A Novel Method of Artificial Bandwidth Extension Using Deep Architecture

Bin Liu, Jianhua Tao, Zhengqi Wen, Ya Li, Danish Bukhari

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190；

liubin@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, zqwen@nlpr.ia.ac.cn, yli@nlpr.ia.ac.cn

10.21437/Interspeech.2015-559

Abstract

This paper presents a novel artificial bandwidth extension (ABE) framework based on deep neural networks (DNNs) with a multiple-layer's deep architecture. It demonstrates the suitability of DNNs for modeling log power spectra of speech signals using the application of ABE. The DNN is used to estimate the log power spectra in the high-band. Two strategies are proposed to improve the performances of the proposed ABE system. First, global variance equalization is proposed to alleviate the over-smoothing issue in generated log spectra. Second, rich acoustic features in the low-band are considered to improve the construction of the log power spectra in the high-band. Experimental results demonstrate that the proposed framework can achieve significant improvements in both objective and subjective measures over the different baseline methods.

Index Terms: deep neural networks, artificial bandwidth extension, rich acoustic features, global variance equalization

1. Introduction

Although intelligibility of the narrowband speech is acceptable, natural speech contains frequency components beyond the telephone band. The missing frequency band carries spectrally rich information of the speech signal. Upgrading to wideband speech communication requires the thorough structure to be redesigned, which is an economical burden. For this purpose, artificial bandwidth extension (ABE) has been studied widely to upgrade the quality of the conventional narrowband speech [1, 2]. They attempt to regenerate the missing spectral content at the receiver based on narrowband speech input. This paper addresses ABE in the frequency range 4 – 8 kHz, which is denoted as the high-band. The frequency range below 4 kHz is denoted as the low-band.

Most ABE methods use the source-filter model of speech production to estimate wideband spectral and temporal envelopes independently. This model breaks down the problem into two parts, namely the extension of the excitation (source) and the extension of the spectral envelope (filter). As it is stated in [2], an extension of the envelope has a 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. The spectral envelope parameters are estimated using features extracted from the narrowband speech. Features indicating the low-band spectral shape are typically used and additional frequency-domain and time-domain features are often utilized [1]. ABE methods have been used with spectral vectors [3], mel-frequency cepstral coefficients (MFCC) [4], and line spectral frequencies (LSF) [5].

Enbom and Kleijn use vector quantization to estimate the spectral envelope of the wideband signal [6]. Wideband features and narrowband features are grouped together and trained, followed by the construction of the codebook using the LBG algorithm. Kim and Park accomplish the goal of ABE using Gaussian mixture model (GMM) [7]. Similarly, narrowband features and wideband features are fused, trained with GMM using expectation maximization. Further approaches utilizing GMM in ABE include adjusting the temporal envelope and gain of the high-band based on GMM-estimated parameters [3] and using GMM for both high-band prediction and denoising of low-band features [8]. Jax and Vary suggest the usage of HMM for wideband feature estimation [2]. They build a supervised HMM structure based on VQ clusters, and exploit the statistics of the HMM state sequence in the MMSE estimation of wideband spectra. A straightforward ABE method was described and evaluated in [9]. The algorithm is robust and computationally inexpensive. In addition, the Sum-product networks (SPNs) which can be interpreted as a neural network representing an inference machine is applied to ABE in [10]. The SPNs is regarded as observation models in HMMs modeling. The resulting log-spectrograms exhibit an improved speech quality and significant improvement over state-of-the-art methods. It is referred to as the baseline method in this work.

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. The basic strategy is to train a deep network with greedy layer wised pre-training plus fine tuning [12]. The restrict Boltzmann machine (RBM) was widely used to build a deep belief network (DBN) [13] in speech recognition [14] and speech enhancement [13].

In this study, we present a novel ABE framework and propose to learn the complex mapping function from narrowband speech to wideband speech with nonlinear deep neural network regression models. We demonstrate the suitability of DNNs for modeling log power spectra of speech signals using in the application of ABE. The narrowband log power spectra are calculated from input speech and the DNNs are used to estimate the log power spectra in the high-band. Two strategies are proposed to further improve the quality of wideband speech and generalization capability of DNNs. First, global variance equalization is proposed to alleviate the over-smoothing issue in DNN-based ABE system. Second, rich acoustic features are considered to predict the log power spectra in the high-band to improve performance.

The rest of this paper is organized as follows: the detail of the proposed approach is introduced in Section 2. The evaluation results are shown in Section 3. Conclusions will be elaborated in the Section 4.

2. Proposed ABE framework

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

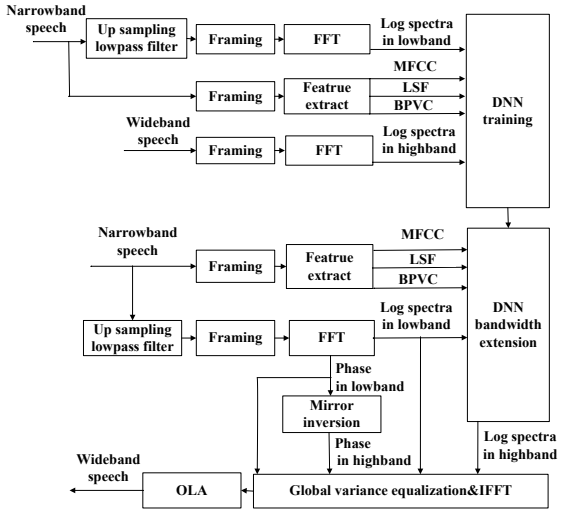

Figure 1: Block diagram of the proposed method

There are total three parts in the proposed method which includes features extraction, DNNs training with rich acoustic features 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 DNNs training, a regression DNNs are trained for mapping from the narrowband speech features to the wideband speech features. So the log power spectra in the high-band could be predicted according to DNNs. The global variance equalization is applied to post-processing for predicted spectra and solve over-smoothing problem. 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 16 ms frames with 8 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 in the high-band for DNNs training.

For DNNs, it not only fuses the shallow advantages of all acoustic features together naturally, but is also able to incorporate the deep regularity of the acoustic features [15]. We apply DNNs to combine the multiple-features and reconstruct the log power spectra in the high-band for the stronger information fusion ability. To better show the advantages of the feature combination techniques, we extract different acoustic features besides log power spectra from the narrowband speech. The selected acoustic features include MFCC, LSP and band pass voicing coefficient (BPVC) [16]. The above mentioned features show high correlation to the log power spectra. The dimensions of the different acoustic features are listed in Table I.

Table I. *The attributes of the features*

ID	Feature	dimension
1	log power spectra (low-band)	129
2	MFCC	12
3	LSF	10
4	BPVC	5
5	log power spectra (high-band)	128

2.2. DNNs training with rich acoustic features

The DNNs training, starting with a randomly initialized network, typically finds poor local minima [17]. Hence, as in [18], we firstly learn a deep generative model of narrowband speech features by a stacking of multiple Restricted Boltzmann Machines (RBMs) [19]. The first one is a Gaussian-Bernoulli RBM that has one visible layer of linear variables, connected to a hidden layer. Then a pile of Bernoulli-Bernoulli RBMs can be stacked behind the Gaussian-Bernoulli RBM. Afterwards, they can be trained layer-by-layer in an unsupervised greedy fashion [17]. During this step, an objective criterion, called contrastive divergence (CD), is used to update the parameters of each RBM [19]. Back-propagation (BP) algorithm with the minimum mean squared error (MMSE) object function between the target and predicted log-power spectra in the high-band is used to train the DNNs. A stochastic gradient descent algorithm is performed in mini-batches with multiple epochs to improve learning convergence as follows,

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(X_n^{'d}(W^{\zeta},b^{\zeta}) - X_n^{d})^2 \quad (1)$$

where E is the mean squared error, $X_n^{'d}(W^{\zeta},b^{\zeta})$ and X_n^{d} denote the *d-th* predicted and target frequency bins of the log power spectra at sample index n, respectively, with N representing the mini-batch size, D being the size of the log power spectra vector, (W^{ζ},b^{ζ}) denoting the weights and bias parameter to be learned at the $\zeta - th$ layer, with L indicating the total number of hidden layers and $L+1$ representing the output layer. Then the updated estimate of the weight W and bias b, with the learning rate λ, can be computed iteratively in the following:

$$(W^{\zeta},b^{\zeta}) \leftarrow (W^{\zeta},b^{\zeta}) - \lambda\frac{\partial E}{\partial(W^{\zeta},b^{\zeta})}, 1 \le \zeta \le L+1 \quad (2)$$

The DNNs are capable of capturing the context information with rich features by concatenating them into a long feature vector for the DNNs learning. The output features of DNNs should be transformed back as follows:

$$X_n^{'}(d) = X_n(d) \times v(d) + m(d) \quad (3)$$

where *m(d)* and *v(d)* are the *d-th* component of the mean and variance of the output feature. $X_n^{'}(d)$ represents the reconstructed log power spectra in the high-band.

The over-smoothing problem causes the degradation of performance on the reconstructed log power spectra. Global variance equalization between the global variance of the estimated and reference wideband speech features is applied to alleviate this problem. It is demonstrated that the use of global variance information could significantly improve the performance in voice conversion [20] and speech enhancement

[21] respectively. The global variance of the estimated wideband speech feature is defined as [20]:

$$GV(d) = \frac{1}{M}\sum_{n=1}^{M}(X_n^{'}(d) - \frac{1}{M}\sum_{n=1}^{M}X_n^{'}(d))^2 \quad (4)$$

where $X_n^{'}(d)$ is the *d-th* component of a DNNs output vector at the *n-th* frame and *M* is the total number of speech frame in the training set. The global variance of the reference wideband speech features can be calculated in a similar way. Fig. 2 shows the global variances of the estimated and reference log power spectra of wideband speech in the high-band. It can be observed that the global variances of the estimated wideband speech features were smaller.

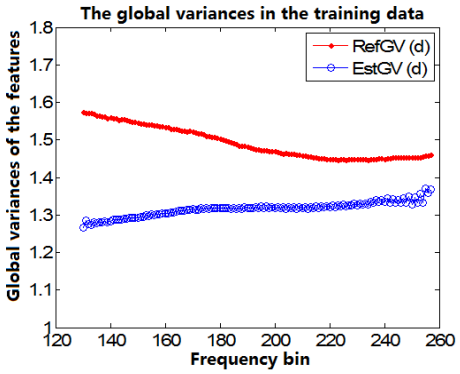

Fig. 2. The global variances on the training set.

To address the over-smoothing problem, a global equalization factor $\alpha(d)$ is applied as follows [21]:

$$\alpha(d) = \sqrt{\frac{GV_{ref}(d)}{GV_{est}(d)}} \quad (5)$$

where $GV_{ref}(d)$ and $GV_{est}(d)$ represented the *d-th* dimension of the global variance of the reference features and the estimation features, respectively. The output features of DNNs should be transformed back as follows:

$$X_n^{'}(d) = X_n(d) \times v(d) \times \alpha(d) + m(d) \quad (6)$$

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 as the post-processing. This step could effectively sharpen the formant peaks in the high-band especially for female speech which could significantly improve the overall listening quality.

2.3. Wideband 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 the high-band and original frequency spectrum in the low-band. IFFT and Overlap-and-Add (OLA) are performed to get the reconstructed wideband speech.

3. Experiments and result analysis

3.1. Data and evaluation methods

In this section, we evaluate the proposed approach on ABE task. The 5000 utterances selected randomly from the TIMIT database [22] were used for the DNNs training. Another 1000 randomly selected utterances from the TIMIT database were used to confirm model parameters. We compared our proposed ABE algorithm with three different baseline algorithms on the GRID corpus [23], where we used the test speakers with numbers 1, 2, 18, and 20, referred to as s1, s2, s18, and s20, respectively. The test set for our algorithm and different baseline algorithms is the same. In our evaluations, we down-sample the speech signals in order to obtain narrowband sets, where the wideband and narrowband sets build a parallel corpus. Feature extraction is performed for each 32 ms with 50% overlap. The first baseline is the method proposed in [2], based on the vocal tract filter model using linear prediction, and referred as HMM-ABE in the following experiment. The second baseline is the method proposed in [10], based on the GMM-HMM with 256 components with diagonal covariance matrices and clustering of log spectra using a codebook size of 64. We refer as GMM-ABE to this baseline. The third baseline is the method proposed in [10], based on the SPN-HMM. A GMM can be formulated as an SPN with a single sum node [24]. We refer as SPN-ABE to this baseline.

The evaluation of ABE system is performed with three distinct objective metrics. The frequency weighted segmental SNR (fwSNRseg) [25] and Itakura-Saito distance [26] 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 with respect to the original wideband log power spectra in the high-band. In addition, a set of subjective evaluations are also performed. We have performed a subjective preference comparison test to evaluate the proposed ABE system.

3.2. Evaluation of DNNs configure

The number of epoch for each layer of RBM pre-training was 10. The learning rate of pre-training was 0.005. As for the fine tuning, learning rate was set at 0.008 for the first 10 epochs, and then decreased by 10% after every epoch. Total number of epoch was 50. The mini-batch size was set to 128. Input features were normalized to zero mean and unit variance.

Fig. 3 shows the average LSD results on the test set using input features with multiple frames expansion, ranging from 1 to 11 frames at a two-frame increment and different the number of hidden layers on DNNs model. The input features are log power spectra in the low-band. Other configurations of the DNNs were three hidden layers, 3072 hidden units. It is clear that the longer frames the DNNs were fed with, the better the performance could be achieved. But the over long frames also made the DNNs 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 performance was improved greatly while the number of hidden layer was increasing.

Table II presents the average LSD results of combining the different multiple-features on DNNs model training with post-processing and without post-processing. Other configurations of the DNNs were 3 hidden layers, 3072 hidden units and 11 frames expansion. The performance was

improved greatly while the rich acoustic features were considered due to DNNs model is also able to incorporate the deep regularity of the acoustic features. The average LSD could be decreased using global variance equalization.

The configurations of the DNNs were 3 hidden layers, 3072 hidden units and 11 frames expansion in propose method.

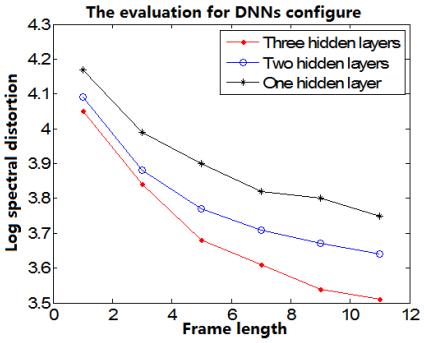

Fig. 3. The LSD for different configuration in DNNs

Table II. *The LSD with different input features*

Additional Features	LSD (without post-processing)	LSD (with post-processing)
None	3.51	3.42
MFCC	3.37	3.31
LSF	3.39	3.34
BPVC	3.44	3.38
MFCC+LSF+BPVC	3.33	3.25

3.3. Overall evaluation

3.3.1. Objective test evaluation

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

Table III. *The objective test for different ABE methods*

ABE method	fwSNRseg	Itakura-Saito distance
HMM-ABE	9.83	40.14
GMM-ABE	13.46	5.87
SPN-ABE	14.69	3.14
Proposed ABE	24.96	3.03

3.3.2. Subjective test evaluation

We have performed a subjective preference test to evaluate the proposed ABE system. During the 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 8 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 Fig 4, Fig 5 and Fig 6 indicate that, speech synthesized with the proposed method outperforms the speech synthesized with the different baseline methods significantly.

3.4. Discussion

We introduce novel log power spectra estimation for the ABE using deep architecture. Subjective evaluation indicates that proposed ABE yields a brighter sound and produce more clear than three different baseline methods. We demonstrated that proposed ABE is a promising regression model for speech, applying them to ABE. Motivated by the success of DNNs on the voice conversion and speech enhancement, we used DNNs 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.

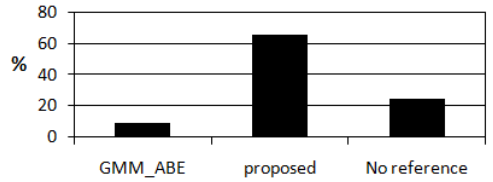

Fig. 4. Subjective test for GMM-ABE and proposed ABE

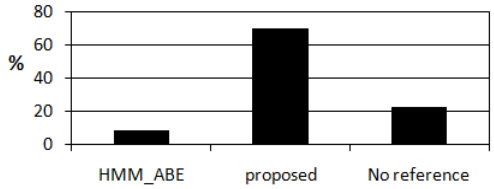

Fig. 5. Subjective test for HMM-ABE and proposed ABE

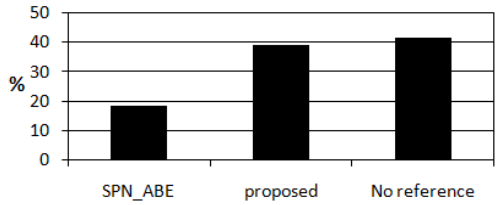

Fig. 6. Subjective test for SPE-ABE and proposed ABE

4. Conclusions

In this paper, a DNN-based framework for ABE is proposed. Among the various DNNs configurations, the rich acoustic features are crucial to learn the complex structure of the mapping function between narrowband speech features and wideband speech features. It was found that the application of more acoustic context information improves the system performance. The global variance equalization was effective in solving the over-smoothing problem of the reconstructed log power spectra in the high-band. Compared with the different baseline methods, the proposed framework can achieve significant improvements in both objective and subjective measures over the state-of-the-art.

In future studies, we will also consider training DNNs model respectively according different phoneme 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, No.61332017, No.61375027, No.61273288, No.61233009, No.61203258), and the Major Program for the National Social Science Fund of China (13&ZD189).

6. References

[1] Cheng, Y. M., O'Shaughnessy, D., & Mermelstein, P. (1994). Statistical recovery of wideband speech from narrowband speech. Speech and Audio Processing, IEEE Transactions on, 2(4), 544-548.

[2] Jax, P., & Vary, P. (2003). On artificial bandwidth extension of telephone speech. *Signal Processing*, *83*(8), 1707-1719.

[3] Park, K. Y., & Kim, H. S. (2000). Narrowband to wideband conversion of speech using GMM based transformation. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (Vol. 3, pp. 1843-1846).

[4] A. H. Nour-Eldin and P. Kabal, "Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech," in Proc. ICASSP, 2009, pp. 4001–4004.

[5] Qian, Y., & Kabal, P. (2003, September). Dual-mode wideband speech recovery from narrowband speech. In *INTERSPEECH*.

[6] Enbom, N., & Kleijn, W. B. (1999). Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients. In *Speech Coding Proceedings, 1999 IEEE Workshop on* (pp. 171-173). IEEE.

[7] K.Y. Park and H.S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'00, vol. 3, pp. 1843–1846, 2000.

[8] Seltzer, M. L., Acero, A., & Droppo, J. (2005, September). Robust bandwidth extension of noise-corrupted narrowband speech. In *INTERSPEECH* (pp. 1509-1512).

[9] Pulakka, H., Laaksonen, L., Vainio, M., Pohjalainen, J., & Alku, P. (2008). Evaluation of an artificial speech bandwidth extension method in three languages. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(6), 1124-1137.

[10] Peharz, R., Kapeller, G., Mowlaee, P., & Pernkopf, F. (2014, May). Modeling speech with sum-product networks: Application to bandwidth extension. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 3699-3703). IEEE.

[11] Hinton, G., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

[12] Erhan, D., Bengio, Y., Courville, A., Manzagol, P. A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning?. *The Journal of Machine Learning Research*, *11*, 625-660.

[13] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). An experimental study on speech enhancement based on deep neural networks. *Signal Processing Letters, IEEE*, *21*(1), 65-68.

[14] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, *29*(6), 82-97.

[15] Zhang, X. L., & Wu, J. (2013). Deep belief networks based voice activity detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, *21*(4), 697-710.

[16] Supplee, L. M., Cohn, R. P., Collura, J. S., & McCree, A. V. (1997, April). MELP: the new federal standard at 2400 bps. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (Vol. 2, pp. 1591-1594). IEEE.

[17] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504-507.

[18] Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A. R., & Hinton, G. E. (2010, September). Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech* (pp. 1692-1695).

[19] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), 1-127.

[20] Toda, T., Black, A. W., & Tokuda, K. (2005, March). Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter. In *ICASSP (1)* (pp. 9-12).

[21] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. A Regression Approach to Speech Enhancement Based on Deep Neural Networks.

[22] Garofolo, J. S. (1988). Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, *107*.

[23] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2421-2424.

[24] Peharz, R., Geiger, B. C., & Pernkopf, F. (2013). Greedy part-wise learning of sum-product networks. In *Machine Learning and Knowledge Discovery in Databases* (pp. 612-627). Springer Berlin Heidelberg.

[25] Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(1), 229-238.

[26] Itakura, F., & Saito, S. (1968, August). Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics* (Vol. 17, pp. C17-C20). pp. C17–C20.