IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 January 2023, accepted 23 February 2023, date of publication 1 March 2023, date of current version 7 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250820

RESEARCH ARTICLE

A Novel Jointly Optimized Cooperative DAE-DNN Approach Based on a New Multi-Target Step-Wise Learning for Speech Enhancement

MATIN PASHAIAN, SANAZ SEYEDIN, (Senior Member, IEEE), AND SEYED MOHAMMAD AHADI

Speech Processing Research Laboratory, Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran 1591634311, Iran

Corresponding author: Sanaz Seyedin (sseyedin@aut.ac.ir)

This work was supported in part by the Iran National Science Foundation (INSF) under Grant 97014206.

ABSTRACT In this paper, we present a new supervised speech enhancement approach based on the cooperative structure of deep autoencoders (DAEs) as generative models and deep neural networks (DNN). The DAE is used as a nonlinear alternative to nonnegative matrix factorization (NMF) for the extraction of harmonic structures and encoded features of the noise, clean and noisy signals, and a DNN is deployed as a nonlinear mapper. We introduce a deep network imitating NMF in a nonlinear manner to overcome the problems of a simple linear model, such as performance degradation in non-stationary environments. Compared to combinatorial NMF and DNN methods, we perform all the decomposition, enhancement, and reconstruction processes in a nonlinear framework via a suitable cooperative structure of encoder, DNN, and decoders, and jointly optimize them. We also propose a supervised hierarchical multi-target training approach, performed in two steps, such that the DNN not only predicts the low-level encoded features as primary targets but it also predicts the high-level actual spectral signals as secondary targets. The first step acts as a pretraining for the second step which improves the learning strategy. Moreover, to exploit a more discriminative model for noise reduction, a DNN-based noise classification and fusion strategy (NCF) is also proposed. The experiments on TIMIT dataset reveal that the proposed methods outperform the previous approaches and achieve an average perceptual evaluation of speech quality (PESQ) improvement of up to about 0.3 for speech enhancement.

INDEX TERMS Deep autoencoders (DAEs), deep neural network (DNN), joint optimization, nonnegative matrix factorization (NMF), speech enhancement.

I. INTRODUCTION

The goal of a speech enhancement problem is to recover the desired speech from a noisy speech. Speech enhancement algorithms can be categorized into supervised and unsupervised techniques. The unsupervised algorithms such as spectral subtraction [1], Wiener filter [2], [3], Kalman filtering [4], and minimum mean-square-error (MMSE) estimator [5] are based on the probabilistic models of noise and speech. Inaccurate estimation of the noise statistical information is a drawback of these methods. In contrast, supervised algorithms applied to speech enhancement, such as nonnegative matrix factorization (NMF) [6], [7], [8], sparse coding [9], [10], deep neural networks (DNN) [11], [12], and deep autoencoder (DAE) [13], [14] need a set of training data to learn a structure to be applied in an unknown situation. In the presence of an adequate amount of training data, the performance of supervised algorithms can be better than the unsupervised ones [6], [11], [13].

In recent years, deep learning-based methods have been extensively studied and have significantly improved enhancement performance over conventional approaches [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22],

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

[23], [24]. Deep learning networks have significant ability in complex function mapping. The approach in [22] is an on-the-fly speech enhancement approach in which a recurrent variational autoencoder (RVAE) and an NMF model are considered for speech signal and noise, respectively. Noise signal parameters are approximated during the test time via a variational expectation-maximization algorithm (VEM) to perform the enhancement. Reference [23] proposes a two-branch convolutional neural network (CNN) model with some interaction modules between the two branches of speech and noise to help each other. A typical spectral DNN-based speech enhancement model commonly maps the speech-noise mixture features into the magnitude spectrograms of the sources directly [25], [26]. In [17] and [18], a convolutional-recurrent network (CRN) is used for magnitude and complex spectral mapping, respectively. In another approach, DNN can be used for mapping the mixture into a specific spectral ideal mask as a gain that describes the proportion of speech and noise in the mixed-signal [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], such as the ideal binary mask (IBM), ideal ratio mask (IRM), phase-sensitive mask (PSM), and complex ideal ratio mask (cIRM). References [15] and [16] used a long short-term memory (LSTM) network for mask estimation and noise suppression. In [33], where the IRM and PSM are the training targets, the time-frequency (T-F) components are differentially weighted to consider the T-F energy distribution of speech. The ideal masks or the original magnitude spectrograms that are used as targets have prominent spectro-temporal structures due to the speech production process [32]. However, one problem of using deep networks in a typical deep learning-based speech enhancement application is that it does not consider the sparsity and intrinsic harmonic structure of speech.

Along with the development of deep learning-based methods in speech enhancement, some model-based methods such as NMF and sparse coding have also grown significantly. This is mostly due to their ability in representing the structure of a signal as a model. Unlike Fourier or wavelet analysis, these methods have an adaptive representation with data and could create the structural representation of speech and noise using supervised techniques [6], [9], [38]. In these methods, a given data matrix is approximated by a simple linear product of a base matrix as a dictionary and a coefficient matrix through a dictionary learning process in which the fundamental patterns of the given data matrix are extracted by minimizing a specific cost function. While using these methods for spectral speech enhancement, the magnitude spectrogram of a contaminated signal is typically decomposed into the weighted linear combinations of the trained base matrices of the speech and noise. The basic assumption in these methods is the orthogonality conditions between the base matrices of the separate sources. However, this condition is not met in many cases, which leads to errors in the estimation of the coefficient matrices and the separation of the sources. If the underlying assumptions are met, these methods perform well.

Some research works combined DNN with NMF or its variants to overcome their shortcomings and improve their performance [39], [40], [41], [42], [43], [44], [45], [46]. In [39], a DNN is learned to map the spectral magnitude of the noisy signal to a set of concatenated NMF activation coefficients of speech and noise. Then, the corresponding activation part of each source is separated from the estimated activation matrix of the observed noisy signal. Next, they are linearly combined with the corresponding trained basis matrix manually to estimate the magnitude spectrograms of the main signals. Then, a Wiener filtering technique is applied as a post-processing step to limit the sum of the estimated speech and noise magnitudes to the mixture magnitude. Moreover, in [46], instead of directly estimating the original magnitude spectrograms or mask values, the estimation of the NMF activation matrix of an IRM is set as the DNN target. Then, the estimated activation matrix of the IRM is multiplied by the corresponding learned basis matrix to reconstruct the IRM. Finally, this estimated IRM is used to separate the speech from the mixture. In these works, the NMF basis matrices capture the patterns of the target sources to some extent, and the DNN approximates the corresponding activation matrices. However, these two stages are performed separately and may lead to double estimation errors. Furthermore, the target of DNN is not the actual objective of the separation, and it just estimates an intermediate objective produced by NMF.

Some works have gone further and proposed a jointly combinational model of DNN and NMF [40], [41], or other variants like Convolutive NMF (CNMF) [42]. They integrate the learned NMF or CNMF basis matrices of speech and noise into a DNN as an extra layer. Then, the DNN directly estimates the actual targets of speech and noise from the mixed signal. On the other hand, in [43], the speech and noise activation coefficients are estimated by applying NMF on the noisy magnitude spectrum and using the concatenated learned speech and noise bases. Then the noisy activation coefficients composed of the concatenated activation coefficients of speech and noise are used as DNN input training features. The main spectral magnitude of speech is then estimated through the modified cost function. In [44] and [45], the DNN is also used for mapping the NMF activation coefficients of the noisy speech to an IBM mask [44] or a new soft mask [45].

In some other works, DNN and NMF are combined in two separate stages to improve the enhancement quality [47], [48], [49], [50], such that the DNN is first used for separation and then, NMF performs the enhancement [47], [48], [49], [50], or vice versa [50]. Grais et al. evaluated both of them and compared them with the combination of two DNNs, one for separation and one for enhancement [50]. Williamson et al. used the NMF and sparse reconstruction as a post-processing step to further enhance the source separated by the binary, soft, and ratio masks in [47], [48], and [49], respectively. Here, a DNN is first used for mask estimation; then, the speech separated by the mask is represented as a

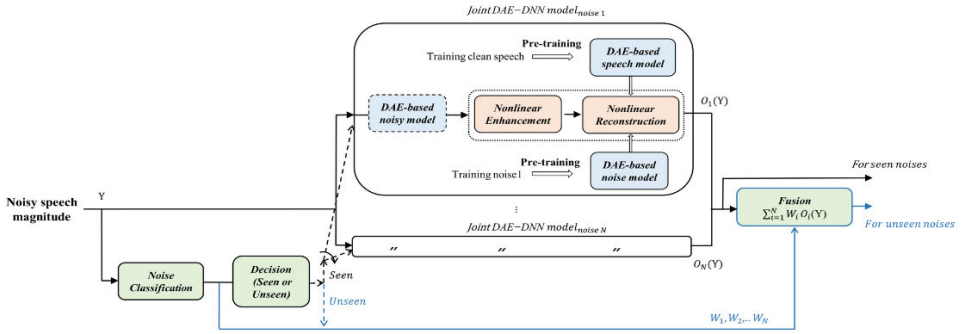

FIGURE 1. The overall system block diagram. W_is are the weights corresponding to the classifier accuracies in each noise type. For seen noises, we choose the best joint DAE-DNN model based on the noise classification decision, while a fusion strategy is required for unseen cases.

linear or sparse linear combination of the base vectors from the pre-trained clean basis matrix. In comparison with [47], in [51] and [52], a DNN is used in the second stage to act as NMF and estimate the NMF activation coefficients of the clean speech from the masked speech. Then, by linearly combining the estimated clean activation coefficients and the pre-trained basis spectra manually outside of the network, the clean speech is approximated. It has been revealed that the use of DNN in estimating the NMF activation coefficients generates less unwanted disturbance and improves the performance more than the NMF method itself [46]. In all of these works, the NMF learned bases and the corresponding activation coefficients used in the DNN structure are extracted from a linear operation. The linear operations are combined with nonlinear ones and either the nonlinear DNN is forced to act as a linear NMF decomposition and map the mixture into the linear activation coefficients directly [39] or its effect is implicit and the main signals are used as the DNN target output [40], [41], [42].

Autoencoders (AEs) are suitable for sparse coding and dictionary learning due to the mapping of the input to a low-dimensional space in the bottleneck layer [53], [54], [55], [56]. In [55], the bottleneck features are used for speech recognition. In [56], the authors used the components extracted from the encoded hidden layer of an AE trained by a complex noisy speech spectrogram as the input for another denoising AE with a complex mask target.

Model-based techniques are beneficial due to the incorporation of prior information. However, they are sensitive to inaccuracies in model knowledge which can lead to poor performance in complicated real systems with dynamic behavior [57]. These methods are more successful for structured interferences. Furthermore, these algorithms mostly necessitate costly inference making them difficult to be applied in real-environment speech-based applications [32].

Data-driven techniques such as DNNs lead to better performance in case they can be given prior knowledge of the signal [57], such as the signal structure. Deep networks have the advantage of working in non-stationary environments better, but the problem of being trapped in local minima is the main disadvantage. This may be solved by adding some prior knowledge such as those extracted through NMF or AE. Given enough training data, this method has also been found to generalize well. Furthermore, the system works in a frame-by-frame manner, and inference is quick, allowing real-time implementation [32].

A. OUR WORK AND CONTRIBUTIONS

In traditional DNN-based speech enhancement which is a direct mapping from noisy features to actual targets, the learning process is difficult due to the variation and contamination of features in different noisy conditions. However, using the structural features which are relatively invariant in various auditory conditions can lead to regulated learning and more robust mapping [46]. Nonlinear models are believed to be able to extract and encode the basic structure of the signal with higher clarity than a linear NMF model. Hence in this paper, we propose a novel supervised method that, as shown in Fig. 1, uses sparse nonlinear generative DAEs models to provide prior knowledge for deep networks. Thus, it could create a more suitable integrated nonlinear framework to improve the results. To this end, we propose exploiting DAEs for nonlinear feature extraction and dictionary learning while the whole decoder portions could be considered as nonlinear dictionaries. In our designed model, the DNN works for nonlinear enhancement and three DAE-based models are used for feature and structure extraction. Speech and noise DAEs are used to create the corresponding decoders and encoded features which are respectively used as the reconstruction part and the intermediate output target of DNN. Also, by using the encoder portion of noisy DAE, the noisy encoded feature is produced. However, when the encoded noisy feature is used as the DNN input feature for enhancement, the dashed-line noisy DAE is applied. Otherwise, when we use the main noisy spectrum as the input feature, the mentioned noisy DAE is not applied. The sparse encoded features are extracted in a nonlinear manner by DAEs, and are fed to the DNN as input (noisy encoded features) and/or target output features (speech and noise encoded features) for nonlinear mapping. Reconstruction of the objective speech and noise signals is

performed by using the relevant pre-trained decoders instead of linear multiplication of NMF basis matrices. Moreover, to further consider the extracted prominent structure of speech components in DNN, the DNN training is performed in two steps. Firstly, we pre-train the DNN to predict the speech and noise DAEs' encoded features as primary targets. Then, we fine-tune the joint model of the pre-trained DNN (enhancement part) and decoders (reconstruction part) with the actual targets of speech and noise signals to improve the results. Thus, in the fine-tuning stage, all pre-trained layers continue to be learned with the new targets and cost function. Another issue in speech enhancement problems is learning a general model with different kinds of noises. In this paper, as shown in Fig. 1, a DNN-based noise classification and fusion strategy (NCF) is applied to choose either one, or an appropriate combination of the learned noise-specific models for treating each input noisy speech. According to Fig. 1, the noise type is first detected in the decision block. Then, for seen noises, the related joint model of DAEs and DNN, trained with that kind of noise, is employed for the feature extraction, enhancement, and reconstruction stages. Meanwhile, for unseen noises, a weighted average of the outputs from multiple models is computed. The weights are based on the classification rates. The advantages of the suggested NCF strategy are improving the generalization, accelerating the convergence, and reducing the probability of local minima traps. The suggested fusion strategy, using different weights of the classifier, shows its effectiveness in unseen noise conditions when that noise type was not seen in the training phase, and thus, one specific model cannot accurately follow the signal.

Our main contributions compared to earlier works are:

- We propose a joint DAE-DNN model, which is a joint and nonlinear sparse equivalent of NMF-DNN previously suggested in [39], [40], [41], and [42]. Therefore, we propose a DAE-based nonlinear replacement for NMF for use in the feature extraction (signal decomposition) and signal reconstruction stages. Also, the nonlinear sparse encoded features of speech and noise, and jointly the objective speech and noise magnitudes are estimated nonlinearly from the noisy feature by using an appropriate DNN integrated with decoder layers as nonlinear reconstruction layers (the joint effort of DAE and DNN). Thus, the joint optimization of the whole cooperative DAE-DNN model is performed. This is contrary to [39], in which NMF and DNN operated independently, and also the linear NMF activation coefficients were set as DNN targets, and the NMF reconstruction was applied manually outside of the DNN to reconstruct the main objective signals. This different strategy also goes for [40], [41], and [42] where the activation coefficients did not directly affect the DNN and the objective signals were directly optimized by DNN through an extra linear NMF reconstruction layer.
- In comparison with [39], [40], [41], [42], [43], [44], and [45], where the NMF activation coefficients were just used as DNN output [39], [40], [41], [42] or input features [43], [44], [45], we investigate the use of NMF as well as DAE-based sparse encoded features in both DNN input and output parts. Thus, the proposed model contains two parts, namely input-related and output-related ones. In [40], [41], [42], and [43], the learning process contained a direct mapping from the noisy signal to the main separation targets without any direct injection of knowledge about the activation coefficients in prior layers. Contrary to [40], [41], [42], and [43], we propose step-wise learning. This involves a two-step training approach with mapping the noisy feature to the sparse encoded features of speech and noise in the first step, and then to the main objective signals in a hierarchical framework. The proposed hierarchical approach helps reduce the local minima problem which could lead to better results.
- In [43], the concatenated speech and noise NMF activation coefficients were considered as noisy activation coefficients and used as DNN input features for further enhancement while ignoring the overlap between speech and noise bases. Unlike the mentioned strategy in [43], in this paper, in the input-related proposed models, the linear and nonlinear sparse encoded noisy features and the base models are extracted directly from the main noisy signal to consider any possible correlation between the speech and noise, and then, are used as DNN input features. Then, extracting the speech and noise encoded features from the noisy encoded features is performed by a DNN in a nonlinear framework rather than the conventional linear separation scheme in the NMF-based enhancement method. This nonlinear framework could represent the speech signal properties more appropriately.
- We compensate for the necessary orthogonality conditions between the linear basis matrices of speech and noise in the NMF-based speech enhancement: the equivalent nonlinear modeling and also mapping of the noisy signal to the target encoded features is carried out by DNN as a regression model. Deep hidden layers incorporate and learn the inter-dependencies and tolerate the mutual coherency between the speech and noise dictionaries to estimate the sparse enough encoded features.
- We do a multi-target estimation of the encoded features and the main signals of speech and noise in a single DNN: it jointly estimates the speech and noise encoded features and the main signals hierarchically at the related output layer based on the individual related loss function. This further considers the complementary and correlative specifications of speech and noise.
- We suggest the appropriate incorporation of the noise-specific models with a noise classification and fusion strategy (NCF).
- We provide an applicable and suitable model under limited data and limited processing capability conditions via the proposed gradual learning and noise classification

strategies. Obviously, with larger amounts of data and larger processing capacities, many challenges may disappear.
- We present a well-suited idea in a simple but effective structure that can be applicable even in other complicated networks which may be suggested in the future.

The remainder of this paper is coordinated as follows: In Section II, the basic problem of speech separation is introduced. Section III reviews the speech separation approach based on NMF. Section IV discusses the proposed nonlinear sparse deep model, composed of extracting the intrinsic spectral structure of the signal (feature extraction stage) and the cooperative DAE-DNN model. In Section V, the experimental results are reported and discussed. Finally, the conclusion and future work are given in Section VI.

II. PROBLEM DESCRIPTION

For a speech signal contaminated by additive noise, i.e. $y(i) = s(i) + n(i)$, where i is the sample index, the short-time Fourier transform (STFT) magnitude spectrogram without considering the speech-noise cross-term is approximated as follows [58]:

$$|Y(f,t)| \approx |S(f,t)| + |N(f,t)| \tag{1}$$

where $|\cdot|$ is the absolute value operator and f and t are the indices of frequency and time, respectively. Also, $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{N} \in \mathrm{R}_{\geq 0}^{F\times T}$ are the magnitude spectra of the noisy signal, clean speech, and noise, respectively. Also, F and T are the frequency bins and time frame numbers, respectively. For the sake of simplicity, we may show (1) as $\mathbf{Y} \approx \mathbf{S} + \mathbf{N}$.

The function of speech enhancement in presence of additive noise $n(i)$ is to acquire an estimate $\widehat{S}(i)$ of clean speech $s(i)$ from a noisy one $y(i)$. In practice, the magnitude spectrum of clean speech is usually approximated and then combined with the noisy phase spectrum. Then, the time-domain signal of the clean speech is reconstructed by using inverse STFT (ISTFT). However in some studies, phase spectrum enhancement is also considered for its importance in perceptual quality [31], [59], but its efficacy is under question. Therefore, many studies have good separation results while reconstructing with a noisy phase [11]. Hence, in this work, only the magnitude enhancement is performed.

III. SPEECH SEPARATION BASED ON NMF

NMF [8] is one of the linear model-based techniques that can be used to separate speech from the speech-noise mixture signal. In this method, a nonnegative matrix, which is usually the magnitude spectrogram of a signal $\mathbf{X}\in\mathrm{R}_{\geq 0}^{F\times T}$, is factorized as a product of a nonnegative basis matrix $\mathbf{W}_x\in\mathrm{R}_{\geq 0}^{F\times K}$ (where $K \leq F$) and an activation matrix $\mathbf{H}_x\in\mathrm{R}_{\geq 0}^{K\times T}$ according to (2). K, T, and F are the number of basis vectors, time frames, and frequency bins, respectively. The basis matrix includes the basic patterns of $\mathbf{X}$, and the activation matrix represents the related amounts as coefficients that linearly combine the basis vectors to approximate $\mathbf{X}$. $\mathbf{X}$ can be the clean speech $\mathbf{S}$ or noise signal $\mathbf{N}$.

$$\mathbf{X} \simeq \mathbf{W}_x\mathbf{H}_x \tag{2}$$

$\mathbf{W}_x$ and $\mathbf{H}_x$ are approximated by iteratively minimizing the distance between $\mathbf{X}$ and $\mathbf{W}_x\mathbf{H}_x$ using the Kullback-Leibler (KL) divergence [8]. In the training stage, $\mathbf{W}_x$ and $\mathbf{H}_x$ are usually randomly initialized and obtained using the iterative multiplicative update rules. Then, $\mathbf{H}_x$ is discarded and $\mathbf{W}_x$ is kept constant for the separation stage. Once $\mathbf{W}_x$ is found, the STFT magnitude matrix of a test signal is estimated as the product of the fixed trained basis matrix $\mathbf{W}_x$ and a new activation matrix that is computed by (2).

In the case of combining two sources such as speech separation from the noisy one, the basis matrices of the sources should be known in advance. Therefore, in the training phase, according to (2), NMF is applied to each source and $\mathbf{W}_s\in\mathrm{R}_{\geq 0}^{F\times Ks}$ and $\mathbf{W}_n\in\mathrm{R}_{\geq 0}^{F\times Kn}$ are trained individually for the speech and noise signals. Then, the basis matrix $\mathbf{W}_y$ of the noisy speech is formed by their concatenation ($[\mathbf{W}_s\mathbf{W}_n]$). K_s and K_n denote the speech and noise basis vectors sizes, respectively, and $K_s + K_n = K$. In the separation stage, by applying the NMF decomposition to the trained $[\mathbf{W}_s\mathbf{W}_n]$, the activation $\widehat{\mathbf{H}}_y$ is extracted as the best approximation of the mixture as below. $\widehat{\mathbf{H}}_y$, each part of which corresponds to a specific source ($\begin{bmatrix}\widehat{\mathbf{H}}_s\in\mathrm{R}_{\geq 0}^{Ks\times T}\\ \widehat{\mathbf{H}}_n\in\mathrm{R}_{\geq 0}^{Kn\times T}\end{bmatrix}$), is obtained using the KL cost function and the multiplicative update rules [8].

$$\begin{aligned}\mathbf{Y}_{test} \simeq \mathbf{W}_y\widehat{\mathbf{H}}_y = [\mathbf{W}_s\mathbf{W}_n]\begin{bmatrix}\widehat{\mathbf{H}}_s\\ \widehat{\mathbf{H}}_n\end{bmatrix} &= \mathbf{W}_s\widehat{\mathbf{H}}_s + \mathbf{W}_n\widehat{\mathbf{H}}_n\\ &= \widehat{\mathbf{S}} + \widehat{\mathbf{N}} \quad \widehat{\mathbf{S}},\ \widehat{\mathbf{N}} \in R_{\geq 0}^{F\times T}\end{aligned} \tag{3}$$

According to (3), an estimation of the magnitude spectrum of each source is calculated by multiplication of the related trained basis matrix and its corresponding activation matrix by subdividing the estimated $\widehat{\mathbf{H}}_y$ into two sections of $\widehat{\mathbf{H}}_s$ and $\widehat{\mathbf{H}}_n$. For further smoothing and obtaining the final separated magnitudes, a Wiener filter composed of the approximated speech and noise magnitudes $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{N}}$ is multiplied by the magnitude spectrum of the mixed-signal to restrict the sum of the approximated speech and noise to the mixed-signal [6], [25]. Then, by transforming the result to the time domain using the noisy phase and ISTFT, the speech and noise waveforms are reconstructed. Although this method is easy to implement, it has some problems. For example, each basis matrix $\mathbf{W}_x$ is learned individually, and the relation $\mathbf{W}_x\widehat{\mathbf{H}}_x = \widehat{\mathbf{X}}$ in (3) is only valid when the orthogonality conditions between the base matrices of speech and noise are met [7], [60]. In this paper, instead of using the concatenated bases, a DNN learns the mapping of the noisy mixture to the speech and noise activations to consider the overlaps between the bases of speech and noise, and then, it estimates them in a nonlinear manner.

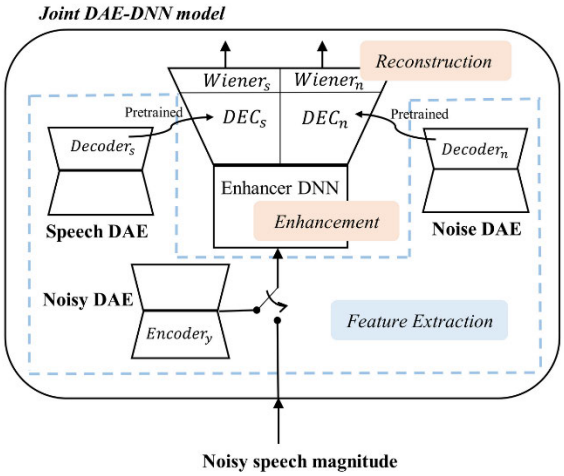

FIGURE 2. Details of the joint DAE-DNN model shown as one box in Fig. 1. The feature extraction part (dashed lines) is composed of three DAEs. The decoder portions of the pre-trained speech and noise DAEs are first integrated into DNN as reconstruction layers, and then, they are updated during the joint learning of DNN-DEC (fine-tuning).

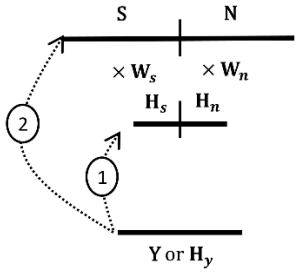

FIGURE 3. The proposed joint two-step mapping of NMF-DNN. The DNN is first learned to map the noisy magnitude (or the noisy activation coefficient) to the speech and noise activation coefficients as the pre-training step, and then, to the main magnitude spectra.

IV. THE PROPOSED NONLINEAR SPARSE DEEP MODEL

In this paper, we propose the joint DAE-DNN model shown in Fig. 2 to capture the structured prominent patterns through the nonlinear mapping in the DAE's encoder and decoder layers. In the proposed model, we suggest exploiting the sparse nonlinear features as powerful prior knowledge for the deep network to address the problems of DNNs such as local minima and a large amount of training data. This proposed method is much different from the conventional DNN that only concentrates on the nonlinear mapping between the mixture and the target clean speech without considering the harmonic and sparse structure of the signal. In other words, we have designed an appropriate structure of the DAEs and DNN to extract the sparse nonlinear properties and inject them into the deep network.

A. EXTRACTING THE SPECTRAL INTRINSIC STRUCTURE OF SIGNAL (FEATURE EXTRACTION)

The exploited DAEs are suitable models for the extraction of data structure because of the dimension reduction in latent representation to acquire informative features and also for capturing the structural patterns in bases or deep layers. The use of encoded representations as DNN input or output features (rather than the signal itself) may lead to the reduction of training complexity. Also, injecting the basic knowledge into the network is a more appropriate initialization for DNN and leads to learning more valuable characteristics of the signal. While most feature extraction methods rely on human knowledge and tuning, this approach is simple, effective, and able to extract useful speech and noise information without human intervention. In the linear NMF-DNN models, the features are extracted by NMF, while they are found by DAE in the suggested nonlinear DAE-DNN models.

- ***DAE***: dimension reduction in DAE is obtained either by reducing the number of hidden nodes or by imposing a sparsity constraint in case of having a large number of nodes [53]. Hence, the network is forced to learn a compact representation of the input data in the bottleneck layer, and thus, the structure of the data is discovered. In this work, in addition to reducing the number of nodes, we also apply a sparsity constraint to achieve more compact representations. In DAEs, the first part acts as an ''encoder'' network that converts the high-dimensional input data into a low-dimensional encoded representation, and the second part is a ''decoder'' network that reconstructs an approximation of the input data using the encoded representation.

- ***NMF***: by applying the NMF method individually to the magnitude spectra of the clean speech and noise, the basis matrices $\mathbf{W}_s$ and $\mathbf{W}_n$ are captured. Then, in a joint model, they are integrated into the DNN output part as an extra layer which is a deterministic layer and there are no connective weights to be trained. In a separate model, the corresponding activation matrices are set as output features, and the basis matrices are applied separately outside of the DNN. We extend the use of NMF-based features in the DNN network in both DNN input and output parts. For the input part, unlike [43], due to considering any possible correlation between speech and noise, $\mathbf{W}_s$ and $\mathbf{W}_n$ are not concatenated as $\mathbf{W}_y$. Instead, $\mathbf{W}_y$ is learned directly by performing NMF on the magnitude spectrogram of the noisy speech. Then, the corresponding activation matrix $\mathbf{H}_y$ is used as an input feature. The DNN itself should map $\mathbf{H}_y$ to $\mathbf{H}_s$ and $\mathbf{H}_n$(in the implemented separate model or to $\mathbf{S}$ and $\mathbf{N}$ in the joint model) in a nonlinear manner and separate $\mathbf{H}_s$(or $\mathbf{S}$). We also propose a two-step training in the complementary NMF and DNN models. Thus, according to Fig. 3, we first directly map the input features (noisy magnitude spectrum $\mathbf{Y}$ or $\mathbf{H}_y$, based on the input feature configuration) to the activation output layer as a pre-training step ($\mathbf{H}_s$, $\mathbf{H}_n$). Then, we jointly map the input to the main spectrum output layer ($\mathbf{S}$, $\mathbf{N}$).

B. COOPERATIVE DAE-DNN MODEL

A detailed view of the proposed cooperative DAE-DNN model which is shown in Fig. 2 is composed of nonlinear

feature extraction (shown by dashed lines), enhancement, and reconstruction parts. It is represented as one box in Fig. 1. In the feature extraction part, three DAEs are used for nonlinear dictionary learning of speech, noise, and noisy signals and for extracting the structural encoded representations (features). The noisy DAE is used when instead of the main noisy spectrum, the encoded noisy feature is applied as the DNN (called enhancer DNN) input feature for enhancement. The DAE-based structure extraction is joined with the enhancer DNN. The enhancer DNN utilizes the extracted patterns found in speech and noise DAEs bottleneck representations as its output targets and the related pre-trained decoder layers as its extra reconstruction layers. The reconstruction part also contains the Wiener filtering layer which is deterministic with no connection weights that needed to be trained. We only use the output of this layer to calculate the error and train the weights of the enhancer DNN. This is believed to lead to an improved estimation of the speech/noise encoded representations and also the main spectral signals through the nonlinear deep and hierarchical structure of the network. Therefore, both feature extraction and enhancement are carried out in a joint and nonlinear fashion which is considered as a point of strength for this cooperative approach. In this model, both speech and noise sources are estimated simultaneously by using their spectral magnitudes or their encoded sparse features as the DNN targets. Because of the complementarity of different sources in the mixed-signal, modeling all sources in one model can lead to the improvement of the separation performance [25], [61].

It should be noted that the reason for using a simple DNN in our proposed models is to provide a fair comparison with the baseline complementary NMF-DNN models [39], [40], [41], [43], and also due to the simple structure of the DNN network. However, since our proposed approach is a fundamental idea based on suggesting an appropriate cooperative model following the characteristics of speech and noise signals, and is independent of the individual DNN structure, the same strategy could be used for more complicated networks such as CNN, CRN, and LSTM to improve their results as well in future. Nevertheless, the applicability of complex networks is limited due to their more power consumption, more complexity of their hardware, and latency. Thus, since the novel fundamental idea proposed in our structure leads to good improvement despite the simple DNN network and outperforms similar strategies, we have focused on DNN in this paper.

- **The training and testing phases:**

A detailed view of the proposed cooperative DAE-DNN model is shown in Fig. 4. The model implementation consists of training and testing phases. The proposed architecture in this figure (top part) is trained individually for each noise type. We have the same approach for different noises to generate noise-specific dictionaries and joint models. Then, in the testing phase, as shown in Fig. 4 (bottom part), based on the decision block result, we will find the final results either by selecting one of the learned DAE-DNN models for seen noises or by using the suggested fusion strategy for unseen noises. As shown in Fig. 4 (top part), for the training phase of our joint two-step DAE-DNN model, we subsequently perform the following stages:

Stage *A*: the speech, noise, and noisy DAEs are individually trained with the corresponding magnitude spectra. The extracted encoded representations ($\mathbf{E}_s, \mathbf{E}_n, \mathbf{E}_y$) and the speech and noise decoders will be used in the next stages *B* and *C*.

Stage *B*: the enhancer DNN is learned to map the main noisy magnitude ($\mathbf{Y}$) or the noisy encoded feature ($\mathbf{E}_y$, the output of the noisy encoder) to the extracted encoded features of speech and noise ($\mathbf{E}_s$, $\mathbf{E}_n$, layer J). Based on the proposed step-wise training, the mapping of the joint model input to the main output layer is performed in two steps: first to the encoded output layer (circle 1, stage *B*) and then, to the main output layer (circle 2, stage *C*).

Stage *C*: stages *A* and *B* are used as pre-training for stage *C* (fine-tuning). The pre-trained enhancer DNN, the decoders, and also the Wiener-type filtering layer are integrated and form a unified framework called the joint DAE-DNN model. The decoders and the Wiener filtering layers, as the reconstruction layers, follow the primary output layer (or the encoded output layer) of the enhancer DNN to directly estimate the magnitudes of the main signals. Therefore, in this stage, the joint DAE-DNN model is trained (fine-tuned) with a new objective and cost function (circle 2). This means that the pre-trained decoder layers of speech and noise are used as initializations for the reconstruction layers of the enhancer DNN. These are updated along with the enhancer DNN layers during the joint learning (circle 2) to extract the main spectral speech and noise signals from the noisy features. Hence, in step 2 (circle 2), the integrated pre-trained parts (the decoders (stage *A*) and the trained enhancer DNN with the objective encoded features (circle 1, stage *B*)) continue to learn and incrementally adapt to the actual speech and noise targets. In the end, the final learned model and parameters are kept fixed for the testing phase. Besides these stages, a classifier DNN is also trained with the noisy mixtures to classify three training noise types. The models' configuration and setup parameters are given in the experiments Section.

In the testing phase, shown in the lower part of Fig. 4, firstly, the learned classifier DNN is used to predict the similarity percentages of a test noisy mixture magnitude ($\mathbf{Y}_{test}$) to the *N* training noise types for each frame, which are averaged as the classification rates of that mixture. Then, similar to Fig. 1, in the decision block, the seen/unseen noise category is detected. This is performed in such a way that if one of the predicted classification rates is greater than a high threshold (set to 0.9), that noisy mixture is considered as a seen mixture, and only the related detected model is used for enhancement. However, for unseen noises, we apply a fusion strategy to obtain the final result by weighted averaging the outputs of the *N*-learned DAE-DNN models (which are related to the different training noise types *i* with the total number *N*). In other words, for an input noisy

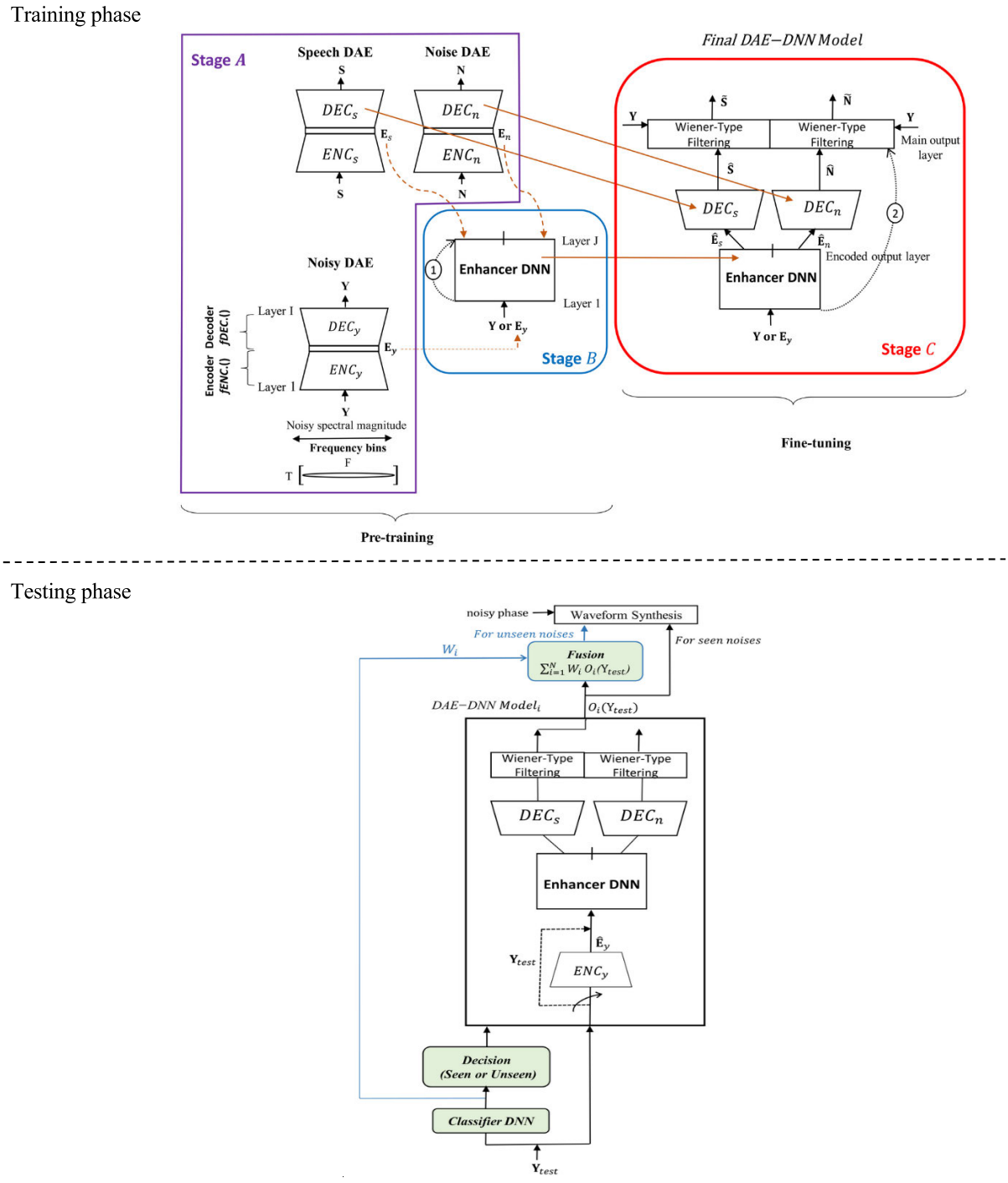

FIGURE 4. The training and testing phases of the detailed DAE-DNN architecture. In the training phase (top part), stages A, B, and C are performed in order so that the extracted encoded features of stage A are used in stage B, and then, the learned models of both stages A and B are used as pre-training for stage C. Stage C is our fine-tuned (updated) joint model. Circles 1 and 2 indicate the two-step mapping. This whole architecture is trained for each noise type (as one model box in Fig. 1), and the same approach is repeated for different noise types. In the testing phase (bottom part), after the detection of noise type in the decision block, either one (for seen noises) or a combination (for unseen noises) of the learned DAE-DNN models (the final model of stage C) via the suggested fusion strategy is used for enhancement.

speech, the outputs of different noise-specific models are weighted by the related predicted classification rates and are linearly combined as an estimate of the clean speech magnitude.

Then, in each learned DAE-DNN model (*DAE-DNN Model*$_i$), in the input-related methods, the noisy encoded feature ($\widehat{\mathbf{E}}_y$) is first estimated from the spectral magnitude of the observed noisy speech ($\mathbf{Y}_{test}$). This is carried out by

using the encoder portion of the trained noisy DAE as the input to the enhancer DNN. Otherwise, the noisy magnitude spectrum $\mathbf{Y}_{test}$ is used as the main input. The *DAE-DNN Model*$_i$ is the final learned joint model for each noise type i (stage C in the training phase). It is composed of the updated trained enhancer DNN and its extra integrated decoder and filtering layers. It is used to estimate the encoded features at the encoded output layer as well as the objective speech and noise magnitudes at the main output layer from the main noisy magnitude (or from the noisy encoded feature, based on the different configurations of the input feature). Finally, the time waveform is obtained by using the ISTFT and the noisy phase. In the unjointed model (separate model), the speech and noise encoded features are only estimated by the enhancer DNN of stage B. Then, the main signals are separately approximated outside of the network by using the fixed pre-trained decoders of stage A and Wiener filtering.

The mapping functions of the noisy DAE (f) and the enhancer DNN (g) in the case of the encoded features mapping are as below:

$$\begin{aligned} &\mathbf{h}_i = f(\mathbf{h}_{i-1}) = \sigma(\mathbf{W}_i\mathbf{h}_{i-1} + \mathbf{b}_i) \quad 1 \le i \le I, \\ &\mathbf{h}_0 = \mathbf{Y}, \mathbf{h}_I = \widehat{\mathbf{Y}} = f_{\mathrm{DEC}}(f_{ENC}(\mathbf{Y})) = f_{\mathrm{DEC}}(\mathbf{E}_y) \end{aligned} \tag{4}$$

$$\begin{aligned} &\mathbf{X}_j = g\left(\mathbf{X}_{j-1}\right) = \sigma\left(\mathbf{W}_j^*\mathbf{X}_{j-1} + \mathbf{b}_j^*\right) \quad 1 \le j \le J, \\ &\mathbf{X}_0 = \mathbf{E}_y, \mathbf{X}_J = \left[\widehat{\mathbf{E}}_s\widehat{\mathbf{E}}_n\right] \end{aligned} \tag{5}$$

where $\mathbf{W}_i$, $\mathbf{b}_i$ and $\mathbf{W}_j^*$, $\mathbf{b}_j^*$ are the weights and biases of the DAE and DNN networks, respectively. The spectral magnitudes $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$ are the input and output of the noisy DAE, respectively. They can be replaced by $\mathbf{S}$ and $\widehat{\mathbf{S}}$, or $\mathbf{N}$ and $\widehat{\mathbf{N}}$, for the clean and noise DAEs, respectively. $\mathbf{E}_y$ represents the bottleneck feature (encoded representation) of the noisy DAE. The noisy decoder is discarded once the noisy DAE is trained. $\widehat{\mathbf{E}}_s$ and $\widehat{\mathbf{E}}_n$ represent the enhancer DNN-estimated encoded features of speech and noise, respectively. f_{ENC} and f_{DEC} are the mapping functions of the encoder and decoder parts of DAEs, respectively. $\sigma(\cdot)$ is the nonlinear activation function, i and j are the layers indices of DAEs and DNN, and I and J are their total number of layers, respectively.

DAEs and DNN are trained and the parameters of the networks, e.g., $\mathbf{W}_i$, $\mathbf{b}_i$, $\mathbf{W}_j^*$, and $\mathbf{b}_j^*$, are computed by using the back-propagation algorithm and a mean-squared error (MSE)-based cost function to increase the similarity and minimize the distance between the predicted output and the corresponding target. The sparsity constraint is applied to the DAEs' hidden layers and added to the cost function as:

$$L_{Cost}(DAE\_s) = \left\|\mathbf{S} - \widehat{\mathbf{S}}\right\|_2^2 + \|\mathbf{E}_s\|_1 \tag{6}$$

where $||\cdot||_1$ denotes the l_1-norm as an approximation of l_0-norm which is NP-hard. Here $L_{Cost}(DAE\_s)$ refers to the speech DAE cost function. By dividing the estimated output of the encoded output layer of the enhancer DNN into two parts for speech and noise, and applying the related decoders, the speech and noise are estimated as follows:

$$\widehat{\mathbf{S}} = \mathrm{DEC}_s(\widehat{\mathbf{E}}_s), \widehat{\mathbf{N}} = \mathrm{DEC}_n(\widehat{\mathbf{E}}_n) \tag{7}$$

Then we have:

$$\mathbf{Y} \simeq \widehat{\mathbf{S}} + \widehat{\mathbf{N}} = \mathrm{DEC}_s(\widehat{\mathbf{E}}_s) + \mathrm{DEC}_n(\widehat{\mathbf{E}}_n) \tag{8}$$

After making a Wiener-type filter from $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{N}}$ similar to the IRM and applying it to the noisy magnitude spectrum, we obtain the final estimations of the speech and noise magnitudes as below:

$$\begin{aligned} \tilde{\mathbf{S}} &= \frac{\widehat{\mathbf{S}}^2}{\widehat{\mathbf{S}}^2 + \widehat{\mathbf{N}}^2} \otimes \mathbf{Y} \\ \tilde{\mathbf{N}} &= \frac{\widehat{\mathbf{N}}^2}{\widehat{\mathbf{S}}^2 + \widehat{\mathbf{N}}^2} \otimes \mathbf{Y} \end{aligned} \tag{9}$$

where the division and multiplication operators are element-wise.

- **Different input-to-output mappings**

In this paper, we call joint one-step and two-step methods as *J1* and *J2,* respectively. *EN* and *DE* indicate the encoder and decoder, respectively. In *DNN-DE-Sep*, the enhancer DNN explicitly maps the noisy spectral magnitude to the speech and noise-encoded features ($Y \rightarrow E_sE_n$). Then, the speech and noise signals are manually reconstructed by using the corresponding fixed trained decoders outside of the DNN (enhancer DNN). Finally, the Wiener-type gains in (9) calculated from the reconstructed speech and noise are applied to smooth the results. In *DNN-DE-J1*, the DNN, decoders, and Wiener-type filters are realized as a joint model, while the effect of DAE-encoded features is implicit. This means that the encoded features are not directly used as the DNN targets and the output of the DNN encoded output layer can be considered as the encoded variables estimates, and the main target signals are directly reconstructed by the DNN in one step through the integrated decoders ($Y \rightarrow SN$). In *DNN-DE-J2*, we perform mapping the noisy speech to the clean and noise spectral magnitudes in two steps: first explicitly mapping the input to the encoded features ($Y \rightarrow E_sE_n$), and then mapping the input to the main targets ($Y \rightarrow SN$). The first step acts as a pre-training for the second step.

EN-DNN-DE-Sep, EN-DNN-DE-J1, and *EN-DNN-DE-J2* approaches are the same as before, but just instead of the noisy spectral magnitude, the encoded representation of the noisy DAE is used as the DNN input feature. In *EN-DNN-DE-Sep*, the DNN maps the noisy encoded feature to the clean and noise-encoded features ($E_y \rightarrow E_sE_n$), and like the equivalent output-related method, the reconstruction of the main speech and noise signals is performed manually by using the corresponding decoders. In *EN-DNN-DE-J1*, the DNN directly maps the noisy encoded features to the clean and noise spectral magnitudes in one step ($E_y \rightarrow SN$) whereas, in *EN-DNN-DE-J2*, it is done gradually in two steps ($E_y \rightarrow E_sE_n$, $E_y \rightarrow SN$). To investigate the effect of the DNN input and output features, in three methods of *DNN-DE-Sep*, *DNN-DE-J1*, and *DNN-DE-J2*, we evaluate the output features. In three methods of *EN-DNN-DE-Sep*, *EN-DNN-DE-J1*, and *EN-DNN-DE-J2*, the input feature is also evaluated. All the above six DAE-based methods are

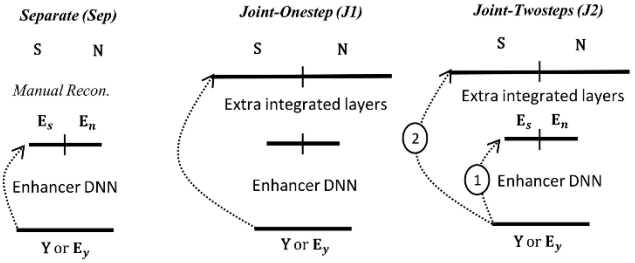

FIGURE 5. Different input-output mappings of the enhancer DNN in the proposed model. E characters are the DAEs encoded features.

compared with their linear NMF counterparts, namely *DNN-NMF-Sep, DNN-NMF-J1, DNN-NMF-J2* (related to output part), and *NMF-DNN-NMF-Sep, NMF-DNN-NMF-J1, NMF-DNN-NMF-J2* (related to input part)), resulting in 12 implemented methods. The overall schematic of the different mapping configurations (*Sep*, *J1,* and *J2*) of the enhancer DNN input-output is shown in Fig. 5. In methods with the "*Sep*" label, the spectral reconstruction part is performed separately outside of the network compared to other methods in which it is integrated and jointly optimized.

The DNN-NMF-Sep, DNN-NMF-J1, and NMF-DNN-NMF-Sep-J1 methods are the same as the main ideas of [39], [41], and [43], respectively, but with our setup parameters and dataset. The other eight mentioned methods are the proposed ones. In DNN-NMF-Sep [39] as described earlier in Section I, the NMF activation coefficients of speech and noise are estimated by the DNN from the noisy magnitude. Then, in a separate process outside of the DNN, the main spectral signals are approximated by the linear product of the estimated coefficients with the pre-learned NMF basis matrices. In DNN-NMF-J1 [41], which is the joint model of [39], the NMF basis matrices and a Wiener-type filtering layer are integrated into the DNN to reconstruct the objective signals. However, in this method, the structural activation features are not directly applied as DNN targets and cannot directly participate in the learning process. The NMF-DNN-NMF-Sep [43] is also similar to [39], but with the input of the noisy activation coefficient instead of the noisy magnitude spectrum.

V. EXPERIMENTS

For evaluation of the proposed and baseline methods, we compare the performances of the proposed DAE-DNN approaches with their equivalent linear NMF-DNN methods [39], [41], [43] in speech enhancement. Furthermore, the baseline NMF [60], DNN [25], LSTM with the IRM target (LSTM-Mask) [15], [16], CRN with the spectral magnitude target (CRN-Mag) [17], RVAE-VEM [22] and SN-Net [23] approaches are considered for comparisons. In the baseline DNN method [25] and CRN-Mag [17], the mixed-signal is directly mapped to the objective signals. While, in the LSTM-Mask method, it is mapped to the IRM mask values which are multiplied by the mixed-signal to estimate the objective signals. We first implement [39] and [41], which are the separate and joint complementary models of NMF-DNN, respectively (*DNN-NMF-Sep*, *DNN-NMF-J1*) where the noisy magnitude spectrum is the input feature. Also similar to [43], we implement "*NMF-DNN-NMF-Sep*, *NMF-DNN-NMF-J1"* in which the NMF activation coefficients are used as the input features. Then, our proposed joint two-step training is performed by setting the NMF activation coefficients as primary targets, as well as the main speech and noise signals as secondary targets (*DNN-NMF-J2* or *NMF-DNN-NMF-J2,* as shown in Fig. 3).

Moreover, in addition to the two-step process, for comparison purposes, a multi-target one-step approach is implemented in the DAE-DNN model in which the two losses (MSEs) for the actual targets and the intermediate encoded features are combined (*DNN-DE-J1-combined loss*).

A. DATASET AND SETUP PARAMETERS

The TIMIT corpus [62], containing 630 different speakers is used as the speech dataset and the NOISEX-92 corpus [63], including 15 common noise types of quasi-and non-stationary noises, as well as the *Freesound* data [64] are used as noise datasets. Our training set is formed by randomly selecting 200 clean speech utterances from the TIMIT training sector and randomly adding *babble*, *factory*, and *machinegun* noises from the NOISEX-92 at SNRs from −5 to 20 dB with steps of 5 dB, resulting in 3600 (200 signals×3 noises×6 SNRs) noisy speech and pairs of speech and noise utterances in the training set. The same training set with a 10% validation split is used for evaluating all the proposed and comparison methods. We also conducted the experiments with six types of training noises (*babble*, *factory*, *machinegun*, *buccaneer2*, *leopard*, *destroyerengine*), but the results did not improve over using only three types despite using more noise types. This could improve our claim about the good generalization of our proposed approach due to its cooperative and fusion strategy. However, using a very large number of various types of noises might eventually help improve this generalization since that would lead to a weak unseen condition. In fact, most properties of the noises would have naturally been seen in the training phase in this condition. Thus, we have not focused our paper on this case. We have formed the testing set by randomly choosing 60 clean speech utterances from the TIMIT testing sector. Then, we added the noises of the training set as seen noises, and also the real-world recorded *factorymachine* and *windshieldrain* noises from the *Freesound* data at SNRs of -5 to 10 dB as unseen noises not seen during the training phase. We added noise at different SNRs following the procedure in [65]. N in Fig. 4 is 3 due to our three training noise types.

For setting the speech and noise encoded features as DNN output labels (targets) and matching the speech and noise frame numbers, the clean speech utterance is repeated to match with the noise and noisy frame numbers.

The magnitude spectra are extracted by applying 512-point STFT to the waveforms sampled at $16kHz$ and framed with 512-point (32-ms) frame length. The frames are generated by applying a 512-point (32ms) Hamming window and using a 128-point (8ms) shift size. By cutting off the symmetrical parts of the STFT coefficients, the dimension of the magnitude spectra is $257\times$ *number of time-frames*.

B. CONFIGURATIONS

- NMF configurations

The NMF speech and noise basis numbers are empirically set to be 100. Then, according to the magnitude spectra dimensions, the size of the speech and noise basis matrices is 257×100 (frequency bins$\times$basis number). The maximum number of NMF iterations is set at 50. We acquire the general Wn for all training noises by applying the NMF to the concatenated spectral magnitudes of different noises.

- Networks configurations

The enhancer DNN is composed of four hidden layers with 1024 neurons, and Leaky rectified linear units (LReLU) [66] with α =0.1 ($f(x) = max(\alpha x, x)$) as the activation function of the hidden layers to overcome the ''dying ReLu'' problem. The activation function of the output layer is linear for the main spectral targets. For the output encoded targets (pre-training step), ReLU ($f(x) = max(0, x)$) is used as the activation function due to the nonnegativity of the activation coefficients.

The classifier DNN contains two hidden layers of 1024 neurons with ReLU functions and one output layer of 3 neurons with the softmax activation function for the classification of three noise types. The softmax output is a probability distribution in the range of [0,1] with a total sum of 1 which is usually used for a multi-classification problem. The batch normalization is applied to hidden layers for faster training convergence.

The LSTM-Mask model includes two LSTM layers and a fully connected (FC) layer having 1024 neurons with LReLU activations, and a fully connected output layer with 257 units for mask estimation.

The DAE bottleneck layer node number is empirically set to 100 based on the NMF rank. The size of the clean and noisy DAEs is 257-1024-512-100-512-1024-257 and the noise DAE is 257-512-512-100-512-512-257. The activation functions of the DAE layers are similar to the enhancer DNN, and the sparsity constraints are also applied to the hidden layers. The encoded layer and the main output layer of the enhancer DNN have $100 \times 2 = 200$ and 257×2=514 nodes, respectively. The DAEs, enhancer DNN, and LSTM use the MSE, and the classifier DNN uses the cross-entropy loss function. Also, they are trained by the Adam optimizer [67] with an initial learning rate of 0.001 and a maximum epoch of 100. In the two-step processing, the number of epochs for each step is 100. Also, the early stopping based on the minimum validation loss is used in all models to avoid overfitting.

TABLE 1. Results of noise classification.

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

The structure and the parameters of CRN, RVAE-VEM, and SN-Net are set in the same way as in [17], [22], and [23], respectively, except for the dataset and the maximum number of epochs, which is set to 100 according to our settings. However, for a fair comparison, in the implementation of the SN-Net [23], we omit the merging module and the phase spectrum as additional input since we only use the magnitude spectrum in our methods. The merging module which is used in [23] as a final module combines the outputs of the speech and noise branches in the time domain. Since it is not used in any of the methods in this paper, and could also be applied in all of them, we do not apply it here for a fair comparison. The CRN model [17] is composed of CNN encoder-decoder and LSTM layers.

C. EVALUATION METRICS

Three objective metrics of perceptual evaluation of speech quality (PESQ) [68], short-time objective intelligibility (STOI) [69], and frequency-weighted segmental SNR (SNR_{fw}) [70], [71] are used for the enhancement performance evaluation. PESQ reflects speech quality [68], whereas the STOI metric indicates speech intelligibility. The range of the PESQ score is [-0.5, 4.5] and is calculated by comparing the enhanced speech with the corresponding clean speech. The STOI score is in the range of [0, 1] and measures the correlation of the enhanced and clean speech short-time temporal envelopes. The higher values of PESQ and STOI show better quality and intelligibility, respectively. SNR_{fw} is computed using a weighting function based on a dynamic frequency weighted signal-to-noise ratio and measures a generalized short-time performance. The metrics improvement is also computed which is the difference between the metric score of the enhanced and clean speech with the score of the noisy and clean speech (such as $gPESQ = PESQ(\tilde{s}, s) - PESQ(y, s)$).

D. EXPERIMENTAL RESULTS

The results are presented in the seen and unseen noise conditions on the testing set. Firstly, the accuracy results of the classifier DNN for the classification of three noises are illustrated in Table 1. It indicates that the accuracy is sufficient for noise classification. The unseen *factorymachine* and *windshieldrain* noises have been detected by the classifier DNN with the approximate classification rates of (0.64, 0.22, 0.14) and (0.80, 0.18, 0.02), respectively. Therefore, the above rates have been used to weigh the outputs of the models related to the three classes in speech enhancement.

TABLE 2. The average results of previous and proposed methods over different seen noise types and input SNRs.

No.	Models	PESQ	STOI	SNR_{fw}	No.	Models	PESQ	STOI	SNR_{fw}	
-	Noisy	2.10	0.71	9.30	-	Noisy	2.10	0.71	9.30	Previous
1	NMF [60]	2.42	0.73	6.69	4	CRN-Mag [17]	2.74	0.84	10.47	
2	DNN [25]	2.66	0.80	7.58	5	RVAE-VEM [22]	2.41	0.76	9.74	
3	LSTM-Mask [15], [16]	2.85	0.86	13.19	6	SN-Net [23]	2.47	0.76	9.83	
	Related to the DNN output part					Related to the DNN input/output part				
7	DNN-NMF-Sep [39]	2.89	0.84	11.82	13	NMF-DNN-NMF-Sep [43]	2.89	0.84	11.97	
8	DNN-NMF-J1 [41]	3.07	0.86	13.45	14	NMF-DNN-NMF-J1 [43]	2.99	0.85	13.39	
9	DNN-NMF-J2	3.16	0.88	15.04	15	NMF-DNN-NMF-J2	3.08	0.87	14.93	Proposed
10	DNN-DE-Sep	3.08	0.87	14.61	16	EN-DNN-DE-Sep	3.22	0.88	14.17	
11	DNN-DE-J1	3.18	0.88	15.17	17	EN-DNN-DE-J1	3.20	0.88	15.48	
12	DNN-DE-J2	**3.22**	**0.89**	**15.59**	18	EN-DNN-DE-J2	**3.23**	**0.89**	**15.68**	

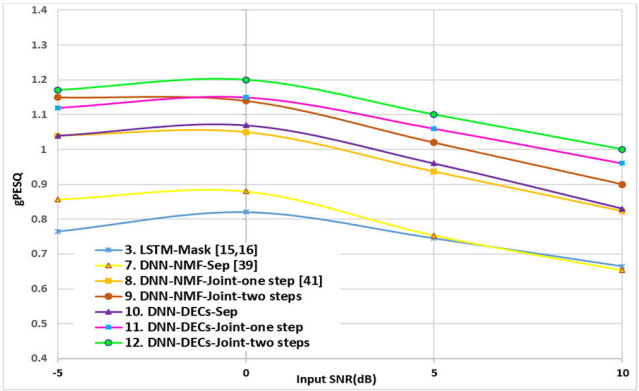

FIGURE 6. The average improvement (gain) of the PESQ score for various previous and proposed methods at different input SNRs.

The models related to the DNN output part are *DNN-NMF-Sep* [39], *DNN-NMF-J1* [41], *DNN-NMF-J2* (related to the NMF coefficients) and *DNN-DE-Sep, DNN-DE-J1, DNN-DE-J2* (related to the DAE-encoded features). Those also related to the input part are *NMF-DNN-NMF-Sep, NMF-DNN-NMF-J1* [43], *NMF-DNN-NMF-J2* (related to the NMF coefficients), and *EN-DNN-DE-Sep*, *EN-DNN-DE-J1, EN-DNN-DE-J2* (related to the DAE-encoded features). The overall performance of the various proposed and baseline methods in the form of the average metrics results over different seen noise types and input SNRs are given in Table 2. Bold scores indicate the method with the best result. The previous and the proposed methods are also distinguished in Table 2. It should be noted that all noise types are unseen for RVAE-VEM [22] method as it does not consider the noise information in the training phase. Nevertheless, we show its average results for our three seen noises in Table 2 only for comparison.

To provide a better comparison, the average improvements of the PESQ score (gPESQ) over various seen noise types are reported in Fig. 6 at different input SNRs and for the output-related methods and also LSTM-Mask as an example.

Also, the average results of the multi-target/combined loss approach in the DAE-DNN model (*DNN-DE-J1-combinedloss*) on seen noises are shown in Table 3 in

TABLE 3. The average results of the multi-target/combined loss approach and the proposed one and two-step DAE-DNN methods.

Models	PESQ	STOI	SNR_{fw}
DNN-DE-J1-combined loss	3.11	0.87	13.94
DNN-DE-J1	3.18	0.88	15.17
DNN-DE-J2	**3.22**	**0.89**	**15.59**

comparison with the proposed one and two-step approaches. Its configuration is the same as *DNN-DE-J1*. As expected, since the types of the two targets (the actual signal and the encoded feature) are different, simultaneously predicting them and updating the model weights based on a combined two-loss function is subject to error. Therefore, it shows no better performance than the proposed two-step approach which is step-wise learning based on the respective target and loss function at each step. Also, as shown in Table 3, the proposed one-step approach (*DNN-DE-J1*) outperforms "*DNN-DE-J1-combinedloss*". We believe that directly estimating only one type of target which is based on one type of loss function in the non-combined case (*DNN-DE-J1*), could be the reason for this result.

For evaluation of the generalization abilities in unseen noise conditions, the average results of various methods over different unseen noise types and input SNRs are shown in Table 4. Bold scores indicate the method with the best result.

To investigate the performance of speech enhancement in more types of unseen noises and draw better conclusions, the average PESQ results of two other unseen noises, *restaurant,* and *street*, from the Aurora-2 database [65] over different input SNRs are reported in Fig. 7 for our proposed *DNN-DE-J2* and *EN-DNN-DE-J2* methods as well as the baseline LSTM-Mask method [15], [16]. The classification rates of these new unseen noises are estimated at about (0.11, 0.71, 0.18) and (0.18, 0.28, 0.54), respectively by the classifier DNN.

Also, to compare the properties of different seen and unseen noises, the spectrograms of two of the seen and unseen noises are shown in Fig. 8 as examples, and the spectrograms

TABLE 4. The average results OF previous and proposed methods over different unseen noise types and input SNRs.

No.	Models	PESQ	STOI	SNR_{fw}	No.	Models	PESQ	STOI	SNR_{fw}	
-	Noisy	2.08	0.74	6.11	-	Noisy	2.08	0.74	6.11	Previous
1	NMF [60]	2.15	0.71	4.53	4	CRN-Mag [17]	2.24	0.78	7.74	
2	DNN [25]	2.24	0.76	6.85	5	RVAE-VEM [22]	2.31	0.77	7.84	
3	LSTM-Mask [15], [16]	2.32	0.78	8.71	6	SN-Net [23]	2.23	0.76	7.14	
	Related to the DNN output part					Related to the DNN input/output part				
7	DNN-NMF-Sep [39]	2.23	0.77	7.67	13	NMF-DNN-NMF-Sep [43]	2.18	0.76	7.41	
8	DNN-NMF-J1 [41]	2.30	0.78	8.35	14	NMF-DNN-NMF-J1 [43]	2.27	0.77	7.64	
9	DNN-NMF-J2	2.37	0.79	8.38	15	NMF-DNN-NMF-J2	2.36	0.78	8.67	
10	DNN-DE-Sep	2.35	0.78	8.98	16	EN-DNN-DE-Sep	2.36	0.78	8.96	Proposed
11	DNN-DE-J1	2.40	0.79	9.63	17	EN-DNN-DE-J1	2.39	0.79	**10.03**	
12	DNN-DE-J2	**2.43**	**0.80**	**10.24**	18	EN-DNN-DE-J2	**2.42**	**0.80**	9.44	

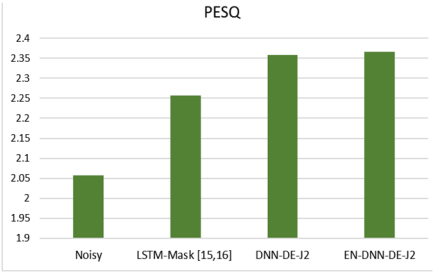

FIGURE 7. The average PESQ results of restaurant and street unseen noises over different input SNRs.

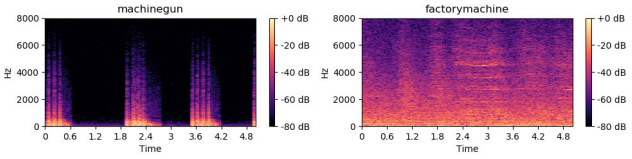

FIGURE 8. The spectrograms of seen machinegun and unseen factorymachine noises.

of other used noises are given in Appendix. As could be observed, they have different properties and structures.

E. DISCUSSION

As can be seen in Table 2, the proposed DAE-DNN models in the output category (row numbers 10, 11, and 12) outperform their equivalent NMF-DNN models (row numbers 7, 8, and 9, respectively), and also the baseline NMF [60], DNN [25], LSTM-Mask [15], [16], CRN-Mag [17], RVAE-VEM [22], and SN-Net [23] methods. This is also observed for the input category methods; i.e., row numbers 16, 17, and 18 from 13, 14, and 15, respectively. This demonstrates the power of the prior knowledge extracted by DAEs compared to NMF, and also better learning of DNN in more structured features. The proposed DAE-based feature extraction has more capacity for learning and extracting the structural patterns compared to NMF-based methods because of the deep and nonlinear structure. Besides, in NMF-DNN models, the nonlinear DNN is forced to act as a linear NMF decomposition and map the mixture into the linear activation coefficients; whereas in our DAE-DNN models, DNN performs similar to an encoder network and estimates the objective encoded features from the mixture in a nonlinear strategy. According to Table 2, each approach also shows a rather high PESQ improvement and some improvements in terms of STOI and SNR_{fw} compared to the related previous one in each colored section (light-colored are related to the NMF-based methods and dark-colored are related to the DAE-based methods), and also baselines. In other words, within the NMF or DAE-based models, whether in the output or input/output category, the "joint-optimized-two-step (J2)" models (rows 12, 18) outperform the "joint-optimized-one step (J1)" models (rows 11, 17, respectively), and subsequently, the "one-step" from the "separate reconstruction (Sep)" as primary models (rows 10, 16, respectively).

As also shown in Fig. 6, the gPESQ values of our DAE-DNN models (No. 10, 11, 12) are higher than the equivalent NMF-DNN models (No. 7, 8, 9, respectively) and also the LSTM-Mask method (No. 3). This can be attributed to the joint effects of the nonlinear feature extraction by DAEs and nonlinear enhancement by DNN, i.e., better extraction of structures by the DAEs compared to the NMF and better learning and enhancement by the DNN with the cooperation of the extracted structural features. However, in some cases, the improvement results of the STOI and SNR_{fw} scores are slightly different. Also, according to Table 2, the performance of each method in the input/output category compared to its equivalent in the output one shows that using the encoded features as input features in addition to output ones does not provide much performance improvement. This could be due to the separate use of the encoder in the DNN input, i.e., in the form of two consecutive models.

According to Table 4, in the unseen conditions, not surprisingly, the overall performances of all methods decrease compared to the seen conditions. However, we have a relatively

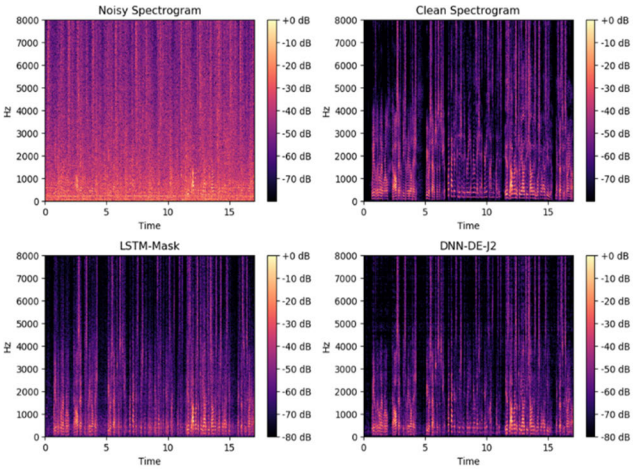

FIGURE 9. Top: The magnitude spectra of the noisy speech with factory noise at -5dB SNR (left) and the clean speech (right); Bottom: the enhanced speech by LSTM-Mask (left) and by DNN-DE-J2 (right).

growing trend in the performance of each method compared to the related previous one in most cases. The DAE-based methods present more improvement compared to the corresponding NMF-based ones and within them, in most cases, "two-steps (J2)" produce the best results. Suggesting the fusion strategy in the unseen conditions is another reason for improving the generalization of the proposed methods in these cases although a limited number of noises have been used in the training phase. Also, Fig. 7 demonstrates that the trend of the results for other unseen noises is similar to those in Table 4, and thus, shows the effectiveness of the suggested classifier-fusion strategy. Fig. 8, which shows spectrograms of a seen and an unseen noise, proves the claim that the properties of the selected unseen noises are different enough to generalize the results of the proposed fusion approach. In fact, the main point is that these unseen noises have various time-frequency properties much different from seen ones, and when the proposed methods work on these randomly selected non-stationary noises, they could work on other ones too.

Finally, the magnitude spectra of the enhanced speech by the proposed two-step DAE-DNN method (*DNN-DE-J2)* and LSTM-Mask are shown in Fig. 9, as examples. As it clearly shows, *DNN-DE-J2* preserves more speech structures (harmonics) and can eliminate more noise components compared to LSTM-Mask. This is mainly due to the power of DAE in structure extraction and its hierarchical joint integration with DNN.

In summary, some of the main reasons for the superiority of the proposed methods over the compared methods in addition to using the NCF approach and noise-specific models, are as follows:

- Compared to NMF [60], DNN [25], LSTM-Mask [15], [16], and CRN-Mag [17], we use a joint cooperative model of DAEs for structure extraction and DNN for enhancement. In DNN [25], LSTM-Mask [15], [16], CRN-Mag [17], RVAE-VEM [22], and SN-Net [23], the traditional spectral features are used and the sparsity and the intrinsic structures of the signal are not considered in the learning process. However, in our methods, such features and nonlinear dictionaries as the signal harmonic structures are extracted in a pre-training stage and incorporated into the network architecture and also as structural features in another training stage. Moreover, in our two-step approach, mapping of the input to output features is hierarchical such that learning the structural encoded features is performed first. Then, in the second step of training (fine-tuning), the learned network, along with its integrated structural reconstruction layers, is re-trained with new objectives to learn the actual separation signals.

- Compared with RVAE-VEM [22], we use a deep generative noise model instead of a linear NMF noise model. Also, unlike [22], which is an on-the-fly enhancement, our work is a supervised approach. It means that in [22], the noise information is not considered in the training phase, and the estimation of the noise model parameters in different noise conditions is only carried out on the fly at the test time from the observed noisy speech. This is a difficult task in the absence of any observation of the noise properties in advance, due to the variational properties of different noises. By contrast, in our model, characteristics of some noises, which may also be shared among different noises, are seen in the training phase. It should also be noted that the concept of fine-tuning in [22] is different from our work, in that it is applied in the testing phase such that the parameters of the clean encoder are fine-tuned with the test noisy speech, and the enhancement is performed via the VEM algorithm. While, it means retraining (refitting) the pre-trained parts with the training noisy data and the new training targets in the last hierarchy in our work.

- In [23], unlike our approach, the whole architecture is trained once with the noisy signal. Hence, it makes the training and mapping difficult, and the structures of speech and noise decoders cannot be extracted well. This may be the reason for the lower enhancement results obtained from the implementation of this method, especially under our limited data. Also, [23] does not involve a noisy encoder so that the noisy encoded feature can be obtained first, and then, the speech and noise encoded features can be extracted from it in the following layers of the model. In a way, in the two encoders in [23], an effort is made to separately extract the speech and noise encoded features from the noisy signal at once.

- Compared with *DNN-NMF-Sep/J1* [39], [41] or *NMF-DNN-NMF-Sep* [43], we use the nonlinear equivalent of NMF by DAE (*DNN-DE-Sep* or *EN-DNN-DE-Sep*), one-step joint learning (*DNN-DE-J1* or *EN-DNN-DE-J1*) and also two-step mapping (*DNN-DE-J2* or *EN-DNN-DE-J2*) in which the encoded structural features implicitly affect learning by directly applying them as the primary targets of the DNN; while the main signals are also jointly optimized as the secondary ones. The two-step mapping in the complementary DNN-NMF models is also performed (*DNN-NMF-J2* or *NMF-DNN-NMF-J2*).

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed jointly-optimized cooperative DAE-DNN models for speech enhancement to jointly take advantage of DNN and DAE's capabilities. DAEs are used for nonlinear feature extraction and dictionary learning of noise, clean, and noisy speech signals. Then, to exploit the structural patterns of signals in the separation, the learned nonlinear dictionaries (decoders) of speech and noise are integrated into the DNN as extra layers. Also in some cases, the extracted encoded features are fed to the DNN as input (the noisy encoded feature) and target output (the clean and noise encoded features) to make use of their nonlinear mapping also useful for enhancement. Furthermore, the DNN mapping for converting the noisy signal into speech and noise signals is performed in three categories named "Separate (Sep)", "Joint-one-step (J1)" and "Joint-two-step (J2)". The "J2" methods show the best performance due to their original mapping to the encoded features, as a pre-training step, and the final mapping to the main targets, as a fine-tuning step. Also, we proposed the use of DNN-based noise classification and fusion strategy which led to deploying the proper noise-specific models and improved the generalization in the noise reduction process. According to the experimental results, our proposed models gave a considerable improvement in speech separation and outperformed the conventional methods as well as the complementary NMF and DNN models, CRN, and LSTM-based enhancement in earlier studies.

For future works, we propose further extensions of the DAEs and enhancer DNN models to such more powerful networks as CNNs and LSTM.

APPENDIX

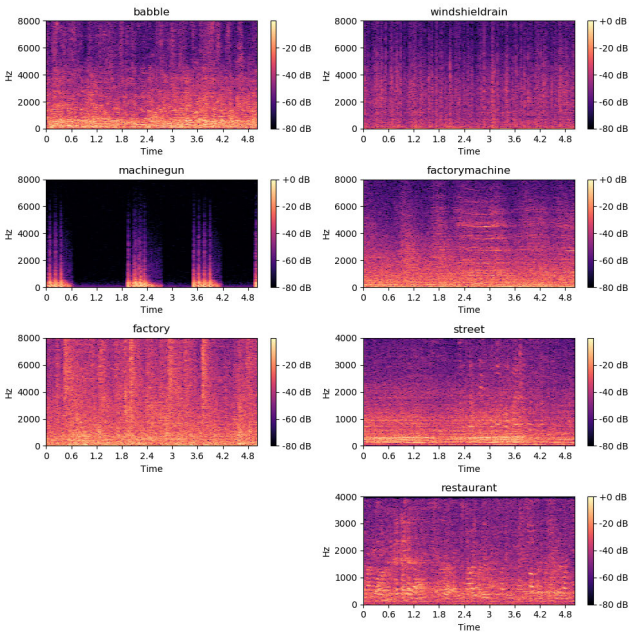

FIGURE 10. The spectrograms of all seen (left) and unseen noises (right).

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] P. Scalart and J. V. Filho, "Speech enhancement based on *a priori* signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 629–632.

[3] P. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.

[4] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 764–773, May 2006.

[5] 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. ASSP-32, no. 6, pp. 1109–1122, Dec. 1984.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[7] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4029–4032.

[8] D. D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, no. 2000, Jan. 2000, pp. 556–562.

[9] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 6, pp. 1698–1712, Aug. 2012.

[10] Y. He, G. Sun, and J. Han, "Optimization of learned dictionary for sparse coding in speech processing," *Neurocomputing*, vol. 173, pp. 471–482, Jan. 2016.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[12] Y. Wang, "Supervised speech separation using deep neural networks," Ph.D. dissertation, Dept. Comput. Sci. Eng., Ohio State Univ., Columbus, OH, USA, 2015.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, Aug. 2013, pp. 436–440.

[14] S.-S. Wang, H.-T. Hwang, Y.-H. Lai, Y. Tsao, X. Lu, H.-M. Wang, and B. Su, "Improving denoising auto-encoder based speech enhancement with the speech parameter generation algorithm," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2015, pp. 365–369.

[15] 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 J. Adv. Signal Process.*, vol. 2020, no. 1, pp. 1–26, Dec. 2020.

[16] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, p. 4705, Jun. 2017.

[17] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, Sep. 2018, pp. 3229–3233.

[18] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6865–6869.

[19] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4390–4394.

[20] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proc. Interspeech*, Oct. 2020, pp. 2477–2481.

[21] S. Routray and Q. Mao, "Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network," *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101270.

[22] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 371–375.

[23] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 14549–14557.
[24] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7867–7871.
[25] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.
[26] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3734–3738.
[27] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7092–7096.
[28] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.
[29] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7857–7861.
[30] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[31] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5220–5224.
[32] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[33] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.
[34] S. Abdullah, M. Zamani, and A. Demosthenous, "Towards more efficient DNN-based speech enhancement using quantized correlation mask," *IEEE Access*, vol. 9, pp. 24350–24362, 2021.
[35] D. Sowjanya, S. Sivapatham, D. A. Kar, and V. Mladenovic, "Mask estimation using phase information and inter-channel correlation for speech enhancement," *Circuits, Syst. Signal Process.*, vol. 41, pp. 1–19, Jul. 2022.
[36] M. Hasannezhad, H. Yu, W.-P. Zhu, and B. Champagne, "PACDNN: A phase-aware composite deep neural network for speech enhancement," *Speech Commun.*, vol. 136, pp. 1–13, Jan. 2022.
[37] Q. Zhang, Q. Song, Z. Ni, A. Nicolson, and H. Li, "Time-frequency attention for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7852–7856.
[38] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement," *IET Signal Process.*, vol. 9, no. 7, pp. 537–545, Sep. 2015.
[39] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 229–233, Feb. 2015.
[40] S. Nie, S. Liang, H. Li, X. Zhang, Z. Yang, W. J. Liu, and L. K. Dong, "Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 469–473.
[41] S. Nie, S. Liang, W. Liu, X. Zhang, and J. Tao, "Deep learning based speech separation via NMF-style reconstructions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2043–2055, Nov. 2018.
[42] H. Li, S. Nie, X. Zhang, and H. Zhang, "Jointly optimizing activation coefficients of convolutive NMF using DNN for speech separation," in *Proc. Interspeech*, Sep. 2016, pp. 550–554.
[43] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 499–503.
[44] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2145–2149.
[45] H. Jia, W. Wang, and S. Mei, "Combining adaptive sparse NMF feature extraction and soft mask to optimize DNN for speech enhancement," *Appl. Acoust.*, vol. 171, Jan. 2021, Art. no. 107666.
[46] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6107–6111.
[47] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 892–902, Aug. 2014.
[48] D. S. Williamson, Y. Wang, and D. Wang, "A sparse representation approach for perceptual quality improvement of separated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7015–7019.
[49] D. S. Williamson, Y. Wang, and D. Wang, "A two-stage approach for improving the perceptual quality of separated speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 7084–7088.
[50] 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 Trans. Audio, Speech, Language Process.*, vol. 25, no. 9, pp. 1773–1783, Sep. 2017.
[51] D. S. Williamson, Y. Wang, and D. Wang, "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," *J. Acoust. Soc. Amer.*, vol. 138, no. 3, pp. 1399–1407, Sep. 2015.
[52] D. S. Williamson, Y. Wang, and D. Wang, "Deep neural networks for estimating speech model activations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5113–5117.
[53] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[54] H. Zhang, H. Liu, R. Song, and F. Sun, "Nonlinear dictionary learning based deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3771–3776.
[55] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 4581–4585.
[56] D. S. Williamson, "Monaural speech separation using a phase-aware deep denoising auto encoder," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.
[57] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," Sep. 2022, *arXiv:2012.08405v3*.
[58] Y. Zhang, "Modulation domain processing and speech phase spectrum in speech enhancement," Ph.D. dissertation, Dept. Comput. Sci., Missouri Univ. Columbia, New York, NY, USA, 2012.
[59] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 708–712.
[60] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Proc. 17th Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2011, pp. 1–6.
[61] S. Nie, W. Xue, S. Liang, X. Zhang, W. Liu, L. Qiao, and J. Li, "Joint optimization of recurrent networks exploiting source auto-regression for source separation," in *Proc. Interspeech*, Sep. 2015, pp. 3308–3311.
[62] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1–1.1," NASA STI/Recon, Washington, DC, USA, Tech. Rep. N, 1993, vol. 93.
[63] 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.
[64] *Sounds Data*. Accessed: Jul. 2020. [Online]. Available: https://freesound.org/
[65] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP)*, Oct. 2000, pp. 1–8.
[66] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Jun. 2011, pp. 315–323.
[67] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–42.

[68] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2001, pp. 749–752.
[69] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio. Speech. Language Process.*, vol. 19, no. 7, pp. 2125–2136, Dec. 2011.
[70] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Dec. 2008.
[71] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1978, pp. 586–590.

MATIN PASHAIAN received the B.Sc. and M.Sc. degrees in electronic engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2011 and 2013, respectively. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Amirkabir University of Technology, Tehran. Her main research interests include machine and deep learning, speech processing, enhancement, and separation.

SANAZ SEYEDIN (Senior Member, IEEE) received the B.Sc. degree in electronics engineering from the Amirkabir University of Technology, Tehran, Iran, in 2001, the M.Sc. degree in electronics engineering from the Iran University of Science and Technology, Tehran, in 2005, and the Ph.D. degree in speech recognition from the Amirkabir University of Technology, in 2010. She is currently an Assistant Professor with the Department of Electrical Engineering, Amirkabir University of Technology, teaching both undergraduate and graduate courses. Her research interests include machine learning and AI, signal processing (audio, speech, image, and biological signals), compressive sensing and sparse coding, and source separation.

SEYED MOHAMMAD AHADI received the B.Sc. and M.Sc. degrees in electronics from the Department of Electrical Engineering, Amirkabir University of Technology, and the Ph.D. degree in engineering from the University of Cambridge, Cambridge, U.K. He was a Professor of electronics with the Department of Electrical Engineering, Amirkabir University of Technology. His research interests include speech processing, including speech recognition, speech enhancement, and robustness in speech processing, image and video processing, watermarking of multimedia signals, biomedical signal processing, and the application of machine learning in signal processing.