Hindawi
IET Signal Processing
Volume 2024, Article ID 8881007, 10 pages
https://doi.org/10.1049/2024/8881007

Research Article

Speech Enhancement Using Joint DNN-NMF Model Learned with Multi-Objective Frequency Differential Spectrum Loss Function

Matin Pashaian and Sanaz Seyedin

Speech Processing Research Lab, Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran

Correspondence should be addressed to Sanaz Seyedin; sseyedin@aut.ac.ir

Received 9 July 2023; Revised 17 December 2023; Accepted 5 January 2024; Published 24 January 2024

Academic Editor: Richard Dansereau

We propose a multi-objective joint model of non-negative matrix factorization (NMF) and deep neural network (DNN) with a new loss function for speech enhancement. The proposed loss function (L_{MOFD}) is a weighted combination of a frequency differential spectrum mean squared error (MSE)-based loss function (L_{FD}) and a multi-objective MSE loss function (L_{MO}). The conventional MSE loss function computes the discrepancy between the estimated speech and clean speech across all frequencies, disregarding the process of changing amplitude in the frequency domain which contains valuable information. The differential spectrum representation retains spectral peaks that carry important information. Using this representation helps to ensure that this information in the speech signal is reserved. Also, on the other hand, noise spectra typically have a flat shape and as the differential operation makes the flat spectral partly close to zero, the differential spectrum is resistant to noises with smooth structures. Thus, we propose using a frequency-differentiated loss function that considers the magnitude spectrum differentiations between the neighboring frequency bins in each time frame. This approach maintains the spectrum variations of the objective signal in the frequency domain, which can effectively reduce the noise deterioration effects. The multi-objective MSE term (L_{MO}) is a combined two-loss function related to the NMF coefficients which are the intermediate output targets, and the original spectral signals as the actual output targets. The use of encoded NMF coefficients as low-dimensional structural features for DNN serves as prior knowledge and helps the learning process. L_{MO} is used beside L_{FD} to take advantage of both the properties of the original and the differential spectrum in the training loss function. Moreover, a DNN-based noise classification and fusion strategy (NCF) is proposed to exploit a discriminative model for noise reduction. The experiments reveal the improvements of the proposed approach compared to the previous methods.

1. Introduction

Speech enhancement is the task of separating the target speech from unwanted noises. Speech enhancement methods generally include statistical and data-driven learning-based methods. The statistical approaches such as the minimum mean-square error (MMSE) method [1] and Wiener filtering [2] are based on the statistical models of speech and noise. Non-negative matrix factorization (NMF) a well-known method in this category has been recently used a lot in speech separation [3]. By NMF, a speech or noise signal can be decomposed into a non-negative basis matrix and an activation matrix. Then, for speech enhancement applications, in the testing phase, the learned concatenated basis matrices of speech and noise are used for an unknown noisy speech to estimate the related activation matrices. The estimated activations are multiplied by the related learned basis matrices to approximate the speech and noise sources. In addition, extracting noise-robust features is another approach for reducing the noise effects of speech signals [4]. Lately, data-driven learning-based methods such as deep learning have also been widely used for complex mapping modeling such as learning the nonlinear mapping of noisy speech to clean speech for applications of speech enhancement and

speech recognition [5–10]. Training targets in data-driven methods are mostly the spectral magnitude of sources directly (mapping-based targets), or the spectral masks (masking-based targets) which are the gain values that represent T–F energy ratios of each source to the mixture and are then multiplied with the mixture of speech and noise to estimate each of them [11, 12].

Moreover, in some research works, NMF or its extended versions are combined with deep neural networks (DNNs) to improve performance [13–20]. In Kang et al.'s [13] study, mapping of the spectral magnitude of the noisy speech to the NMF activation coefficients of speech and noise is performed by a DNN. Then, the related estimated coefficients are multiplied with the corresponding learned basis matrix outside of DNN separately to approximate the actual signals. In Vu et al.'s [17] and Jia et al.'s [19] studies, instead of the main noisy spectrum, the noisy activation matrix which is the concatenated activation matrices of speech and noise is used as the DNN input noisy feature. Furthermore, in Wang and Wang's [20] study, NMF is first applied to an ideal ratio mask (IRM) and it is decomposed into a basis matrix and an activation matrix. Then, instead of directly predicting a mask as the DNN target, the related activation coefficients are estimated by DNN as an intermediate target. Then, the estimated activation matrix and the learned basis matrix of IRM are linearly combined outside of DNN to reconstruct the IRM. Afterward, the estimated IRM separates the desired speech from the noisy mixture. On the other hand, in Williamson et al.'s [21–23] and Grais et al.'s [24] studies, DNN and NMF have combined in two subsequence separate stages, so that DNN in the first stage is applied for the separation purpose and then, NMF in the second stage for enhancement, or vice versa. In Williamson et al.'s [21–23] studies, the NMF reconstruction is used as a postprocessing step to enhance the separated speech by the mask estimated by DNN. In Williamson et al.'s [25, 26] studies compared with Williamson et al.'s [21] study, a DNN is used in the second stage as an NMF alternative to estimate the activation matrix of clean speech from the first masked speech. Then, the estimated coefficients are multiplied with the pretrained basis matrix separately outside of the DNN to acquire the enhanced speech.

However, in the mentioned approaches, the NMF and DNN processes are carried out separately. Also, the DNN does not directly estimate the main targets but only estimates an intermediate target, which is the NMF activation coefficients. Therefore, in Nie et al.'s [14, 15] and Li et al.'s [16] studies, the NMF and DNN processes are jointly combined, so that the learned NMF bases are integrated into the DNN as an extra layer. Then, the main objective signals are directly estimated by the DNN. However, in these methods, the activation coefficients do not have a direct effect on the DNN learning process and are not directly optimized by DNN. Hence in this paper, it is suggested that the activation coefficients be used in the network as prior knowledge in a multi-objective multi-loss training approach so that the extracted activation coefficients be injected at an intermediate output (prior) layer of DNN as a direct target and in the loss function while the original signals are also estimated by the DNN at the main output layer simultaneously.

The training loss function is also a remarkable subject in speech enhancement algorithms. The traditional MSE function is widely used as the training loss function in spectral speech enhancement. However, the spectral changes are not considered in the MSE. Due to the unique characteristics of each individual's sound source and vocal tract, the pitch frequency, the difference between frequency bins, and the process of changing amplitude in the frequency domain are different for each frame. Consequently, incorporating the process of changes between frequency bins in a time frame into the loss function can improve network learning and the performance of speech and noise separation. In addition, as described in Chen et al.'s [27] study and according to our observations, in the differentiated spectrum representation, the spectral peaks that carry valuable information are kept almost intact and the smooth parts of the spectrum become zero. Thus, to take an account the peak movements in the frequency domain in addition to the pitch information, we propose a frequency-differentiated loss function which is added to the multi-objective MSE loss function in this paper. In the MSE terms, the multi-objective training is applied so that in addition to the estimation of the actual signals in an MSE, the related NMF activation coefficients are also considered in another MSE function.

Also, another issue of interest is that most of the speech enhancement models are trained with a pool of different types of noises, and then these general models are used in the testing phase for enhancing each observed noisy speech. In this paper, to have better improvement, our joint models are exclusively learned for each type of training noise, and in the testing phase, using a noise classification and fusion approach (NCF), one or a suitable combination of the multiple learned models is used to enhance each detected noise.

The organization of this paper is structured as follows: in Section 2, an overview of NMF-based speech enhancement is given. In Section 3, the proposed system, including the Jnt-DNN-NMF model, the proposed loss function, and the noise classification and fusion approach will be explained. In Section 4, the experimental setup and results are presented. Finally, the conclusion is provided in Section 5.

2. NMF-Based Speech Enhancement

In the NMF approach, a non-negative data matrix, which in our work is the magnitude spectrum $X \in R_{\geq 0}^{F \times T}$, is decomposed into a non-negative basis matrix $\mathrm{B}_x \in R_{\geq 0}^{F \times K}$ $(K \leq F)$ and an activation matrix $\mathrm{H}_x \in R_{\geq 0}^{K \times T}$ according to Equation (1). K, T, and F represent the number of basis vectors (columns of B_x), time frames, and frequency bins, respectively. The basic structures of X are captured in the basis matrix. X can be the clean speech S, noisy speech Y, or noise N:

$$X \simeq B_x H_x \,. \tag{1}$$

Kullback–Leibler (KL) divergence as one of the multiplicative update rules is used to extract B_x and H_x matrices by iteratively minimizing the error between the observed signal X and its reconstruction $B_x H_x$ as follows:

$$\min_{B,H>0} \mathscr{D}_{\mathrm{KL}}(X \parallel BH), \tag{2}$$

$$\begin{aligned} H_x &\leftarrow H_x \otimes \frac{B_x^T \frac{X}{B_x H_x}}{B_x^T 1} \\ B_x &\leftarrow B_x \otimes \frac{\frac{X}{B_x H_x} H_x^T}{1 H_x^T}, \end{aligned} \tag{3}$$

where $\mathscr{D}$ is Euclidean distance and 1 is an $F \times T$ matrix with all elements equal to one. By assuming additive noise (i.e., $y(i) = s(i) + n(i)$, i is the sample index) and without considering the speech-noise cross-term we have $Y(f,t) \approx S(f,t) + N(f,t)$ in the spectrum domain. Y, S, and $N \in R_{\geq 0}^{F \times T}$ are the noisy, clean, and noise spectral magnitudes, respectively. f and t are the frequency and time indices. In using NMF for speech enhancement, in the training phase, B_x and H_x for clean speech and noise are usually randomly initialized and then obtained using the iterative multiplicative update rules. H_x is discarded and B_x is held fixed for the enhancement stage. The noisy basis matrix B_y is formed by concatenating the trained basis matrices of clean and noise ($B_y = [B_s B_n] \in R_{\geq 0}^{F \times (Ks+Kn)}$). Then in the testing phase, the magnitude of a test noisy speech is approximated as a product of the fixed B_y matrix and a new activation matrix $\widehat{H}_y = [\widehat{H}_s^T \widehat{H}_n^T]^T \in R_{\geq 0}^{(Ks+Kn) \times T}$ which is calculated iteratively by Equation (3). Finally, the estimated speech and noise magnitudes are obtained as follows:

$$\widehat{S} = B_s \widehat{H}_s, \ \widehat{N} = B_n \widehat{H}_n \quad \widehat{S}, \widehat{N} \in R_{\geq 0}^{F \times T}. \tag{4}$$

3. Proposed System

Since the effect of phase enhancement is not significant in speech improvement, we only use the short-time Fourier transform (STFT) magnitude spectrum of the framed signals for enhancement. As shown in Figure 1, the proposed system is performed in two phases of training and testing. The training phase includes the sections of NMF training, Jnt-DNN-NMF training, and classifier DNN training. Jnt-DNN-NMF is the joint cooperative model of DNN and NMF which will be explained in Section 3.1. The testing phase contains the classifier DNN prediction and the Jnt-DNN-NMF prediction for the test data. The NMF and the Jnt-DNN-NMF training parts are two consecutive stages, the NMF training is pretraining for the Jnt-DNN-NMF training (so that the results of the NMF training are used in the Jnt-DNN-NMF training as a pretraining stage). This will be described in Section 3.1. It should be noted that NMF and Jnt-DNN-NMF training are performed for each training noise type to produce the noise-specific Jnt-DNN-NMF models which will be used in the testing phase (the dashed boxes in the bottom part of Figure 1). In other words, the repeated dashed boxes in Figure 1 are the learned Jnt-DNN-NMF models related to the N training noise types and have the same approach as dashed box1 (for noise1). The classifier DNN training is performed with different noisy speech magnitudes as input and N output class labels. In the testing phase, the noise type (matched or mismatched) of each input noisy speech is detected based on the classifier results (Section 3.3). Then according to Figure 1, after predictions made by N different Jnt-DNN-NMF models, in the fusion block, only one corresponding detected model is used for enhancement of each matched noise. However, for mismatched noises, a weighted combination of outputs of N models is regarded as enhanced speech. Finally, an inverse STFT followed by the overlap-add method is applied to reconstruct the waveform of the desired signal using the estimated magnitude and noisy phase. It should be noted that in the training phase, models are trained using noise-specific data which is the smaller dataset, and in the testing phase, multiple models are instantaneously and parallelly applied to the input noisy speech, so the computations are light.

3.1. Jnt-DNN-NMF Model. According to Figure 1, at first, in the NMF training stage, the structures in the magnitude spectra of the speech and noise sources are captured by applying the NMF inference for speech and each noise type independently as a feature and structure extraction process. So, the corresponding activation coefficients and basis matrices are obtained. In such a way that bases are trained first and then coefficients are extracted with the fixed bases. Then, as shown in Figure 1, the extracted NMF activation coefficients and basis matrices are employed in the next stage (the Jnt-DNN-NMF training stage). The extracted activation coefficients (H_s, H_n) are directly served as the primary target features for the DNN (dashed lines) while the spectral magnitude of the noisy speech is as input (Y). The trained basis matrices are integrated into the DNN as an additional layer named the NMF reconstruction layer. The DNN together with the integrated NMF reconstruction and Wiener-like filtering layers form the multi-objective Jnt-DNN-NMF model to jointly optimize the main spectral magnitudes in the main output layer and the related NMF coefficients in the coefficients output layer. So, in the Jnt-DNN-NMF training stage, the joint model is trained with the noisy magnitude Y as input and the multi-objective targets of the activation coefficients at the coefficient output layer and the main speech and noise magnitudes at the main output layer using the proposed loss functions (Section 3.2). The mapping function of the DNN (g) is as follows:

$$\begin{aligned} &L_j = g(L_{j-1}) = \sigma\left(W_j^* L_{j-1} + b_j^*\right) 1 \leq j \leq J \\ &L_0 = Y, L_J = \left[\widehat{H}_s \widehat{H}_n\right], \end{aligned} \tag{5}$$

where W_j^* and b_j^* are the weights and biases of the DNN, respectively. J is the index of the coefficient output layer. $\widehat{H}_s$ and $\widehat{H}_n$ represent the estimated activation coefficients of speech and noise, respectively. In the NMF reconstruction layer, the speech and noise basis matrices are multiplied by the estimated coefficients, and then through a Wiener

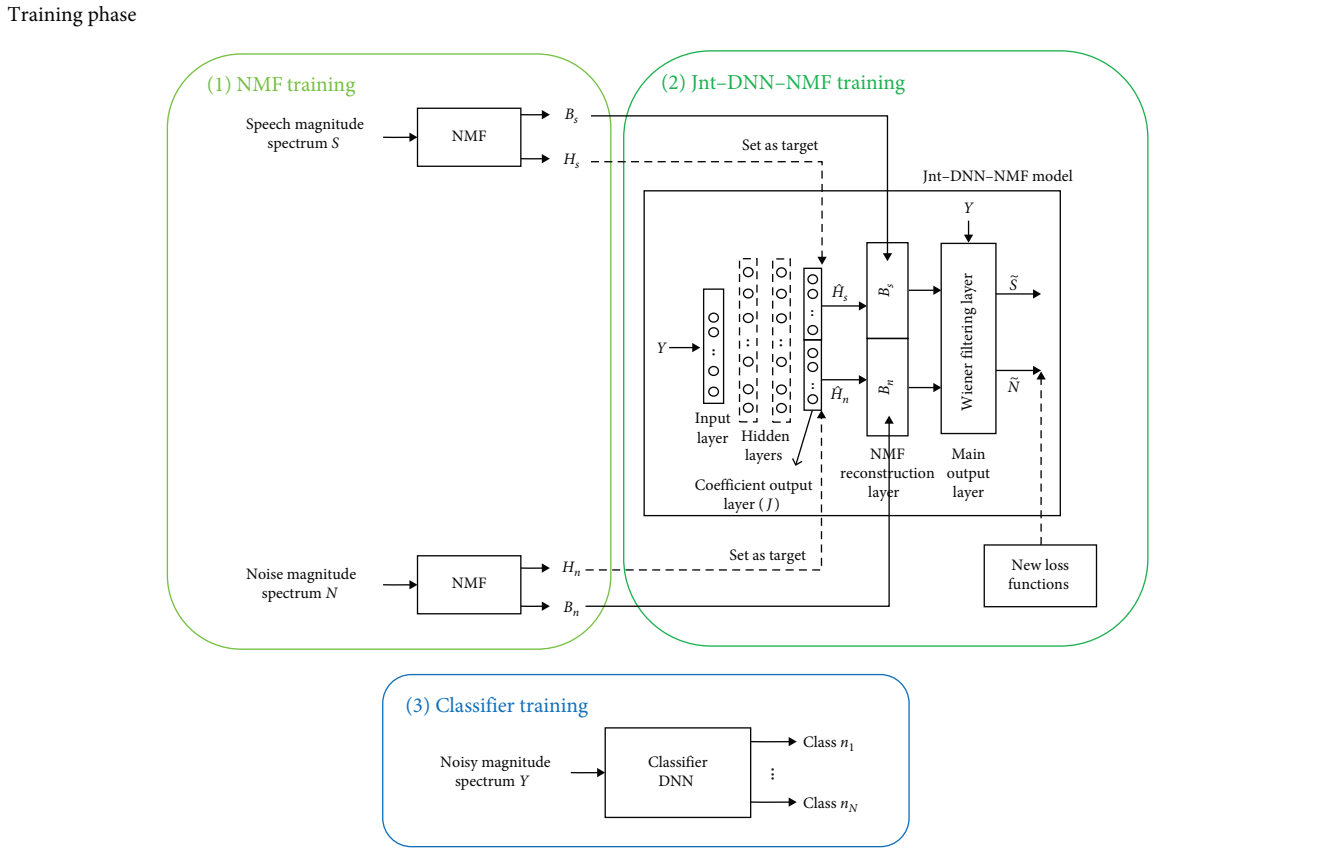

FIGURE 1: Block diagram of the proposed system including training and testing phases. The NMF and Jnt-DNN-NMF training is repeated for each noise type. The classification and fusion blocks are applied in the testing phase to select one model (for matched noises) or combine the output of N models that have already been trained for N training noise types (for mismatched noises). Each dashed box in the testing phase is the learned model for each training noise type.

filtering layer, the final speech and noise estimates are achieved as follows:

$$\tilde{S} = \frac{\left(B_s\hat{H}_s\right)^2}{\left(B_s\hat{H}_s\right)^2 + \left(B_n\hat{H}_n\right)^2} \otimes Y, \tag{6}$$

$$\tilde{N} = \frac{\left(B_n\hat{H}_n\right)^2}{\left(B_s\hat{H}_s\right)^2 + \left(B_n\hat{H}_n\right)^2} \otimes Y. \tag{7}$$

$\tilde{S}$ and $\tilde{N}$ are the final estimated speech and noise magnitudes. The division and multiplication operations are element-wise. Jnt-DNN-NMF is trained with the proposed

loss functions (Section 3.2), and the weights and bias parameters are computed by the backpropagation algorithm.

3.2. Proposed Loss Functions. In most traditional DNN-based speech enhancement methods, the learning process contains a direct mapping from the noisy signal to the actual separation targets without the use and direct influence of the structural features as prior knowledge on DNN and in the training process. So, in our Jnt-DNN-NMF model, we first propose a multi-objective combined loss function (L_{MO}) that not only optimizes the actual spectral signals of speech and noise but also the intermediate activation coefficients as follows:

$$L_{\mathrm{MO}} = \frac{1}{F}\sum_{f=1}^{F}\left(C(t,f) - \widetilde{C}(t,f)\right)^2 + \frac{1}{F}\sum_{f=1}^{F}\left(H(t,f) - \widehat{H}(t,f)\right)^2, \tag{8}$$

where C is the concatenated speech and noise spectral magnitudes ($[S\,N]$) at each time step t and frequency bin f, and $\tilde{C}$ is its estimated version ($[\tilde{S}\,\tilde{N}]$). H is the concatenated NMF activation coefficients of speech and noise ($[H_s\,H_n]$) and $\widehat{H}$ is its estimated version ($[\widehat{H}_s\,\widehat{H}_n]$). Also, F is the total number of frequency bins.

Then, to consider the spectrum changes in the frequency domain, according to Equation (9), we use a frequency-differentiated loss function (L_{FD}) which calculates the amplitude differences between the neighboring frequency bins in each frame. Using this function allows the network to gain a better understanding of frequency characteristics and changes in the frequency domain. It calculates the MSE between the target signal and the estimated signal concerning frequency changes in each frame.

$$L_{\mathrm{FD}} = \frac{1}{F}\sum_{f=1}^{F}\sum_{i=1}^{M}\left[\left(C(t,f+i) - C(t,f-i)\right) - \left(\widetilde{C}(t,f+i) - \widetilde{C}(t,f-i)\right)\right]^2. \tag{9}$$

M is the number of neighboring frequency bins for a frequency bin that are involved in the calculation of the cost function for that frequency bin for each frame. Then, as shown in Equation (10), we propose the MSE-based multi-objective frequency-differentiated loss function (named as L_{MOFD}) which is a weighted combination of the frequency-differentiated loss function L_{FD} and a multi-objective combined loss function (the last two terms). These terms are the MSEs related to the objective spectral signals and the NMF activation coefficients, respectively. This leads to the simultaneous optimization of the encoded output features and the original spectral signals jointly in a single model. Indeed, the encoded features directly affect the learning process by considering a separate optimization term in the overall loss function L_{MOFD}. Therefore, the joint model is trained based on two types of targets at the related output layer:

$$L_{\mathrm{MOFD}} = \alpha_1 L_{\mathrm{FD}} + \frac{\alpha_2}{F}\sum_{f=1}^{F}\left(C(t,f) - \widetilde{C}(t,f)\right)^2 + \frac{1}{F}\sum_{f=1}^{F}\left(H(t,f) - \widehat{H}(t,f)\right)^2, \tag{10}$$

where α_1 and α_2 are the weight parameters for L_{FD} and the first MSE, respectively.

3.3. Noise Classification and Fusion Approach. According to Figure 1, in the testing phase, first, to judge the noise type, a classifier DNN which has already been learned to classify the N training noisy types is used to estimate the similarity rates of each observed noisy speech to the training noise classes. The noise type (matched or mismatched) is diagnosed such that if one of the estimated rates is greater than a high threshold (set to 0.90), that noisy speech is regarded as one of the training noisy mixtures i.e., a matched condition, otherwise, it is a mismatched condition. Then, in the fusion block, for matched noises, the enhanced speech is obtained from the output of only one learned model corresponding to the detected noise. However, for mismatched noises, the final result is calculated based on a weighted combination of the outputs of multiple models, where the weights are the corresponding classification rates.

4. Experimental Setup and Results

The performance of the proposed system is compared with the following methods:

(i) NMF [28]: the explained NMF-based speech enhancement in Section 2.

(ii) DNN-Mag [29]: the traditional DNN-based speech enhancement where a DNN is used to map the spectral magnitude of noisy speech to the spectral magnitude of clean speech.

(iii) LSTM-Mask [7, 30]: a long short-term memory (LSTM) network maps the noisy speech magnitude to the IRM mask values. Then, the estimated mask values are multiplied by the noisy speech to estimate the sources.

(iv) CRN-Mag [31]: a convolutional–recurrent network (CRN) is used with the mapping-based magnitude target. CRN is composed of CNN encoder–decoder and LSTM layers and its architecture is set similar to [31].

(v) DNN-NMF-Sep [13]: a separate combinatorial model of DNN and NMF where the DNN maps the noisy speech to the NMF activation coefficients and the reconstruction of the main objective signals is separately performed outside of DNN.

(vi) Jnt-DNN-NMF [15]: a joint combinatorial model of DNN and NMF where the DNN optimizes the objective signals. However, the activation coefficients do not directly incorporate into the DNN structure and learning process.

We denote our proposed Jnt-DNN-NMF model with two loss functions of the multi-objective loss function L_{MO} and the multi-objective frequency-differentiated loss function L_{MOFD} as "Jnt-DNN-NMF-MO" and "Jnt-DNN-NMF-MOFD," respectively.

The proposed and comparison methods are trained and evaluated on the TIMIT dataset [32] which consists of 6,300 different utterances. We randomly select 200 clean speech utterances from the training set of TIMIT and are corrupted with *babble*, *factory*, and *machinegun* noises from the NOISEX-92 corpus [33] at SNRs −5 to 20 dB with steps of 5 dB. Our test set includes different 60 utterances from the test set of TIMIT which are corrupted with the training noises as matched noises and the real-world recorded *factorymachine* and *windshieldrain* noises from the *Freesound* data as mismatched noises at −5 to 10 dB SNRs. The baselines are trained and evaluated with the same training and testing datasets used for the proposed methods, respectively.

We use a 512-point STFT for the waveforms sampled at 16 kHz and framed using a 512-sample (32 ms) frame length, 512-sample (32 ms) Hamming window, and 128 shift samples (8 ms). The symmetric part of the STFT coefficients is cut off, so the dimension of our spectral magnitude matrices is 257 × frame numbers.

4.1. DNN and NMF Parameters. The NMF ranks of speech and noise basis matrices in all the baseline and proposed NMF-based methods are empirically set at 100 each ($Ks, Kn = 100$). So, the size of the basis matrices is 257 × 100 (frequency bins × bases numbers). The maximum NMF iteration number is set to 50.

The architecture of the used DNN in all the baseline and proposed models includes four hidden layers with 1,024 units for a fair comparison. It should be noted that the main idea of the baseline DNN-Mag, DNN-NMF-Sep, and Jnt-DNN-NMF methods are, respectively, from Kang et al.'s [13], Nie et al.'s [15], and Huang et al.'s [29] studies, while the network topology and configurations are set according to our proposed models for a fair comparison. In all methods, the input layer includes 257 nodes due to the size of the noisy magnitude spectrum. The coefficient output layer due to the activation labels has $100 \times 2 = 200$ nodes and the main output layer for the main spectral magnitudes labels contains $257 \times 2 = 514$ nodes. The activation functions of the hidden layers and the main output layer are leaky rectified linear units (LReLU) [34] with $\alpha = 0.1$ ($f(x) = \max(\alpha x, x)$) and linear, respectively. The activation function of the coefficient output layer is ReLU ($f(x) = \max(0, x)$) due to the non-negativity of the activation coefficients. The classifier DNN has two hidden layers of 1,024 units with the ReLU function and one output layer of three units with the softmax activation function for three classes. The softmax output is a probability distribution in the [0,1] range with a total sum of 1. The batch normalization is used after each hidden layer for faster training convergence. The baseline LSTM-Mask network has two LSTM layers of 3,072 LReLU units and a fully connected (FC) layer of 1,024 LReLU units, and a fully connected output layer of 257 linear units for mask values prediction. This configuration is according to the LSTM part in [7] and the LSTM-IRM method [30] which was used as a comparison method in Strake et al.'s [7] study. However, instead of 425 nodes in Strake et al.'s [7] study, here the nodes of the two LSTM layers are experimentally set to 3,072 to have better results.

Table 1: Noise classification results.

Actual class	Predicted class		
	Factory	Babble	Machinegun
Factory	0.94	0.03	0.03
Babble	0.03	0.90	0.07
Machinegun	0.03	0.07	0.90

The classifier DNN uses the cross-entropy loss function and the proposed model uses L_{MO} and L_{MOFD} loss functions. All networks are trained by the Adam optimizer [35] with an initial learning rate of 0.001 and a maximum epoch of 100. The weights of α_1 and α_2 in Equation (10) and the M parameter in Equation (9) are experimentally set to 2.3, 0.1, and 2, respectively. The N in Figure 1 is equal to 3 due to the three training noise classes. Moreover, to avoid overfitting, the early stopping method which stops the learning process based on the minimum validation loss is used in all models.

4.2. Results and Discussion. This section explains the results of the proposed methods and baselines evaluated by three metrics of perceptual evaluation of speech quality (PESQ) [36], short-time objective intelligibility (STOI) [37], and frequency-weighted segmental SNR (SNR_{fw}) [38, 39].

First, the classification results of the training noise types are displayed in Table 1 which indicates an appropriate classification. The classifier DNN identifies the mismatched noises of *factorymachine* and *windshieldrain* with the rates of (0.64, 0.22, 0.14) and (0.80, 0.18, 0.02), respectively. So, based on these prediction ratios (w_1, w_2, w_3 in Figure 1), the proportional contribution of the corresponding models is used for enhancement.

The average improvements of the PESQ metric (gPESQ), STOI, and SNR_{fw} results of different methods over matched noise types and for each input SNR are displayed in Figures 2–4, respectively.

As can be seen in Figures 2–4, the proposed Jnt-DNN-NMF-MO and Jnt-DNN-NMF-MOFD methods outperform the baseline comparative methods. The superiority of Jnt-DNN-NMF-MO results over Jnt-DNN-NMF [15] in terms of three metrics is due to the use of extracted speech and noise NMF activation coefficients as direct intermediate targets by DNN, which are structural features and act as prior knowledge for DNN training. In fact, incorporating these coefficients in addition to the main signals into the loss function has led to improvement. The superior performance of Jnt-DNN-NMF-MO over other baseline methods is due to the joint learning of the integrated model of DNN and NMF bases and the use of the structural NMF features for DNN. According to Figures 2–4, in terms of three metrics, the Jnt-DNN-NMF-MOFD method outperforms the Jnt-DNN-NMF-MO and also the baselines, which demonstrates the strength of the proposed

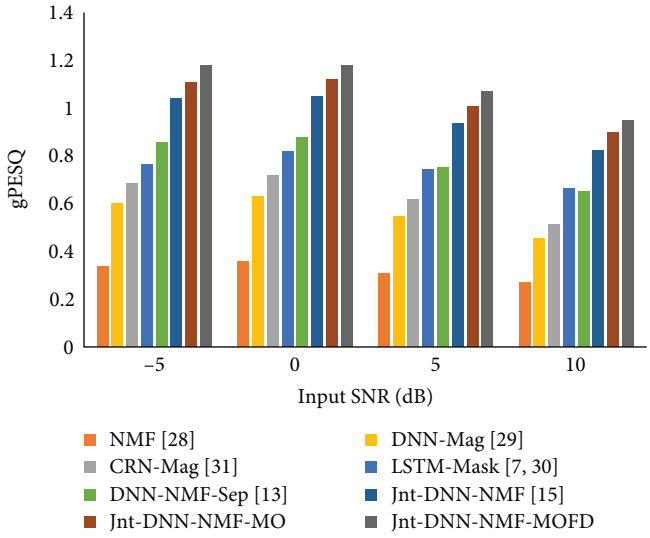

FIGURE 2: The average gPESQ results over matched noises at different input SNRs.

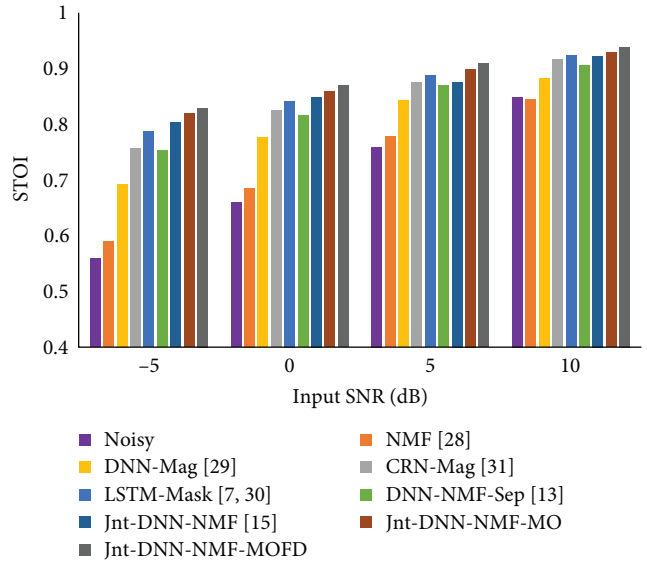

FIGURE 3: The average STOI results over matched noises at different input SNRs.

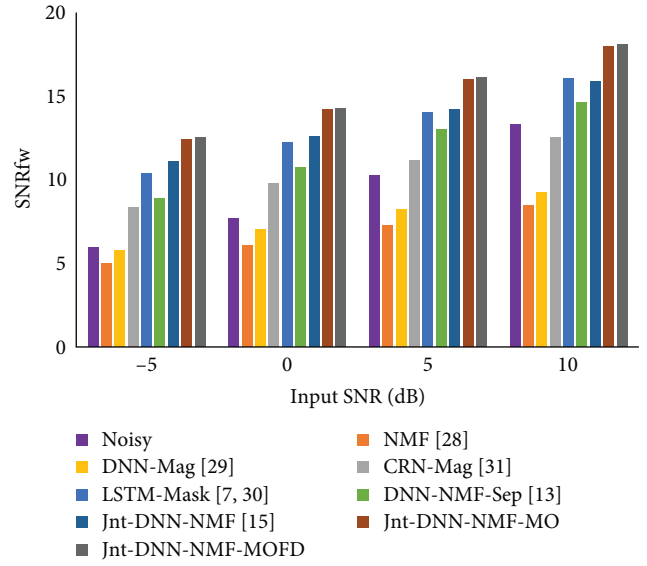

FIGURE 4: The average SNR_{fw} results over matched noises at different input SNRs.

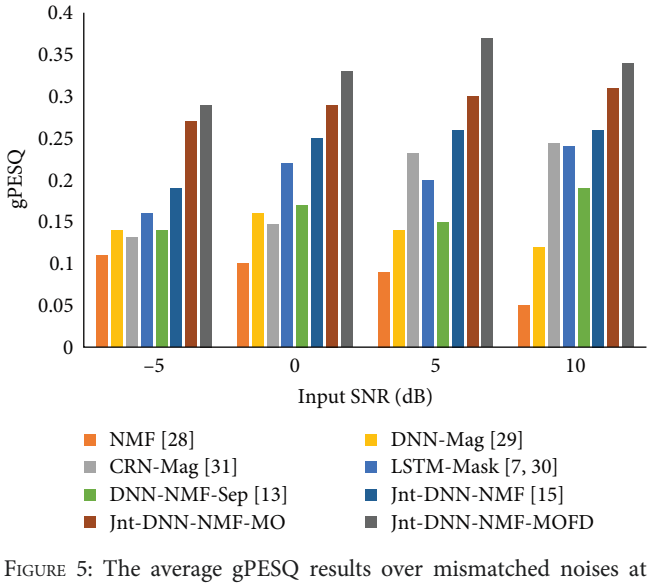

FIGURE 5: The average gPESQ results over mismatched noises at different input SNRs.

frequency-differentiated loss function. By using L_{MOFD}, the frequency dynamics are learned better than using L_{MO}.

The results in Figure 2 show that the proposed methods (the last two methods) have higher PESQ improvements at each input SNR than the previous methods. The proposed Jnt-DNN-NMF-MOFD method reaches about 0.15 more average PESQ improvement compared to the best baseline (Jnt-DNN-NMF [15]) and 1.10 over the noisy speech. Also, the increase of STOI and SNR_{fw} scores for the proposed methods is more than the baselines (Figures 3 and 4). Better extraction of frequency characteristics in the Jnt-DNN-NMF-MOFD method has led to improved performance.

Furthermore, to evaluate the generalization ability, the average gPESQ results of the proposed and comparison methods over the mismatched noise types are given in Figure 5 for each input SNR. According to this figure, the results of the proposed methods are better compared to others. It indicates an average PESQ improvement of 0.11 higher than the best baseline (Jnt-DNN-NMF [15]).

To examine the enhancement performance in more mismatched noise types, the average gPESQ results of the proposed Jnt-DNN-NMF-MOFD method and the baseline LSTM-Mask method [7, 30] over two other mismatched noises, *restaurant* and *street*, are given in Figure 6 for each input SNR. The

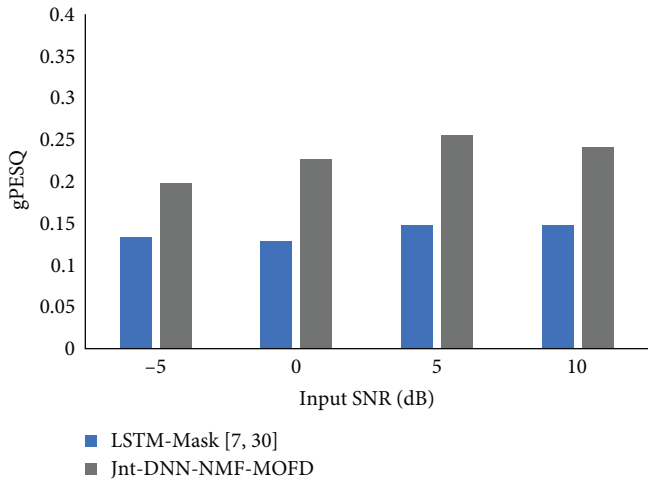

Figure 6: The average gPESQ results over the *restaurant* and *street* mismatched noises at different input SNRs.

Table 2: The noise classification (NC) performance for the proposed Jnt-DNN-NMF-MOFD model.

	Matched noises			Mismatched noises		
	PESQ	STOI	SNR_{fw}	PESQ	STOI	SNR_{fw}
Noisy	2.10	0.70	9.30	2.08	0.74	6.11
Without NC	3.10	0.87	13.96	2.35	0.78	8.73
With NC	3.20	0.88	15.27	2.42	0.79	9.51

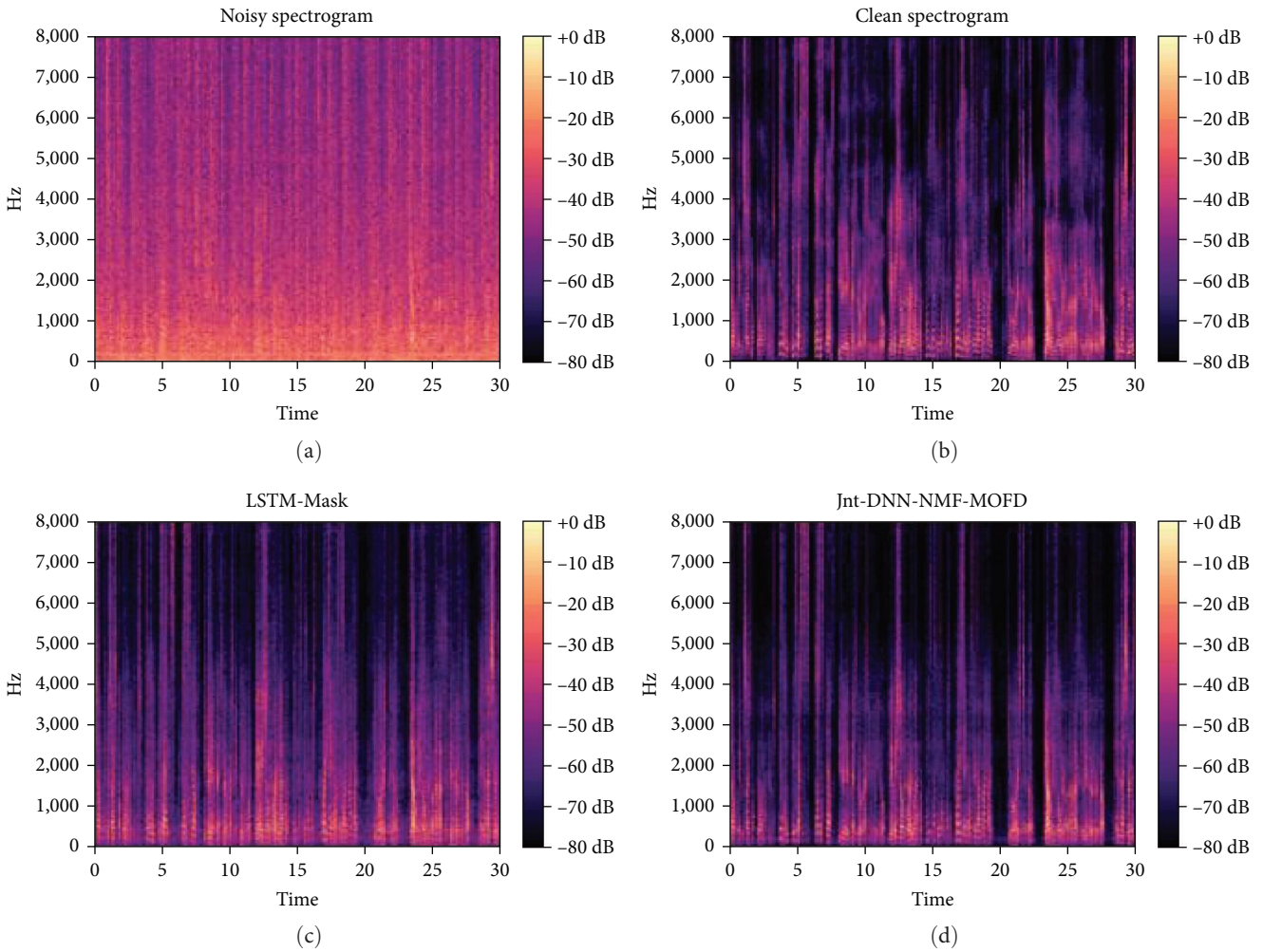

Figure 7: The magnitude spectrograms of noisy speech contaminated with factory noise at −5 dB SNR (a) and clean speech (b); the estimated speech by LSTM-Mask (c) and by the proposed Jnt-DNN-NMF-MOFD method (d).

restaurant and *street* noises are given from the Aurora-2 database [40]. These noises have different properties and structures from the previous matched and mismatched noises. The estimated classification rates of the *restaurant* and *street* noises are approximately (0.11, 0.71, 0.18) and (0.18, 0.28, 0.54), respectively. According to this figure, the performance trend is similar to the previous mismatched noises indicating the effectiveness of the suggested approach and the generalization capability.

Moreover, to investigate the noise classification (NC) performance, the average results of the final proposed Jnt-DNN-NMF-MOFD method with and without NC are reported in Table 2 for matched and mismatched noises. Better results *with NC* compared to *without NC* indicate the benefit of using the noise-specific models and the fusion strategy.

Finally, the magnitude spectrograms of the estimated speech by the proposed Jnt-DNN-NMF-MOFD method and the baseline LSTM-Mask method are given in Figure 7, as examples. As can be seen, Jnt-DNN-NMF-MOFD restores speech components and removes noise parts better than LSTM-Mask. One reason for this result is due to the joint cooperation of NMF and DNN and the direct effect of the NMF activation coefficients on DNN as structural intermediate target features. The joint estimation of the actual spectral targets and the activation coefficients by DNN as multi-objective joint learning, and also consideration of the frequency domain spectrum changes in the loss function are the main reasons for this result.

5. Conclusion

We proposed a joint multi-objective model of NMF and DNN with new loss functions for speech enhancement. In the proposed multi-objective loss function (L_{MO}), the NMF activation coefficients are estimated simultaneously with the objective spectral signals by the DNN. Setting the NMF activation coefficients as a direct target of DNN and integration of the NMF speech and noise bases and wiener filters with the DNN layers leads to further improvement. It is due to the extraction of the harmonic structures by NMF and the direct incorporation of the extracted structural characteristics into the DNN structure. Then, to consider and maintain the frequency domain changes of speech and noise spectrums, we proposed a frequency-differentiated loss function (L_{FD}) that considers the spectrum differences between the adjacent frequency bins. Finally, to improve the enhancement results, we proposed a multi-objective frequency differentiated loss function (L_{MOFD}) to optimize the Jnt-DNN-NMF model which is a weighted combination of the frequency-differentiated loss function and two MSEs related to the actual spectral signals and the NMF activation coefficients.

Data Availability

Research data are not shared.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported in part by Iran National Science Foundation (INSF) under grant no. 97014206.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas et al., "Speech enhancement with an adaptive Wiener filter," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 53–64, 2014.

[3] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

[4] S. Seyedin and M. Ahadi, "Feature extraction based on DCT and MVDR spectral estimation for robust speech recognition," in *2008 9th International Conference on Signal Processing*, pp. 605–608, IEEE, Beijing, China, October 2008.

[5] M. Pashaian, S. Seyedin, and S. M. Ahadi, "A novel jointly optimized cooperative DAE-DNN approach based on a new multi-target step-wise learning for speech enhancement," *IEEE Access*, vol. 11, pp. 21669–21685, 2023.

[6] Y. Wang, J. Han, T. Zhang, and D. Qing, "Speech enhancement from fused features based on deep neural network and gated recurrent unit network," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, pp. 1–19, 2021.

[7] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, pp. 1–26, 2020.

[8] R. Safari, S. M. Ahadi, and S. Seyedin, "Modular dynamic deep denoising autoencoder for speech enhancement," in *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 254–259, IEEE, Mashhad, Iran, 2017.

[9] K. Wang, W. Lu, P. Liu, J. Yao, and H. Li, "Multi-stage attention network for monaural speech enhancement," *IET Signal Processing*, vol. 17, no. 3, 2023.

[10] S. Alisamir, S. M. Ahadi, and S. Seyedin, "An end-to-end deep learning model to recognize Farsi speech from raw input," in *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 1–5, IEEE, Tehran, Iran, 2018.

[11] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[12] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[13] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[14] S. Nie, S. Liang, H. Li et al., "Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 469–473, IEEE, Shanghai, China, 2016.

[15] S. Nie, S. Liang, W. Liu, X. Zhang, and J. Tao, "Deep learning based speech separation via NMF-style reconstructions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2043–2055, 2018.

[16] H. Li, S. Nie, X. Zhang, and H. Zhang, "Jointly optimizing activation coefficients of convolutive NMF using DNN for speech separation," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, pp. 550–554, ISCA, 2016.

[17] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 499–503, IEEE, Shanghai, China, 2016.

[18] H.-W. Tseng, M. Hong, and Z.-Q. Luo, "Combining sparse NMF with deep neural network: a new classification-based approach for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2145–2149, IEEE, South Brisbane, QLD, Australia, 2015.

[19] H. Jia, W. Wang, and S. Mei, "Combining adaptive sparse NMF feature extraction and soft mask to optimize DNN for speech enhancement," *Applied Acoustics*, vol. 171, Article ID 107666, 2021.

[20] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6107–6111, IEEE, Florence, Italy, 2014.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 892–902, 2014.

[22] D. S. Williamson, Y. Wang, and D. Wang, "A sparse representation approach for perceptual quality improvement of separated speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7015–7019, IEEE, 2013.

[23] D. S. Williamson, Y. Wang, and D. Wang, "A two-stage approach for improving the perceptual quality of separated speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7084–7088, IEEE, Florence, Italy, 2014.

[24] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1773–1783, 2017.

[25] D. S. Williamson, Y. Wang, and D. Wang, "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1399–1407, 2015.

[26] D. S. Williamson, Y. Wang, and D. Wang, "Deep neural networks for estimating speech model activations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5113–5117, IEEE, South Brisbane, QLD, Australia, 2015.

[27] J. Chen, K. K. Paliwal, and S. Nakamura, "Cepstrum derived from differentiated power spectrum for robust speech recognition," *Speech Communication*, vol. 41, no. 2-3, pp. 469–484, 2003.

[28] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, IEEE, Corfu, Greece, 2011.

[29] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1562–1566, IEEE, Florence, Italy, 2014.

[30] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[31] Y. Shi, W. Yuan, S. Hu, and Y. Lou, "A convolutional recurrent neural network for real-time speech enhancement," in *International Speech Communication Association (Interspeech)*, pp. 3229–3233, 2018.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[33] 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[34] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, PMLR, 2011.

[35] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, University of Amsterdam, 2014.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, IEEE, USA, 2001.

[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[39] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 586–590, 1978.

[40] D. J. B. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *6th International Conference on Spoken Language Processing*, ICSLP, China, 2000.