INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech Bandwidth Extension Using Bottleneck Features and Deep Recurrent Neural Networks

Yu Gu, Zhen-Hua Ling, Li-Rong Dai

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

Abstract

This paper presents a novel method for speech bandwidth extension (BWE) using deep structured neural networks. In order to utilize linguistic information during the prediction of high-frequency spectral components, the bottleneck (BN) features derived from a deep neural network (DNN)-based state classifier for narrowband speech are employed as auxiliary input. Furthermore, recurrent neural networks (RNNs) incorporating long short-term memory (LSTM) cells are adopted to model the complex mapping relationship between the feature sequences describing low-frequency and high-frequency spectra. Experimental results show that the BWE method proposed in this paper can achieve better performance than the conventional method based on Gaussian mixture models (GMMs) and the state-of-the-art approach based on DNNs in both objective and subjective tests.

Index Terms: speech bandwidth extension, deep neural networks, recurrent neural networks, long short-term memory, bottleneck features

1. Introduction

The bandwidth of speech signal in the existing public switching telephone network (PSTN) is less than 4kHz. The absence of high-frequency component leads to degraded speech quality and intelligibility compared with its wideband counterpart. Therefore, the speech bandwidth extension (BWE) technology has been studied in order to improve the quality and intelligibility of narrowband speech by restoring its missing high-frequency component.

Various algorithms have been proposed for the BWE task during last decades including some simple methods such as codebook mapping [1], linear mapping [2] and rule-based spectrum folding, and some more complicated statistical approaches using GMMs [3, 4, 5] and hidden Markov models (HMMs) [6, 7, 8]. Nevertheless, these methods usually employ low-dimensional spectral features such as line spectral pairs or mel-cepstral coefficients for spectral representation, which fail to represent the details of spectral envelopes. These methods also suffer from the over-smoothing effect and muffled speech quality due to inaccurate acoustic modeling.

Recently, deep learning has emerged as a new branch of machine learning. In contrast to the conventional acoustic modeling methods, deep learning techniques have significantly improved the naturalness, intelligibility and quality of the generated speech in various speech generation tasks, such as speech enhancement, voice conversion and text-to-speech synthesis [9]. Different kinds of stochastic neural networks such as restricted Boltzmann machines (RBMs), bidirectional associative memories [10] and DNNs with different structures [10, 11, 12, 13, 14] have also been utilized in BWE to replace GMMs or HMMs to model the sophisticated and non-linear mapping relationship from low-frequency speech parameters to high-frequency ones. Because DNNs have better ability of modeling high-dimensional observations with cross-dimension correlations, these methods can use the raw and high-dimensional spectral envelopes or magnitude spectra directly rather than their low-dimensional derivatives. Therefore, more spectral details can be preserved when reconstructing high-frequency spectra. The experimental results of previous work [10, 11, 12, 13] have shown that DNN-based BWE methods can effectively alleviate the over-smoothing effect and improve the quality of BWE outputs compared with GMM-based ones.

Existing BWE approaches including the DNN-based ones focus on modeling the intrinsic correlation or mapping relationship between the acoustic parameters of the input narrowband speech and the missing high-frequency component. The linguistic information that the input narrowband speech contains is always ignored during BWE. However, the characteristics and energy distributions of high-frequency spectra are intrinsically phone-dependent. Such linguistic information could be beneficial to the reconstruction of high-frequency spectral components. Therefore the BN features [15] generated from a DNN-based state classifier for narrowband speech are employed to improve the performance of the DNN-based BWE approach in this paper. The BN features can be regarded as a compact representation of the linguistic information extracted from the input narrowband speech. Meanwhile, motivated by the success of RNNs in speech generation tasks such as text-to-speech synthesis [16, 17], deep RNNs with stacked layers of LSTM cells [18] are adopted to model the temporal dependencies among the sequences of low-frequency and high-frequency spectral features. RNNs are also expected to alleviate the discontinuity due to the frame-independent mapping functions in feed-forward DNNs. Experimental results demonstrate that the proposed BWE approach outperforms the conventional DNN-based one in both objective and subjective evaluations.

This paper is organized as follows. In Section 2, the conventional DNN-based BWE approach is briefly reviewed. Section 3 describes the BN features and RNNs with LSTM cells utilized in this paper. Section 4 introduces our proposed methods in detail. Section 5 presents the experimental results and conclusions are drawn in Section 6.

2. DNN-based approach to BWE

A DNN-based BWE approach [11] has been proposed to estimate the spectral mapping function from narrowband to

http://dx.doi.org/10.21437/Interspeech.2016-678

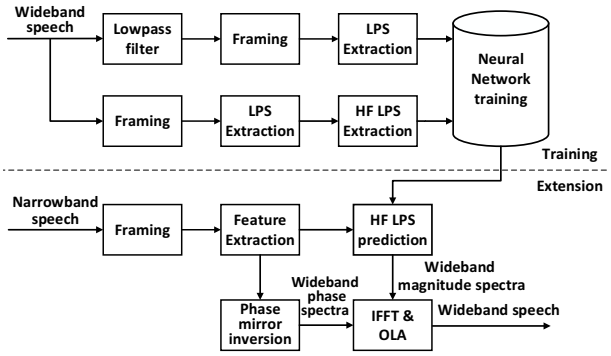

Figure 1: Block diagram of the DNN-based approach, where HF and LPS stand for "high-frequency" and "log power spectra" respectively.

wideband speech. The flowchart of the DNN-based BWE framework is shown in Figure 1. At training phase, the parallel narrowband speech was generated by down-sampling the wideband speech in training database through a lowpass filter as illustrated in Figure 1. Log power spectra derived from the short time Fourier transform (STFT) [19] of narrowband speech and its corresponding high-frequency components were extracted respectively. A regression DNN was then estimated to map the spectral features of the narrowband speech towards the ones of the high-frequency band frame-by-frame. During the DNN training procedure, unsupervised pre-training by stacking multiple RBMs was first performed [20]. Then the supervised back-propagation algorithm [21] was conducted to fine-tune the DNN parameters under minimum mean squared error (MMSE) criterion. At the stage of restoration, the log power spectra of wideband speech were reconstructed by concatenating the input log power spectra of narrowband speech and the high-frequency log power spectra predicted using the trained DNN. The phase spectra of wideband speech was roughly estimated from the phase spectra of narrowband speech by mirror inversion. Finally, inverse FFT (IFFT) and overlap-add (OLA) algorithm were conducted to reconstruct the wideband waveforms according to the extended magnitude and phase spectra.

3. BN features and deep RNNs

3.1. BN features

Bottleneck features were initially proposed for automatic speech recognition [15, 22, 23]. They are extracted using a specially structured DNN as depicted in Figure 2, which contains a narrow bottleneck layer with smaller number of hidden units compared to the size of other hidden layers. The inputs to this neural network are acoustic features such as mel-frequency cepstral coefficients (MFCCs) and the output layer estimates the posterior probability of HMM states. After the DNN is trained using cross-entropy (CE) criterion, it acts as a BN feature extractor by removing the network parameters from the BN layer to the output layer. The bottleneck layer produces a constriction in the neural network and converts the input information pertinent to the classification of HMM states into a low-dimensional representation. Therefore, BN features can be regarded as a compact, flexible, and non-linear representation of both acoustic and linguistic information [15]. Some similar BN features extracted from regression DNNs have also been utilized in statistical speech synthesis [24, 25].

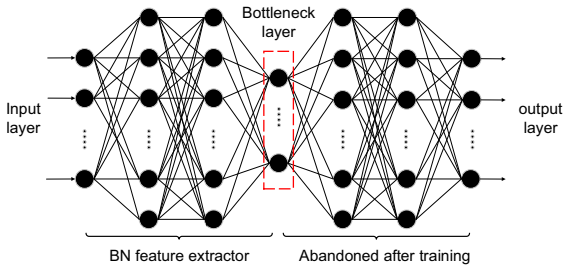

Figure 2: Illustration of a DNN for extracting BN features.

3.2. Deep RNNs with LSTM cells

A recurrent neural network (RNN) is a dynamic neural network where there are cyclical connections among hidden nodes, which is different from feed-forward neural networks. Benefiting from such recurrent connections, the output vector can not only depend on current input vector but also on the history of input vector sequence. RNNs provide better ability of processing dynamic and temporal information than DNNs. Therefore, they are suitable for modeling speech signals and generating acoustic features [16, 17]. For a regression RNN with only one layer of hidden units, given the input vector sequence $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ of T frames, its hidden state vector sequence $\boldsymbol{h} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ and output vector sequence $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T]$ can be formulated as

$$\boldsymbol{h}_t = \mathcal{H}\left(\boldsymbol{W}_{xh}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h\right), \quad (1)$$

$$\boldsymbol{y}_t = \boldsymbol{W}_{hy}\boldsymbol{h}_t + \boldsymbol{b}_y, \quad (2)$$

where $\mathcal{H}$ is the non-linear activation function of the hidden layer, $\boldsymbol{W}_{xh}$,$\boldsymbol{W}_{hh}$ and $\boldsymbol{W}_{hy}$ are the input-hidden, hidden-hidden and hidden-output weight matrices respectively, and $\boldsymbol{b}_h$,$\boldsymbol{b}_y$ are corresponding bias vectors. A deep RNN (DRNN) can be built up by stacking multiple recurrent hidden layers one on top of another, and the hidden state vector in intermediate n-th hidden layer $\boldsymbol{h}^{(n)} = \left[\boldsymbol{h}_1^{(n)}, \boldsymbol{h}_2^{(n)}, \cdots, \boldsymbol{h}_T^{(n)}\right]$ is calculated as

$$\boldsymbol{h}_t^{(n)} = \mathcal{H}\left(\boldsymbol{W}_{h^{(n-1)}h^{(n)}}\boldsymbol{h}_t^{(n-1)} + \boldsymbol{W}_{h^{(n)}h^{(n)}}\boldsymbol{h}_{t-1}^{(n)} + \boldsymbol{b}_{h^{(n)}}\right). \quad (3)$$

Its parameter set can be estimated from training data using back-propagation through time (BPTT) algorithm [26].

The activation function $\mathcal{H}$ in RNNs is conventionally set as a hyperbolic tangent or a sigmoid function. However, training such conventional RNNs is quite difficult due to the problem of exploding and vanishing gradients when the error is back-propagated through multiple number of layers and time steps [27]. The LSTM architecture [18] as illustrated in Figure 3 has been proposed to address this issue. An LSTM cell is a complex hidden unit accomplished with gating structure. The information flow transmitting iteratively through the network is controlled by the input gate, forget gate, output gate and the cell memory state. Therefore, an RNN using LSTM cells is capable of remembering the information from a long span of time steps. For an LSTM-based RNN, the activation function $\mathcal{H}$ is implemented according to the following equations

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i), \quad (4)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f), \quad (5)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \quad (6)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o), \quad (7)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{c}_t), \quad (8)$$

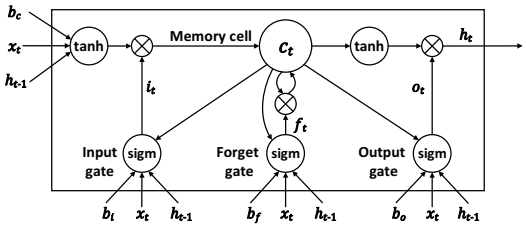

Figure 3: The structure of an LSTM cell.

where σ is the sigmoid function, $*$ denotes an element-wise multiplication, i, f, o and c are the input gate, forget gate, output gate and cell memory respectively.

4. BWE using BN features and DRNNs

Our proposed BWE method using BN features and DRNNs follows the framework similar to the DNN-based one shown in Figure 1. However, the procedures of narrowband feature extraction and neural network training are different from the baseline DNN-based method as shown in Figure 4.

At first, a DNN model with a BN layer for HMM state classification is constructed using a multi-speaker database of narrowband speech with corresponding transcriptions. At the training stage of this DNN, the input features are multi-frame MFCCs and the target outputs are HMM state labels generated by forced alignment using trained GMM-HMM models. After the DNN is well-trained based on CE criterion, it is utilized to form a BN feature extractor and to extract BN feature vectors for all frames of narrowband speech in the training corpus.

Then at frame t, the extracted narrowband BN feature $\boldsymbol{x}_t^{bn}$ and the narrowband log power spectrum $\boldsymbol{x}_t^{lps}$ compose a joint feature vector $\boldsymbol{x}_t = [\boldsymbol{x}_t^{bn\top}, \boldsymbol{x}_t^{lps\top}]^\top$. The log power spectrum is calculated as the logarithmic magnitude spectrum derived from STFT analysis on speech waveforms. $\boldsymbol{x}_t$ acts as an improved input to predict the output feature $\boldsymbol{y}_t = \boldsymbol{y}_t^{lps}$, which is the high-frequency band of the log power spectrum at the t-th frame of wideband speech. Comparing with conventional narrowband log power spectra, the proposed input feature vectors include not only acoustic information but also linguistic information embedded in the BN features. The extra linguistic information is expected to be beneficial to the reconstruction of high-frequency component in BWE systems.

A deep RNN (DRNN) with multiple layers of LSTM cells is utilized in this paper to model the temporal mapping relationship from the input narrowband feature sequence $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ toward the output high-frequency feature sequence $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T]$ as shown in Figure 4. For DRNN training, BPTT algorithm is applied to minimize the mean square error (MSE) between the prediction outputs and the target spectra extracted from training data. At the stage of restoration, the missing high-frequency spectra can be generated sequentially from the sequence of BN features and low power spectra extracted from the input narrowband speech. Given the reconstructed wideband log power spectra, the wideband speech can be generated following the phase generation, IFFT and OLA processes shown in Figure 1.

5. Experiments

5.1. Experimental setup

The TIMIT database [28], which contained English speech from multi-speakers sampled at 16kHz with 16 bits resolution, was

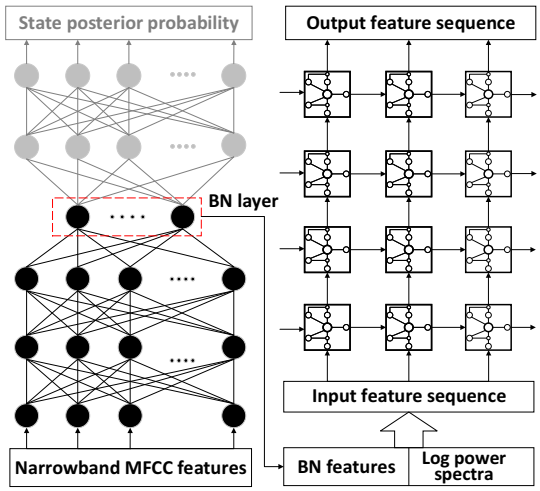

Figure 4: Illustration of the proposed BWE method using BN features and DRNNs.

adopted in our experiments. 3696 utterances were chosen to construct the training set and 1152 utterances were selected as the validation set for model choosing. 192 utterances from the speakers not included in the training set were used as the test set to evaluate the performance of different BWE methods. Parallel narrowband speech at 8kHz was produced by down-sampling the wideband speech at 16kHz in our experiments.

Five BWE systems were constructed using the training dataset for experiments, including:

- ***GMM***: The conventional GMM-based method [3], where the spectra prediction was conducted based on MMSE criterion and the number of mixtures was tuned to be 256 by informal listening tests;
- ***DNN***: The DNN-based method without BN features;
- ***DNN-BN***: The proposed DNN-based method, where log power spectra and BN features were used as inputs;
- ***DRNN***: The proposed DRNN-based method without BN features;
- ***DRNN-BN***: The proposed method using DRNN and BN features.

For extracting log power spectra in all these five systems, the window size of STFT was 320 samples with a shift of 160 samples on the wideband speech, and the narrowband speech took a window size of 160 samples with a shift of 80 samples for consistent frequency resolution. For constructing the BN feature extractor, 11-frames of 39-dimensional narrowband MFCCs were used as the input to a DNN classifier and the output was the posterior probability of 183 HMM states for 61 monophones. The DNN had 6 hidden layers where there were 100 hidden units at the BN layer and 1024 hidden units at other hidden layers. The BN layer was placed on the fifth hidden layer, which was close to the output layer for capturing more linguistic information. According to the MSEs on the validation set for different systems shown in Figure 5, the numbers of hidden layers and hidden units in each hidden layer were set to 5 and 512 in the ***DNN*** system. The ***DNN-BN***, ***DRNN***, ***DRNN-BN*** systems employed the identical network configuration with 4 hidden layers and 1024 hidden units.

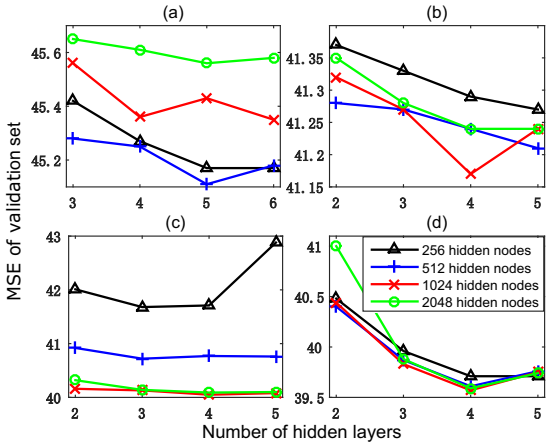

Figure 5: MSEs on the validation set for different model structures in the ***DNN*** (a), ***DNN-BN*** (b), ***DRNN*** (c), and ***DRNN-BN*** (d) systems respectively.

Table 1: LSDs of the wideband speech in the test set reconstructed by different systems.

System	$LSD(dB)$
Narrowband	15.12
GMM	8.50
DNN	6.61
DNN-BN	6.41
DRNN	6.38
DRNN-BN	6.35

5.2. Objective evaluation

The log-spectral distortion (LSD) between the reconstructed wideband speech and the original natural wideband speech is used as the objective quality measure in our experiment. The LSD in dB is defined by

$$LSD = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{2}{N+1}\sum_{k=0}^{N/2}\left(\mathcal{L}S_{tk} - \mathcal{L}\hat{S}_{tk}\right)^2\right]^{\frac{1}{2}}, \quad (9)$$

where $\mathcal{L}S_{tk}$ and $\mathcal{L}\hat{S}_{tk}$ are the log-spectra of natural speech and generated speech at frame t and frequency point k respectively; N is the number of FFT for spectral analysis. Table 1 lists the LSD results of the reconstructed wideband speech in the test set for different BWE systems. The experiment results indicate that all the proposed methods outperformed the conventional GMM-based method and the baseline DNN-based method. The LSDs of the DRNN-based systems were smaller than the DNN-based systems. The ***DRNN-BN*** system achieved the lowest LSD. The results also demonstrate that BN features can help reduce the error of predicting high-frequency spectra, and such improvement is more significant for the DNN-based approach than for the DRNN-based approach.

5.3. Subjective evaluation

Five groups of preference tests were conducted on the crowd-sourcing platform of Amazon Mechanical Turk (`https://www.mturk.com`) to investigate the subjective performance of different BWE systems. In each preference test, the wideband speech of the 20 test utterances randomly selected from the test set were reconstructed by two different systems. Each pair

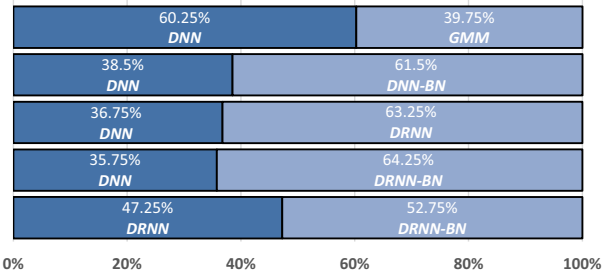

Figure 6: Preference scores among different BWE systems. The p-values of t-test in these tests are 3.5×10^{-5}, 5.4×10^{-6}, 7.2×10^{-8}, 6.2×10^{-9}, and 0.27 respectively.

of generated wideband speech was evaluated in random order by 20 English native listeners. The listeners were required to identify which sentence in each pair sounded better in speech quality.[1] The preference scores of these listening tests are exhibited in Figure 6 with the p-values from t-test. These results demonstrate the DNN-based BWE approach outperformed the conventional method based on GMMs, and the three proposed systems acquired better preference scores than the ***DNN*** system. The subjective results are also consistent with the objective LSD performance. The superiority of the ***DRNN*** system over the ***DNN*** system indicates DRNNs are better than DNNs when modeling the mapping relationship between two feature sequences in the BWE task. The comparison between the ***DNN-BN*** and ***DNN*** systems shows that the BN features are effective to improve the speech quality of reconstructed wideband speech. The proposed ***DRNN-BN*** system took advantages of both DRNNs and BN features and achieved the most significant improvement compared with the baseline DNN-based approach. However, the difference between the *DRNN-BN* and *DRNN* systems are insignificant, which demonstrates the effect of using BN features was decreased when RNNs were adopted for sequence modeling.

6. Conclusions

In this paper, we have proposed a method of using BN features and DRNNs for speech bandwidth extension. Compared with the baseline DNN-based BWE method, employing BN features as input can introduce additional linguistic information into the prediction of high-frequency spectra. DRNNs are good at modeling the temporal dependency and mapping relationship between the acoustic feature sequences corresponding to narrowband speech and high-frequency component. Objective and subjective experimental results show that the proposed method achieved lower LSDs and better preference scores than the DNN-based method. Some other neural networks with more complex structures such as bidirectional DRNNs and hybrid DNN-RNN models will be investigated for BWE in the future.

7. Acknowledgements

This work was partially funded by the National Nature Science Foundation of China (Grant No.61273032), the Electronic Industry Development Fund of Ministry of Industry and Information Technology (Grant No. [2014]425) and National Key Technology Support Program (2014BAK15B05).

[1]Examples of generated speech can be found at `http://home.ustc.edu.cn/~hicolin/demos_IS2016.html`.

8. References

[1] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: Extrapolations of spectral envelop and harmonicity quality of excitation," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*, vol. 3, May 2006, pp. III–III.

[2] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Acoustics, Speech, and Signal Processing (ICASSP), 2001 IEEE International Conference on*, vol. 1. IEEE, pp. 665–668.

[3] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Acoustics, Speech, and Signal Processing (ICASSP), 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1843–1846.

[4] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," in *INTERSPEECH 2014*, 2014, pp. 2489–2493.

[5] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," in *2008 RISP International Workshop on Nonlinear Circuits and Signal Processing*, 2008, pp. 283–286.

[6] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Acoustics, Speech and Signal Processing (ICASSP), 2004 IEEE International Conference on*, vol. 1, May 2004, pp. I–709–12 vol.1.

[7] G.-B. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[8] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[9] 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," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 35–52, May 2015.

[10] Y. Gu and Z.-H. Ling, "Restoring high frequency spectral envelopes using neural networks for speech bandwidth extension," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, July 2015, pp. 1–8.

[11] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4395–4399.

[12] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *INTERSPEECH 2015*, September 2015, pp. 2593–2597.

[13] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *INTERSPEECH 2015*, September 2015, pp. 2598–2602.

[14] K. Li, Z. Huang, Y. Xu, and C. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *INTERSPEECH 2015*, September 2015, pp. 2578–2582.

[15] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH 2011*, 2011, pp. 237–240.

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

[17] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4470–4474.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

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

[21] D. Rummelhart, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 9, pp. 533–536, 1986.

[22] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3377–3381.

[23] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[24] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4460–4464.

[25] Z. Wu and S. King, "Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features," in *INTERSPEECH 2015,*, September 2015, pp. 309–313.

[26] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct 1990.

[27] S. Hochreiter, Y. Bengio, and P. Frasconi, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *Field Guide to Dynamical Recurrent Networks*, J. Kolen and S. Kremer, Eds. IEEE Press, 2001.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.