A Deep Generative Architecture for Postfiltering in Statistical Parametric Speech Synthesis

Ling-Hui Chen, Tuomo Raitio, Cassia Valentini-Botinhao, Zhen-Hua Ling, *Member, IEEE*, and Junichi Yamagishi, *Senior Member, IEEE*

***Abstract*—The generated speech of hidden Markov model (HMM)-based statistical parametric speech synthesis still sounds "muffled." One cause of this degradation in speech quality may be the loss of fine spectral structures. In this paper, we propose to use a deep generative architecture, a deep neural network (DNN) generatively trained, as a postfilter. The network models the conditional probability of the spectrum of natural speech given that of synthetic speech to compensate for such gap between synthetic and natural speech. The proposed probabilistic postfilter is generatively trained by cascading two restricted Boltzmann machines (RBMs) or deep belief networks (DBNs) with one bidirectional associative memory (BAM). We devised two types of DNN postfilters: one operating in the mel-cepstral domain and the other in the higher dimensional spectral domain. We compare these two new data-driven postfilters with other types of postfilters that are currently used in speech synthesis: a fixed mel-cepstral based postfilter, the global variance based parameter generation, and the modulation spectrum-based enhancement. Subjective evaluations using the synthetic voices of a male and female speaker confirmed that the proposed DNN-based postfilter in the spectral domain significantly improved the segmental quality of synthetic speech compared to that with conventional methods.**

***Index Terms*—Deep generative architecture, hidden Markov model (HMM), modulation spectrum, postfilter, segmental quality, speech synthesis.**

I. Introduction

STATISTICAL parametric speech synthesis is one of the most popular methods of speech synthesis due to its flexibility and compact footprint [2]. Statistical parametric speech synthesizers have also been found to be as intelligible as natural human speech several times at the annual evaluation events of corpus-based speech synthesis systems called "Blizzard Challenge" [3]. It is known, however, that synthesized speech generated from statistical models still sounds "muffled" compared to natural speech. 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 improving segmental quality.

Deep neural networks (DNNs) with many hidden layers have been actively investigated to improve the quality of synthetic speech and several significant improvements have been reported. For instance, DNNs have been applied to acoustic modeling [4]. Zen *et al.* [5] used DNN to learn the relationship between input texts and extracted features instead of using decision tree-based state tying. Restricted Boltzmann machines (RBMs) or deep belief networks (DBNs) have been used to model the output probabilities of hidden Markov model (HMM) states instead of Gaussian mixture models (GMMs) [6]. DBNs have also been used to model the joint distribution of linguistic and acoustic features [7]. A hybrid model which combines a DBN with an Gaussian process regression has been used for F0 modeling [8]. In addition, an auto-encoder neural network has also been used to extract low dimensional excitation parameters [9]. Recently, recurrent neural networks (RNNs) with long-short term memories (LSTMs) have been used for prosody modeling [10] and acoustic trajectory modeling [11], [12].

In addition to these above improvements to acoustic modeling, there have also been several successful attempts to improve the segmental quality of synthesized speech *at synthesis time (without changing the acoustic models)*, including postfiltering to enhance spectral peaks [13], [14] and a global variance (GV) parameter generation algorithm that enhances the dynamics within a speech utterance [15]. An interesting approach based on the enhancement of the modulation spectrum (MS) has recently been proposed [16]. The main aim of this method is to enhance the natural frequency modulation in the spectral parameter trajectories. These methods have been demonstrated to improve the quality of synthetic speech based on empirical findings of acoustic differences between natural and synthetic speech, which tend to occur for most speakers.

Another possible way of reducing the gap between the segmental quality of natural and synthetic speech is to learn acoustic differences directly from data. If we have a parallel set of natural and synthetic speech, we can estimate the conditional probability of acoustic differences, i.e., the probability of natural speech given "muffled" synthetic speech. One could

Manuscript received March 30, 2015; revised July 12, 2015; accepted July 18, 2015. Date of publication July 28, 2015; date of current version July 31, 2015. This work was supported by the EU FP7 (FP7/2007-13) project under Grant 287678 (Simple4All), by the Academy of Finland under Grants 256961 and 284671, by the National Natural Science Foundation of China under Grant 61273032), and by EPSRC through Programme Grants EP/I031022/1 (NST) and EP/J002526/1 (CAF). Part of this work was presented at Interspeech 2014 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sin-Horng Chen.

L.-H. Chen is with the National Engineering Laboratory for Speech and Language Information Processing (NELSLIP), University of Science and Technology of China, Hefei 230027, China, and also with iFLYTEK Co., Ltd., Hefei 230088, China (e-mail: chenlh@mail.ustc.edu.cn).

T. Raitio is with Department of Signal Processing and Acoustics, Aalto University, FI-02015 TKK Espoo, Finland (e-mail: tuomo.raitio@aalto.fi).

C. Valentini-Botinhao and J. Yamagishi are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh EH8-9AB, U.K. (e-mail: cvbotinh@inf.ed.ac.uk; jyamagis@inf.ed.ac.uk).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2461448

then model and reconstruct the spectral fine structures through data-driven statistical methods. This is conceptually similar to voice conversion (VC) techniques that take into consideration the conditional probability of parallel speaker pairs [17].

This paper introduces a deep generative architecture as a post-filter [1] to model the conditional probability of acoustic differences. The proposed architecture is a DNN with layer-wise generative training[1]. In voice conversion [18] this is typically done with a Gaussian mixture model (GMM) but a DNN was chosen here instead due to its abilities to handle highly correlated and high-dimensional data, allowing us to conduct spectral shaping directly in the spectral amplitude domain. We compared the proposed method with the GV and the recently proposed MS enhancement as well as the normal spectral peak enhancement filter.

This paper is organized as follows: in Section II we overview the related techniques, and in Section III we explain the proposed DNN-based approach. The experimental conditions and evaluation results are shown in Section IV. Analysis and discussions on the proposed DNN-based postfilter and its relation to other postfilter methods are given in Section V, and the summary of our findings is given in Section VI.

II. Related Techniques

A. Mel-cepstral Postfilter

Statistical averaging of parameters creates trajectories that are overly smooth across frames in the time domain but also within a frame in the spectrum domain. One of the first postfilter techniques applied to statistically generated speech trajectories appeared in [14]. The method was originally presented in [19] to enhance the formant structure in speech coding, but it can also be used to compensate for the overly smooth spectrum in speech synthesis. The method works by modifying the generated mel-cepstral coefficients so that spectrum peaks and valleys are enhanced. The postfilter is controlled by a single parameter, referred to as β. When $\beta = 0$, no postfilter is applied and the degree of formant enhancement increases with increasing β. A similar postfilter for line spectral pairs was also proposed in [13].

B. Global Variance

Another method frequently used for improving the quality of synthetic speech is a parameter generation algorithm that considers GV [15]. In the GV parameter generation algorithm, we define an objective function including HMM's likelihood and a penalty term that reflects the dynamic range of each dimension of the parameter sequence at the utterance level. This penalty term is intended to keep the variance of the generated trajectory as wide as that of the natural speech, while maintaining an appropriate parameter sequence in the sense of maximum likelihood [15]. An extended algorithm that calculates GV in the spectral domain has also been investigated [20].

C. Modulation Spectrum

Short-term spectral analysis is one of the most widely used methods in speech processing. Parameters that characterise the spectral envelopes can be derived in a number of ways, e.g., using fast Fourier transform (FFT), linear prediction, or cepstral analysis, and the changes in the vocal tract shape and also the glottal excitation 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 modeling and averaging thereof [2], [21]. The over-smoothing can be partly alleviated, for example, by using the aforementioned mel-cepstral postfilter [19] or GV [15]. The latter forces the variance of the generated parameter trajectories closer to the variance observed in parameter trajectories of natural speech, but it does not explicitly modify the frequency-dependent modulation characteristics (i.e., the spectral content) of the trajectories. On the contrary, processing in the modulation spectrum (MS) domain, the frequency-dependent temporal modulations of the parameter trajectories can be explicitly enhanced [1], [16].

Enhancement in the modulation spectrum domain was first proposed in [16], and it was also studied in our earlier work [1], which confirmed the results in [16] that the MS enhancement has approximately an equal effect to the quality as GV enhancement.

In this work, we apply the MS enhancement in the mel-cepstral domain (although MS enhancement can be also performed in the high-dimensional spectrum domain). The spectrum of a speech frame is parametrized by the mel-cepstrum [22], resulting in a vector $\boldsymbol{c} = [c_1, c_2, \cdots, c_M]$ of length M, which is the order of the cepstral analysis. Short-term spectral analysis of a speech utterance thus yields a matrix $\boldsymbol{R} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_T^\top]$ of size $M \times T$, where T is the number of frames. The time trajectory of the mth mel-cepstrum is defined as $\boldsymbol{r}_m = [c_{m,1}, c_{m,2}, \ldots, c_{m,T}]$. Finally, the MS of trajectory $\boldsymbol{r}_m$ is defined as:

$$s_{m,f} = \log\left(|\mathcal{F}\{\boldsymbol{r}_m\}|\right), \tag{1}$$

where f is the modulation frequency bin, defined by the number of points in the Fourier analysis used in Eq. (1). The number of points in the Fourier analysis in Eq. (1) must be greater than the 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:

$$s_{m,f} \sim \mathcal{N}\left(\mu_{m,f}, \sigma_{m,f}\right). \tag{2}$$

Fig. 1 illustrates the MS statistics $s_{m,f}$ of natural and synthetic speech over a large 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 [1], [16]. This can be done by the formula [16]:

$$s'_{m,f} = (1-\alpha)s_{m,f} + \alpha\left\{\frac{\sigma_{m,f}^{(N)}}{\sigma_{m,f}^{(S)}}(s_{m,f} - \mu_{m,f}^{(S)}) + \mu_{m,f}^{(N)}\right\}, \tag{3}$$

[1]In the rest of this paper, the proposed deep generative architecture is called DNN for simplification.

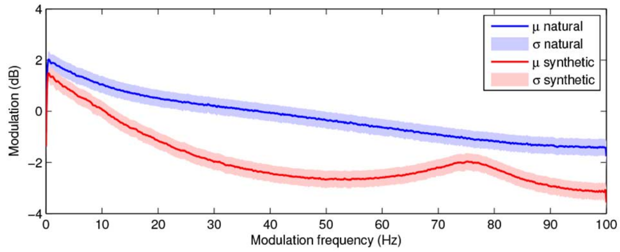

Fig. 1. Modulation spectra of the 16th mel-cepstral coefficient estimated from natural speech and generated from a statistical model.

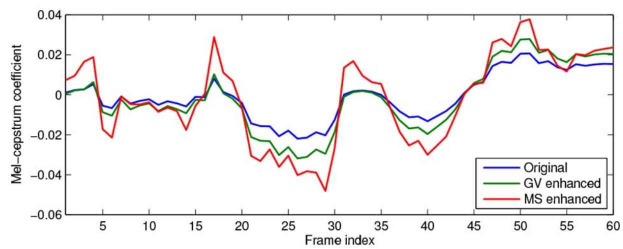

Fig. 2. Illustration of enhancing the 36th mel-cepstral coefficient trajectory by variance scaling (equal scaling across different modulation frequencies) and MS enhancement that can modify the frequency-dependent modulation characteristic of speech.

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. (1) and preserving the original phase:

$$\mathbf{r}'_m = \mathcal{F}^{-1}\{e^{\mathbf{s}'_m + i\phi}\}, \tag{4}$$

where ϕ is the phase of the original parameter trajectory. Fig. 2 illustrates MS enhancement of a mel-cepstrum trajectory.

III. DNN-BASED PROBABILISTIC POSTFILTER

In Section II, we introduced several frequently used postfiltering techniques for enhancing the segmental quality of synthetic speech. However, these techniques were proposed based on empirical findings on the acoustic differences between the spectral features of synthetic and natural speech. There are various acoustic differences between natural and synthetic speech, but each of these techniques mostly deals with only one specific aspect.

In this paper, we proposed a probabilistic postfilter to automatically discover and compensate the acoustic differences observed in the spectral domain. The postfilter is similar to VC in the sense that it converts synthetic spectral features into natural spectral features. However, the conventional approaches for VC, such as the ones based on GMMs and conventional neural networks (NN) [23], still suffer from the over-smoothing problem caused by the statistical averaging of the underlying model. Recently, we have proposed a generatively trained DNN for spectral conversion in VC [24], [25] and showed that it can significantly improve the segmental quality of generated speech. In this paper, we extend this approach to spectral postfiltering for HMM-based parametric speech synthesis.

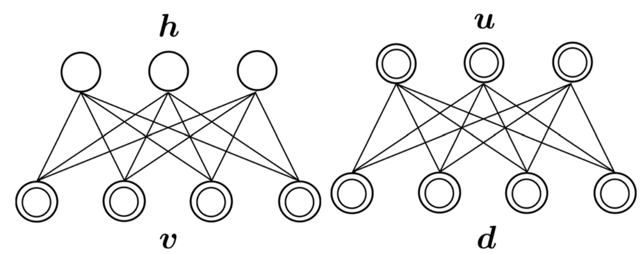

Fig. 3. The graphical model representations for an RBM (left) and a BAM (right). The double circles represent visible units while the single circles represent hidden units.

A. Basic Components

The proposed DNN is composed by three types of generative neural networks: restricted Boltzmann machine (RBM) [26], deep belief network (DBN) and bidirectional associative memory (BAM) [27].

1) Restricted Boltzmann Machine: An RBM is a two layered generative neural network, including a visible layer and a hidden layer, which correspond to visible random variable $\boldsymbol{v}$ and hidden random variable $\boldsymbol{h}$ as can be seen from the left of Fig. 3. Units between different layers are fully connected and there are no connections between units in the same layer.

An RBM is an undirected graphical model that describes a probabilistic distribution defined by an energy function. We assumed that it would obey a Gaussian distribution to model spectral features and hence the Gaussian-Bernoulli RBM (GBRBM) was used. The energy function of a GBRBM is given by

$$E_{\mathrm{RBM}}(\boldsymbol{v},\boldsymbol{h}) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^{V} \frac{v_i}{\sigma_i}\boldsymbol{w}_{i*}\boldsymbol{h} - \boldsymbol{b}^\top \boldsymbol{h}, \tag{5}$$

where v_i is the ith element in the visible random variable vector $\boldsymbol{v}$ and a_i is that in bias vector $\boldsymbol{a}$. Here $\boldsymbol{h}$ is the hidden variable vector, $\boldsymbol{b}$ is the hidden bias vector. $\boldsymbol{w}_{i*}$ is the ith row vector of the weight matrix $\boldsymbol{W}$, and V is the number of units in the visible layer. $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2, \cdots \sigma_V^2\}$ is usually fixed to the diagonal covariance matrix of the training data [28] and is not considered to be a parameter of the model. Therefore the parameter set of an RBM is $\{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$. σ_i has been ignored in the rest of this paper for the sake of simplicity.

The probabilistic distribution of visible random variable $\boldsymbol{v}$ described by an RBM can be written as

$$P(\boldsymbol{v}) = \frac{1}{\mathcal{Z}_{\mathrm{RBM}}} \sum_{\boldsymbol{h}} \exp\{-E_{\mathrm{RBM}}(\boldsymbol{v},\boldsymbol{h})\}, \tag{6}$$

where $\mathcal{Z}_{\mathrm{RBM}} = \sum_{\boldsymbol{h}} \int_{\boldsymbol{v}} \exp\{-E_{\mathrm{RBM}}(\boldsymbol{v},\boldsymbol{h})\} d\boldsymbol{v}$ is the partition function, which is intractable to compute and evaluate. Therefore, the contrastive divergence (CD) algorithm is usually used to estimate the parameters of an RBM [29], [30] and the annealed importance sampling (AIS) algorithm is adopted to approximate the partition function $\mathcal{Z}_{\mathrm{RBM}}$ for model evaluation [31]. RBMs have been proven to be powerful for spectral modeling in statistical parametric speech synthesis [6].

2) Bidirectional Associative Memory: BAM is also a shallow neural network with only two layers, as can be seen in the right of Fig. 3. Both layers in BAM are visible layers without any

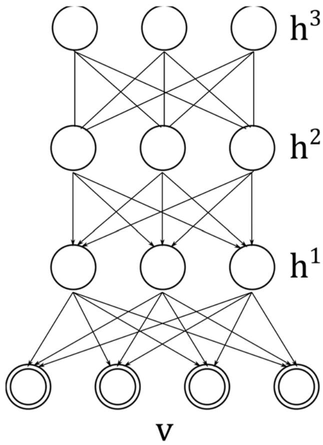

Fig. 4. Graphical representation of a deep belief network with three hidden layers ($\boldsymbol{h}^1$, $\boldsymbol{h}^2$ and $\boldsymbol{h}^3$) and a visible layer ($\boldsymbol{v}$).

hidden layers, which is different from RBM. BAM was originally proposed as a special case of the Hopfield network [32] for information retrieval [27]. Chen *et al.* [1] and Liu *et al.* [33] extended BAM as a generative model whose probabilistic distribution can also be given by an energy function. The energy function for modeling binomial random variables of BAM is given by

$$E_{\mathrm{BAM}}(\boldsymbol{d}, \boldsymbol{u}) = -\boldsymbol{a}^\top \boldsymbol{d} - \boldsymbol{b}^\top \boldsymbol{u} - \boldsymbol{d}^\top \boldsymbol{W} \boldsymbol{u}, \tag{7}$$

where $\boldsymbol{d}$ and $\boldsymbol{u}$ correspond to the binomial random variable vectors in the two visible layers, and $\boldsymbol{a}$ and $\boldsymbol{b}$ are the corresponding bias vectors. The joint distribution over $\boldsymbol{d}$ and $\boldsymbol{u}$ is therefore given by

$$P(\boldsymbol{d}, \boldsymbol{u}) = \frac{1}{\mathcal{Z}_{\mathrm{BAM}}} \exp\{-E_{\mathrm{BAM}}(\boldsymbol{d}, \boldsymbol{u})\}, \tag{8}$$

where $\mathcal{Z}_{\mathrm{BAM}} = \sum_{\boldsymbol{d},\boldsymbol{u}} \exp\{-E_{\mathrm{BAM}}(\boldsymbol{d}, \boldsymbol{u})\}$ is also an intractable partition function. Therefore, following the training method of an RBM, we adopted the CD algorithm to estimate the parameters of BAM [33], which are $\{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$.

3) Deep Belief Network: DBN is another type of neural network-based generative model, but with multiple hidden layers. Fig. 4 shows the graphical structure of a DBN with three hidden layers. The connections between different layers are directed except for the two top hidden layers. The units in the visible layer are Gaussian random variables to enable spectral feature modeling and those in the hidden layers are binomial variables. The probabilistic distribution of a DBN as a generative model, with L hidden layers, can be written as:

$$P(\boldsymbol{v}) = \sum_{\boldsymbol{h}^1,\cdots,\boldsymbol{h}^L} P(\boldsymbol{v}|\boldsymbol{h}^1) \prod_{l=1}^{L-2} P(\boldsymbol{h}^l|\boldsymbol{h}^{l+1}) P(\boldsymbol{h}^{L-1}, \boldsymbol{h}^L), \tag{9}$$

where $\boldsymbol{h}^l = [h_1^l, h_2^l, \dots h_{H_l}^l]^\top$ are the hidden variables in the lth hidden layer, and H_l is the number of hidden units in this layer. The conditional probabilities are given by

$$P(\boldsymbol{v}|\boldsymbol{h}^1) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{W}^1 \boldsymbol{h}^1 + \boldsymbol{b}^1, \boldsymbol{I}), \tag{10}$$

$$P(h_j^{l-1} = 1|\boldsymbol{h}^l) = g(\boldsymbol{w}_{j*}^l \boldsymbol{h}^l + b_j^l), \quad l = 2, \dots, L-1, \tag{11}$$

where $\{\boldsymbol{W}^1, \boldsymbol{b}^1\}$ are the parameters of the first layer, $\boldsymbol{w}_{j*}^l$ is the jth row vector of weight matrix $\boldsymbol{W}^l$ that connects the lth and $l - 1$th layers, b_j^l is the jth element of corresponding bias vector $\boldsymbol{b}^l$, and $g(x) = 1/(1 + e^{-x})$ is the sigmoid activation function. The joint probability of the two top hidden layers is given by BAM Eq. (8), whose energy function is

$$\begin{aligned} E(\boldsymbol{h}^L, \boldsymbol{h}^{L-1}) = &-\boldsymbol{h}^{L\top} \boldsymbol{b}^L - \boldsymbol{h}^{L-1\top} \boldsymbol{b}^{L-1} \\ &- \boldsymbol{h}^{L-1\top} \boldsymbol{W}^L \boldsymbol{h}^L. \end{aligned} \tag{12}$$

The parameters of the DBN, $\{\boldsymbol{W}^l, \boldsymbol{b}^l\}_{l=1}^L$, can be estimated by using a layer-wise greedy learning algorithm initialized by an RBM. Therefore, the DBN has a better ability to describe the probabilistic distribution of visible variables than the RBM [6], [28].

B. Model Training

The right of Fig. 5 outlines the structure of the proposed DNN-based probabilistic postfilter. We can see that it has a symmetric structure, including an input layer, an output layer, and several hidden layers. The inputs and outputs of the DNN are synthetic and natural spectral features. They can be in the form of mel-cepstrum or higher-dimensional spectrum, for example. As we can see from the left of Fig. 5, the proposed DNN-based postfilter is generatively trained layer-by-layer by cascading two RBMs/DBNs with a BAM. The training procedure is conducted in the following four detailed steps:

1) *Acoustic space modeling*: Two generative neural networks are constructed in this first step, the first (θ_x) is for modeling the probabilistic distribution of the synthetic feature space and the second (θ_y) is for modeling that of the natural feature space. The generative neural network here can consist of either RBMs or DBNs. The respective model parameters are

$$\theta_x = \{\boldsymbol{W}_x^l, \boldsymbol{a}_x^l, \boldsymbol{b}_x^l\}_{l=1}^L, \tag{13}$$
$$\theta_y = \{\boldsymbol{W}_y^l, \boldsymbol{a}_y^l, \boldsymbol{b}_y^l\}_{l=1}^L, \tag{14}$$

for the two DBNs with L hidden layers ($L = 1$ for RBMs). The training process for a DBN actually consists of stacking L RBMs, and therefore $\{\boldsymbol{W}_x^l, \boldsymbol{a}_x^l, \boldsymbol{b}_x^l\}$ and $\{\boldsymbol{W}_y^l, \boldsymbol{a}_y^l, \boldsymbol{b}_y^l\}$ correspond to the parameters for the lth RBM of synthetic and natural spectra.

2) *Binary encoding of spectral features*: The estimated RBMs/DBNs may also serve as auto-encoders for spectral features. These auto-encoders can encode the raw spectral features into high-level hidden binary representations [34]. The hidden binary representations are obtained according to the conditional distribution derived from the RBM, e.g., for synthetic spectral features:

$$h_{x,j} \sim P(h_{x,j} = 1|\boldsymbol{x}) = g(\boldsymbol{x}^\top \boldsymbol{w}_{*j} + b_j), \tag{15}$$

where $\boldsymbol{x}$ is the spectral feature, $h_{x,j}$ is the jth dimension of its hidden representation $\boldsymbol{h}_x$, and $\boldsymbol{w}_{*j}$ and $\boldsymbol{b}_j$ are the model parameters related to the jth hidden unit. Because the

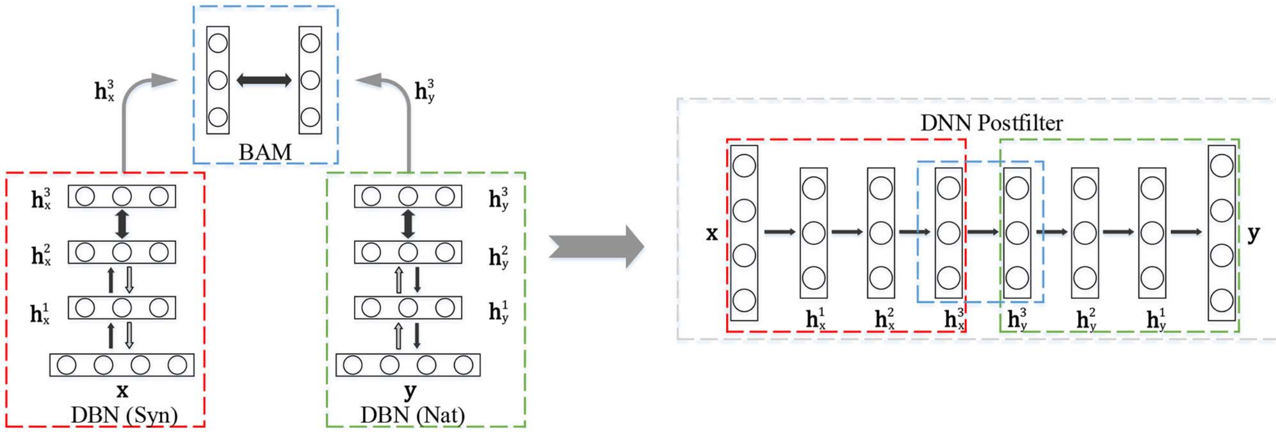

Fig. 5. Structure and training procedure for proposed DNN-based postfilter. The six-hidden-layer DNN is composed of a BAM and two DBNs, with three hidden layers for synthetic and natural speech.

hidden units are conditionally independent of each other, the hidden representations can be sampled conveniently dimension-by-dimension.

The hidden representations for the DBNs are extracted layer-by-layer as the binary code of the DBN auto-encoders [34]. Note that although the directed connections in the DBN are top-down for generation as a decoder in Fig. 4, they can also be bottom-up for extracting hidden variables as an encoder [35].

3) *Joint modeling*: BAM $\theta_h = \{\boldsymbol{W}_h, \boldsymbol{a}_h, \boldsymbol{b}_h\}$ is adopted in the third step to model the joint distribution of hidden variables from the two RBMs/DBNs estimated in step 1. Note that the two RBMs (or the top hidden layers of the DBNs) are trained separately in an unsupervised way in step 1 and the relationship (or acoustic difference) between synthetic and natural speech is captured by a single BAM in this step in high-level hidden space.

4) *Model combination*: The three estimated generative models are combined in the final step by concatenating the two RBMs/DBNs with the BAM. The concatenated model is then converted to a DNN (feed-forward stochastic neural network) with $2L$ hidden layers, as shown in Fig. 5. The parameters of the DNN $\theta = \{\boldsymbol{W}^l, \boldsymbol{b}^l\}_{l=1}^{2L+1}$ are copied from the RBMs/DBNs and BAMs, which are

$$\{\boldsymbol{W}^l, \boldsymbol{b}^l\} = \begin{cases} \{\boldsymbol{W}_x^l, \boldsymbol{b}_x^l\} & l \leq L, \\ \{\boldsymbol{W}_h, \boldsymbol{b}_h\} & l = L+1, \\ \{\boldsymbol{W}_y^{2L+2-l^{\top}}, \boldsymbol{a}_y^{2L+2-l}\} & l > L+1. \end{cases} \quad (16)$$

The parameters of each layer are estimated separately in this training procedure and copied to form a DNN. We did not jointly fine-tune the parameters of all layers. This does not mean that joint fine-tuning is unnecessary. The minimum mean square error (MMSE) criterion is usually used for DNN training in regression tasks, such as those in speech generation. However, previous work in VC has indicated that listeners prefer synthetic speech generated using a network architecture without the fine-tuning over one using the fine-tuning based on the MMSE criterion [25]. Therefore we can assume that this criterion may not be optimal for training a postfilter, either.

This probabilistic postfilter works because of the powerful modeling ability of RBMs/DBNs:

- An RBM is equivalent to a structured GMM with 2^H components. The number of Gaussian components in an RBM can be considerably larger than the number of training samples we can obtain, due to its ability to describe very complicated multimodal distributions of spectral features.
- An RBM is a product of experts (PoE) [36] model that describes a probabilistic distribution with very sharp modes.
- A DBN is a deep extension of an RBM and it is reported that it is a better model for spectral envelopes [6].

C. Spectral Postfiltering

The proposed DNN directly describes a conditional distribution of natural spectral feature $\boldsymbol{y}$ given synthetic spectral feature $\boldsymbol{x}$:

$$\begin{aligned} P(\boldsymbol{y}|\boldsymbol{x}) &= \sum_{\boldsymbol{h}^1,\cdots \boldsymbol{h}^{2L}} P(\boldsymbol{y}, \boldsymbol{h}^1, \cdots \boldsymbol{h}^{2L}|\boldsymbol{x}) \\ &= \sum_{\boldsymbol{h}^1,\cdots \boldsymbol{h}^{2L}} P(\boldsymbol{y}|\boldsymbol{h}^{2L}) \prod_{l=1}^{2L-1} P(\boldsymbol{h}^{l+1}|\boldsymbol{h}^l) P(\boldsymbol{h}^1|\boldsymbol{x}) \\ &\simeq P(\boldsymbol{y}|\boldsymbol{h}^{2L^*}) \prod_{l=1}^{2L-1} P(\boldsymbol{h}^{l+1^*}|\boldsymbol{h}^{l^*}) P(\boldsymbol{h}^{1^*}|\boldsymbol{x}), \end{aligned} \quad (17)$$

where $\boldsymbol{h}^1, \ldots, \boldsymbol{h}^{2L}$ are random variables in the $2L$ hidden layers of the proposed DNN-based postfilter. Here, $P(\boldsymbol{h}^1|\boldsymbol{x})$ and $P(\boldsymbol{h}^l|\boldsymbol{h}^{l-1})$ are multi-variate binomial distributions defined similarly to those in Eq. (15) and

$$P(\boldsymbol{y}|\boldsymbol{h}^{2L}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{W}^{2L+1}\boldsymbol{h}^{2L} + \boldsymbol{b}^{2L+1}, \boldsymbol{I}). \quad (18)$$

We make an approximation in Eq. (17) in order to reduce the computational cost by using the optimal samples for $\boldsymbol{h}^{l^*}$ instead of summing over them. This approximation is reasonable because the models are trained similarly layer-by-layer. The optimal binary samples $\boldsymbol{h}^{l^*}$ are sampled from the conditional distribution according to the maximum probabilities as:

$$\boldsymbol{h}^{1^*} = \arg\max_{\boldsymbol{h}^1} P(\boldsymbol{h}^1|\boldsymbol{x}), \quad (19)$$

$$\boldsymbol{h}^{l^*} = \arg\max_{\boldsymbol{h}^l} P(\boldsymbol{h}^l|\boldsymbol{h}^{l-1}), l = 2, \ldots, 2L. \quad (20)$$

When the mean-field approximation is used here, the proposed DNN is treated exactly the same as a conventional feed-forward neural network.

The input and output spectral features may be composed of multiple frames in practice to capture sequential properties of the feature trajectories. The maximum likelihood parameter generation (MLPG) algorithm [37] is adopted in this case to generate a static feature sequence for synthesizing speech. For example, the output spectral feature sequence, $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_T^\top]^\top$, is generated by

$$\boldsymbol{c}^* = \arg\max_{\boldsymbol{c}} P(\boldsymbol{y}|\boldsymbol{x}), \tag{21}$$

$$\simeq \arg\max_{\boldsymbol{c}} \prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{h}_t^{2L^*}), \tag{22}$$

$$\text{s.t.} \quad \boldsymbol{y} = \boldsymbol{M}\boldsymbol{c}, \tag{23}$$

where $\boldsymbol{M}$ is the matrix that is used to convert the static feature sequence into multiple frame sequence [1]. Note that the conditional distribution in Eq. (18) is a Gaussian distribution with a unit covariance matrix because the training samples are normalized to zero mean and unit variance. Therefore, the conditional distribution needs to be converted into the real distribution before applying the MLPG algorithm. Since the conditional distributions are single Gaussian distributions with a globally shared diagonal covariance matrix, the MLPG in this paper is the same as that in conventional approaches.

IV. Evaluation

This section presents the subjective evaluation and acoustic analysis of synthetic speech processed using various quality-enhancement methods[2]. First, we will describe the text-to-speech voices used in the experiments and the methods we used in evaluations to compensate for over-smoothing. Then, the acoustic analysis in terms of modulation characteristics and spectra is presented, after which we will present the design of the listening test and finally the test results.

A. Voices and Methods

We used a female and a male synthetic voice for the evaluation, both of which were in English. The male voice was created from a high-quality average voice model adapted to 2840 sentences recorded from a British male speaker, which consisted of approximately three hours of speech material. The female voice was built using 4546 sentences recorded from a Scottish female speaker, which comprised approximately four hours of speech.

All data were sampled at 48 kHz. We extracted the following acoustic features at 5 ms intervals: 59 mel-cepstral coefficients, mel scale f_0 and 25 aperiodicity band energies extracted using the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed (STRAIGHT) [38] analysis. 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 f_0 and one for band-aperiodicity. Speech was synthesized in the frequency domain.

TABLE I
Methods That Were Evaluated

NONE	No enhancement
PF	Mel-cepstral postfilter [14]
GV	Global variance [15]
MS	Modulation spectrum in mel-cepstral domain [16]
DNN-MCEP	Deep neural network in mel-cepstral domain
DNN-SPEC	Deep neural network in spectral domain

Table I outlines the methods we evaluated. The parameter β was set to 0.4 to create the PF entry as in [14]. We applied the method of global variance [15] only to the mel-cepstral stream for the GV entry.

The MS of the natural and the synthetic utterances were evaluated using Eqs. (1) and (2) and using mel-cepstrum for representing the spectrum of speech for MS enhancement. The MS was evaluated for each file and each mel-cepstral coefficient trajectory, from which the MS statistics (mean μ and standard deviation σ) were estimated. We used 4096-point Fourier analysis in Eq. (1) in order to exceed the maximum number of frames in an utterance in the database. The synthetic trajectories were enhanced using Eq. (3) based on the statistics that were evaluated. The value of α was set to 0.85 based on the findings by Takamichi *et al.* [16]. The MS enhanced mel-cepstra were then used for synthesizing speech (in the frequency domain).

The input and output of the DNN postfilters were formed by using multiple consecutive frames of spectral features in both mel-cepstral and spectral domains:

- *Mel-cepstral domain* The DNNs were trained with paired synthetic and natural spectral features aligned using the dynamic time warping (DTW) algorithm[3]. Only a DNN with two hidden layers was constructed for the mel-cepstral domain, because we observed that the more hidden layers we used from our preliminary experiments, the worse the generated speech was. There were 2048 hidden units in each hidden layer. The postfilter was only applied to the lower dimensional mel-cepstral coefficients (1–18th mel-cepstral coefficients), which are mostly related to the formants of speech.
- *Spectral domain* The spectral envelopes, which were extracted using STRAIGHT with a fast Fourier transform (FFT) length of 4096, were directly used as the spectral domain features. The dimensionality of the spectral envelopes was 2049. The spectral envelopes were warped into the Bark scale (using a bilinear transform with a warping factor of 0.77 [39]) before the DNNs were trained. Spectral envelopes of synthetic and natural speech were aligned using the alignment paths calculated from their corresponding mel-cepstra. We found from our internal experiments that the generated speech improved as we increased the number of hidden layers. However, DBNs with three hidden layers, which formed a DNN with six

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

[3]It is also possible to obtain such paired data via the forced alignment algorithms. However a preliminary subjective evaluation test showed that the above DTW algorithm was preferred in terms of the quality of synthetic speech.

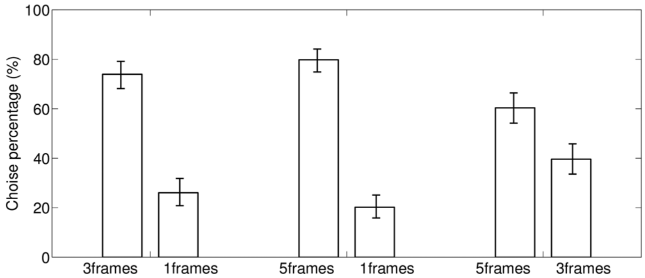

Fig. 6. Preference scores between samples generated with DNN postfilters with one, three and five frames in input/output.

hidden layers, were used to limit the computational costs. There were 2048 hidden units in each hidden layer.

The RBMs, DBNs and BAMs were estimated using the CD algorithm with one-step Gibbs sampling (CD-1). The mini-batch size was set to 10 during training. The learning rate was set to 0.0001 for all models. The momentum and weight decay were also employed to train the models [30]. Two hundred epochs were executed in training the RBMs and DBNs, and 50 epochs were executed in training the BAMs.

B. Listening Experiment: Context Size of DNN Postfilter

We used three consecutive frames for input and output of the DNN postfilter in our previous experiments [1]. We wanted to evaluate the effect of context size in this experiment by varying the number of consecutive frames. We trained DNN postfilters with one, three and five frames as input and output to do this.

We evaluated the quality of the postfilters by three possible paired comparisons. Ten native English speakers participated in the listening test. Each listener compared 120 pairs of speech samples, which were comprised of 40 samples from each of the three paired comparisons.

Fig. 6 provides the breakdown in percentages excluding the no preference option with 95% confidence intervals calculated using a two-tailed binomial test. The scores indicate that the DNN postfilter with five frames was preferable to those with one and three frames. The three-frame system was also preferred over the one-frame system. Although we also built systems with seven and nine frames, no clear differences were perceived between these and the five-frame system, and the model training took much longer. Therefore, we fixed the context size of the DNN-based postfilter to five frames for the experiments in the rest of this paper.

Note that this experiment was conducted in the mel-cepstral domain. The performance of the DNN postfilter in the spectral domain could differ from what we observed in the mel-cepstral domain. However, it was difficult to train the DNN with more than five frames in such a high dimensional space. Therefore, we also fixed the context size of DNN in spectral space to five frames in the experiments.

C. Acoustic Analysis

This section presents the results obtained from acoustic analysis. One interesting aspect to compare is to analyze the modulation characteristics. This is because the proposed DNN-based postfilter uses five frames as input and hence may

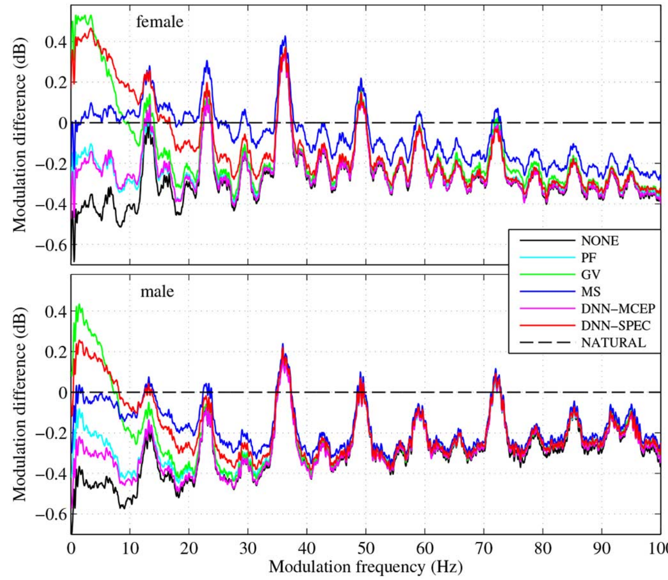

Fig. 7. Average difference in modulation spectrum of mel-cepstra for different systems compared to natural speech for the female (top) and male (bottom) speakers.

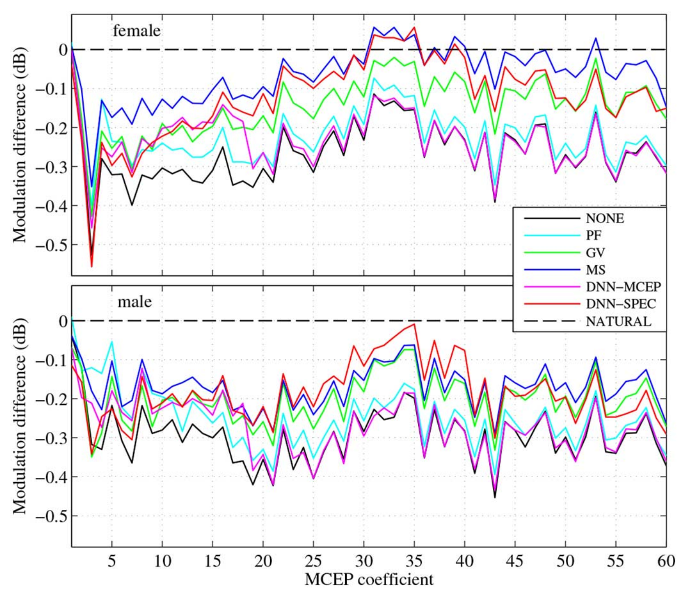

Fig. 8. Average difference in modulation per mel-cepstral coefficient for different systems compared to natural speech for the female (tom) and male (bottom) speakers.

implicitly learn such temporal characteristics without explicitly using modulation spectrum features.

Frame-wise mel-cepstra were evaluated from all the synthetic and natural speech waveforms to study the modulation characteristics of the test speech samples. The average modulation spectra of all systems were then evaluated following the same procedure as that in MS enhancement, which was described in Section II-C. Fig. 7 shows the differences in modulation spectra with respect to natural speech for each method calculated from mel-cepstra and averaged across sentences and all mel-cepstral coefficients for the female and male speakers. The same data are presented in Fig. 8, but they have been plotted separately for each mel-cepstral coefficient and averaged over all modulation frequencies.

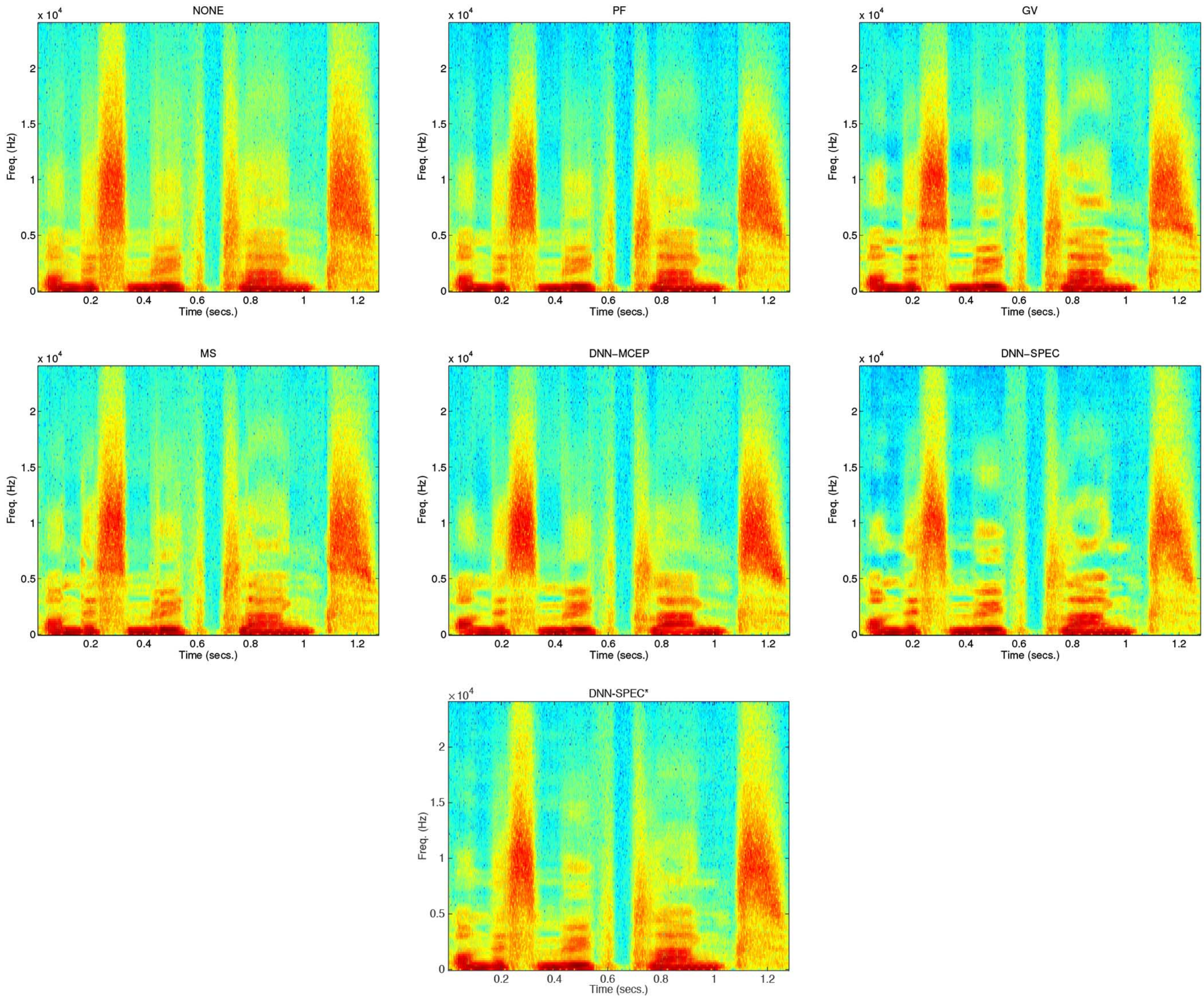

Fig. 9. Spectrogram of utterance "on the smooth planks" generated using baseline system (NONE) and the enhancement methods: PF, GV, MS, DNN-MCEP, DNN-SPEC and DNN-SPEC* (refers here to the DNN-SPEC method but with the mean-field approximation). The female speaker model was used.

Fig. 7 indicates that GV and DNN-SPEC have the highest modulation at low modulation frequencies that represent modulation frequencies that are mostly associated with relatively slow movements of the articulators. Interestingly, the modulation in these two systems is even higher than that in natural speech. Speech with no enhancement (NONE) has the least modulation overall, and the rest of the systems fall between these two extremes. Although the modulation decreases for higher modulation frequencies for all systems, MS enhancement indicates a consistent increase in modulation for all frequencies, especially for the female speaker, thus possibly over-enhancing the higher modulation frequencies.

Fig. 8 indicates that DNN-SPEC provides the largest boost in modulation for mid-quefrency mel-cepstral coefficients, while MS enhancement seems to create the highest overall boost in modulation for each coefficient, probably due to all modulation frequencies being enhanced. Speech with no enhancement (NONE) has the lowest modulation for all mel-cepstral coefficients. However, all systems have less modulation on almost all mel-cepstral coefficients compared to natural speech. Interestingly, the DNN-MCEP that enhanced the coefficients from 1 to 18 shows increase only within these coefficients.

Finally, we present the spectrogram of a test sentence produced by the systems we evaluated here in Fig. 9. We can see that both the formants and the spectral fine structure are more enhanced when using the DNN-SPEC postfilter compared to other methods of enhancement. We also present the spectrogram generated by the proposed postfilter with mean-field sampling for hidden units to show the effectiveness of the proposed sampling method (Eqs. (19) and (20)). Benefiting from direct modeling in the spectral domain, the spectrogram of the DNN-SPEC system has a more detailed spectral structure especially at the high frequencies than the conventional methods of enhancement that operate in the mel-cepstral domain.

D. Listening Experiment: Comparing Postfilters

We evaluated the methods in Table I using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)

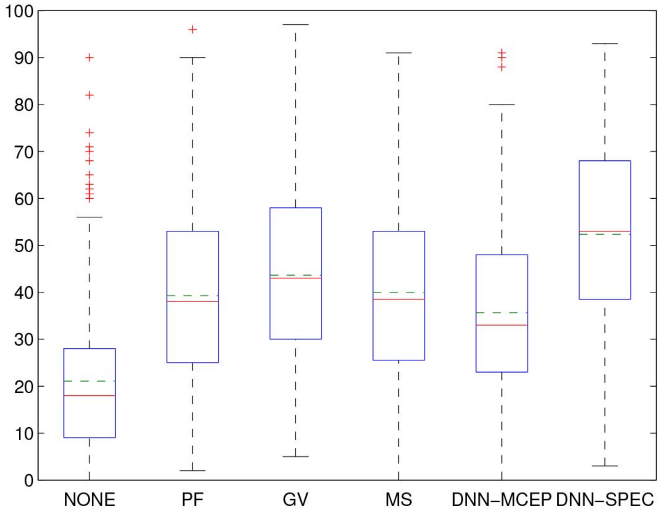

Fig. 10. Results for the male voice: box plots of subjective ratings. Means are represented by solid red lines and medians are represented by dashed green horizontal lines.

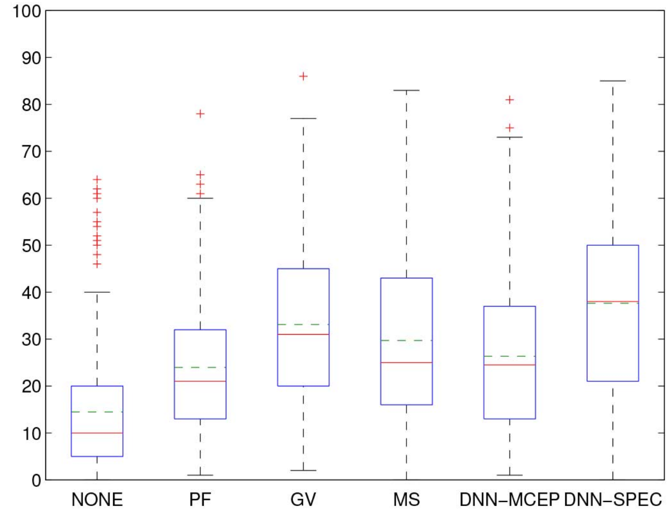

Fig. 11. Results for the female voice: box plots of subjective ratings. Means are represented by solid red lines and medians are represented by dashed green horizontal lines.

methodology [40]. Participants rated stimuli produced by all methods in parallel in the MUSHRA test using a scale from 0 to 100. It was possible for subjects to directly compare the methods and revise scores accordingly in this way. Such tests require reference stimulus to be presented that participants should rate as 100. The reference was natural speech in our tests. The same sentence was used in each comparison apart from the female voice whose natural speech reference was a different sentence as we did not have the recordings of the test sentences used here. Each participant evaluated 10 sentences for the male voice and 10 for the female voice. A set of 60 sentences were balanced across participants so that for every six participants all sentences were rated under all conditions. The sentences were chosen from the first six sets of the Harvard dataset [41], which was a set that was not used to train either of the voices. As 24 native English speakers participated in the listening test, 240 scores were obtained for each method applied to each voice.

E. Results

The distributions of the subjective scores are indicated by the box plots in Fig. 10 and Fig. 11 for the male and female voices.

We performed a series of pairwise t-tests to identify significant differences in mean scores between the methods. We applied the Bonferroni correction to compensate for the large number of comparisons. All pairs were found to be significantly different at a 1% level except (PF, MS), (PF, DNN-MCEP) and (GV, MS) for the male voice and (PF, DNN-MCEP) and (DNN-MCEP, MS) for the female voice according to this procedure.

The results indicate that all the postfiltering methods resulted in better quality of synthetic speech than that without post-processing. GV and MS were the most preferable for the male speaker out of the conventional postfiltering methods, and GV was the most preferable for the female speaker.

Further we can see that the proposed DNN-based postfilter in the mel-cepstral domain performs as well as the conventional mel-cepstral postfilter. Finally, we found that the proposed DNN-based postfilter in the spectral domain produced synthetic speech that was of higher quality than that obtained with any conventional postfilters.

V. Discussion

A. Why did DNN-based Spectral Postfilter Perform Better?

The results presented in Section IV-E indicate that the proposed DNN-based postfilter in the spectral domain produced synthetic speech of significantly higher quality than that obtained with the conventional postfilters. Three possible reasons for this include:

- The DNN was trained directly in the spectral domain rather than in the mel-cepstral domain, and was therefore able to learn spectral fine structures in detail. Note that we did not include GV in the spectral domain in our experiments although it provided good results in a previous study on speech data sampled at 16 kHz [20]. However, it did not work well on the speech data sampled at 48 kHz in our experiments. In contrast, the proposed DNN-based postfilter worked well for speech sampled at both 16 and 48 kHz [1], [42]. The DNN was also able to learn the gap in speech dynamics between synthetic and natural speech in the spectral domain similarly to GV in the spectral domain.
- 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 (2^{2048} in this work) [25]. The RBMs/DBNs are probabilistic models with some beneficial properties, as was discussed in Section III-B. The acoustic differences between synthetic and natural speech are modeled in a high-level binary hidden space. There are fewer patterns in this space than in the original spectral space, and it is therefore easier to compensate for the differences with a single layered BAM.

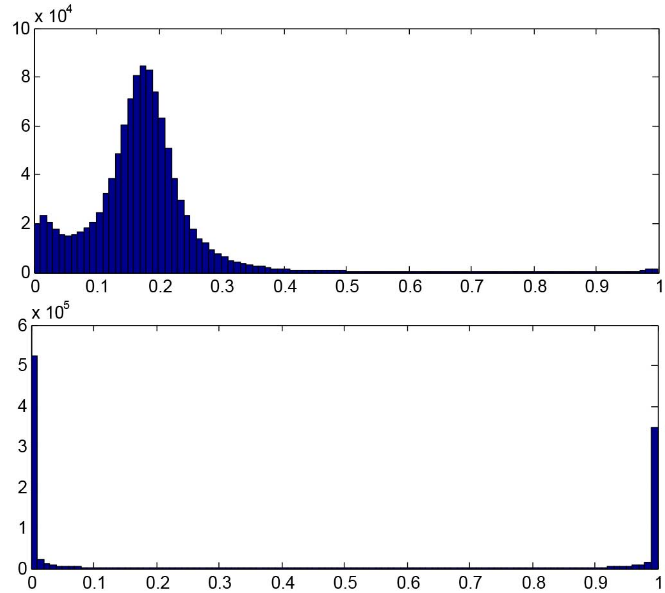

Fig. 12. Histogram of $P(h_j = 1)$ for hidden units from first hidden layer in mel-cepstral domain (top) and spectral domain (bottom).

- The DNN could also learn modulation characteristics since it uses five consecutive frames for mapping and because there is a close relationship between the DNN and MS. The FFT convolution is equivalent to the weighted sum in a network unit of the convolutional DNN [43], and the next deep layer of a DNN trained in the spectrum domain may therefore contain a representation related to MS.

The spectral features were encoded into binary representations by RBMs/DBNs for mapping in the proposed DNN-based postfilter. This is important because the modeling and mapping in a transformed binary space can avoid the statistical averaging effect in the continuous space of original spectra, which is the main cause of the over-smoothing problem in conventional HMM-based statistical parametric speech synthesis.

However, the subjective results indicate that the proposed method is feature sensitive. Although it works well in the spectral domain, it is significantly worse than DNN-SPEC in mel-cepstral domain. However, it is better than the baseline method without any post-processing (NONE). One reason for this is the use of high-dimensional spectra in DNN-SPEC. Another reason could be that the DNN is not well estimated in the mel-cepstral domain. It is vital in the training of the proposed DNN to first generate good binary representations for spectral features using RBMs for estimating higher hidden layers of DBNs and BAM. Each dimension of these binary representations are produced according to the probability of the corresponding unit being one (probability of the unit being "switched on", e.g., $P(h^1_{j,x} = 1|\boldsymbol{x})$ for synthetic speech in Eq. (15)). Fig. 12 presents the histograms for $P(h^1_{j,x} = 1|\boldsymbol{x})$ in the mel-cepstral and spectral domains. The histograms were counted using all 2048 hidden units of a sentence from the training set. We can see a clear 0/1 pattern in histogram of the spectral domain, i.e., the probabilities are either close to zero or close to one. This makes it easy to sample reliable binary representations with many units being one. However, most probabilities are focused on 0.2 in the mel-cepstral domain and very few are close to one. The sampled binary sample is not a good representation of the mel-cepstrum because it was sampled with a very low probability. Therefore we used a mean-field approximation for DNN-MCEP discussed in this paper instead of sampling binary representations. Using mean-field approximation loses the beneficial properties of binary representations in avoiding over-smoothing.

B. Modulation Spectrum

The results suggest that low modulation frequencies are perceptually most significant, and enhancing these improves the quality of synthetic speech. There is still a large gap in modulation spectra at the higher modulation frequencies in comparison to natural speech, but it is not yet clear how much this has perceptual relevance. MS enhancement, which had the highest modulation at high modulation frequencies, did not produce the best quality. However, the higher modulation frequencies, probably linked to the excitation patterns, may still be perceptually important, but simple MS enhancement probably cannot reproduce or enhance the modulation patterns present in natural glottal excitation.

We noticed that the excitation of speech had a significant effect on the modulation characteristics of the estimated spectral parameters in the experiments with MS enhancement. Fig. 2 plots difference in the modulation spectra between 1) parameters estimated from natural speech, and 2) parameters generated from statistical models. However, if the modulation spectrum of the latter is estimated from a synthesized speech waveform instead of the generated parameters, the MS has higher levels of modulation. This is probably due to the excitation of speech that generates additional modulation at higher modulation frequencies. Thus, the difference in modulation spectra between natural and synthetic speech should theoretically be estimated using parameters estimated from natural and synthetic *waveforms* in both cases. Chen *et al.* calculated the difference in MS between parameters estimated from natural speech and parameters generated from statistical models [1], [16], thus ignoring the effect of excitation of synthetic speech. The effect of ignoring synthetic excitation will most likely over-estimate the difference in modulation between natural and synthetic speech and thus higher modulation frequencies will be over-emphasized after MS enhancement, as is shown in Fig. 7. This might degrade speech quality due to the strong, overly fast modulations in the spectral parameters. Due to this issue, Takamichi *et al.* uses low-pass filtering of the MS before enhancement [16] (although it was not explicitly mentioned in the paper), which might explain why MS enhancement performed better in that particular experiment. Despite this previously mentioned issue, the method in [1] (i.e., MS estimated from generated parameters and without low-pass filtering of MS) was used as a reference in this study since it was proven to be successful despite the effect of excitation being ignored. Preliminary experiments on estimating MS from the natural and synthetic speech waveforms indicated that the method is feasible: the higher modulation spectrum is not overly emphasized, but lightly less enhancement will be achieved also in the lower modulation frequencies.

C. Computational Cost

The proposed DNN-based enhancement can be time consuming since the model is applied directly to high-dimensional spectra. For example, applying a sentence with T frames, the computational complexity of this method is $O(NHTL)$, where N is the dimensionality of the spectral envelope, H is the number of units in each hidden layer, and L is the number of hidden layers. 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 $M \ll N$).

We can see that the computational complexity of the proposed DNN-based postfilter is still hundreds of times that of the conventional GV-based approach. This could be a limitation in real time systems. However, the DNN-based postfilter can also be applied to the model parameters of HMMs to accelerate the synthesis process. For example, the mean vector of the spectral stream (mel-cepstrum) of each HMM state can be converted into multiple frames of spectra, and the DNN-based postfilter can be applied to the converted mean vectors. The postfiltered mean vectors can then be converted back to the mel-cepstral domain with dynamic features to replace the corresponding mean vectors of the HMMs. In this case, the computational cost of the synthesis process is exactly the same as that of the conventional method (NONE).

VI. Conclusion

We proposed a data-driven postfilter technique to improve the segmental quality of statistical parametric text-to-speech synthesis. The proposed method uses a DNN to model the conditional probability of the spectrum of natural speech given the spectrum of synthetic speech. We evaluated the proposed postfilter in two different spectral domains: the low dimensional mel-cepstral domain and the full spectrum domain, which we described in correspondence. We found that the full spectral domain DNN-based postfilter significantly improved the segmental quality of synthetic speech by comparing these two variants with existing postfilter techniques. We also compared and evaluated them with conventional methods for both a female and male voice.

Future work will include studies on the DNN-based postfilter in a speaker independent fashion, investigation into long term modulation spectra with LSTM-based RNN in hidden binary space, and also studies on enhancements to modulation spectra using higher-dimensional spectra instead of mel-cepstra.

References

[1] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," in *Proc. Interspeech*, 2014, pp. 1954–1958.

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

[3] A. Black and K. Tokuda, "The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proc. Interspeech*, 2005.

[4] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

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

[6] 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. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2129–2139, Oct. 2013.

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

[8] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-gaussian process hybrid model," in *Proc. ICASSP*, Apr. 2013, pp. 6885–6889.

[9] R. Vishnubhotla, S. Fernandez, and B. Ramabhadran, "An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech," in *Proc. ICASSP*, 2010, pp. 4614–4617.

[10] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 2268–2272.

[11] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[12] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Apr. 2015, pp. 4470–4474.

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

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *Syst. Comput. Jpn.*, vol. 36, no. 12, pp. 43–50, 2005.

[15] 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.

[16] 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, pp. 290–294.

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

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

[19] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on mel-cepstral analysis," in *Proc. ICASSP*, 1995, vol. 1, pp. 33–36.

[20] 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.

[21] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

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

[23] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.

[24] L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Voice conversion using generative trained deep neural networks with multiple frame spectral envelopes," in *Proc. Interspeech*, 2014, pp. 2313–2317.

[25] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[26] 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.

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

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

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

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

[31] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, Univ. of Toronto, Toronto, ON, Canada, 2009.

[32] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Acad. Sci.*, 1982, vol. 79, no. 8, pp. 2554–2558.

[33] 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, pp. 7884–7888.

[34] L. Deng, M. Seltzer, D. Yu, A. Acero, A. rahman Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech*, Sep. 2010.

[35] Y. Bengio, "Learning deep architectures for AI," *Foundations Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[36] G. Hinton, "Products of experts," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, 1999, vol. 1, pp. 825–828.

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

[38] 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 Commun.*, vol. 27, pp. 187–207, 1999.

[39] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis–A unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, vol. 3, pp. 1043–1046.

[40] "Method for the subjective assessment of intermediate quality level of coding systems," ITU Rec. ITU-R BS.1534-1, Int. Telecomm. Union Radiocommunication Assembly. Geneva, Switzerland, Mar. 2003.

[41] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[42] L.-H. Chen, Z.-H. Ling, Y.-Q. Zu, R.-Q. Yan, Y. Jiang, X.-J. Xia, and Y. Wang, "The USTC system for blizzard challenge 2014," in *Proc. Blizzard Challenge Workshop*, Singapore, Sep. 2014.

[43] J. Anden and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

Ling-Hui Chen received the B.E. degree in electronic information engineering, and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively. From April 2010 to October 2010, he was a visiting student at Nagoya Institute of Technology, Japan. He is currently a joint Postdoctoral Researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. His research interests include voice conversion, speech synthesis and machine learning.

Tuomo Raitio received the M.Sc. degree in telecommunication technology from the Helsinki University of Technology, Espoo, Finland, in 2008, and the Ph.D. degree from the Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland, in 2015 on voice source modeling in statistical parametric speech synthesis. He was a visitor at the CSTR, University of Edinburgh, UK, from September 2013 to February 2014. His research interests include speech analysis, modeling, statistical parametric and unit selection speech synthesis, voice source modeling, and voice quality. He is currently working at Apple Inc.

Cassia Valentini-Botinhao graduated from the Federal University of Rio de Janeiro, Brazil, receiving the title of Electronic Engineer in 2006 and received an M.Sc. from the University of Erlangen-Nuremberg in Germany, in 2009, on the program Systems of Information and Multimedia Technology. As a Marie Curie Fellow Cassia obtained her PhD in University of Edinburgh, UK, with the thesis "Intelligibility enhancement of synthetic speech in noise." Her research interests are speech intelligibility and signal processing for speech synthesis. She is a Research Associate at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK.

Zhen-Hua Ling (M'10) received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008, respectively.

From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. From July 2008 to February 2011, he was a joint Postdoctoral Researcher at the University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an Associate Professor at the University of Science and Technology of China. He also worked at the University of Washington, USA, as a Visiting Scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Junichi Yamagishi (SM'13) was awarded a Ph.D. by the Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. He is an Associate Professor at the National Institute of Informatics in Japan. He is also a Senior Research Fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. Since 2006, he has authored and co-authored about 100 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists' Prize from the Minister of Education, Science and Technology in 2010, 2013, and 2014, respectively.