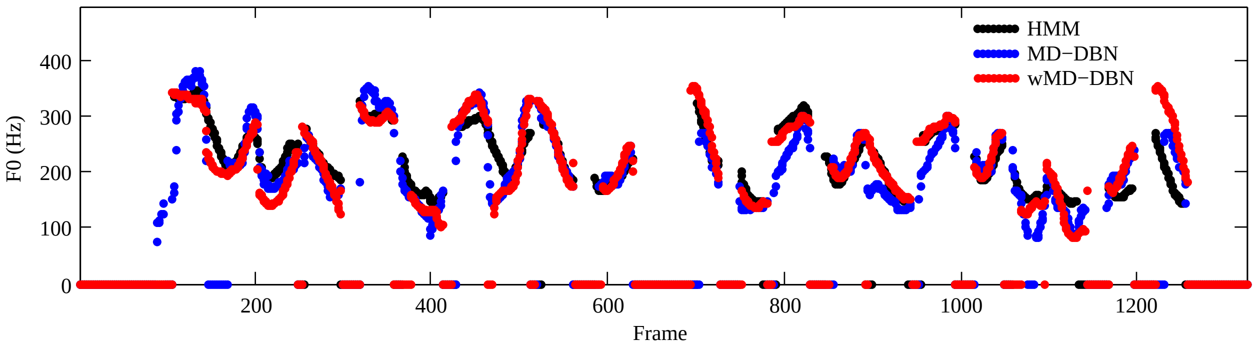

Figure 3: F_0 contour predicted by the wMD-DBN, MD-DBN and HMM systems.

HMM approaches give the reasonable results. The dynamic range of the F_0 from wMD-DBN approach is slightly larger, which may result in a more lively intonation.

The evaluation results for the root mean squared error (RMSE) in F_0 are shown in Figure 4. It can be seen that the RMSE in F_0 from the two systems are close. (i.e, 25.4 for HMM baseline v.s. a minimum of 25.6 for wMD-DBN approach).

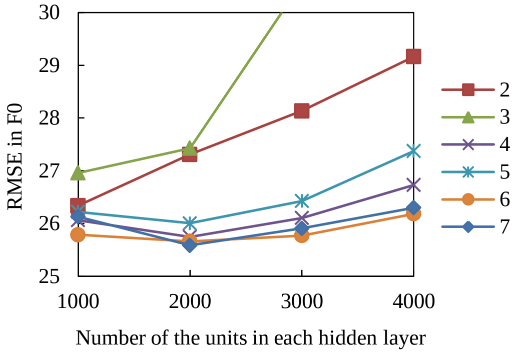

Figure 4: The RMSE of the F_0 predicted by the wMD-DBN-based systems. The number besides each line denotes the number of the hidden layer. The F_0 RMSE of the HMM baseline systems is 25.4.

4.3. Subjective Evaluation

In the subjective evaluation, we first conduct a preference test between the MD-DBN approach and the proposed wMD-DBN approach. 8 experienced listeners are asked to give preference on 15 pairs of the synthesized speech samples generated from these two systems. The samples are played with high quality headphones in quiet office room.

The result preference scores (%) are shown in Table 1. It can be seen that the wMD-DBN-based approach are preferred significantly compared to the MD-DBN approach. Testing feedbacks suggest that a more fluent voice contributes more to the preference.

wMD-DBN	MD-DBN	N/P	p value
47.5	7.5	45.0	$< 10^{-6}$

Table 1: Preference scores (%) between the synthesized speech samples from the wMD-DBN and the MD-DBN-based systems, where N/P denotes “No Preference”. The system which achieves significantly better preference at $p < 0.01$ level are in the bold font.

A Mean Opinion Scoring (MOS) test is also conducted to compare the subjective perception among the three approaches. In this test, 10 experienced listeners are asked to rate the naturalness of 15 utterances using a five-point scale (i.e. 5:excellent, 4:good, 3:fair, 2:poor, 1:bad). The result is shown in Table 2. It can be seen that there is a clear naturalness improvement between the MD-DBN and the proposed wMD-DBN approach, and the performance of the wMD-DBN approach is similar to that of the context-dependent HMM approach.

System	MOS	95% CI
wMD-DNM	3.21	±0.09
MD-DBN	2.93	±0.12
HMM	3.27	±0.14

Table 2: MOS results with the 95% confidence interval (CI), showing that the naturalness of wMD-DBN-based system is similar to that of the context-dependent HMM baseline.

5. Conclusions

This paper investigates the use of an extra weighting vector on the MD-DBN, and proposes the wMD-DBN for speech synthesis. The wMD-DBN tries to provide balanced control for different streams in the acoustic parameter such as the spectrum and F_0. Both the objective evaluation and the subjective evaluation show a performance upgrade by giving F_0 and V/UV flags a higher weight during the model training procedure, and the performance of the wMD-DBN approach is similar to the context-dependent HMM approach. Future directions for the improvement may include applying weighting control to the input contextual factors in order to achieve better prosody.

6. Acknowledgements

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 415511).

7. References

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

[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 *Eurospeech*, 1999, pp. 2347–2350.

[3] J. J. Odell, "The use of context in large vocabulary speech recognition," 1995.

[4] S.-Y. Kang, X.-J. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.

[5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[6] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," in *Proc. ISCA SSW8*, 2013, pp. 261–265.

[7] 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.

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

[9] 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.

[10] H. Zen, "Deep learning in speech synthesis," *Keynote speech given at ISCA SSW8*, 2013.

[11] K. Yu, H. Zen, F. Mairesse, and S. Young, "Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis," *Speech Communication*, vol. 53, no. 6, pp. 914–923, 2011.

[12] 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.

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

[14] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 116.

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

[16] Z. Shuang, S. Kang, Q. Shi, Y. Qin, and L. Cai, "Syllable HMM based mandarin TTS and comparison with concatenative TTS," in *INTERSPEECH*, 2009, pp. 1767–1770.

[17] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *ICSLP*, 1994.

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

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

[20] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.