IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 May 2024, accepted 5 July 2024, date of publication 15 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3427854

RESEARCH ARTICLE

Speech Enhancement Based on a Joint Two-Stage CRN+DNN-DEC Model and a New Constrained Phase-Sensitive Magnitude Ratio Mask

MATIN PASHAIAN AND SANAZ SEYEDIN, (Senior Member, IEEE)
Speech Processing Research Laboratory, Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran 15875-4413, Iran

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

ABSTRACT In this paper, we propose a jointly-optimized stacked-two-stage speech enhancement. In the first stage, a convolutional recurrent network (CRN)-based masking is integrated with the signal analysis (fast Fourier transform (FFT)) and resynthesis (inverse FFT (IFFT)) parts as extra joint layers (FFT-CRN-IFFT). This joint FFT-CRN-IFFT model is used to separate time domain (TD) speech and noise signals. Additionally, we propose new constrained phase-sensitive magnitude ratio masks (cPSIRMs) for speech and noise sources, which are estimated at this stage by the CRN in relation to the ultimate time-domain signals. In the second stage, a deep neural network integrated with the decoder layers of a deep autoencoder (DNN-DEC) is used to further enhance the separated signals and reduce distortions. We also introduce a supervised multi-objective step-wise learning approach to gradually map the input to the main output of the unified two-stage model (CRN+DNN-DEC), through multiple training steps (e.g., a 4-step mapping as our final suggestion). In this approach, the learned layers of each step serve as pre-training for the next step, with the final step fine-tuning the entire integrated end-to-end model. This unified model not only estimates low-level structural features as direct intermediate targets but also high-level signals as main targets. Experimental results show that the proposed approaches achieve up to a 0.6 improvement in the average perceptual evaluation of speech quality (PESQ) compared to the prior methods.

INDEX TERMS Speech enhancement (SE), convolutional recurrent network (CRN)-based masking, phase-sensitive magnitude ratio mask, joint modeling, hierarchical learning.

I. INTRODUCTION

The goal of speech enhancement (SE) is to reduce the noise and recover the desired speech from its noisy counterpart [1], [2]. The informative features of speech signals can be extracted using signal processing techniques [3], [4]. SE algorithms can be categorized into signal processing-based, model-based, and data-driven methods [5]. Spectral subtraction [6], [7] is a popular technique within signal processing-based methods. It works by subtracting the power spectrum of estimated noise from the noisy speech. Another method in this category is Wiener filtering (WF) [8], [9], where the optimal Wiener filter is estimated to minimize the mean square error (MSE) and thus recover the clean speech in the power spectrum. While these methods generally perform well at relatively high signal-to-noise ratios (SNRs), their effectiveness diminishes in low SNRs and non-stationary noisy environments [10].

Model-based methods rely on creating speech and/or noise models using learned priors, showing promising performance in challenging situations. For instance, in nonnegative matrix factorization (NMF) based enhancement [11], a noisy signal is approximated as a weighted sum of nonnegative bases of speech and noise. These methods perform reasonably when the underlying assumptions are satisfied. However, they are often more effective with structured signals, and their generalization capability to unseen noises is usually limited [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Hongli Dong.

In recent years, with the development of machine learning and deep learning as data-driven methods, SE has been approached as a supervised learning problem. In this approach, noisy features are used to train a supervised learning algorithm on massive labeled datasets, such as deep neural networks (DNNs) [13], [14], more complex networks like convolutional recurrent networks (CRNs) [15], [16], [17], [18] and long short-term memory (LSTM) networks [19], [20], advanced architectures like Transformer networks [21], [22], [23], [24], [25], and state-of-the-art methods including TF-GridNet [26] and diffusion-based models [27]. Hence a non-linear function is learned from mapping noisy speech to clean speech without relying on statistical assumptions about the relationship between speech and noise. Data-driven methods, compared to model-based techniques, have the advantage of performing well in situations where analytical models are unknown or too complex. However, they require massive amounts of data and have a high computation burden for training. In contrast, model-based methods provide prior knowledge and additional information without relying heavily on data for learning mapping structures; instead, the data is often used to estimate parameters. However, simple models may struggle to represent intricacies within complex data and fluctuations over time [28]. To leverage the strengths of both approaches, some recent studies have combined them as a hybrid system that uses data-driven inference with model-based prior knowledge (model-based machine learning [28], [29]) as a model-aided network for specific problems.

Training targets in data-driven SE methods mainly fall into two groups: spectral mapping-based and masking-based [5], [30]. In mapping-based techniques, the training target is directly a spectral representation of the desired source. In contrast, in masking-based techniques, the target is a spectral mask gain that represents the Time-Frequency (T-F) ratio of the desired source to the mixture [5], [31], [32], [33], [34], [35]. In masking-based methods, the mask gain is directly estimated by the network from the mixture, eliminating the need for explicit estimation of the unwanted source or SNR [36]. In conventional deep learning-based SE approaches, whether using T-F mask targets or main spectral magnitudes targets, the domain knowledge of frequency-domain (FD) to time-domain (TD) transformation is not incorporated into the learning process. This means that spectral mapping is performed by the deep network, and the TD speech signal is reconstructed outside of the network separately [33]. Besides masking or mapping-based SE methods, end-to-end enhancement (time-domain mapping without resorting to a T-F mask) has recently gained popularity [37], [38], [39], [40], [41], [42]. A potential advantage of this method is considering the phase of noisy signal during signal reconstruction. Compared to masking-based methods, end-to-end approaches often achieve higher PESQ but lower STOI [33]. Inspired by these three types of approaches, we use a masking-based approach in the first stage and a spectral mapping-based in the second stage. Also, we propose an end-to-end network with temporal mapping that incorporates spectral mask estimation within it (as a built-in component) for better speech quality and intelligibility. Furthermore, the proposed mask utilizes both magnitude and phase information in the enhancement process.

In many studies (e.g. [5], [22], [31], [40], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]), various types of masks such as ideal binary mask (IBM) [5], ideal ratio mask (IRM) [5], phase-sensitive mask (PSM) [40], and complex ideal ratio mask (cIRM) [59] have been proposed and used as training targets for speech enhancement. The PSM mask considers phase but only enhances the magnitude, while the cIRM is a complex mask that enhances both the real and imaginary spectrum, thus jointly enhancing magnitude and phase. Wang et al. [5] showed that ratio masking produces speech with better objective quality than binary masking due to the lower sensitivity of predicting ratio values to estimation error compared to binary values. However, residual noise and distortion caused by mask estimation errors remain issues [44], [62]. In response, some works [31], [43], [44], [45], [60], [61], [63], [64], [65] have proposed two-stage masking-based approaches. In these approaches, after speech separation by an ideal mask, another separate processing stage is used to improve enhancement quality [44], [45], [60], reduce distortions, and compensate for mask estimation errors [31], [43], [61]. For instance, [31] combined DNN-masking and NMF (or vice versa) in two separate stages for separation and enhancement purposes, respectively, and compared this with scenarios where either DNN or NMF was used in both stages. In [44], [45], and [60], DNN-based masking was combined with sparse/NMF reconstruction in two sequential separate stages to further improve the quality of the separated speech. In [44], the separated speech by a binary mask was represented as a linear combination of NMF basis vectors from a trained linear speech model (basis matrix). In [45], this approach was applied using a soft mask, and in [60], a ratio mask was used similarly. Williamson et al. in [43] and [61] as a further study of [44], [45], and [60], after applying a DNN-based masking stage, used a DNN in the second stage to estimate the clean NMF activation coefficients (complementary DNN-NMF model) from the masked speech. Then, the clean speech magnitude was separately approximated outside of the DNN by multiplying the estimated activation matrix and the clean NMF basis matrix. In DNN-NMF, the DNN estimates the NMF activations. Notably, [44], [45], [60] estimated masked speech, while [43], [61] approximated raw clean speech. It has been shown that combining a masking approach with a model-based method like NMF as a post-processing stage [44], [45], [60], or with its DNN-based approximation (DNN-NMF) [43], [61] performs better than using a single stage or other two-stage approaches. In all these methods, the DNN-based magnitude masking was used in the first stage, followed by NMF reconstruction [44], [45], [60] or DNN-NMF [43], [61] in the second stage. Additionally, the

two stages were performed separately. In this work, we use a CRN-based phase-aware magnitude masking in the first stage, and the joint DNN-decoder structure (DNN-DEC, similar to our previous work [66]) in the second stage. DNN-DEC is a joint and non-linear alternative to the DNN-NMF. Additionally, the two stages are performed jointly. In sections I-A and I-B, the differences between our work and these two stages methods are explained in detail.

In some other works, the ideal masks and the NMF coefficients were used in a single DNN, such that the DNN mapped the noisy activation coefficients to the binary mask [67] or a new soft mask [68]. In [12], instead of directly predicting the original mask, the NMF activation coefficients of the mask were estimated by the DNN. These coefficients were then separately multiplied by the corresponding learned basis matrix to approximate the original mask. Studies [12], [69] demonstrated that estimating the NMF activation coefficients using DNN performs better than NMF inference in speech separation. In [12], [67], [68], [69], [70], [71], [72], [73], and [74], the linear operations of NMF and the non-linear operations of DNN were complementary within one stage. In [12] and [69], the non-linear DNN was forced to learn the information obtained from the linear NMF operations. In [70], [71], [72], and [74], NMF inference was jointly combined with DNN, contrary to [12] and [69] where they were performed separately (treated independently). In [12] and [69], the linear activation coefficients, which are intermediate targets, were estimated by DNN as the main output, whereas in [70], [71], and [72], they were not the direct target, and the DNN directly estimated the main spectral signals through the integrated NMF bases.

A. RELATED WORKS AND OUR PROPOSED APPROACH

In [43] and [61], as previously described, DNN-based masking with IRM target was performed in the first stage, and in the separate second stage, the approximation of NMF activations by DNN (DNN-NMF) was done. Additionally, the reconstruction of the speech signal was performed separately outside of the DNN by linearly combining the estimated activation coefficients with the NMF bases. In contrast, in this paper, we propose CRN-based masking in the first stage, where the proposed constrained phase-sensitive magnitude ratio masks (cPSIRMs) are estimated by the CRN. Moreover, a joint DNN-DEC structure [66] is applied in the second stage. In DNN-DEC, the decoder layers of the pre-trained deep autoencoders (DAEs) are integrated into the DNN as a non-linear alternative to the NMF basis. The DAE [75], [76] as a data-driven scheme is useful for dimension reduction, compact representation of data, and capturing the structures. In DNN-DEC, unlike DNN-NMF (the second stage of [43] and [61]), the decoders are joined with the DNN as extra layers so that the main signal is optimized as the output training target instead of the output targeting of activation coefficients. The DNN-DEC, which is a joint and non-linear sparse equivalent of DNN-NMF, is used in the second stage of our proposed system as an enhancement stage. It has the capabilities of more powerful extraction of harmonic structures by DAE over NMF in a non-linear way and better enhancement by the DNN incorporating the extracted non-linear structural characteristics.

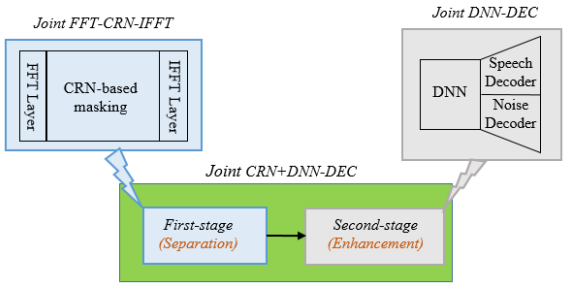


FIGURE 1. The high-level block diagram of the proposed joint CRN+DNN-DEC model. The joint CRN+DNN-DEC model includes the FFT-CRN-IFFT and DNN-DEC blocks.

On the other hand, spectral SE typically uses the short-time Fourier transform (STFT) as a separate front-end FD representation, and the TD signal is reconstructed separately outside of the learning network. In contrast, in our proposed system, the TD signals are directly fed as the input and the target output of the network to help the SE process. This is done by integrating the speech analysis (TD-to-FD transformation, fast Fourier transform (FFT)) and the speech resynthesis (FD-to-TD transformation, inverse fast Fourier transform (IFFT)) parts into the network as extra layers. Furthermore, [77] mentioned the challenge of discerning fundamental speech phones from background noise when using a TD loss function. In other words, an FD loss function has clear discrimination ability and can restore speech with high quality. Hence, for this reason, and also due to the differentiability property of the TD-to-FD transformation, we use an FD loss function to train our first-stage model in which the time-framed estimated signals are converted to the FD in the loss function. To take advantage of both FD and TD information, in the first stage of our system, a CRN is first learned with the proposed cPSIRM mask targets. Then, it is used as a pre-trained model in a new joint model that includes additional FFT and IFFT layers (FFT-CRN-IFFT). Thus, the CRN is updated in the joint model (FFT-CRN-IFFT) according to the new TD objective targets (which are converted to the FD in the loss function). This cooperation between frequency and time information, also incorporating the phase information, leverages domain knowledge of FD to TD conversion (or vice versa) such as spectral properties, which differentiates it from the conventional DNN approach. The overall block diagram of the proposed model (CRN+DNN-DEC) is shown in Fig. 1. The joint stacked model of FFT-CRN-IFFT as a model-aided network (the first stage, separation)and DNN-DEC (the second stage, enhancement) forms the composite CRN+DNN-DEC model which can attenuate more noise and boost the overall performance. According to this figure, the joint FFT-CRN-IFFT model includes the CRN-based

masking integrated with the FFT and IFFT layers. Therefore, the FFT and IFFT are involved in the training process, and the proposed cPSIRM mask values are estimated by the CRN with respect to the final time-domain signals. The joint FFT-CRN-IFFT component is newly proposed and within it the CRN-based masking is critical. Our previously proposed DNN-DEC structure [66] is used as an enhancement stage in the second stage, which is integrated with the first stage. Since DNN-DEC is more potent than DNN-NMF (the second stage of [43] and [61]), it can better correct mask estimation errors. The joint DNN-DEC model comprises a DNN integrated with the speech and noise decoders. It consists of the joint effort of the DAEs in capturing the structures, the DNN in enhancing them, and jointly estimating the main signals using the integrated decoder layers.

Furthermore, the input-to-output mapping (learning) in the CRN+DNN-DEC model is proposed to be performed hierarchically in multiple steps. In our proposed four-step mapping (our ultimate suggestion), the four training steps are as follows: Step 1, the noisy speech spectrum is mapped to the output mask layer; Step 2, the TD noisy speech is mapped to the TD output layer; Step 3, the TD noisy speech is mapped to the encoded layer; and finally, Step4, it is mapped to the main spectral output layer. These mappings, which will be explained in detail in section III-C (Fig. 3c), are done via the related layers in CRN+DNN-DEC. In each new step, the pre-trained layers are updated along with the newly added layers according to the training target and the loss function of that step. Thus, the spectral masks are the primary output targets, the objective TD signals and the encoded representations are the intermediate output targets, and the objective spectral signals are the main output targets. The first three steps act as pre-training for the final training step, which consists of fine-tuning the whole integrated end-to-end model. This leads to a gradual structural learning process and improves performance.

B. OUR CONTRIBUTIONS CONCERNING PREVIOUS WORKS

Overall, the main differences and advancements of this work compared to the earlier ones are:

- In existing CRN-based speech enhancement, the magnitude or complex spectrum of the desired speech is the training target of CRN (spectral mapping-based method). We propose a CRN-based enhancement with mask targets (CRN-based masking), incorporating phase information in addition to the magnitude.
- Inspired by conventional IRM and PSM masks, we propose a new hybrid mask. This mask, named the constrained phase sensitive-magnitude ratio mask (cPSIRM), has limited values like IRM and has phase-difference (PD) information like PSM. Additionally, a phase constraint is applied to modify the PD values and restrict the final enhanced magnitudes. The cPSIRM is provided for both speech and noise signals and is estimated by the CRN.
- In conventional spectral masking-based speech enhancement, TD-to-FD (FFT) and FD-to-TD (IFFT) transformations are not part of the learning process and are performed separately outside the network. In our system, they are integrated as additional layers into the CRN pre-trained with the T-F mask targets. This results in the CRN estimating the mask values with the influence of the final time-domain signals. This leads to the joint estimation of the time-domain signal and the T-F mask (as an intermediate target) within a single network (Joint FFT-CRN-IFFT). This can help the SE process.
- In previous two-stage masking-based speech enhancement approaches [31], [43], [44], [45], [60], [61], DNN-based masking is separately combined with linear NMF or its DNN-based approximation (DNN-NMF) as consecutive enhancement methods. These approaches have the following disadvantages:
  1) The two stages are performed separately.
  2) A T-F mask is the training target of a simple DNN in the first stage, and the final time-domain speech signal is resynthesized separately outside the DNN network. Also, a spectral magnitude mask is used without incorporating phase knowledge.
  3) When using the DNN-based approximation of NMF activations (DNN-NMF) in the second stage, the NMF and DNN work separately. The DNN estimates the activation weights of a linear NMF model as the main output, which are intermediate targets.
- In [31] and [60], a separate combination of DNN-based masking with NMF reconstruction as post-processing is suggested. In [61], the DNN-based masking is combined with DNN-NMF (DNN-IRM+DNN-NMF-Sep) in two consecutive separate stages. In the second stage (DNN-NMF), the DNN and NMF inference are used separately, so that the NMF basis matrix is applied outside the DNN as a linear multiplication operation to reconstruct the main speech signal. In contrast, in the first stage of our proposed model, we use CRN for mask estimation, integrated with FFT and IFFT layers (FFT-CRN-IFFT) to simultaneously estimate the final time-domain signal. In the second stage, we use our joint DNN-DEC structure [66] which is a joint and non-linear sparse equivalent of DNN-NMF (the second stage of [61]). Then, we propose a joint framework of CRN-based masking and DNN-DEC (Joint CRN+DNN-DEC).
- This work extends our previous research [66] by adding the FFT-CRN-IFFT model as the first stage of the system. The joint DNN-DEC model from [66] is used as the second stage in the proposed system, creating a more robust and comprehensive approach.
- In [61], in the second stage, NMF inference and DNN estimates of activations operate independently, and the DNN only predicts the linear activation coefficients as the main output, while they are intermediate targets.

Thus, the main spectral speech signal is approximated manually (separately) outside the DNN by multiplying the estimated activations and the learned NMF speech basis. However, in the second stage of our system (joint DNN-DEC, Fig. 1), properly designed decoders are used as non-linear alternatives to NMF bases to reconstruct the spectral speech and noise signals in a non-linear manner. These decoders are also integrated with the DNN (joint DNN-DEC). Thus, by the cooperation of integrated speech and noise decoders as extra layers of DNN (instead of linear NMF bases), the actual spectral signals are directly estimated by the DNN-DEC. Also, by using hierarchical training, non-linear encoded features are explicitly exploited as direct intermediate targets for the encoded output layer, in addition to using the original signals as the final target for the main output layer.

- This paper proposes a hierarchical three and four-step training approach. In the proposed 4-step mapping (our final suggestion), in step 1, the mapping of the spectral noisy input to the mask output layer is learned through the CRN layers. Then in step 2, the TD noisy input is mapped to the TD speech and noise signals, and in step 3, it is mapped to the sparse encoded features through the pre-trained CRN layers and the extra added layers. Finally, in step 4, the TD noisy input is mapped to the main output layer through all integrated parts (CRN+DNN-DEC model). This approach helps to directly incorporate and maintain the harmonic structures of the spectral masks and the encoded features in learning, and it improves the local minima issue, which could lead to better results.
- The speech and noise NMF activations in [69] and [72] or the DAE's encoded features in [66] are directly estimated from the noisy speech. However, in our model, they are estimated from the separated speech and noise signals using the initial CRN-based masking as a separation stage. Hence, in CRN+DNN-DEC, the DNN performs the regression between the CRN-separated masked signals and their related DAEs-extracted encoded features.
- In this paper, to consider the complementarity between speech and noise, multi-target simultaneous estimation at each step and joint modeling of them is proposed. This means estimating the joint targets of both speech and noise masks, their encoded features, and their actual signals. This is done hierarchically within a single network at the corresponding target output layer based on the related loss function (corresponding training step). This approach shares parameters and exploits the speech and noise correlations in each output layer for better separation.
- In [67] and [68], the NMF activation features and the mask target are used in a single DNN. We use two joint consecutive stages in which the non-linear encoded features are estimated in the second stage from the first masked signals.
- We provide a simple approach with minimal data, facilitating learning through the proposed model-aided network, hierarchical learning, extracting the appropriate compressed features of speech and noise signals, and injecting the encoded features and mask estimation as prior knowledge.

To sum up, the core contributions of this work are:

- We propose a constrained phase-sensitive magnitude ratio mask (cPSIRM) that incorporates both magnitude and phase information.

- Domain knowledge of frequency-domain to time-domain conversion (or vice versa) is integrated into the network.

- We introduce a new pretrain/finetune (step-wise) learning approach (gradual structural learning).

- A joint, fully nonlinear, two-stage separation and enhancement approach is developed.

- The joint FFT-CRN-IFFT component is newly proposed, and within it the CRN-based masking is critical. The development of the cPSIRM mask is a significant contribution to the field.

The rest of this paper is organized as follows: Section II briefly describes the speech enhancement problem. The proposed system is introduced in Section III. Section IV provides evaluations and comparisons between different approaches. Finally, the conclusion is presented in Section V.

II. OVERVIEW OF SPEECH ENHANCEMENT

Given a noisy speech as $y(k) = s(k) + n(k)$, where k is the sample index, the goal of a single-channel speech enhancement problem is to extract an estimate $\hat{s}(k)$ of the desired speech $s(k)$ from a noisy speech $y(k)$. In the short-time Fourier transform (STFT) domain, the corresponding magnitude spectrograms, ignoring the speech-noise cross-term, can be expressed as $Y(f,t) \approx S(f,t) + N(f,t)$, where Y, S, and $N \in \mathrm{R}_{\geq 0}^{F \times T}$ are the magnitude spectrograms of the noisy speech, the clean speech, and the noise, respectively. f and t are the frequency and time indices, and F and T are the total frequency bins and time frames, respectively [78].

In **DNN-based speech enhancement**, the DNN is often employed in two phases: training and testing. In the training phase, the DNN is learned using the training data to map the noisy speech to the desired speech. In the testing phase, the learned DNN estimates the speech from the observed noisy speech [33], [79]. Due to the unbounded values of the raw signals, directly estimating them by DNN (mapping-based separation) is challenging and needs learning a wide dynamic range of values [13], [80]. In contrast, in the masking-based separation, a T-F speech mask that contains the gain values at each T-F unit is usually predicted by the DNN and applied to the observed noisy signal to separate the clean speech from it [5], [33], [61].

In the NMF method, a nonnegative matrix of the signal such as the magnitude spectrum $\mathbf{Y} \in \mathrm{R}_{\geq 0}^{F \times T}$ is decomposed into

a product of a nonnegative basis matrix $\mathbf{W}\in\mathrm{R}_{\geq 0}^{F\times K}$ ($\mathrm{K}\leq\mathrm{F}$) and a nonnegative activation matrix $\mathbf{H}\in\mathrm{R}_{\geq 0}^{K\times T}$ ($\mathbf{Y}\approx\mathbf{WH}$) [81]. K indicates the basis number (columns of $\mathbf{W}$). The basis matrix represents the signal structures, and the activation matrix contains the coefficients that linearly combine the basis vectors to estimate the signal. The matrices $\mathbf{W}$ and $\mathbf{H}$ are found by minimizing a cost function such as Kullback-Leibler (KL) divergence and using a multiplicative update rule. The matrices are typically randomly initialized and then updated using the iterative updating rules [81]. Similar to DNN, **NMF-based speech enhancement** often includes two phases: training and testing. In the training phase, the basis matrices are trained individually from the magnitude spectrograms of the training speech and noise data and kept fixed for the testing phase. In contrast, the related activation matrices are discarded. The trained basis matrices of speech and noise are concatenated as the noisy basis matrix. Then, this larger matrix is used in the testing phase to estimate the noisy activation matrix from the test noisy spectrum ($\mathbf{Y}\simeq[\mathbf{W}_{s-tr}\mathbf{W}_{n-tr}]\begin{bmatrix}\widehat{\mathbf{H}}_s\\ \widehat{\mathbf{H}}_n\end{bmatrix}$). Then, the magnitude spectra of speech ($\widehat{\mathbf{S}}$) and noise ($\widehat{\mathbf{N}}$) are estimated by $\mathbf{W}_{s-tr}\widehat{\mathbf{H}}_s$ and $\mathbf{W}_{n-tr}\widehat{\mathbf{H}}_n$. Finally, a Wiener filter is calculated from the estimated sources and usually applied to the noisy magnitude to obtain the smoothly separated magnitudes of speech and noise. The underlying assumption of this approach is the orthogonality between the bases of speech and noise. However, there are overlaps between them. Thus, the estimation of the test activation using the concatenated bases, which have been separately trained, is accompanied by error. In the **DNN-based approximation of NMF** [12], [61], [69], instead of using the concatenated bases, a DNN learns the mapping of the noisy speech to the activations. Therefore, the DNN estimates them in a non-linear manner.

Similar to the NMF mechanism in signal decomposition, a DAE can approximate a signal by a fundamental model and extract an encoded representation in a non-linear manner. The DAE consists of two deep neural networks: an encoder (f_{ENC}) that maps the input data into the encoded representation and a decoder (f_{DEC}) that reconstructs the data from the encoded representation [75]. The DAE maps the signal to itself through $f_{DEC}(f_{ENC}())$. By restricting the latent space to be lower-dimensional than the input or imposing a regularizing constraint, the model is prevented from learning identity mapping [82]. By this mechanism, the encoder extracts a compact and structural representation of input data while preserving enough information so that the reconstructed data by the decoder is as close as possible to the original data [83], [84]. In a denoising DAE, the noisy speech as input is directly mapped to the clean speech as output [75]. In this paper, to perform enhancement using DAEs, compared to the linear separation in the NMF method, the estimation of the non-linear encoded features of speech and noise from the noisy speech is performed by a DNN in a non-linear manner.

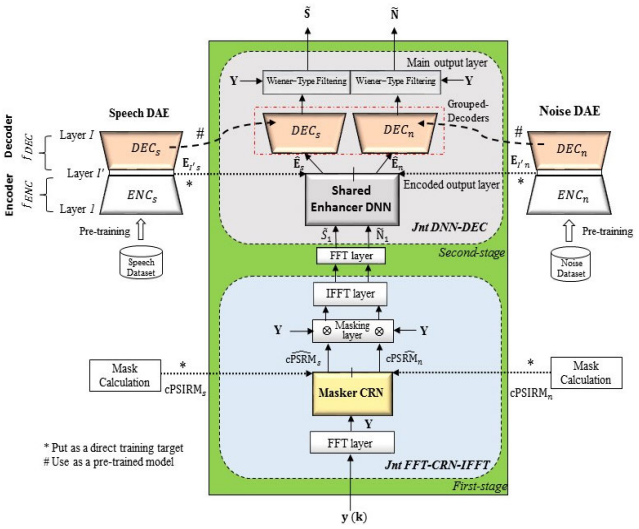


FIGURE 2. The components of the proposed joint CRN+DNN-DEC model. The DAEs and mask calculation blocks are only used in the training phase. The extracted encoded features are directly put as the encoded output target of the enhancer DNN (the dotted lines of E) only in the proposed four-step mapping approaches.

III. THE PROPOSED JOINT CRN+DNN-DEC MODEL

The detailed block diagram of the proposed CRN+DNN-DEC model is shown in Fig. 2. This model consists of the Jnt FFT-CRN-IFFT model (separation stage) and its extra integrated DNN and decoder/Wiener filtering (WF) layers named Jnt DNN-DEC (enhancement stage). The motivation of multiple components is as follows: The FFT-CRN-IFFT part (first stage), which is a masking-based model, is utilized as a separation stage to coarsely distinguish speech and noise signals through a masking approach. To address the limitations of existing time-frequency (T-F) masks, we introduce a new phase-aware mask that effectively incorporates both magnitude and phase information. Additionally, FFT and IFFT transformations are integrated with the masker network (CRN) to embed domain knowledge, such as frequency domain representation of signals, into the learning process. The second stage (based on our previous work) functions as a compensation stage to reduce distortions and improve the quality of the separated speech. These stages are described in Section III-A and Section III-B, respectively. Section III-C presents the proposed different step-wise input-to-output mapping approaches in CRN+DNN-DEC and the training phases of the four-step mapping as our final proposed mapping approach. Multiple training phases are employed to facilitate the learning of the complex (composite) model in a step-wise manner, using a new pre-train and fine-tune approach. This strategy helps overcome the vanishing gradient issue and aids the learning process. Initially, parts of the large CRN+DNN-DEC model undergo pre-training with relevant targets. Subsequently, with the addition of new components (layers), the entire integrated network (CRN+DNN-DEC) is retrained in a fine-tuning phase based on new training targets. This step-wise approach contrasts

with the conventional method of pre-training and fine-tuning, where the entire network is first trained on large datasets and then fine-tuned for more specific datasets. By gradually evolving the network structure and training targets, this method enables continuous optimization and enhances adaptability. In some proposed mapping approaches, the aim is to inject the encoded structural features and mask values directly into the network as training targets, acting as prior knowledge and aiding the learning process.

A. JOINT FFT-CRN-IFFT (SEPARATION STAGE)

The separation stage includes the unified FFT-CRN-IFFT model in which the input and output are the TD-framed signals. The FFT layer is located after the input layer, and the IFFT layer is before the output layer of FFT-CRN-IFFT. As shown in Fig. 2, this model comprises the masker CRN and its integrated masking layer (CRN-based masking), FFT layer, and IFFT layer. Thus, the speech synthesis (FFT), the mask estimation, and the speech resynthesis (IFFT) are integrated. The FFT-CRN-IFFT incorporates the domain knowledge of FD to TD conversion and estimates the mask values concerning the main objectives of separation and the final time-domain signals. The mask values are calculated in the mask calculation blocks (Fig. 2) and put as training targets for the masker CRN.

1) THE PROPOSED CONSTRAINED PHASE-SENSITIVE MAGNITUDE RATIO MASKS (cPSIRMs)

Despite that the cIRM [59] and PSM [40] contain phase information, they have unbounded values which can be destructive for gradient descent-based supervised learning [85], so they are compressed. In [40], the PSM is directly truncated to between 0 and 1, which changes the mask [85], and in [59], the cIRM is compressed by using the hyperbolic tangent. However, compression causes the mask values not to have direct interaction with the signal spectrum and not to represent it clearly. Also, according to findings in [86], [87], [88], and [89], the imaginary part of the cIRM has random patterns and no learnable structure. Reference [89] showed that there is no information in the imaginary part of the cIRM estimated by the DNN, and surprisingly, the network can not estimate it. Thus, to effectively incorporate both magnitude and phase information we propose a constrained phase-sensitive IRM (cPSIRM), which is a hybrid mask of magnitude ratio and PD information. It is estimated by a CRN for both speech and noise signals (CRN-estimated cPSIRM_s and cPSIRM_n). The speech PD is the phase difference between noisy speech and clean speech ($\alpha_1 = \alpha_y(t,f) - \alpha_s(t,f)$), and the noise PD is the phase difference between noisy speech and noise ($\alpha_2 = \alpha_y(t,f) - \alpha_n(t,f)$). According to [51], in high SNRs, α_y spectrum is almost similar to α_s, so $\cos(\alpha_1) \approx 1$. Conversely, in low SNRs, α_1 spectrum values are uncertain and may be random. About α_2, in low SNRs, the difference between α_y and α_n is insignificant, so $\cos(\alpha_2) \approx 1$, and in high SNRs, it tends to be random. These distinctions and characteristics and the available structures in the $|\alpha_1|$ and $|\alpha_2|$ spectra are valuable enough to consider them as a training target of a deep learning-based network. Inspired by the PSM idea, we use $\cos(\alpha_1)$ and $\cos(\alpha_2)$ as training targets which are PD gains for speech signal (PDG_s) and noise signal (PDG_n), respectively. However, as these PDGs are applied on the noisy magnitude along with the related magnitude ratio mask, by investigating the PDG_s and PDG_n values in different PDs, they need to be restricted to estimate the speech and noise signals correctly. According to the triangle rule (Eq. (1)), for PDG_s in the case of $\cos(\alpha_1) < 0$ ($\pi/2 < |\alpha_1| < \pi$), the noise is dominant, and its magnitude squared is greater than the sum of the noisy and clean magnitude squared. Thus, the PDG_s is better to be zero.

$$|N|^2 = |Y|^2 + |S|^2 - 2\,|Y|\,|S|\cos(\alpha_1) \tag{1}$$

A similar condition is established in $\cos(\alpha_2) < 0$ for PDG_n so that the noisy magnitude is smaller than the clean magnitude [90], [91]. Therefore, it is necessary to impose limitations on PDG_s and PDG_n. According to Eq. (2), under the nonnegativity condition, these values are bounded to zero, similar to the rectified linear unit (ReLU) function [92] ($f(x) = max(x, 0)$). Thus, we define the speech and noise-constrained PD gains ($cPDG_s$, $cPDG_n$) and their related magnitude ratio masks (IRM_s, IRM_n) as follows:

$$\begin{aligned} cPDG_s(t,f) &= ReLU(\cos(\alpha_1)) \\ cPDG_n(t,f) &= ReLU(\cos(\alpha_2)) \\ IRM_s(t,f) &= \frac{S(t,f)}{S(t,f)+N(t,f)} \\ IRM_n(t,f) &= \frac{N(t,f)}{S(t,f)+N(t,f)} \end{aligned} \tag{2}$$

$S(t,f)$ and $N(t,f)$ are the speech and noise magnitude spectra at each T-F unit, respectively. Then, the speech and noise cPSIRMs ($cPSIRM_s$, $cPSIRM_n$) are obtained as the product of the related IRM and cPDG as follows (mask calculation blocks in Fig. 2):

$$\begin{aligned} cPSIRM_s(t,f) &= IRM_s(t,f) \times cPDG_s(t,f) \\ &= \frac{S(t,f)}{S(t,f)+N(t,f)} ReLU(\cos(\alpha_1)) \\ cPSIRM_n(t,f) &= IRM_n(t,f) \times cPDG_n(t,f) \\ &= \frac{N(t,f)}{S(t,f)+N(t,f)} ReLU(\cos(\alpha_2)) \end{aligned} \tag{3}$$

Instead of the hard labeling of T-F units in IBMs [5], the proposed cPSIRM is the bounded soft mask that assigns a value between zero and one on each T-F unit. The CRN-estimated cPSIRMs are applied to the noisy magnitude spectrum Y in the integrated masking layer, and the speech and noise are primarily separated as follows:

$$\tilde{\mathbf{S}}_\mathbf{1} = \widehat{cPSIRM}_s \times \mathbf{Y}, \tilde{\mathbf{N}}_\mathbf{1} = \widehat{cPSIRM}_n \times \mathbf{Y} \tag{4}$$

where $\times$ indicates the element-wise multiplication. $\tilde{\mathbf{S}}_\mathbf{1}$ and $\tilde{\mathbf{N}}_1$ are the masked speech and noise estimates, respectively.

This stage (separation) can be considered part of the feature extraction for the second stage (enhancement).

B. JOINT DNN-DEC (ENHANCEMENT STAGE)

According to Fig. 2, the structures in the magnitude spectrograms and the compact representations of speech and noise signals are captured using the related DAEs. The sparse encoded features are extracted from the output activations of the DAEs bottleneck layers and the structural base models from the pre-trained decoder portions. According to Fig. 2, the pre-trained decoders are integrated with DNN layers (DNN-DEC). In capturing the encoded features, in addition to limiting the number of bottleneck layer nodes, a sparsity constraint is also imposed on the latent representation to make it more compressed. The mapping function of the speech and noise DAEs layers (f_i) is as follows:

$$\begin{aligned}
\mathbf{h}_i &= f_i\left(\mathbf{h}_{i-1}\right) = \sigma\left(\mathbf{W}_i\mathbf{h}_{i-1} + \mathbf{b}_i\right) \quad 1 \le i \le I,\\
\mathbf{h}_{I'} &= f_{I'}\left(\ldots f_2\left(f_1\left(\mathbf{h}_0\right)\right)\right) = f_{\mathrm{ENC}}\left(\mathbf{h}_0\right),\\
\mathbf{h}_I &= f_I\left(\ldots f_{I'+2}\left(f_{I'+1}\left(\mathbf{h}_{I'}\right)\right)\right) = f_{\mathrm{DEC}}\left(\mathbf{h}_{I'}\right),\\
\mathbf{h}_0 &= \mathbf{SorN}, \quad \mathbf{h}_{I'} = \mathbf{E}_{I's} \text{ or } \mathbf{E}_{I'n}, \ \mathbf{h}_I = \hat{\mathbf{S}}or\hat{\mathbf{N}},\\
\hat{\mathbf{S}}or\hat{\mathbf{N}} &= f_{\mathrm{DEC}}(f_{\mathrm{ENC}}(\mathbf{SorN})) = f_{\mathrm{DEC}}\left(\mathbf{E}_{I's} \text{ or } \mathbf{E}_{I'n}\right) \qquad (5)
\end{aligned}$$

where f_{ENC} and f_{DEC} are the encoder and decoder mapping functions, respectively. $\mathbf{W}_i$ and $\mathbf{b}_i$ are the DAE weight and bias parameters, respectively. The spectral magnitudes $\mathbf{S}/\hat{\mathbf{S}}$ and $\mathbf{N}/\hat{\mathbf{N}}$ are the input/output of speech DAE and noise DAE, respectively. $\mathbf{E}_{I's}$ indicate the speech-encoded representation and $\mathbf{E}_{I'n}$ shows the noise-encoded representation. $\sigma(\cdot)$ is the non-linear activation function. i is the DAEs layers index with the total number of I, so that i=0 is the input layer, i=I' is the bottleneck layer and i=I is the output layer. $\mathbf{h}_i$ indicates the output activation of the hidden layer $i = 1$ to $I - 1$.

In our four-step mapping, which will be explained in Section III-C, the extracted encoded features ($\mathbf{E}_{I's}$, $\mathbf{E}_{I'n}$), which are the outputs of the pre-trained encoders, are used as a direct intermediate training target for the encoded output layer of the shared enhancer DNN (the dotted lines of E in Fig. 2). This work, similar to prior knowledge, leads to the incorporation of more structural features and improves enhancement results. The encoded features and the main spectral signals are jointly estimated through the enhancer DNN and its extra integrated grouped decoders and WF layers as reconstruction layers (Jnt DNN-DEC). The enhancer DNN maps the masked signals (the separated speech $\tilde{S}_1$ and noise $\tilde{N}_1$) to the corresponding encoded features. The masking process allows the estimation of the speech and noise encoded features to be captured from the separated speech and noise signals. This is in contrast to our previous work [66], where they were directly estimated from the input noisy speech. This results in acquiring more accurate estimates and reduces the noise residue. Then, the speech and noise decoders are applied to the estimated encoded features. Finally, by using the WF layers, the final speech and noise estimates are approximated as follows:

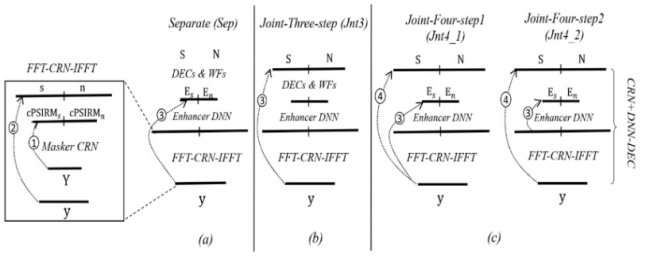


FIGURE 3. Different input-to-output training mappings in CRN+DNN-DEC: (a) Separate reconstruction (CRN+DNN-DEC-Sep); (b) Joint-Three-step (CRN+DNN-DEC-Jnt3); (c) Joint Four-step1 (CRN+DNN-DEC-Jnt4_1), Joint-Four-step2 (CRN+DNN-DEC-Jnt4_2). Steps 1 and 2 are included in all (a), (b), and (c) methods. The learned parts of prior steps serve as initials for the next training step. In "Sep", the decoders are applied separately outside the network.

$$\begin{aligned}
\tilde{\mathbf{S}} &= \frac{\left(DEC_s(\hat{\mathbf{E}}_s)\right)^2}{\left(DEC_s(\hat{\mathbf{E}}_s)\right)^2 + \left(DEC_n(\hat{\mathbf{E}}_n)\right)^2} \times \mathbf{Y}\\
\tilde{\mathbf{N}} &= \frac{\left(DEC_n(\hat{\mathbf{E}}_n)\right)^2}{\left(DEC_s(\hat{\mathbf{E}}_s)\right)^2 + \left(DEC_n(\hat{\mathbf{E}}_n)\right)^2} \times \mathbf{Y} \qquad (6)
\end{aligned}$$

The division operation is element-wise. $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{N}}$ are the enhanced versions of the speech and noise magnitudes, respectively. Thus, the masked signals are mapped to the corresponding encoded features, and then jointly to the main speech and noise signals through the Jnt DNN-DEC layers.

C. DIFFERENT INPUT-TO-OUTPUT TRAINING MAPPINGS IN CRN+DNN-DEC

The proposed different input-to-output mappings in the CRN+DNN-DEC model as shown in Fig. 3 include "*Separate reconstruction (Sep)*", "*Joint-Three-step (Jnt3)*", and "*Joint Four-step1 (Jnt4_1)/Joint-Four-step2 (Jnt4_2)*". In "*Sep*", we have an integrated FFT-CRN-IFFT+DNN model, and the decoders are applied separately outside of the model. While in "*Jnt3*" and " *Jnt4_1/ Jnt4_2*", the decoders are integrated with the FFT-CRN-IFFT+DNN model and we have the joint CRN+DNN-DEC model. First, we define the first and second mappings (training process), which are done in FFT-CRN-IFFT (the left part in Fig. 3). **In the first mapping (circle1)**, the noisy spectral magnitude is mapped to the mask values through the masker CRN layers ($Y \rightarrow [cPSIRM_s cPSIRM_n]$) and via $Loss1$(Eq. (7)).

$$Loss1 = \left\|\mathbf{cPSIRM}_s - \widehat{\mathbf{cPSIRM}}_s\right\|_2^2 + \left\|\mathbf{cPSIRM}_n - \widehat{\mathbf{cPSIRM}}_n\right\|_2^2 \qquad (7)$$

In the second mapping (circle2),the noisy time frames are mapped to the clean and noise frames through FFT-CRN-IFFT and via $Loss2_{TDsig}$ (Eq. (8)).

$$Loss2_{TDsig} = \frac{1}{K}\sum_{k=1}^{K}\left|\left[\left|\mathrm{FFT}(\hat{s}(k))\right|\left|\mathrm{FFT}(\hat{n}(k))\right|\right] - \left[\left|\mathrm{FFT}(s(k))\right|\left|\mathrm{FFT}(n(k))\right|\right]\right| \qquad (8)$$

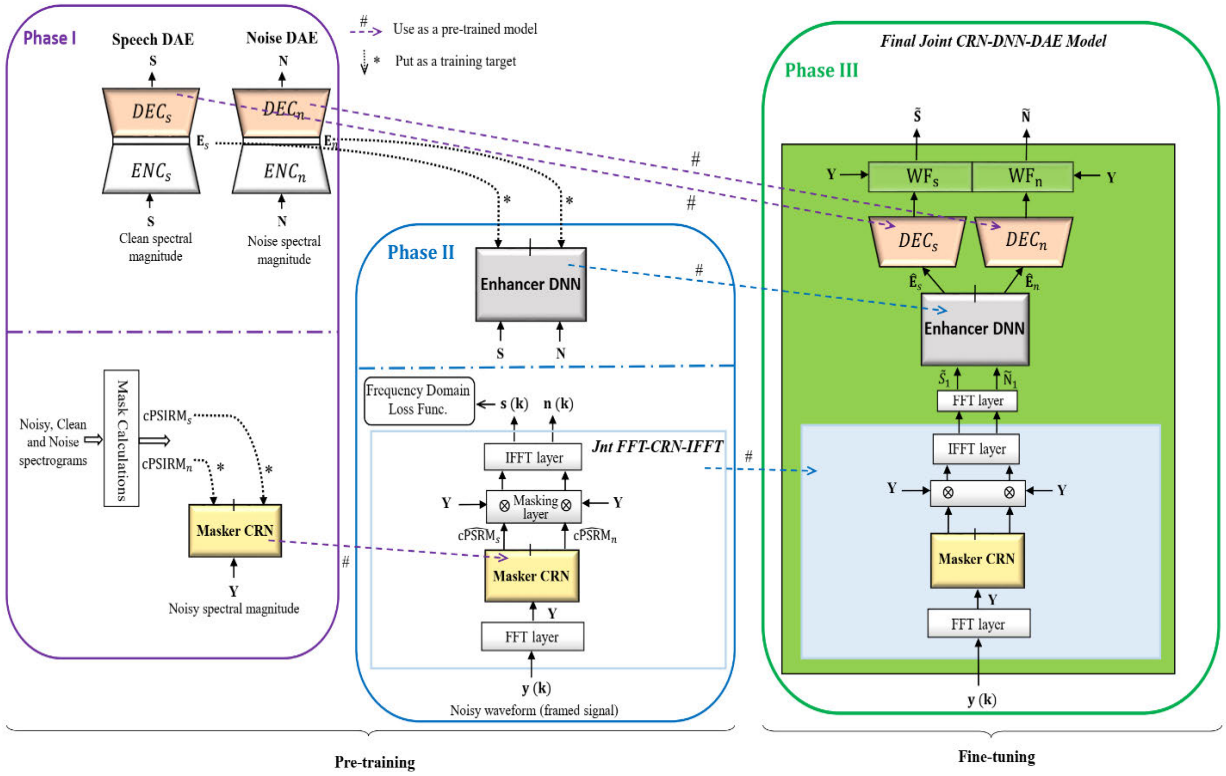


FIGURE 4. The detailed training phases of our joint four-step CRN+DNN-DEC model (Joint-Four-step2 (Jnt4_2)). Phases I, II, and III are done sequentially so that the extracted encoded features of phase I are put as a direct target for the enhancer DNN of phase II. Also, the learned models of phases I and II are used as pre-trained layers for phase III, which is our fine-tuned joint end-to-end model. The dashed and dotted arrows between phases indicate use as pre-training models and placement as output training targets, respectively.

where $|\cdot|$ denotes the absolute value operation (magnitude). $s(k)$ and $n(k)$ are the framed speech and noise signals with size K (frame samples). Reference [77] expressed that an FD loss function has a clear discrimination ability, and can restore speech with higher quality than a TD loss function. In addition to this reason, the TD-to-FD transformation is differentiable. Hence, we suggest using an FD loss function to train our FFT-CRN-IFFT model, whose input/output are time-framed signals ($Loss2_{TDsig}$ in Eq. (8)). Thus, we perform an extra operation of converting the estimated time-framed signals to the FD (FFT) in the loss function at the training phase. Then, according to Eq. (8), the MAE is computed between the estimated and the actual clean and noise spectral magnitudes. **In the "*Separate (Sep)*"case (Fig. 3a)**, after applying the first and second mappings (FFT-CRN-IFFT), the noisy frames are mapped to the encoded features of speech and noise through the integrated FFT-CRN-IFFT and enhancer DNN (Fig. 3a, circle3, $y\rightarrow [E_s E_n]$), and via *Loss3* (Eq. (9)). Then, by separately applying the learned speech and noise decoders and the WFs outside of the network (FFT-CRN-IFFT+DNN) on the estimated encoded features, the speech, and noise spectral magnitudes (S, N) are reconstructed manually.

$$Loss3 = \left\|\mathbf{E}_s - \hat{\mathbf{E}}_s\right\|_2^2 + \left\|\mathbf{E}_n - \hat{\mathbf{E}}_n\right\|_2^2 \tag{9}$$

In the "*Joint-Three-step (Jnt3)*" case (Fig. 3b), the grouped decoders and WFs (reconstruction layers) are integrated into the enhancer DNN so that the objective speech and noise spectral magnitudes (S, N)play the role of the main output targets. While the encoded features are not directly targeted by the enhancer DNN as an intermediate output target. Thus, in this method, after the first two steps (the left part in Fig. 3),the framed noisy speech is directly mapped to the main output layer through the joint model of FFT-CRN-IFFT, enhancer DNN, and reconstruction layers (Fig. 3b, circle3, $y\rightarrow SN$), and via *Loss4* (Eq. (10)).

$$Loss4 = \left\|\mathbf{S} - \tilde{\mathbf{S}}\right\|_2^2 + \left\|\mathbf{N} - \tilde{\mathbf{N}}\right\|_2^2 \tag{10}$$

In the case of the joint four-step mapping (Fig. 3c), as our ultimate suggestion, the input-to-output mapping of the joint CRN+DNN-DEC model is performed in two different ways. **In the first way (*Joint-Four-step1*, Fig. 3c left part),** the hierarchical four-step mapping is as follows: **In step 1**, the CRN is trained with the noisy spectrum as input and the proposed speech and noise masks as output targets. **In step 2**, the FFT-CRN-IFFT is trained with the TD noisy speech as input and the TD speech and noise signals as output. **In step 3**, the integrated FFT-CRN-IFFT and DNN layers are trained with the TD noisy speech as input and the encoded features of speech and noise as output. Finally, **in step 4**,

the integrated FFT-CRN-IFFT, DNN, and decoders (unified CRN+DNN-DEC model) are trained with the TD noisy speech as input and the main spectral speech and noise signals as output. In *Joint-Four-step1*, the integrated FFT-CRN-IFFT and enhancer DNN (step 3) is the pre-trained part. **In the second way (Fig. 3c right part, *Joint-Four-step2*)**, instead of mapping the input noisy speech to the encoded layer in *Joint-Four-step1*(Fig. 3c left part, circle3, $y \rightarrow [E_s E_n]$), the separated masked signals are mapped to the encoded layer by the enhancer DNN (Fig. 3c right part, circle3, $[\tilde{S}_1 \tilde{N}_1] \rightarrow [E_s E_n]$). In *Joint-Four-step2*, FFT-CRN-IFFT (steps 1 and 2) and enhancer DNN (step 3) are separately pre-trained components. Then, the framed noisy speech is mapped to the main output layer through the unified CRN+DNN-DEC model (step 4), while the learned components of the prior steps operate as pre-trained parts. In other words, the initial mappings act as pre-trainings for the final mapping.

1) THE TRAINING PHASES OF THE FOUR-STEP MAPPING IN CRN+DNN-DEC

As shown in Fig. 4, the training phases of our second four-step mapping approach (*Joint-Four-step2*) in the CRN+DNN-DEC model as one of the proposed mappings (Fig. 3c, right part) are as follows:

In phase I, the speech DAE is trained with the clean speech spectral magnitude as the input and output feature, and the noise DAE is similarly trained with the noise spectral magnitude. Simultaneously, a Convolutional Recurrent Network (CRN) is trained and the complex Ideal Ratio Masks (cPSIRMs) of speech and noise (computed in the Mask Computation Block (Section III-A)) Also, the masker CRN is trained with the noisy speech spectral magnitude as the input and the speech and noise cPSIRMs calculated in the mask calculation block (section III-A) as the output targets. Then, we use the learned CRN and decoders as pre-trained models for phase II and phase III, respectively. We retrain them in the new phases along with the other integrated layers by the training targets of that phase. In Fig. 4, the dashed arrows between the phases indicate the use of the learned models of the previous phases as the pre-trained models for the next ones. They are updated in the new phases by the newobjectives. Also, the dotted arrows depict that the extracted features are used as an explicit training target, i.e., the extracted encoded features are used as direct targets for the enhancer DNN, and the calculated masks are put as training targets for the masker CRN.

In phase II, the enhancer DNN is trained with the DAEs-extracted encoded features as output targets. Also, the Jnt FFT-CRN-IFFT model (section III-A) is trained with the framed noisy speech as input and the framed speech and noise signals as output targets through the frequency-domain loss function ($Loss2_{TDsig}$ (Eq. 8)). Therefore, the CRN layers learned with the T-F mask targets (phase I) are now updated according to the ultimate time-domain signals, so that the mask values are estimated regarding the final time-domain output. The FFT and IFFT layers are deterministic and are not changed during training. Finally, the learned enhancer DNN and Jnt FFT-CRN-IFFT are used as the pre-trained constituent components for the composite CRN+DNN-DEC model (phase III).

In phase III, the unified CRN+DNN-DEC model, which consists of the pre-trained FFT-CRN-IFFT, enhancer DNN, grouped decoders, and some additional deterministic layers, is trained end-to-end with the output target of speech and noise spectral magnitudes with the spectral magnitudes of speech and noise as the output targets and through *Loss4* (Eq. (10)). The deterministic layers (WFs, FFT/IFFT, and masking layers) do not have connection weights and do not require learning. The pre-trained weights are used as the initial values for the new joint model and continue to be learned during the joint training.

Therefore, the training process of the CRN and DNN components of CRN+DNN-DEC is as follows:

According to Fig. 3 (circle1) and Fig. 4, in the first training process, the CRN is trained with the noisy magnitude spectrum as input, and the cPSIRM values (Eq. (3)) as output targets through *Loss1* (Eq. (7)). The CRN weights are then updated in the joint models along with the other integrated layers in other training mappings (circles 2, 3, 4 in Fig. 3) and via the related loss functions (section III-C).

In Separate reconstruction (Fig. 3a), the DNN layers, along with the FFT-CRN-IFFT layers, are trained with the noisy speech as input and the concatenated speech and noise encoded features as the output targets of the integrated FFT-CRN-IFFT+DNN model (circle 3 in Fig. 3a). In Joint-Three-step (Fig. 3b), the DNN, which is a component within the joint CRN+DNN-DEC model, is trained along with the other integrated layers based on the final spectral speech and noise signals (which are the output training targets of the joint CRN+DNN-DEC model). Therefore, the resulting outputs of the decoder layers are used to compute the error metric for optimizing the DNN weights. By calculating and propagating the output error through the decoder layers to the DNN layers, the DNN parameters are tuned. In Joint Four-step1 (left part of Fig. 3c), at first, similar to the Fig. 3a approach, the DNN, along with the FFT-CRN-IFFT, is trained (pre-trained) with the encoded features as output target(circle 3 in the left part of Fig. 3c), then similar to the Fig. 3b approach, the DNN layers are updated based on the final spectral speech and noise signals (circle 4 in the left part of Fig. 3c). In Joint-Four-step2 (right part of Fig. 3c and Fig. 4), the DNN is first trained (pre-trained) with the speech and noise magnitudes as input, and the encoded features as output target (circle 3 in right part of Fig. 3c and Phase II in Fig. 4). Then its layers are updated based on the final speech and noise signals similar to Joint-Three-step (Fig. 3b) and Joint Four-step1 (circle 4 in left part of Fig. 3c).

IV. PERFORMANCE EVALUATION

The performance of the proposed CRN+DNN-DEC methods (*"CRN+DNN-DEC-Sep"*, *"CRN+DNN-DEC-Jnt3"*, and *"CRN+DNN-DEC-Jnt4_1, CRN+DNN-DEC-Jnt4_2"*)

is compared with DNN-DEC [66] and other baseline methods. In the " *CRN+DNN-DEC-Sep*" approach, the decoders and Wieners are applied separately outside the network compared to others)labeled by *"Jnt3"* and *"Jnt4"*) in which they are integrated into the DNN and jointly optimized (Fig. 3).

The DNN-DEC [66] is the main comparison method, as this work is an extension of [66], mainly by introducing an extra CRN masking network (Jnt FFT-CRN-IFFT). The DNN-IRM+DNN-NMF-Sep [61] method has also been used as another comparison method since it is the most relevant previous work. As a side note, we note that DNN-IRM+DNN-NMF-Sep [61] without the first stage (i.e., DNN-NMF-Sep) was already presented in [66] as a compared method. The DNN with the IRM target (DNN-IRM) [5], the LSTM with IRM target (LSTM-IRM) [19], [20], and the CRN with the magnitude target (CRN-Mag) [15], [16], [17], [18] are also implemented as additional comparison methods. The CRN is trained with the cPSIRM target (CRN-cPSIRM) as a proposed approach. The CRN is also performed with the IRM target (CRN-IRM) to compare with CRN-cPSIRM and investigate the performance of the proposed cPSIRM mask. The cPSIRMs are as defined in Eq. (3). The estimation of the IRMs and PDGs is also performed by two separate CRNs; however, the results do not differ much from the combined-mask estimation by one CRN (CRN -cPSIRM). In DNN/CRN/LSTM-IRM methods, the speech and noise IRM masks, which are the training targets, are estimated from the mixture by the DNN/CRN/LSTM networks and are applied on the mixture separately outside the network to approximate the main signals. On the other hand, a comparison with a transformer-based approach [25] and a current state-of-the-art method, the diffusion-based model [27], is performed to further assess the performance of the proposed approach. The work in [25] is a multi-head self-attention network (MHANet) that we implemented on our dataset. It should be noted that performing this model with a mask target gave better results, and thus, for a fair comparison, we performed it with the IRM target (named MHANet [25]-IRM). Also, as our work is not causal speech enhancement, instead of the masked multi-head self-attention block in [25], we used the traditional MHA block. MHANet, which is similar to a Transformer's encoder, includes 6 stacked encoder layers with 1024 nodes ($d_{model} = 1024$), 2 heads, and a dropout rate of 0.1. The research in [27] explored the application of diffusion-based generative models for speech enhancement and dereverberation. This research, named Score-based Generative Model for Speech Enhancement (SGMSE+), builds on earlier works by utilizing a stochastic differential equation framework to improve speech quality. Unlike traditional conditional generation tasks, this method initiates the reverse diffusion process from a mixture of noisy speech and Gaussian noise, rather than from pure Gaussian noise. We use its two pre-trained models with VoiceBank-DEMAND and WSJ0-CHiME3 for speech enhancement. For a fair comparison, they were performed on our unseen noisy signals, so their average results are reported in the unseen part.

A. DATASET AND MEASURES SETUP

1) DATASET DESCRIPTION

The TIMIT database [93], which includes the utterances of 630 male and female speakers, is used as the speech dataset. Similar to [66] as the main comparison method and for a fair comparison, for training, 200 clean speech utterances were randomly selected from the TIMIT training dataset and mixed with *babble*, *factory*, and *machinegun* noises from the NOISEX-92 DB [94] at SNRs from −5 to 20 dB with 5 dB steps. The validation split was set to 10% to achieve validation data. For testing, 60 clean speech utterances were randomly selected from the TIMIT testing dataset and mixed with the above noises as seen noises. Additionally, they were mixed with the real recorded *factorymachine* and *windshieldrain* noises from the *freesound* website (freesound.org) at SNRs of −5, 5, 0, and 10 dB as unseen noises. The same training and testing set was used for all the proposed and comparison methods.

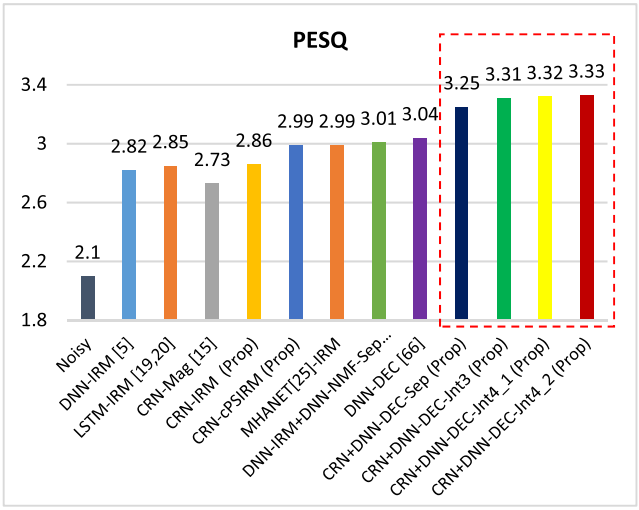
The magnitude spectrograms were obtained using a 512-point (32*ms*) Hamming window, a 128-point (8*ms*) shift size, a 512-point (32 *ms*) frame length, and a 512-point STFT. Thus, the frame size K in Eq. (8) is 512. By cutting the symmetric parts of the STFT coefficients, the dimensions of the spectrograms are $257 \times$ *time-frame numbers*.

2) NMF SETTINGS

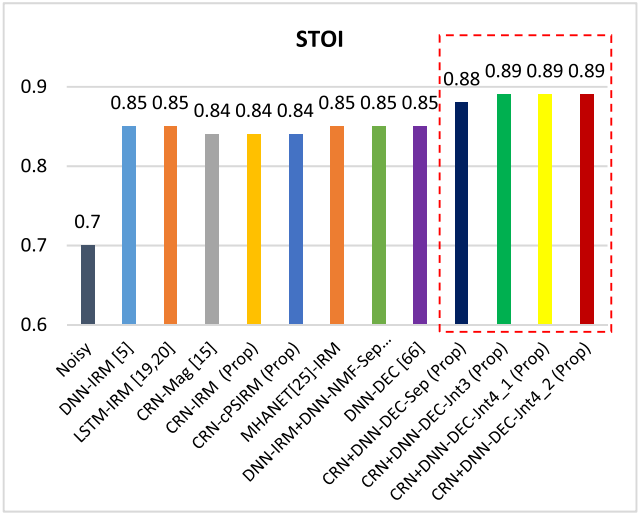
The number of speech and noise bases is empirically set to 100 each. Therefore, the dimensions of the basis and activation matrices are 257×100 (frequency bins $\times$ basis numbers) and $100 \times$ *time-frame numbers,* respectively. The NMF is applied to the concatenated magnitude spectrograms of all training noises to obtain the overall $\mathbf{W}_n$. The NMF maximum iteration number is set to 50.

3) NETWORKS SETTINGS

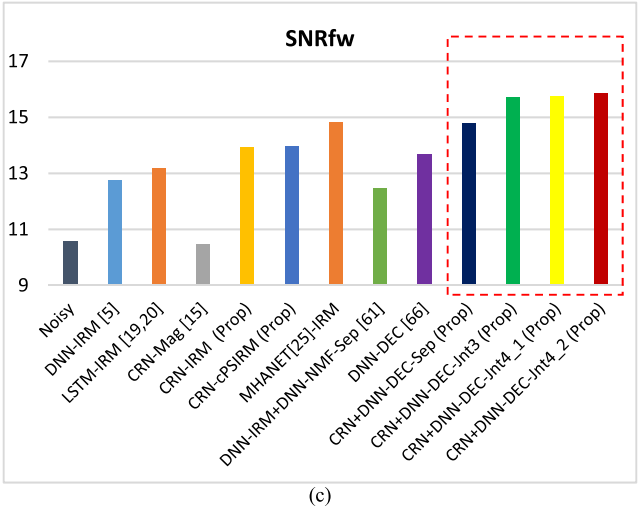
For a fair comparison, the used DNN in all baseline and proposed models has four hidden layers of 1024 nodes. The architecture of speech and noise DAEs is empirically set to 257-1024-512-100-512-1024-257 and 257-512-512-100-512-512-257, respectively. In the CRN+DNN-DEC model, the encoded output layer, which is related to the encoded features of speech and noise signals, includes 100×2=200 nodes. The main output layer, which is associated with the main spectral speech and noise signals, has 257×2=514 nodes. The mask layer contains 257×2=514 nodes due to the spectral mask dimensions. The TD output layer has 512×2=1024 nodes due to the frame size. The DAEs and enhancer DNN use Leaky rectified linear units (LReLU) [96] with $\alpha = 0.1$ ($f(x) = max(\alpha x, x)$) as the activation function for the hidden layers to address the "dying ReLU" issue. These networks use the linear activation function for the



(a)



(b)



(c)

FIGURE 5. Comparison of PESQ (a), STOI (b), and SNRfw (c) scores for each averaged over the seen noise types and input SNRs.

output layer. The activation function of the encoded output layer is set to ReLU when used as a direct output target.

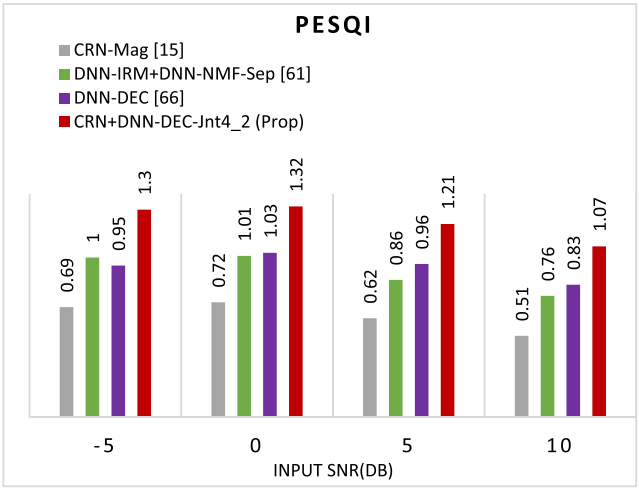


FIGURE 6. The average PESQI results at each input SNR for the proposed four-step CRN+DNN-DEC method over DNN-DEC [66], DNN-IRM+DNN-NMF-Sep [61], and CRN-Mag [15] methods as examples.

The CRN consists of the convolutional neural network (CNN) encoder-decoder and LSTM. The CRN configuration in all CRN-used methods is based on [15] and [18]. Unlike the kernel size of 2 × 3 (time × frequency) in [15], similar to [18], we use 1 × 3 kernels, without changing the performance. The LSTM-IRM model has two LSTM layers of 3072 LReLU units and one fully connected (FC) layer, including 1024 nodes with LReLU activations. It also has an FC output layer with 257 nodes for mask estimation. This setting is based on the LSTM in [20] and the LSTM-IRM method [19], which was used as a comparison method in [20]. However in [20], LSTM layers have 425 nodes, we experimentally use more nodes for better results.

The networks are trained by the Adam optimizer [97] with an initial learning rate of 0.001 and a maximum epoch of 100. Batch normalization is also used to accelerate learning and avoid local minima issues. The weights and bias parameters of networks are computed by using the backpropagation algorithm.

4) LOSS FUNCTIONS

The mean-square error (MSE) and mean absolute error (MAE)-based loss functions as defined in Eq. (7)-(10) are optimized to minimize the distance between the predicted output and the corresponding target in each related training mapping in CRN+DNN-DEC model. The masker CRN is trained by $Loss1$ (Eq. (7)). $Loss2_{TDsig}$ (Eq. (8)) is used for training the joint FFT-CRN-IFFT model. $Loss3$ (Eq. (9)) and $Loss4$ (Eq. (10)) are related to the encoded and the main output layer of CRN+DNN-DEC, respectively. L_{DAE} (Eq. (11)) is used for DAE training (e.g., speech DAE in Eq. (11)). It includes an MSE term and a sparsity regularization term in the form of l_1-norm ($||.||_1$) as an approximation of l_0-norm which is NP-hard. The sparsity constraint is applied to the hidden representations activities so fewer nodes would "fire"

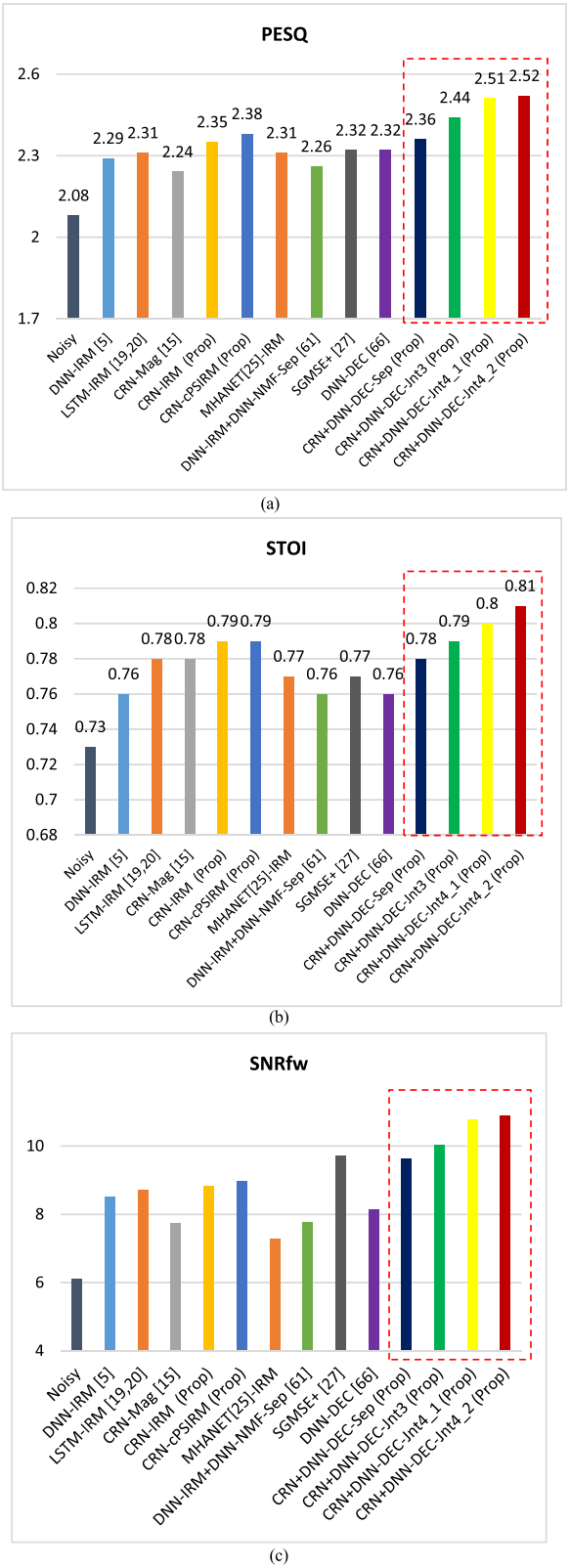


FIGURE 7. Comparison of PESQ (a), STOI (b), and SNRfw (c) scores averaged over the unseen noise types and input SNRs.

at a given time. $||.||_2$ denotes the l_2-norm.

$$L_{DAE} = \|\mathbf{S} - f_{DEC} f_{ENC}(\mathbf{S})\|_2^2 + \|\mathbf{E}_s\|_1 \qquad (11)$$

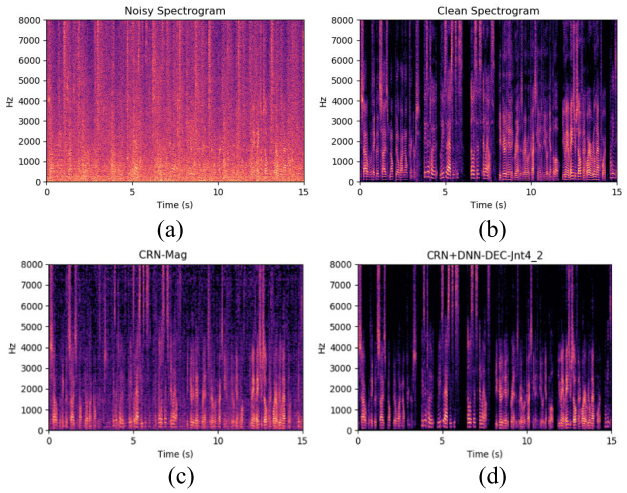


FIGURE 8. The magnitude spectrograms of different signals: (a) Noisy speech with factory noise at −5 dB SNR; (b) Clean speech; (c) Speech enhanced by CRN-Mag [15]; (d) Speech enhanced by the proposed CRN+DNN-DEC-Jnt4_2 approach.

5) METRICS

The evaluations were done using the perceptual evaluation of speech quality (PESQ) [98], short-time objective intelligibility (STOI) [99], and frequency-weighted segmental SNR (SNR_{fw}) [100], [101]. Higher values indicate better performance. The PESQ score ranges from −0.5 to 4.5 and measures speech quality [100]. The STOI range is [0, 1] and reflects speech intelligibility. The SNR_{fw} measures a generalized short-time performance. Furthermore, we also calculate the improvement of the PESQ metric (PESQI) as the difference between the PESQ scores of the enhanced and noisy speech versus the clean speech ($PESQI = PESQ(\tilde{s}, s) - PESQ(y, s)$).

B. RESULTS AND DISCUSSION

The performance of the proposed CRN+DNN-DEC models, CRN-IRM, and CRN-cPSIRM is evaluated on the testing set in the seen and unseen noise conditions. The DNN-DEC [66], DNN-IRM+DNN-NMF-Sep [61], DNN-IRM [5], LSTM-IRM [19], [20], and CRN-Mag [15] methods are evaluated on our dataset as comparison methods.

1) RESULTS

The average metrics results of all methods over different seen noises and input SNRs are presented in Fig. 5. The average PESQI results are also given in Fig. 6 at each input SNR for the proposed four-step CRN+DNN-DEC approach and the DNN-DEC [66], DNN-IRM+DNN-NMF-Sep [61], and CRN-Mag [15] methods as comparisons.

We present the average metrics results of all methods over the different unseen noises and input SNRs in Fig. 7.

In the end, we illustrate spectrograms of the enhanced speech by *CRN+DNN-DEC-Jnt4_2* and *CRN-Mag* in Fig. 8 as examples. As can be observed, the proposed

CRN+DNN-DEC-Jnt4_2 method (Fig. 8d) improves speech with high quality and restores more harmonic structures. The reconstructed speech by *CRN-Mag* [15] (Fig. 8c) contains more noise components.

In addition, to assess the practical feasibility of the proposed model, the execution time was evaluated. The execution time of our final model on an NVIDIA GeForce GTX 1080 8GB graphics processing unit (GPU) for a test noisy speech with a duration of 3 minutes and 22 seconds is 3.6 seconds, averaged over 5 trials. When using an Intel Core i5-7600 @ 3.50GHz central processing unit (CPU), the execution time is about 68.4 seconds. These results indicate that the proposed model is fast enough and suitable for practical applications.

2) DISCUSSION

As shown in Fig. 5 (a, b, c), the proposed CRN+DNN-DEC models (indicated by the red dashed box) outperform the average results of DNN-DEC [66] for seen noise types in terms of three metrics. This indicates that applying Jnt FFT-CRN-IFFT as a preliminary separation stage can significantly lead to better distinguishing speech and noise components in the subsequence enhancement stage (DNN-DEC), which is joined with the first stage. Jnt FFT-CRN-IFFT includes CRN-based cPSIRM masking, which is the fundamental reason for the superiority. Also, our methods outperform DNN-IRM+DNN-NMF-Sep [61] because of our use of Jnt FFT-CRN-IFFT in the first stage versus DNN-IRM in [61] and our use of non-linear Jnt DNN-DEC layers in the second stage versus DNN-NMF-Sep in [61]. The superiority of Jnt DNN-DEC over DNN-NMF-Sep represents the better ability of the decoders in capturing structures and learning patterns in comparison to the NMF basis matrix due to their non-linear and deep layers and jointing them with the DNN. Furthermore, this represents the better learning of DNN on more structural patterns and features extracted by DAE compared to NMF. We explained more detailed points in section I-A.

The proposed CRN+DNN-DEC models also offer a considerable improvement over DNN-IRM [5], LSTM-IRM [19], [20], CRN-Mag [15], CRN-IRM, CRN-cPSIRM, and MHANet [25]-IRM. This is mainly due to the joint hierarchical efforts of the CRN-based cPSIRM masking for separation, the DAEs for spectral structure extraction, and the DNN for enhancement. The improved results of our CRN+DNN-DEC methods over the CRN-cPSIRM show the effect of the second enhancement stage (DNN-DEC layers) following the masking stage to compensate for the mask estimation errors. The better performance of CRN-cPSIRM over CRN-IRM indicates the superiority of the proposed phase-sensitive mask over the IRM due to the appropriate incorporation of both magnitude and phase information. In summary, for each seen noise, different CRN+DNN-DEC approaches, and among them, the four-step mappings produce the best results. Indeed, according to Fig. 5a, in CRN+DNN-DEC models, going from the method labeled by "*Sep*" to "*Jnt3*" and then to "*Jnt4*", the PESQ score increases for each noise. This performance shows that the injection of the base structures as basic knowledge into the DNN in the form of the joint extra integrated layers (*Jnt3* versus *Sep*) and the direct targeting of the encoded features by DNN (*Jnt4* versus *Jnt3*) leads to improved performance. In terms of STOI (Fig. 5b) and SNR_{fw}(Fig. 5c), the results improve from "*Sep*" to "*Jnt3*", although "*Jnt3*" and " *Jnt4*" have almost the same results. According to these figures, the two ways of four-step mapping (*Jnt4_1* and *Jnt4_2* explained in Fig. 3c) get nearly the same results. This result indicates that the direct mapping of the FFT-CRN-IFFT output to the encoded features does not differ much from mapping the noisy speech. We can also see in Fig. 6 that the average PESQI result of the proposed four-step CRN+DNN-DEC model is considerably higher than DNN-DEC [66], DNN-IRM+DNN-NMF-Sep [61], and CRN-Mag [15] at each input SNR. In the unseen noise conditions (Fig. 7a, b, c), the improvement of scores is naturally less than the seen noises. In most cases, the performance of the proposed CRN+DNN-DEC methods is improved over DNN-DEC and other baseline methods, especially the state-of-the-art SGMSE+ method. Similar to seen noises, within the CRN+DNN-DEC models, "*Jnt4_1*"/"*Jnt4_2*" have better results than "*Jnt3*", and likewise, "*Jnt3*" performs better than "*Sep*" in three metrics.

V. CONCLUSION

In this work, we proposed the joint cascaded two-stage CRN+DNN-DEC model to jointly exploit the CRN-based masking, DAEs-based structure extraction, and DNN-based enhancement in noise elimination. In the CRN-based masking part, we proposed the estimation of a constrained phase-sensitive magnitude ratio mask (cPSIRM) to consider both magnitude and phase information for better enhancement results. The CRN-based masking integrated with the FFT and IFFT layers (FFT-CRN-IFFT) was applied for speech/noise separation. The DAEs extract the non-linear sparse encoded representations (features) and the structural patterns (non-linear dictionaries) of speech and noise signals. The DNN further distinguishes between the separated signals and suppresses the residual interferences in collaboration with the DAEs-extracted structures (decoder layers as non-linear bases) (DNN-DEC). The input-to-output mapping in the CRN+DNN-DEC model was proposed to perform in three forms: "Separate", "Joint-three-step" and "Joint-four-step". The four-step mappings presented the best results due to the explicit effect of the knowledge injected into the system by placing them as a direct target and through step-wise (gradual) learning. In other words, mappings to the mask and encoded output layers (as intermediate output layers) act as pre-training steps for the final mapping to the main output layer, which is fine-tuning the whole unified model. Thus, the unified model not only estimates the low-level structural features as direct intermediate targets but also estimates the high-level signals as main targets. It should be noted that the proposed step-wise learning approach is a suitable method for use in large networks to facilitate learning. The

experimental results showed that our proposed CRN+DNN-DEC approaches can further improve noise suppression performance and perform better than the prior methods. One of the constraints of the proposed model is its robustness and applicability in diverse speech-processing environments mainly due to the hardware limitations in applying several noise types, which could be considered for future work. Also, future work could involve conducting experiments to demonstrate the effectiveness of the CRN+DNN-DEC model and the constrained phase-sensitive magnitude ratio mask in real-world speech enhancement scenarios.

REFERENCES

[1] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey, "Fundamentals, present and future perspectives of speech enhancement," *Int. J. Speech Technol.*, vol. 24, no. 4, pp. 883–901, Dec. 2021.

[2] A. Chaudhari and S. B. Dhonde, "A review on speech enhancement techniques," in *Proc. Int. Conf. Pervasive Comput. (ICPC)*, Jan. 2015, pp. 1–3.

[3] S. Seyedin and M. Ahadi, "Feature extraction based on DCT and MVDR spectral estimation for robust speech recognition," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 605–608.

[4] S. Alisamir, S. M. Ahadi, and S. Seyedin, "An end-to-end deep learning model to recognize Farsi speech from raw input," in *Proc. 4th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS)*, Dec. 2018, pp. 1–5.

[5] 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.

[6] 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.

[7] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Proc. Comput. Sci.*, vol. 54, pp. 574–584, Jan. 2015.

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

[9] N. Upadhyay and R. K. Jaiswal, "Single channel speech enhancement: Using Wiener filtering with recursive noise estimation," *Proc. Comput. Sci.*, vol. 84, pp. 22–30, Jan. 2016.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5751–5755.

[17] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 380–390, Nov. 2020, doi: 10.1109/taslp.2019.2955276.

[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] 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.

[20] 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.

[21] M. Ye and H. Wan, "Improved transformer-based dual-path network with amplitude and complex domain feature fusion for speech enhancement," *Entropy*, vol. 25, no. 2, p. 228, Jan. 2023.

[22] Q. Zhang, X. Qian, Z. Ni, A. Nicolson, E. Ambikairajah, and H. Li, "A time-frequency attention module for neural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 462–475, Nov. 2023, doi: 10.1109/TASLP.2022.3225649.

[23] K. Wang, W. Lu, P. Liu, J. Yao, and H. Li, "Multi-stage attention network for monaural speech enhancement," *IET Signal Process.*, vol. 17, no. 3, pp. 1–15, Mar. 2023, doi: 10.1049/sil2.12182.

[24] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient transformer-based speech enhancement using long frames and STFT magnitudes," in *Proc. Interspeech*, Sep. 2022, pp. 2948–2952.

[25] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Commun.*, vol. 125, pp. 80–96, Dec. 2020.

[26] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3221–3236, Aug. 2023, doi: 10.1109/TASLP.2023.3304482.

[27] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, Jun. 2023, doi: 10.1109/TASLP.2023.3285241.

[28] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning: Key approaches and design guidelines," in *Proc. IEEE Data Sci. Learn. Workshop (DSLW)*, Jun. 2021, pp. 1–6.

[29] N. Shlezinger and Y. C. Eldar, "Model-based deep learning," *Found. Trends Signal Process.*, vol. 17, no. 4, pp. 291–416, 2023.

[30] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: State of the art analysis," *Artif. Intell. Rev.*, vol. 56, no. S3, pp. 3651–3703, Dec. 2023.

[31] 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.

[32] Y. Luo, J. Wang, L. Xu, and L. Yang, "Multi-stream gated and pyramidal temporal convolutional neural networks for audio-visual speech separation in multi-talker environments," in *Proc. Interspeech*, Aug. 2021, pp. 1104–1108.

[33] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[34] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components loss for neural networks in mask-based speech enhancement," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 24, Dec. 2021.

[35] L. Zhou, X. Chen, C. Wu, Q. Zhong, X. Cheng, and Y. Tang, "Speech enhancement via mask-mapping based residual dense network," *Comput., Mater. Continua*, vol. 74, no. 1, pp. 1259–1277, 2023.

[36] Y. Kang, N. Zheng, and Q. Meng, "Deep learning-based speech enhancement with a loss trading off the speech distortion and the noise residue for cochlear implants," *Frontiers Med.*, vol. 8, pp. 1–13, Nov. 2021, doi: 10.3389/fmed.2021.740123.

[37] 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.

[38] Z.-Q. Wang, J. Le Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018, pp. 2708–2712.

[39] Y. Koizumi, N. Harada, Y. Haneda, Y. Hioka, and K. Kobayashi, "End-to-end sound source enhancement using deep neural network in the modified discrete cosine transform domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 706–710.

[40] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le 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.

[41] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 684–688.

[42] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 686–690.

[43] 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.

[44] 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. 7034–7038.

[45] 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.

[46] 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.

[47] 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.

[48] D. Sowjanya, S. Sivapatham, A. Kar, and V. Mladenovic, "Mask estimation using phase information and inter-channel correlation for speech enhancement," *Circuits, Syst., Signal Process.*, vol. 41, no. 7, pp. 4117–4135, Jul. 2022.

[49] 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.

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

[51] X. Wang and C. Bao, "Mask estimation incorporating phase-sensitive information for speech enhancement," *Appl. Acoust.*, vol. 156, pp. 101–112, Dec. 2019.

[52] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "Speech enhancement with phase sensitive mask estimation using a novel hybrid neural network," *IEEE Open J. Signal Process.*, vol. 2, pp. 136–150, 2021.

[53] 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.

[54] S. Balasubramanian, R. Rajavel, and A. Kar, "Estimation of ideal binary mask for audio-visual monaural speech enhancement," *Circuits, Syst., Signal Process.*, vol. 42, no. 9, pp. 5313–5337, Sep. 2023.

[55] S. Sivapatham, A. Kar, R. Bodile, V. Mladenovic, and P. Sooraksa, "A deep neural network-correlation phase sensitive mask based estimation to improve speech intelligibility," *Appl. Acoust.*, vol. 212, Sep. 2023, Art. no. 109592.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] E. M. Grais and H. Erdogan, "Spectro-temporal post-enhancement using MMSE estimation in NMF based single-channel source separation," in *Proc. Interspeech*, Aug. 2013, pp. 3279–3283.

[63] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 239–243.

[64] A. A. Nair and K. Koishida, "Cascaded time + time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7153–7157.

[65] Z. Zhang, L. Zhang, X. Zhuang, Y. Qian, and M. Wang, "Supervised attention multi-scale temporal convolutional network for monaural speech enhancement," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, p. 20, Apr. 2024.

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

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] M. Pashaian and S. Seyedin, "Speech enhancement using joint DNN-NMF model learned with multi-objective frequency differential spectrum loss function," *IET Signal Process.*, vol. 2024, pp. 1–10, Jan. 2024.

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

[76] 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.

[77] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[78] Y. Zhang, "'Modulation domain processing and speech phase spectrum in speech enhancement," Ph.D. thesis, Dept. Comp. Sci., Univ. Missouri, Columbia, MO, USA, 2012.

[79] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[80] 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.

[81] 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.
[82] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proc. IEEE*, vol. 111, pp. 465–499, Dec. 2020, doi: 10.1109/JPROC.2023.3247480.
[83] A. Ng, "Lecture notes sparse autoencoder," *CS294A*, vol. 72, pp. 1–19, Jan. 2011.
[84] X. Lu, S. Matsuda, C. Hori, and H. Kashioka, "Speech restoration based on deep learning autoencoder with layer-wised pretraining," in *Proc. Interspeech*, Sep. 2012, pp. 1504–1507.
[85] Z. Wang, X. Wang, X. Li, Q. Fu, and Y. Yan, "Oracle performance investigation of the ideal masks," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.
[86] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 9458–9465.
[87] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 764–768.
[88] S. Xia, H. Li, and X. Zhang, "Using optimal ratio mask as training target for supervised speech separation," 2017, *arXiv:1709.00917*.
[89] M. Hasannezhad, "'Speech enhancement with improved deep learning methods," Ph.D. thesis, Dept. Elect. Comput. Eng., Concordia Univ., Montreal, QC, Canada, 2021.
[90] T. Hasan and M. K. Hasan, "MMSE estimator for speech enhancement considering the constructive and destructive interference of noise," *IET Signal Process.*, vol. 4, no. 1, p. 1, 2010.
[91] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," *IEE Proc. Vis., Image, Signal Process.*, vol. 147, no. 3, pp. 247–253, 2000.
[92] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[93] 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, Washington, DC, USA, STI/Recon, Tech. Rep. 93, 1993.
[94] 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.
[95] (Jul. 2020). *Sounds Data*. [Online]. Available: https://freesound.org/
[96] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013.
[97] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.
[98] 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.*, Oct. 2001, pp. 749–752.
[99] 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, Sep. 2011.
[100] 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, Jan. 2008.
[101] 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 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 Iran University of Science and Technology, Tehran, in 2005, and the Ph.D. degree in speech recognition from 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.