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

A Novel Research to Artificial Bandwidth Extension Based on Deep BLSTM Recurrent Neural Networks and Exemplar-based Sparse Representation

Bin Liu [1], *Jianhua Tao* [1,2]

[1] National Laboratory of Pattern Recognition,
[2] CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190；
liubin@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn

Abstract

This paper presents a two stages artificial bandwidth extension (ABE) framework which combine deep bidirectional Long Short Term Memory (BLSTM) recurrent neural network with exemplar-based sparse representation to estimate missing frequency band. It demonstrates the suitability of proposed method for modeling log power spectra of speech signals in ABE. The BLSTM-RNN which can capture information from anywhere in the feature sequence is used to estimate the log power spectra in the high-band firstly and the exemplar-based sparse representation which could alleviate the over-smoothing problem is applied to generated log power spectra in the second stage. In addition, rich acoustic features in the low-band are considered to reduce the reconstruction error. Experimental results demonstrate that the proposed framework can achieve significant improvements in both objective and subjective measures over the different baseline methods.

Index Terms: BLSTM-RNN, artificial bandwidth extension, rich acoustic features, exemplar-based sparse representation

1. Introduction

Although intelligibility of the narrowband speech is acceptable, wideband speech contains frequency components beyond the telephone band. The missing frequency band carries spectrally rich information for the speech signal. Upgrading to wideband speech communication requires the thorough structure to be redesigned, which is a huge burden. For this purpose, artificial bandwidth extension (ABE) has been studied widely to improve quality of the narrowband speech [1, 2]. They attempt to regenerate the missing spectral content at the receiver based on narrowband speech input.

Most of ABE methods use the source-filter model of speech production to estimate wideband spectral envelopes and excitation signal independently. As it is stated in [2], an extension of the spectral envelope has greater contribution to the perceived speech quality compared to the extension of the excitation. Therefore, more emphasis is given to the extension of the spectral envelope. However, the vocoder will lead to degradation of speech quality in source-filter model based ABE. To solve this problem, log power spectra and phase spectra are extended from the narrowband speech. The log power spectra in the high-band are estimated using features extracted from the narrowband speech [3]. Features indicate the low-band spectral shape are typically used which include spectral vectors [4], mel-frequency cepstral coefficients (MFCC) [5] and line spectral frequencies (LSF) [6].

A large number of ABE model are already proposed. Enbom and Kleijn use vector quantization (VQ) to the spectral envelope of the wideband signal [7]. The generated spectral envelope may be discontinuous in VQ method. Kim and Park accomplish the goal of ABE using Gaussian mixture model (GMM) [8]. GMM will lead to over-smoothing problem. Jax and Vary suggest the usage of HMM for wideband feature estimation [2] which the temporal trajectory of speech parameters could be captured. A straightforward ABE method based on sum product network (SPN) was described and evaluated in [9]. The SPN is regarded as observation models in HMMs modeling. The algorithm is robust and computationally inexpensive [10]. Deep learning has emerged as a new area of machine learning research [11]. It can discover the underlying regularity of multiple features, and have strong generalization abilities than shallow models [12]. The restrict Boltzmann machine (RBM) was used to build a deep belief network (DBN) [13] in ABE. The log power spectra [14] and spectral envelope [15] in the high-band could be regenerated based on RBM-DBN. The Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) model [16, 17] was devised to better find and exploit long-range context using special memory cells compared to RBM-DBN. It can also be stacked together in multiple layers and to have deep structure in space. It has been shown to efficiently model a self-learned amount of feature-level context and to be highly beneficial to voice conversion [18] and speech synthesis [19]. Therefore, BLSTM-RNN could be suitable for speech generation problems. The exemplar-based sparse representation has been proposed to model the high-resolution spectra directly for voice conversion [20] and speech enhancement [21]. This method assumes that a target spectrogram can be generated from a small set of basis target spectra through a weighted linear combination. In this way, the target spectrogram is generated from the real target speech exemplars rather than generated from model parameters. It is an effective method as post-processing step in speech generation [22].

In this study, we present a novel ABE method to learn the complex mapping relationship from narrowband speech to wideband speech. We demonstrate the suitability of proposed method for modeling log power spectra of speech signals using in the application of ABE. We propose to use two stages to perform ABE. In the first stage the BLSTM-RNN is used to produce the log power spectra in the high-band. In the second stage, the exemplar-based sparse representation is applied to generated log power spectra. Our two-stage approach is compared to different baseline approaches, including SPN, RBM-DBN and BLSTM-RNN.

http://dx.doi.org/10.21437/Interspeech.2016-772

2. Proposed method

In this section, we firstly introduce the framework of the proposed ABE method. Subsequently, the further details are presented. The flowchart of proposed algorithm is shown in Figure 1.

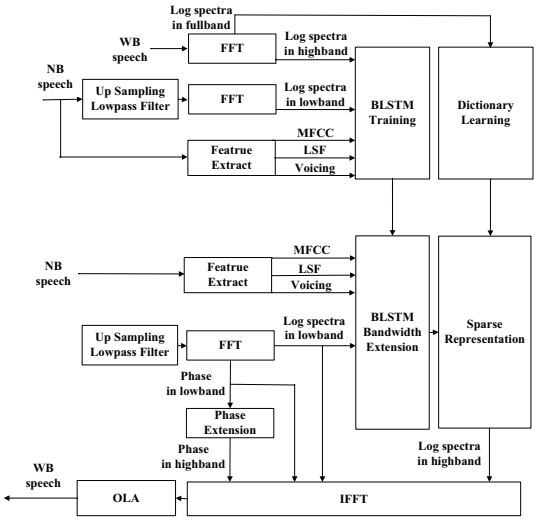

Figure 1: *Block diagram of the proposed method*

There are total four parts in the proposed method which includes features extraction, BLSTM-RNN training with rich acoustic features, exemplar-based sparse representation and wideband speech reconstruction. In features extraction stage, the original speech is divided into overlapping frames; the different features are extracted for each frame. For BLSTM-RNN training, a regression BLSTM-RNN is trained for mapping from the narrowband speech features to the wideband speech features. The exemplar-based sparse representation is applied to generated log power spectra in the high-band to alleviate the over-smoothing problem. Therefore, the log power spectra in the high-band could be predicted in two stages. The speech signal is reconstructed according to predicted frequency spectra in the high-band and original frequency spectra in the low-band.

2.1. Features extraction

The input signal is narrowband speech sampled at 8 kHz. It is up sampled to the sampling rate of 16 kHz, prefiltered with a low pass filter, and windowed into 32 ms frames with 16 ms overlap using a Hamming window. The up sampled signal is used for extracting both log power spectra and phase spectra in the low-band. The target signal is wideband speech sampled at 16 kHz and it is used for extracting log power spectra which the high-band are used for BLSTM-RNN model training and the full-band are applied to dictionary learning.

We apply BLSTM-RNN to combine the multiple-features and reconstruct the log power spectra in the high-band for the stronger information fusion ability. The selected acoustic features include MFCC, LSP and band pass voicing coefficient (BPVC) [23]. The above mentioned features show high correlation to the log power spectra. The dimensions of the different acoustic features are referred in [3].

2.2. BLSTM-RNN training with rich features

The Deep BLSTM-RNN can model the deep representation of long-span acoustic features for ABE. A BLSTM layer consists of a number of recurrently connected such memory blocks which could solve gradient vanish and gradient expansion problem. Each block contains the connected memory cells and three multiplicative units, which could respectively provide write, read, reset operation for the cells. The surrounding network can only interact with the memory cells via the gates. Two separate recurrent hidden layers are operating in opposite directions, thus providing access to long-range context in both input directions. The rich acoustic features are selected as the input in BLSTM-RNN and transformed according to equation (1) and (2):

$$\overrightarrow{h_t^1} = \mathcal{H}\left(W_{\underset{xh^1}{\rightarrow}} x_t + W_{\underset{h^1h^1}{\rightarrow\rightarrow}} \overrightarrow{h_{t-1}^1} + b_{\underset{h^1}{\rightarrow}}\right) \quad (1)$$

$$\overleftarrow{h_t^1} = \mathcal{H}\left(W_{\underset{xh^1}{\leftarrow}} x_t + W_{\underset{h^1h^1}{\leftarrow\leftarrow}} \overleftarrow{h_{t+1}^1} + b_{\underset{h^1}{\leftarrow}}\right) \quad (2)$$

where x represents the acoustic features in the low-band, $\overrightarrow{h}$ and $\overleftarrow{h}$ are hidden vector for forward sequence and backward sequence respectively, $\mathcal{H}$ is the activation function of hidden layer, W is the weight matrix, b is the bias vectors, t indicates frame index, and 1 indicates the first hidden layer.

Deep bidirectional RNN can be established by stacking multiple RNN hidden layers on top of each other and transform the input sequence. The iterative process is:

$$\overrightarrow{h_t^n} = \mathcal{H}\left(W_{\underset{h^{n-1}h^n}{\rightarrow\rightarrow}} \overrightarrow{h_t^{n-1}} + W_{\underset{h^nh^n}{\rightarrow\rightarrow}} \overrightarrow{h_{t-1}^n} + b_{\underset{h^n}{\rightarrow}}\right) \quad (3)$$

$$\overleftarrow{h_t^n} = \mathcal{H}\left(W_{\underset{h^{n-1}h^n}{\leftarrow\leftarrow}} \overleftarrow{h_t^{n-1}} + W_{\underset{h^nh^n}{\leftarrow\leftarrow}} \overleftarrow{h_{t+1}^n} + b_{\underset{h^n}{\leftarrow}}\right) \quad (4)$$

$$y_t = W_{\underset{h^Ny}{\rightarrow}} \overrightarrow{h_t^N} + W_{\underset{h^Ny}{\leftarrow}} \overleftarrow{h_t^N} + b_y \quad (5)$$

where N represents the number of hidden layers, y is the log power spectra in the high-band, and n indicates hidden layers index.

The weights of BLSTM-RNN are trained by using pairs of input features x_t and output features y_t extracted from training data to minimize the errors between the mapped output from the given input and the target output. The Back-propagation through time (BPTT) algorithm is applied to both forward hidden nodes and backward hidden nodes, and back-propagates layer by layer. The weight gradients are computed over the entire utterance. The output features of BLSTM-RNN should be transformed back as follow:

$$y_t'(d) = y_t(d) \times v(d) \times \alpha(d) + m(d) \quad (6)$$

where *m(d)* and *v(d)* are the *d-th* component of the mean and variance of the output feature, $\alpha(d)$ could be used to lift the variance of the reconstructed log power spectra proposed in [3], and $y_t'(d)$ represents the reconstructed log power spectra in the high-band.

The effective learning capability of BLSTM-RNN is expected to benefit ABE. Deep-layered architectures can represent high level representation of input features [24] and BLSTM-RNN can capture information from anywhere in the feature sequence. Therefore, the functions can be compactly represented with deep BLSTM-RNN which can outperform both shallow architecture model and RBM-DBN model.

2.3. The exemplar-based sparse representation

The generated log power spectra based on BLSTM-RNN inevitably contain distortion that may negatively affect the perceptual quality of the speech signal. To combat this effect, a sparse representation approach is used to reconstruct the log

power spectra in the high-band. This method could model high-resolution spectra for spectral details, it has been proposed for voice conversion [20] and speech enhancement [21].

The generated wideband speech can be represented as a sparse linear combination of basis vectors [25]. Each segment of spectra can be modeled independently using the same dictionary. Given a dictionary $D \in R^{N\times K}$ and activation matrix $A \in R^{K\times T}$, the log power spectra $Z \in R^{N\times T}$, are approximated as $D \times A$. Letting A=[a_1, a_2,… , a_T], Z=[z_1, z_2,… , z_T], D=[d_1, d_2,… , d_K], then

$$[z_1, z_2 \ldots\ldots z_T] \approx [d_1, d_2 \ldots\ldots d_K][a_1, a_2 \ldots\ldots a_T] \quad (7)$$

where a_t represents the sparse coefficient vector, d_k is basis vector in the dictionary, z_t indicates the log power spectra of wideband speech which covers with original low-band and generated high-band by BLSTM-RNN, T is the number of frame, K is the number of basis vectors, and N is the dimension of log power spectra. N is set to 257 in this paper. The important steps in sparse representations are then to define the dictionary, compute the activation matrix for a given log power spectra.

The dictionary D is generated by the concatenation of log power spectra of original wideband speech. The dictionary training problem is solved via minimization equation (8), where one starts with initial guesses for the dictionaries.

$$\min \sum_{t=1}^{N} \| z_t' - Da_t \|_2^2 \quad s.t. \ \| a_t \|_0 \le L \quad (8)$$

where z_t' indicates the log power spectra of original wideband speech and N represents the number of frame in training data. Subsequently, one alternates between the following two steps: (1) the sparse coefficient vectors are updated for fixed dictionaries using the orthogonal matching pursuit (OMP) algorithm [26]; (2) the dictionaries are updated using the K-SVD algorithm [27]. Given D and Z, A is found by solving the following equation,

$$a_t = \arg\min_a \| z_t - Da \|_2^2 \quad s.t. \ \| a \|_0 \le L, 1 \le t \le T \quad (9)$$

where L is a parameter that controls sparseness called sparsity and T indicates the number of frames in one utterance. Since the goal is to approximate the log power spectra of predicted speech by BLSTM-RNN using sparse representations, the only constraint for L is that it is much smaller than the number of basis vectors in the dictionary. In practice, the OMP algorithm is adopted to estimate the activation matrix A. With the parameters solved, the log power spectra of wideband speech are approximated; then the log power spectra in low-band is replaced with original log power spectra and the log power spectra in high-band is reserved. The full-band is considered in the dictionary due to the strong dependency between low-band and high-band. Generated log power spectra based on BLSTM-RNN is replaced with a sparse linear combination of the log power spectra from original wideband speech to alleviate the over-smoothing problem.

2.4. WB speech reconstruction

To synthesize a time-signal from the bandwidth extended log power spectra, we need to associate a phase to the estimated magnitude spectra. The extension of the phase spectra has a minor role compared to the extension of the amplitude spectra in improving the perceived speech quality. In order to recover phase information for ABE, we employ a simple, yet effective, phase mirroring inversion method for the extension of the phase spectrum. The wideband phase is estimated from up sampled narrowband phase spectra via mirroring inversion.

We reconstructed speech signal according to predicted frequency spectrum in high-band and original frequency spectrum in low-band. IFFT and Overlap-and-Add (OLA) are performed to get the reconstructed WB speech.

3. Experiments and result analysis

3.1. Data and analysis methodology

In this section, we evaluate the proposed approach on ABE task. For BLSTM-RNN training, the 5000 utterances selected randomly from the TIMIT database were used for training and another 1000 randomly selected utterances from the TIMIT database [28] were used to optimum model parameter. For exemplar-based sparse representation, the 1000 utterances were selected randomly to train the dictionary from the TIMIT database and another 1000 randomly selected utterances from the TIMIT database were used to optimum hyper parameter. We compared our proposed ABE algorithm with three different baseline algorithms on the GRID corpus [29], where we used the test speakers with numbers 1, 2, 18, and 20. The test set for our algorithm and different baseline algorithms is the same. Feature extraction is performed for each 32 ms with 50% overlap. The first baseline is the method proposed in [10], based on the SPN-HMM. We refer as SPN to this baseline. The second baseline is based on the RBM-DBN model [3], and referred as RBM-DBN in the following experiment. The third baseline is based on the BLSTM-RNN which the sparse representations are not considered. We refer as BLSTM.

The evaluation of ABE system is performed with three distinct objective metrics. The frequency weighted segmental SNR (fwSNRseg) [30], Itakura-Saito distance (IS) [31] and Log Likelihood Ratio (LLR) [32] are employed to compare the synthesized wideband speech to the original wideband speech. The logarithmic spectral distortion (LSD) is used to evaluate the estimated log power spectra in the high-band. In addition, we have performed a subjective preference comparison test to evaluate the different ABE system.

3.2. The evaluation of hyper parameter configure

For BLSTM-RNN, learning rate was set at 0.0005 for the first 10 epochs, and then decreased by 10% after every epoch. Total number of epoch was 20. There are two hidden layers and the number of hidden unit is 1024 for each direction. The number of cell is set to 800. Input features of BLSTM-RNN were normalized to zero mean and unit variance.

Figure 2 (left) shows the average LSD results on the test set using input features with different hidden units (256, 512, 1024 and 2048) and different the number of hidden layers on BLSTM-RNN. The input features include different acoustic parameter proposed in [3]. It is clear that the more hidden units the BLSTM-RNN were fed with, the better the performance could be achieved. But the more nodes also made the BLSTM-RNN structure more complicated to learn in training. Poor results were obtained if there is one hidden layer, which was a kind of shallow model, indicating that the deep layer structure is very important to obtain a more generalized model. The optimum performance was obtained while the number of hidden layer was two. For BLSTM-RNN, the reconstructed error is lower compared with the RBM-DBN because it can capture information from anywhere.

Figure 2 (right) shows the average LSD results on the test set using different features in the dictionary with different sparsity on exemplar-based sparse representation method. We could obtain the lowest reconstruction error while the full-band features are considered. It could be explained that there is strong correlation among different frequency band and the features in low-band could contribute to predict the features in high-band. The reconstruction error is largest while the low-band features are considered due to it is a shallow model and the low-band and high-band share the same activation matrix. The optimum performance was obtained while the sparsity is set to 10.

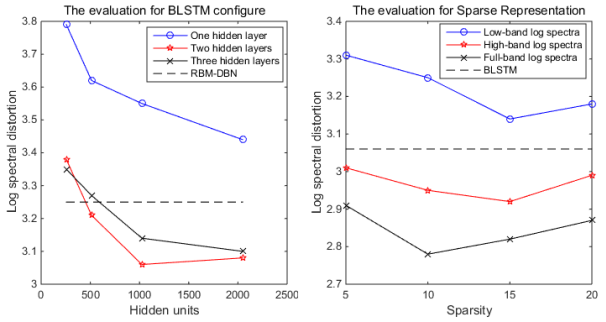

Figure 2: *The evaluation of hyper parameter configure, left: BLSTM configure, right: sparse representation configure*

3.3. Overall evaluation

3.3.1. Objective test evaluation

This test is used to measure the objective quality of estimated speech. The fwSNRseg, IS distance and LLR are adopted. Tables 1 show the performance of all four ABE methods respectively. The proposed method always performs best and there is highest fwSNRseg, lowest Itakura-Saito distance and Log Likelihood Ratio.

Table 1. *The objective test for different ABE method*

ABE method	fwSNRseg	LLR	IS
SPN	14.69	0.83	3.14
RBM-DBN	24.96	0.75	3.03
BLSTM	26.91	0.72	2.91
Proposed ABE	27.69	0.67	2.78

3.3.2. Subjective test evaluation

During the subjective test, the subjects are asked to indicate their preference for each given ABE test pair where the scale corresponds to prefer A, no preference and prefer B. The subjective preference test includes 10 listeners, who compared 20 sentence pairs randomly chosen from test database. The proposed method is compared with three different baseline methods respectively. Test results, given in Figure 3 indicate that, speech synthesized with the proposed method outperforms the speech synthesized with the different baseline methods significantly. Proposed ABE yields a brighter sound and produce more clear than three different baseline methods.

Figure 4 gives an example of one female utterance. The spectrograms of the bandwidth-extended speech, using the proposed method is shown in Figure 4 (right); the spectrograms of the actual wideband speech are also shown in Figure 4 (left) to facilitate visual comparison. We observe that the proposed algorithm recovers the missing high-frequency part of the input narrowband spectrogram reasonably well.

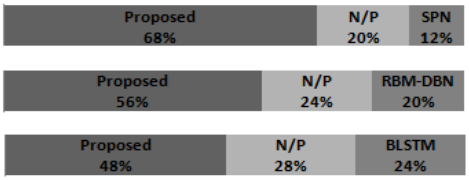

Figure 3: *Subjective test for different ABE method*

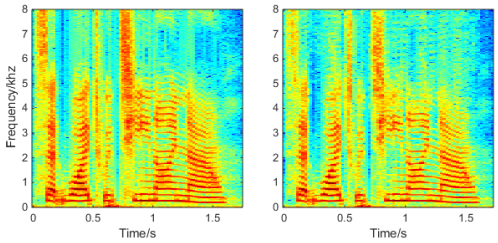

Figure 4: *Spectrogram of one female test utterance, left: original signal, right: reconstructed signal*

3.4. Discussion

We introduce novel ABE which combines BLSTM-RNN with exemplar-based sparse representation to estimate log power spectra. Deep-layered architectures can represent high level representation of input features and BLSTM-RNN can capture information from anywhere in the feature sequence. The exemplar-based sparse representation could alleviate the over-smoothing problem. Motivated by the success of two stages speech generation on the speech enhancement, we combine BLSTM-RNN and exemplar-based sparse representation to reconstruct log power spectra in the high-band. The resulting system clearly improves the state-of-the-art both in subjective performance evaluation and objective performance evaluation. We demonstrated that proposed ABE is a promising regression model for speech, applying them to ABE.

4. Conclusions

In this paper, two stages ABE which combined BLSTM-RNN and exemplar-based sparse representation are proposed. Among the various BLSTM-RNN configurations, the deep architecture is crucial to learn the complex structure of the mapping function from narrowband speech to wideband speech. It was found that the BLSTM-RNN which can capture information from anywhere improves the system performance. In addition, the exemplar-based sparse representation was effective in solving over-smoothing problem of the reconstructed log power spectra. Compared with the different baseline method, the proposed framework achieves significant improvements in both objective and subjective measures.

In future studies, we would design the real-time ABE system. We also will consider training the model respectively according different phone classification. In addition, the phase spectra bandwidth extension will also be investigated.

5. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No. 61305003), the Strategic Priority Research Program of the CAS (GrantXDB02080006) and the Major Program for the National Social Science Fund of China (13&ZD189).

6. References

[1] Y. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech," *IEEE Transactions on Speech & Audio Processing,* vol.2, no.4, pp. 544-548, 1992.

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

[3] B. Liu, J. Tao, and Z. Wen, et al., "A Novel Method of Artificial Bandwidth Extension Using Deep Architecture," in *Annual Conference of the International Speech Communication Association, Interspeech*, 2015, pp. 2598–2602.

[4] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation." in *icassp IEEE Computer Society*, 2000, pp. 1843-1846.

[5] A. H. Nour-Eldin and P. Kabal, "Combining frontend-based memory with MFCC features for Bandwidth Extension of narrowband speech." in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2009, pp.4001-4004.

[6] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech." in *Annual Conference of the International Speech Communication Association, Interspeech*, 2003, pp. 1433--1436.

[7] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients." in *Speech Coding Proceedings,* 1999, pp. 171-173.

[8] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech" in *Annual Conference of the International Speech Communication Association, Interspeech*, 2005, pp. 1509-1512.

[9] H. Pulakka, L. Laaksonen, and M. Vainio, et al., "Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages." *IEEE Transactions on Audio Speech & Language Processing*, vol.16, no.6, pp. 1124-1137, 2008.

[10] R. Peharz, G. Kapeller, and P. Mowlaee, et al., "Modeling speech with sum-product networks: Application to bandwidth extension." in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2014, pp. 3699-3703.

[11] G. E. Hinton, S. Osindero and Y. W. The, "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation*, vol. 18, no.7, pp. 1527-54, 2006.

[12] D. Erhan, Y. Bengio, A. Courville, et al., "Why Does Unsupervised Pre-training Help Deep Learning?" *Journal of Machine Learning Research*, vol.3, no.6, pp. 625-660, 2010.

[13] Y. Wang, S. Zhao, and D. Qu, et al., "Using conditional restricted Boltzmann machines for spectral envelope modeling in speech bandwidth extension", in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2016, pp. 5930-5934.

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

[15] Y. Wang, S. Zhao, and W. Liu, et al., "Speech bandwidth expansion based on deep neural network," in *Annual Conference of the International Speech Communication Association, Interspeech*, 2015, pp. 2593–2597.

[16] A. Graves, "Long Short-Term Memory", *Neural Computation*, vol.9, no. 8, pp. 1735-80, 1997.

[17] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, et al., "Learning precise timing with lstm recurrent networks." *Journal of Machine Learning Research,* vol.3, no.1, pp. 115-143, 2003.

[18] L. Sun, S. Kang and K. Li, et al., "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks", in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2015, pp. 4869-4873.

[19] Y. Fan, Y. Qian and F. Xie et al., "TTS Synthesis with Bidirectional LSTM based Recurrent Neural Network," in *Annual Conference of the International Speech Communication Association, Interspeech*, 2014, pp. 1964–1968.

[20] Z. Wu, E. S. Chng and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization." *Multimedia Tools & Applications,* vol.74, no.22, pp. 9943-9958, 2015.

[21] D. S. Williamson, Y. Wang and D. Wang, "A two-stage approach for improving the perceptual quality of separated speech." in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2014, pp. 7034 - 7038.

[22] D. S. Williamson, Y. Wang and D. Wang, "A sparse representation approach for perceptual quality improvement of separated speech." in *International Conference on Acoustics, Speech, & Signal Processing, ICASSP*, 2014, pp. 7015-7019.

[23] L. Supplee, R. Cohn, and J. Collura, et al., "MELP: the new Federal Standard at 2400 bps." in *International Conference on Acoustics,*1997, pp. 1591-1594.

[24] Y. Bengio. "Learning Deep Architectures for AI." *Foundations & Trends® in Machine Learning,* vol.2, no.1, pp.1-127, 2009.

[25] E. Michael and A. Michal, "Image denoising via sparse and redundant representations over learned dictionaries." *IEEE Transactions on Image Processing,* vol.15, no.12, pp. 3736-45, 2007.

[26] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. on Info. Theory, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[27] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. on Sig. Proc., vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[28] J. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", *National Institute of Standards and Technology Gaithersburgh MD*, 1988.

[29] C. Martin, B. Jon, and C. Stuart, et al., "An audio-visual corpus for speech perception and automatic speech recognition", *Journal of the Acoustical Society of America,* vol.120, no.5Pt1, pp. 2421-2424, 2006.

[30] Y. Hu, and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Transactions on Audio Speech & Language Processing*, vol.16, no.1, pp. 229-238, 2008.

[31] F. Itakura and S. Saito, S, "Analysis synthesis telephony based on the maximum likelihood method", International Congress on Acoustics, 1968, pp. 17–20.

[32] S. R. Quackenbush, "Objective Measures of Speech Quality", Prentice Hall Advanced Reference Series, Englewood Cliffs, NJ, Georgia Institute of Technology, 1988.