NMF-based Speech Enhancement Incorporating Deep Neural Network

Tae Gyoon Kang[1]*, Kisoo Kwon*[1]*, Jong Won Shin*[2]*, and Nam Soo Kim*[1]

[1]Department of Electrical and Computer Engineering and INMC,
Seoul National University, Korea
[2]School of Information and Communications,
Gwangju Institute of Science and Technology, Korea
{tgkang, kskwon}@hi.snu.ac.kr, jwshin@gist.ac.kr, nkim@snu.ac.kr

Abstract

Recently, lots of algorithms using machine learning approaches have been proposed in the speech enhancement area. One of the most well-known approaches is the non-negative matrix factorization (NMF) -based one which analyzes noisy speech with speech and noise bases. However, NMF-based algorithms have difficulties in estimating speech and noise encoding vectors when their subspaces overlap. In this paper, we propose a novel speech enhancement algorithm which uses deep neural network (DNN) to improve the encoding vector estimation of the NMF-based technique. A DNN is trained to represent the mapping from noisy speech to corresponding encoding vectors. The quality of the enhanced speech from the proposed NMF-based scheme adopting DNN-based encoding vector estimation is compared with that from the conventional NMF-based technique. The experimental results showed that the proposed speech enhancement algorithm outperformed the conventional NMF-based speech enhancement technique.

Index Terms: speech enhancement, noise suppression, nonnegative matrix factorization, deep neural network

1. Introduction

Over the last few decades, many speech enhancement algorithms which suppress noise from noisy speech have been proposed. Statistical model-based approach including minimum mean-square error estimators of the spectral amplitudes and log spectral amplitudes [1]-[3] is one of the most popular approaches. Although this approach has been shown to be successful to reduce stationary noises from the noisy speech, the performance of this family of algorithms degrades in nonstationary noise environments. This may be mainly because the estimation of noise characteristics requires an assumption that the noise does not change rapidly over time. A number of attempts have been made to overcome this difficulty within the framework of statistical model-based approach [4]-[6], but couldn't resolve this issue completely.

Data-driven approaches deal with the speech enhancement task in different aspects without a stationary assumption. In this approach, the speech and noise model are trained to learn the characteristics of speech and noise from a training database. Non-negative matrix factorization (NMF) algorithm [7]-[11] is one of the most well-known algorithms that are based on a dictionary learning approach. The NMF algorithm represents the speech and noise spectral subspaces using non-negative basis matrices. By factorizing noisy speech data vector with these basis matrices, the speech and noise components are described by the product of each basis matrix and a corresponding encoding vector.

Speech enhancement algorithm based on NMF analysis assumes that the subspaces of speech and noise are almost orthogonal with each other. However, the subspaces of speech and noise often overlap which leads to degraded performance for estimated speech or noise components. Adding an orthogonal constraint to the objective function as in [12] would not resolve the issue as long as the basis matrices of speech and noise are trained independently.

In this paper, we propose a speech enhancement algorithm which uses deep neural networks (DNNs) to estimate the encoding vectors of the NMF algorithm. The DNN is chosen in this paper due to its ability to learn complicated mapping between the input and output vectors [13]-[15]. After the NMF algorithm is applied to obtain the speech and noise bases, a DNN learns the function between noisy speech spectra and corresponding speech and noise encoding vectors. Experimental results showed that the performance of NMF-based speech enhancement algorithm can be improved by adopting DNN for encoding vector estimation.

2. Speech enhancement based on NMF algorithm

In the NMF analysis, an $l_f \times l_d$ dimensional matrix V is described by the product of W and C as follows:

$$V \approx WC \tag{1}$$

where W is a nonnegative $l_f \times k$ dimensional matrix and C is a nonnegative $k \times l_d$ dimensional matrix. For the rest of the paper, W is denoted as a basis matrix and C is denoted as an encoding matrix. W and C are iteratively updated while minimizing the objective function $D(V|WC)$ which measures the distance between an input matrix V and a multiplication of the basis and encoding matrices WC. The Euclidean distance can be one example of the objective function. At each iteration of the multiplicative rule for Euclidean distance, W and C are updated as follows [8]:

$$C \leftarrow C \otimes \frac{W'V}{W'WC} \tag{2}$$

$$W \leftarrow W \otimes \frac{VC'}{WCC'} \tag{3}$$

where $\otimes$ and $\frac{(\cdot)}{(\cdot)}$ mean element-wise multiplication and division operations between matrices or vectors and the prime denotes the transpose of a matrix or a vector.

In the NMF analysis, each column of W denotes a basis vector for V. Since each column of C represents the corresponding column of V with a weighted sum of basis vectors in V independently, the basis matrix from the NMF algorithm can effectively represent the non-stationary characteristics of a data matrix. By applying NMF algorithm into speech enhancement task, the speech and noise parameters can be obtained without the stationary assumption.

Speech enhancement using NMF algorithm consists of training and test stages. In the training stage, speech and noise spectral magnitude from training database are analyzed by the NMF algorithm. Let us denote an $l_f \times l_S$ dimensional speech spectral magnitude matrix as S and an $l_f \times l_N$ dimensional noise spectral magnitude matrix as N where l_f denotes a number of frequency bins and l_S, l_N denotes a number of speech and noise frames respectively. The NMF algorithm finds the speech and noise models $\{W_S, C_S\}$ and $\{W_N, C_N\}$ from S and N respectively through an iterative procedure such as (2) and (3).

In the test stage, each noisy speech spectral magnitude vector $\mathbf{x}$ is separated into speech and noise components by NMF algorithm with a concatenated basis matrix of W_S and W_N [9]-[11]. The concatenated basis matrix W for speech and noise is formed by a simple concatenation as follows:

$$W = [W_S \quad W_N]\,. \tag{4}$$

By applying (2) iteratively with fixed W, a corresponding encoding data vector $\mathbf{c}(\mathbf{x})$ in which each element has information for each speech or noise basis from $\mathbf{x}$ is obtained. From W and $\mathbf{c}(\mathbf{x})$, $\mathbf{x}$ can be factorized as

$$\begin{aligned} \mathbf{x} &= W\mathbf{c}(\mathbf{x}) && (5)\\ &= [W_S \quad W_N] \begin{bmatrix} \mathbf{c}_S(\mathbf{x}) \\ \mathbf{c}_N(\mathbf{x}) \end{bmatrix} && (6) \end{aligned}$$

where $\mathbf{c}_S(\mathbf{x})$ and $\mathbf{c}_N(\mathbf{x})$ are encoding vectors for speech and noise basis matrices given by $\mathbf{x}$.

The estimated spectral magnitudes of speech and noise can be obtained by multiplying corresponding basis matrices and encoding vectors, i.e., $W_S\mathbf{c}_S(\mathbf{x})$ and $W_N\mathbf{c}_N(\mathbf{x})$, respectively. Instead of using $W_S\mathbf{c}_S(\mathbf{x})$ as the estimated speech spectral magnitude $\mathbf{s}$ directly, the gain function similar to Wiener filter is usually applied to increase the speech quality [9]-[11]. Let us denote the estimates for speech and noise spectral magnitude vectors from NMF algorithm as $\mathbf{p}_S(\mathbf{x})$ and $\mathbf{p}_N(\mathbf{x})$. From these parameters, the gain function for $\mathbf{x}$ is obtained and $\mathbf{s}$ is derived as follows [11]:

$$\begin{aligned} \mathbf{p}_S(\mathbf{x}) &= W_S\mathbf{c}_S(\mathbf{x}) && (7)\\ \mathbf{p}_N(\mathbf{x}) &= W_N\mathbf{c}_N(\mathbf{x}) && (8)\\ \hat{\mathbf{s}} &= \frac{(\mathbf{p}_S(\mathbf{x}))^m}{(\mathbf{p}_S(\mathbf{x}))^m + (\mathbf{p}_N(\mathbf{x}))^m} \otimes \mathbf{x} && (9) \end{aligned}$$

where m is a positive constant. Using $\hat{\mathbf{s}}$ and the phase information from noisy speech, the estimated speech components is reconstructed to time domain.

Though the speech enhancement algorithm using NMF technique is simple and easy to implement, it has limitation when the subspaces of speech and noise overlap. In conventional NMF algorithm, this overlap in subspaces induces quality degradation of enhanced speech since the speech and noise basis matrices are trained independently without considering the correlation between the speech and noise basis matrices.

3. NMF-based speech enhancement employing DNN

A major difficulty in speech enhancement using NMF algorithm is estimating encoding vectors from $\mathbf{x}$ with a concatenated basis matrix. When some form of the orthogonality condition between speech and noise is not satisfied, minimizing the objective function does not guarantee an optimal solution for speech and noise encoding vectors. That is, when a data vector from the speech or noise can be possibly represented by a linear combination of basis vectors corresponding to the other source, the NMF-based speech enhancement algorithm is likely to fail since the speech and noise components can be misjudged to the other components.

To separate speech and noise components in this condition, a novel model structure is adopted to learn the function from input noisy speech data vectors to the optimal encoding vectors with a set of the training data. Under this framework, the problem of estimating the encoding vectors for speech and noise can be treated as a regression task where the input is a noisy speech vector and the output is encoding vectors of speech and noise. In this paper, DNN is chosen to learn the function between these variables.

DNNs are known to describe complicated functions or mappings more effectively than shallow neural networks. However, when the parameters of neural networks are randomly initialized, DNNs show a slightly worse performance than that of shallow neural networks. A breakthrough in DNN was made with the introduction of the stacked restricted Boltzmann machines accompanied with greedy layer-wise unsupervised learning in order to initialize the DNN parameters. After this pretraining stage, a supervised learning algorithm using backpropagation and stochastic gradient descent is carried out to fine-tune the weights of the DNN. The detailed procedure about pretraining and fine-tuning stages is described in [15], [16].

The proposed speech enhancement algorithm consists of NMF training, DNN training, and enhancement stages. In the NMF training stage, conventional NMF technique is applied to speech and noise separately. The basis matrices for speech and noise W_S and W_N obtained in this stage are used in the following stages while the corresponding encoding matrices are discarded. This procedure is illustrated in Figure 1. (a).

To train DNN, the input and output transcription of DNN are generated in DNN training stage. Firstly, for each speech and noise vector $\mathbf{s}_t$ and $\mathbf{n}_t$ where the subscript t denotes data for DNN training stage, the conventional NMF technique is applied with only W_S and W_N, respectively, to find the speech and noise encoding vectors $\mathbf{c}_{S,t}$ and $\mathbf{c}_{N,t}$ which are not affected by the ambiguity caused by overlapped subspaces of basis matrices.

After the pairs of $\mathbf{s}_t, \mathbf{c}_{S,t}$ and $\mathbf{n}_t, \mathbf{c}_{N,t}$ are obtained, we artificially generate a noisy speech vector $\mathbf{x}_t$ as the input and a corresponding encoding vector $\mathbf{c}_t$ as the output transcription with a set of arbitrary weights $\{\alpha_{S,t}, \alpha_{N,t} > 0\}$ as follows:

$$\begin{aligned} \mathbf{x}_t &= \alpha_{S,t}\mathbf{s}_t + \alpha_{N,t}\mathbf{n}_t && (10)\\ \mathbf{c}_t &= \begin{bmatrix} \alpha_{S,t}\mathbf{c}_{S,t} \\ \alpha_{N,t}\mathbf{c}_{N,t} \end{bmatrix}. && (11) \end{aligned}$$

After generating a collection of the input and output transcription vectors $\mathbf{x}_t$ and $\mathbf{c}_t$, we train a DNN where $\mathbf{x}_t$ defined in (10) is fed to the input layer of the neural network and $\mathbf{c}_t$ defined in (11) is applied to the output layer of the neural network as a transcription. Before $\mathbf{x}_t$ and $\mathbf{c}_t$ are fed to the DNN, they are

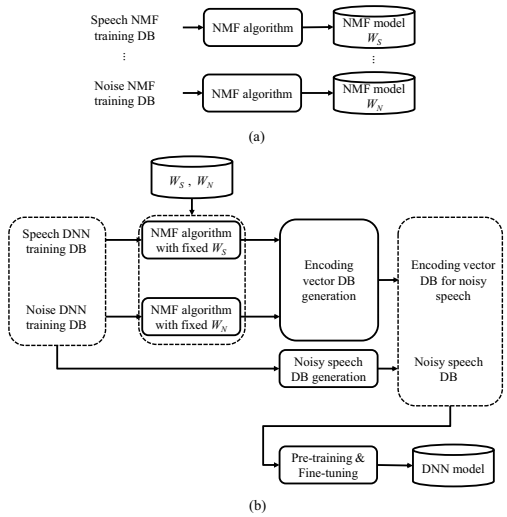

Figure 1: The block diagrams of NMF and DNN training stages.

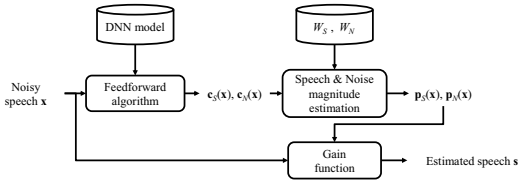

Figure 2: Overall block diagram of the enhancement stage for the proposed NMF-based speech enhancement employing deep neural network.

normalized to have values in the range (0,1). Multiplying non-negative weight coefficients $\alpha_{S,t}$, $\alpha_{N,t}$ and normalizing $\mathbf{x}_t, \mathbf{c}_t$ does not hamper the inter-relationship among them since the NMF algorithm is scale-independent when the Euclidean distance is used as an objective function. The detailed block diagram of the DNN training stage is shown in Figure 1. (b).

In the enhancement stage, a noisy speech vector $\mathbf{x}$ is fed to the DNN with standard feedforward algorithm. In feedforward algorithm, the output vector is computed as follows: Suppose that the hidden layer vector of $(k-1)-$th layer which is given by $\mathbf{x}$ is denoted as $\mathbf{h}^{k-1}(\mathbf{x})$. With the bias vector $\mathbf{b}_k$ and weight matrix A_k for $k-$th hidden layer of the DNN, $\mathbf{h}^k(\mathbf{x})$ is determined as follows:

$$\mathbf{h}^k(\mathbf{x}) = sigm(A^k \mathbf{h}^{k-1}(\mathbf{x}) + \mathbf{b}^k) \tag{12}$$

where $sigm(\mathbf{a})$ refers a element-wise sigmoid function of a vector $\mathbf{a}$. By iterating this procedure with the number of hidden layer, the output layer vector $\mathbf{o}(\mathbf{x})$ is obtained from the DNN. $\mathbf{o}(\mathbf{x})$ is re-scaled to $\mathbf{c}(\mathbf{x})$ for which the reconstructed noisy speech spectral magnitude $W\mathbf{c}(\mathbf{x})$ has the same l_1-norm with $\mathbf{x}$. With the estimated $\mathbf{c}(x)$, the speech and noise spectrum vectors $\mathbf{p}_S(\mathbf{x})$, $\mathbf{p}_N(\mathbf{x})$ and corresponding speech estimate $\mathbf{s}$ is obtained with (7)-(9). Figure 2 shows the block diagram of speech enhancement stage which utilizes trained NMF and DNN models.

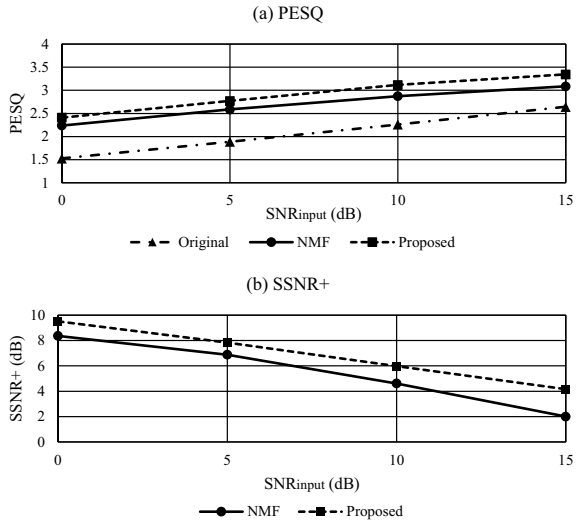

Figure 3: PESQ scores and SSNR+ values of the original and enhanced speeches for white noise environment

4. Experiments

In order to evaluate the performance of the proposed algorithm, we compared the quality of the enhanced speech obtained from the proposed algorithm with those from the conventional NMF-based speech separation algorithm [11]. We conducted experiments with TIMIT database (DB) as the clean speech DB and the NOISEX-92 DB as the noise DB [17]. The white and factory noises were added to the clean speech with the SNRs from 0 to 15 dB. Each utterance was sampled at 16kHz and the 512-dimensional Hamming window with 75% overlap was used. The value of m in (9) was determined experimentally to 2.

The basis matrix for speech W_S was trained with the 10000 frames of the clean speech data and the basis matrix for each noise type W_N was trained with the 9000 frames of corresponding noise data. The number of basis vectors for speech and noise was fixed to 40 each.

To train the DNN, the noisy speech utterances were generated with the signal-to-noise ratios (SNRs) from -5 to 20 dB. The number of noisy speech frame and corresponding encoding vectors for speech and noise to train the DNN for each noise type was 52300. The DNN was built by stacking three hidden layers with 400 nodes each. We ran 30 epochs for pre-training the hidden layer and 300 epochs to fine-tune the DNN. We used mini-batch training which samples 100 frames of training data to organize mini-batch.

The performances of the conventional NMF-based algorithm and proposed algorithm were tested with the ten utterances from 5 male and 5 female speakers. The perceptual evaluation of speech quality (PESQ) score [18] and segmental SNR improvement (SSNR+) [19] was used to measure the quality of enhanced speech.

Figures 3 and 4 show the PESQ scores and SSNR+ values of the enhanced speech when the speech is corrupted by the white and factory noise respectively. In these figure, the performance of the proposed algorithm outperformed the conventional NMF-based algorithm in all conditions. This results showed that NMF-based speech enhancement with proposed DNN-based encoding vector estimation scheme outperformed

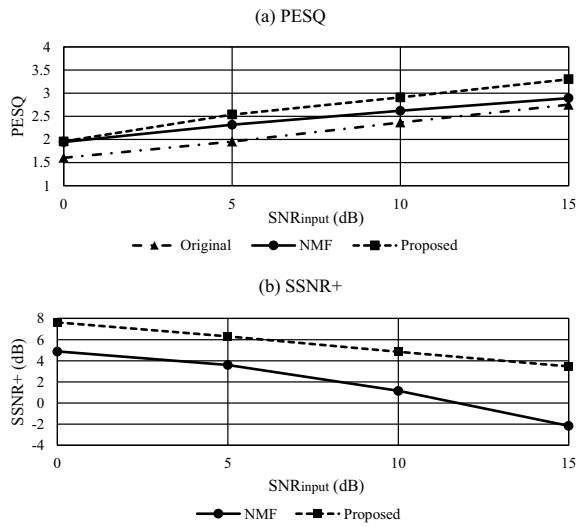

Figure 4: PESQ scores and SSNR+ values of the original and enhanced speeches for factory noise environment

that with conventional update rule.

Comparing the results in Figures 3 and 4, the performance improvement was higher in factory noise. One of the reasons may be the fact that the subspace of the factory noise is more complicated than that of the white noise since the factory noise is composed by many acoustic sources.

5. Conclusions

In this paper, a novel approach to improve the NMF-based speech enhancement algorithm with DNN has been proposed. DNN is trained to model the mapping from the noisy speech vectors to the encoding vectors of speech and noise to enhance the performance of NMF analysis when the subspaces that speech and noise components span overlap. Through the experiments, it is shown that the proposed speech enhancement algorithm outperformed the conventional NMF-based algorithm. The future work will include the adaptation for DNN model to reliably estimate the speech from the noisy spectrum when the noise type is not known.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1019) supervised by the NIPA (National IT Industry Promotion Agency).

7. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 32, No. 6, pp. 1109-1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 2, pp. 443-445, Apr. 1985.

[3] N.S. Kim and J. -H. Chang, "Spectral enhancement based on global soft decision, *IEEE Signal Process. Lett.*, Vol.7, No.5, pp.108110, May 2000.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, Vol. 9, No. 5, pp. 504-512, Jul. 2001.

[5] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, Vol. 9, No. 1, pp. 12-15, Jan. 2002.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Process.*, Vol. 11, No. 5, pp. 466-475, Sep. 2003.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, Vol. 401, pp. 788-791, Oct. 1999.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Inform. Process. Syst.*, Vol. 13, pp. 556-562, Nov. 2001.

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. of IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, Las Vegas, 2008, pp. 4029-4032.

[10] F. Weninger, J. Geiger, M. Wllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. 1st Int. Workshop on Mach. Listening in Multisource Environments (CHiME)*, pp. 24-29, Sep. 2011.

[11] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Int. Conf. on Digital Signal Process.*, Corfu, 2011, pp. 1-6.

[12] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Hong Kong, 2008, pp. 1828-1832.

[13] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 20, No. 1, pp. 14-22, Jan. 2012.

[14] XL. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 21, No. 4, pp. 697-710, Apr. 2013.

[15] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, Vol. 18, No. 7, pp. 1527-1554, Jul. 2006.

[16] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Technical University of Denmark, 2012. (http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6284)

[17] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II.NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, Vol. 12, No. 3, pp. 247-251, Jul. 1993.

[18] ITU, "Perceptual evalulation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Rec. P. 862, 2000.

[19] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, 1988.