IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 16, 2020, accepted August 26, 2020, date of publication September 1, 2020, date of current version September 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021061

On Learning Spectral Masking for Single Channel Speech Enhancement Using Feedforward and Recurrent Neural Networks

NASIR SALEEM[1,2], MUHAMMAD IRFAN KHATTAK[1], MUATH AL-HASAN[3], (Senior Member, IEEE), AND ABDUL BASEER QAZI[4], (Member, IEEE)

[1]Department of Electrical Engineering, University of Engineering and Technology, Peshawar 25000, Pakistan
[2]Department of Electrical Engineering, Faculty of Engineering and Technology (FET), Gomal University, Dera Ismail Khan 29050, Pakistan
[3]College of Engineering, Al Ain University, Al Ain, United Arab Emirates
[4]Department of Software Engineering, Bahria University, Islamabad 44000, Pakistan

Corresponding author: Nasir Saleem (nasirsaleem@gu.edu.pk)

This work was supported in part by the Abu-Dhabi Department of Education and Knowledge (ADEK) Award for Research Excellence 2019 under Grant AARE19-245.

ABSTRACT Human speech in real-world environments is typically degraded by the background noise. They have a negative impact on perceptual speech quality and intelligibility which causes performance degradation in various speech-related technological applications, such as hearing aids and automatic speech recognition systems. It also degrades the original phase of the clean speech and introduces perceptual disturbance which leads to the negative impacts on the quality of speech. Therefore, speech enhancement must vigilantly be dealt with in everyday listening environments. In this article, speech enhancement is performed using supervised learning of spectral masking. Deep neural networks (DNN) and recurrent neural networks (RNN) are trained to learn the spectral masking from the magnitude spectrograms of the degraded speech. An iterative procedure is adopted as a post-processing step to deal with the noisy phase. Additionally, an intelligibility improvement filter is also used to incorporate the critical band importance function weights where higher weights contribute more towards intelligibility. Systematic experiments demonstrated that the proposed approaches greatly attenuated the background noise. Also, they led to large improvements of the perceived speech quality and intelligibility, as well as automatic speech recognition. In experiments, TIMIT database is used. The STOI is improved by 17.6% over the noisy speech. Also, SDR and PESQ are improved by 5.22dB and 19% over the noisy speech utterances. These comparisons showed that the proposed speech enhancement approaches outperformed the related speech enhancement methods.

INDEX TERMS Deep neural network (DNN), recurrent neural network (RNN), speech enhancement, spectral masking, speech quality, speech intelligibility, supervised learning, background noise.

I. PROBLEM STATEMENT

Speech enhancement aims to improve the intelligibility and quality of the noisy speech. The conventional unsupervised speech enhancement method improves the quality but fail to improve the intelligibility in nonstationary background noises. Moreover, most of the speech enhancement methods use the noisy phase during reconstruction of the enhanced speech. It is vital in the various speech-related applications to design a robust method that has the ability to improve the speech intelligibility and quality as well as deal with the noisy phase for the better results. This article is based on the supervised learning of the spectral-masking for speech enhancement using DNN and RNN frameworks. Since, spectral phase has impacts on speech quality; we have used a post-processing step to deal with the noisy phase during time-domain speech recovery for improved quality.

II. INTRODUCTION

The objective of single-channel speech enhancement is to suppress the background noise components and recover the

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

components of clean speech from the noisy version with improved perceptual quality and intelligibility. The speech enhancement algorithms are primarily used to improve the voice quality of a real-time speech communication system, pre-recorded multimedia contents, to increase the accuracy of automatic speech recognition (ASR) systems and hearing aids. Previously, many unsupervised speech enhancement methods were suggested such as the spectral subtraction [1] and its variants [2], [3], Wiener filtering [2] and its variants [4], [5], as well as the minimum mean square error (MMSE) estimator [6] and its variants [7], [8]. Although the aforesaid speech enhancement methods are apt for many real-time speech-related applications since they present a small computational complexity, but their performance remains poor for many real-world acoustic environments where they fail to track the power spectral density of an extremely non-stationary background noise. To surmount this issue, the supervised learning-based speech enhancement methods have been opted and trained with a large quantity of the training data in presence of different background noises [9], [10]. Regression, spectral-mapping and spectral masking-based deep neural networks are among the most successful methods in single-channel speech enhancement tasks [11]–[15].

In a DNN-based speech enhancement task, the relation between input and target features is not linear; therefore, network architecture composed of multiple layers with non-linear activation functions are more suitable for speech enhancement [12] rather than shallow neural networks. In addition, to completely confine the temporal dynamics of the speech signal, feed-forward DNN and recurrent neural network structures have been opted. Particularly, single-channel speech enhancement-based on feedforward and recurrent neural networks have shown considerable performance gains compared to shallow neural networks and conventional unsupervised speech enhancement methods. Furthermore, network architecture types, training-targets and associated objective functions are vital concerns for deep learning-based speech enhancement [16]. The learning approaches for speech enhancement are grouped into two groups. In the first group, the training procedure is carried out in a direct spectral-mapping rule and the output clean spectral features are mapped from the input noisy spectral features; however, it is observed that the estimated spectra have a tendency to be over-smoothed [11], [12]. The second and successful group is that of spectral-masking. A number of learning approaches have recently been proposed for estimating spectral-masks with confirmed notable results [17]–[20]. Few of the recent related work regarding supervised speech enhancement is available in [21]–[23].

Spectral masking-based learning approaches map from a noisy speech signal to a time-frequency mask and the gain parameters are multiplied to the noisy magnitude spectra to obtain a noise suppressed enhanced speech signal. Spectral masking usually estimates the ideal binary mask (IBM) [24], where a time-frequency unit is assigned a binary 1, if the signal-to-noise ratio (SNR) within the unit exceeds a local criterion (0dB), implying speech dominance. Otherwise, a time-frequency unit is assigned a binary 0, implying noise dominance. Another popular spectral mask is the ideal ratio mask (IRM) [25], where a time-frequency unit is assigned a ratio of clean and noisy speech energies. The spectral magnitude mask, called ideal amplitude mask (IAM) is defined on the short-time Fourier transform magnitudes of clean speech and noisy speech. Unlike IRM, IAM is not upper-bounded by 1. To get enhanced speech, we apply the estimate of IAM to the spectral magnitudes of noisy speech, and resynthesize the enhanced speech. Gaussian mixture models (GMM) are used to learn the distribution of speech and noise dominant time-frequency units and for developing a Bayesian classifier for IBM estimation [26]. Multilayer perceptron is employed using one hidden layer to estimate the IBM which showed encouraging results in reverberant situations [27]. Support vector machines (SVM) are used to estimate time-frequency mask which delivered more factual classification results compared to the GMM-based classifiers [28]. For the first time, GMMs are used to compute posterior probabilities of speech dominance in time-frequency units and SVMs are trained with novel features to estimate the IBM [29]. The presented approach generalized significantly to an ample range of SNRs. Motivated by the deep hidden structure with several layers, DNN was used for the binary classification for the first time to separate a speech from the mixtures [30] and it significantly outperformed the earlier speech separation methods. A number of training-targets were examined and IRM was suggested to be preferred over IBM while dealing with speech quality [16]. DNN and RNN frameworks were used to minimize the reconstruction loss associated with the spectra of two premixed speakers by lodging IRM into the loss function [31], [32] and named as signal approximation [32]. The proposed method expressed substantial performance gain over NMF-based approaches. The signal approximation was deemed an optimization objective function and suggested long short-term memory (LSTM) into RNN architecture which outperformed DNN methods [33]. Signal approximation is further extended to the phase-sensitive mask and LSTM is used for speech denoising [33], [34]. Complex ideal ratio mask (cIRM) is proposed for speech enhancement. DNN-based cIRM learnt the real and imaginary parts of the complex spectra together instead of learning the magnitude spectra only [35]. This method significantly improved perceptual speech quality.

In this article, spectral masking-based learning approaches are used to construct three time-frequency masks: IRM, IBM and IAM using DNN and RNN architectures. During the training procedure, the mask approximation is used as a loss function. Critical band importance functions are used during training to further improve the performance of DNN and RNN architectures in terms of speech intelligibility and perceptual speech quality. Since background noise degrades the original phase of clean speech; therefore, it introduces perceptual disturbance which leads to negative impacts on

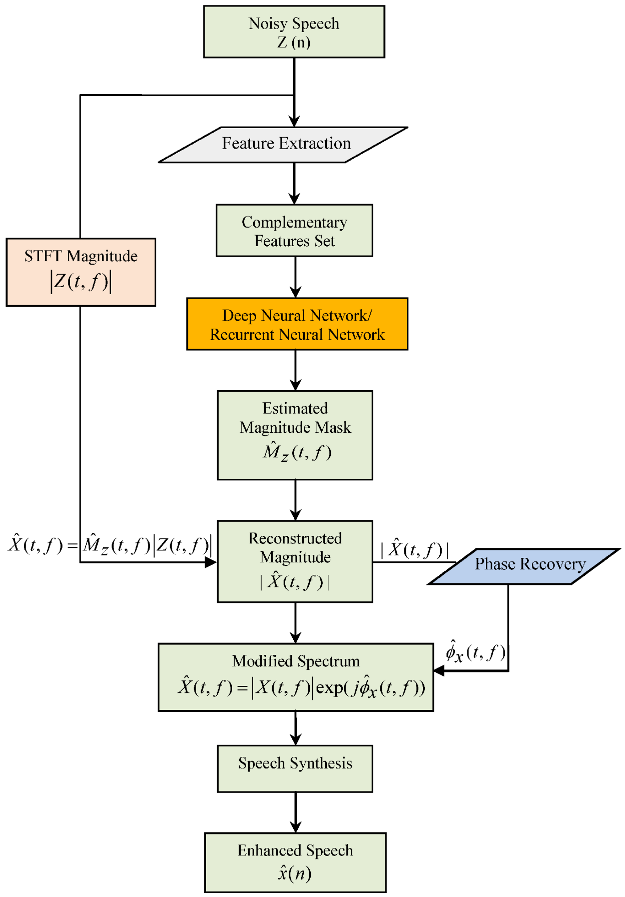

FIGURE 1. The flow diagram of the proposed speech enhancement.

speech. To avoid these degradations, an iterative procedure is adopted as a post-processing step. Figure 1 shows the flow diagram of the proposed speech enhancement method. The main contributions of this study are drawn as:

i. Spectral masking-based learning methods are developed using DNN and RNN architectures to enhance speech in noisy backgrounds which notably improve perceptual quality and intelligibility. In the proposed methods, we have constructed three time-frequency masks including IRM, IBM and IAM. In literature, we can find many DNN-based IRM, IBM and IAM construction, however; very few studies are available that have constructed such time-frequency masks using RNN frameworks.
ii. Critical band importance functions and their weights are used in the training procedures to further improve the perceptual quality and intelligibility of the noisy speech. The weights of the functions are directly applied to the clean training data using an intelligibility improvement filter and the testing process revealed enhanced speech with the filter.
iii. Most speech enhancement methods use the noisy phase for the reconstruction of the enhanced speech. We have addressed this vital problem in proposed method. We have adopted a widely used iterative procedure called as Griffin-Lim Algorithm (GLA) to deal with the noisy phase during time-domain speech reconstruction. This contribution has notably improved the performance of speech enhancement in terms of the perceptual speech quality and intelligibility.
iv. Less computational complexity and fast convergence is achieved by the proposed methods as compared to the baseline feedforward-DNN and RNN frameworks. The baseline and the proposed feedforward-DNNs have used same number of layers, quantity of the neurons in the hidden and visible layers. Similarly baseline and the proposed RNNs used same LSTM units. However, we achieved better speech quality and intelligibility. The reason for fast network convergence (less loss function) is adaptation of critical band weights which are directly applied to the clean training data.
v. Automatic speech recognition systems are usually tested with magnitude-only spectrums. Our proposed methods with both magnitude and phase processing improved the ASR performance in adverse noisy conditions.

The remaining paper is organized as follows. Spectral masking-based speech enhancement and loss functions are discussed in Section II. Experiments are presented in Section III. Results and analysis are presented in Section IV. Finally, the discussion and conclusions are presented in the Section V.

III. SPECTRAL MASKING-BASED SPEECH ENHANCEMENT AND LOSS FUNCTIONS

Mostly in speech enhancement, the enhancement of noisy speech $z(n)$ is performed in the time-frequency domain by applying the short time Fourier transform (STFT). Since the time-domain speech signal is a real-valued signal, and only considers $X = X(t,f) \in C^{L\times(K/2+1)}$, where L and K indicate the frame number and the size of discrete Fourier transform (DFT). In time-frequency domain, the magnitude spectrum of the enhanced speech signal $|\hat{X}(t,f)|$ can be achieved via following time-frequency masking procedure:

$$|\hat{X}(t,f)| = \hat{M}_x(t,f) \otimes |Z(t,f)| \tag{1}$$

where, $\hat{M}_x(t,f)$ denotes the intended time-frequency mask and $|Z(t,f)| = |X(t,f)| + |D(t,f)|$ denotes the magnitude of noisy speech which is the sum of the clean speech and noise signal in *t-th* frame and *f-th* frequency bin, respectively. We have used 20 ms with 75% overlapping in the proposed methods. The foundation for making 20 ms frame length comes from the quasistationarity assumption. We want the speech analysis frame to be stationary. As a result, in a too large analysis frame, the signal will become nonstationary. In supervised spectral masking-based learning tasks, the loss function is typically formulated to predict the masking parameters that can effectively restore the components of the clean speech by suppressing the undesired background noise

TABLE 1. Various time-frequency masks with dynamic ranges.

Time-Frequency Mask Type	Mathematical Formula	Dynamic Range
Ideal Amplitude Mask (IAM)	$M_x^{IAM}(t,f) = \lvert X(t,f)\rvert / \lvert Z(t,f)\rvert$	$R \geq 0$
Ideal Ratio Mask (IRM)	$M_x^{IRM}(t,f) = \left(\lvert X(t,f)\rvert^2 / \lvert X(t,f)\rvert^2 + \lvert D(t,f)\rvert^2\right)^{\alpha}$	$[0,1]$
Ideal Binary Mask (IBM)	$M_x^{IBM}(t,f) = \begin{cases} 1, \mathrm{SNR}_{t,f} \geq \kappa \\ 0, \mathrm{Otherwise} \end{cases}$	$\{0,1\}$

components in all time-frequency units. The time-domain enhanced speech is then recovered by applying the inverse STFT (*i*STFT), as illustrated in Fig. 2. A different approach, called the spectral mapping, directly learns the mapping rule from the spectral features of noisy speech to clean speech. But, spectral-masking is identified to be more successful than spectral mapping since the time-frequency mask typically has a bounded dynamic range; hence, it achieves faster convergence speed [16]. In deep learning, there exist many approaches to learn a time-frequency mask depending on the training-target or the optimization-domain. In mask approximation, the time-frequency mask is estimated such that the mean square error (MSE) with the predefined time-frequency mask is minimized, is given as:

$$MSE_{MA} = \frac{1}{2K}\sum_{t=1}^{K-1}\left[(M_x(t,f) - \hat{M}_x(t,f))^2\right] \tag{2}$$

where $\hat{M}_x$ and M_x denote the estimated and the predefined reference time-frequency mask, respectively. The time-frequency mask can be derived in various forms, as given in Table 1. Signal approximation is an alternative approach introduced in a study [32]. In this approximation approach, the time-frequency mask is estimated in such a way that the estimated speech is closest to the reference clean speech. Magnitude spectra approximation [32] is a kind of signal approximation in which the optimization is achieved in the magnitude spectra domain, given as:

$$MSE_{MSA} = \frac{1}{2K}\sum_{t=1}^{K-1}\left[|Z|^2 (M_x(t,f) - \hat{M}_x(t,f))^2\right] \tag{3}$$

where e indicates element-wise, Hadamard product.

A. ACOUSTIC FEATURES

A set of acoustic features is extracted from the input speech at frame level where frame length and frame shift are set to 20 ms and 10 ms, respectively [16]–[18]. The set of acoustic features contains 15-D Amplitude Modulation Spectrogram (AMS), 31-D Mel-Frequency Cepstral Coefficients (MFCC), 13-D Relative Spectral Transformed Perceptual Linear Prediction Coefficients (RASTA-PLP) and 64-D Gammatone Filter-bank Energies (GFE). The GFE features are extracted from the Cochleagram, a time-frequency representation typically used in computational auditory scene analysis (CASA) [36]. Cochleagram representation describes the working of the human auditory system. Cochleagram is computed by using a 64-channel gammatone filterbank. Additionally, delta and double-delta feature coefficients are calculated and appended with all raw acoustic features. Acoustic feature extraction has been done by using RASTAMAT toolbox. In spectral-domain, each frame can be represented as a vector, given as:

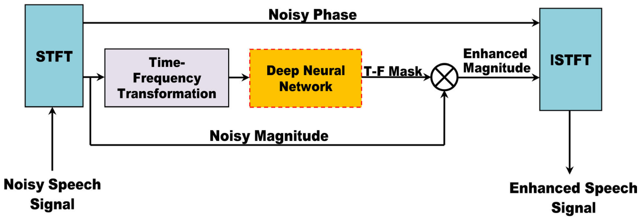

FIGURE 2. Spectral-masking based deep neural network for single-channel speech enhancement.

$$x(t) = [X(t,1), X(t,2), X(t,3), \ldots., X(t,N)]^T \tag{4}$$

An auto-regressive moving average filter (second order) is used to flatten the temporal trajectories of acoustic features as it improves the speech enhancement performance [16]. To include the temporal information, a context window of two prior and two future frames are used, hence resulting in 1845-D (369-D × 5 = 1845-D) feature vector. The feature vector before applying to neural network is given as:

$$\tilde{x}(t) = [x(t-d), \ldots., x(t), \ldots..., X(t+d)]^T \tag{5}$$

where, d denotes the neighboring frames on each side and

T denotes transpose operator. Zero mean and unit variance normalization have been applied to all feature vectors before applying to train neural networks. The acoustic features extraction procedure is illustrated in Fig. 3.

B. NETWORK ARCHITECTURES

In this study, feedforward-DNN and RNN networks are used as spectral-masking learning approaches. Afterward, feedforward-DNN will be denoted by DNN. The network architectures of both networks are described in this section.

DNNs are selective learning machines and have shown to perform exceptionally well in the speech enhancement task [37]–[41]. The DNN architecture consists of five layers; an input layer, three hidden layers, and an output layer.

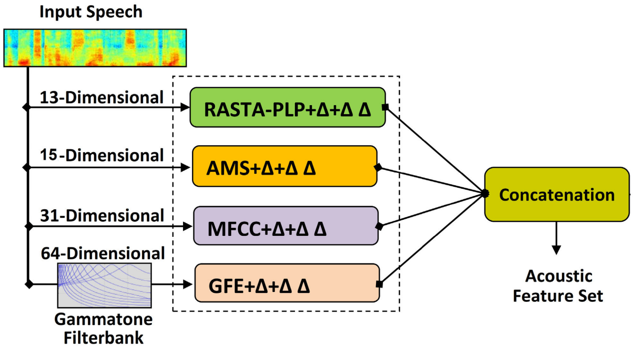

FIGURE 3. The procedure of acoustic features extraction.

The size of the input layer is 1845 units, that is, 246*5 = 1230, including 246-D acoustic features and features window of 5 frames. Each hidden layer consists of 1024 hidden units and the output layer contains 257 visible units. From the input layer to output layer, the proposed DNN has 5 × 246, 1024, 1024, 1024 and 257 units, respectively. Backpropagation and dropout regularization [42], [43] are used in training. The adaptive gradient descent algorithm [44] with a momentum parameter γ is used to optimize DNN. 512 samples batch size is used. The scaling factor for adaptive gradient descent is set to 0.0010 and the learning rate is reduced linearly from 0.06 to 0.002. 100 epochs are used during the process. For the first few epochs, the γ is fixed at 0.5 and the rate is increased to 0.8 for remaining epochs. The MSE loss function using mask approximation is used. The loss optimization curves at 100 epochs are shown in Fig. 4. The rectified linear unit (ReLU) activation converts a weighted sum of the inputs to a model neuron to the neuron's output. Recent practice shows that a moderately deep MLP with ReLU can be effectively trained with the large training data without unsupervised pretraining. Therefore, ReLU is used as activation function in all hidden layers and sigmoid activation function is used in the output layer. The reason for selecting the sigmoid as output activation function is the dynamic range [0 1]. It is used for models where to predict output probabilities, since probability exists between 0 and 1. Also, the dynamic range of T-F masks exists between 0 and 1. This function is differentiable and a slope of sigmoid curve is obtainable at any two pints. The sigmoid function is monotonic but its derivative is not. The functions are:

$$\begin{aligned} ReLU : f(k) &= \max(0, k) \\ \text{Sigmoid} : f(k) &= \frac{1}{1+e^{k}} \end{aligned} \tag{6}$$

On the other hand, the RNN architecture contains an input layer, three LSTM layers consists of 256 hidden units, and a fully connected output layer with 64 sigmoid units. The Adaptive Gradient Descent (AGD) algorithm is adopted for network training. The learning rate, epochs and batch-size are set to 0.001, 100 and 1024, respectively. The AGD is adopted to minimize loss function. From the input layer to output layer, the proposed RNN has 5 x 246, 256, 256, 256 and 257 units, respectively. The training procedures for feedforward-DNN and RNN are illustrated in Fig. 5. Both learning approaches are used in this study to estimate three masks: IRM, IBM, and IAM respectively.

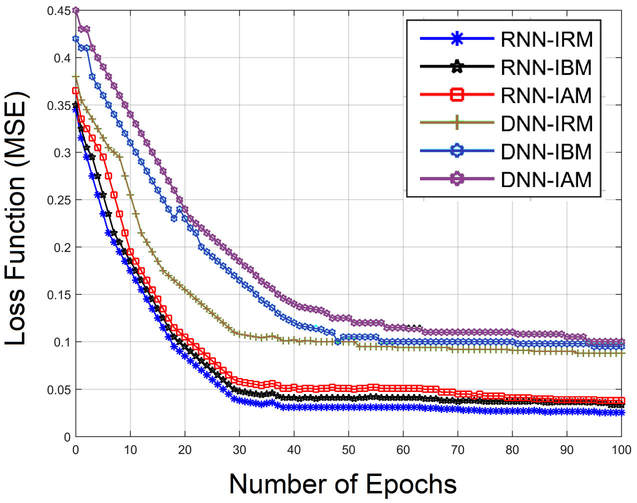

FIGURE 4. Loss optimization curves using mask approximation-based MSE.

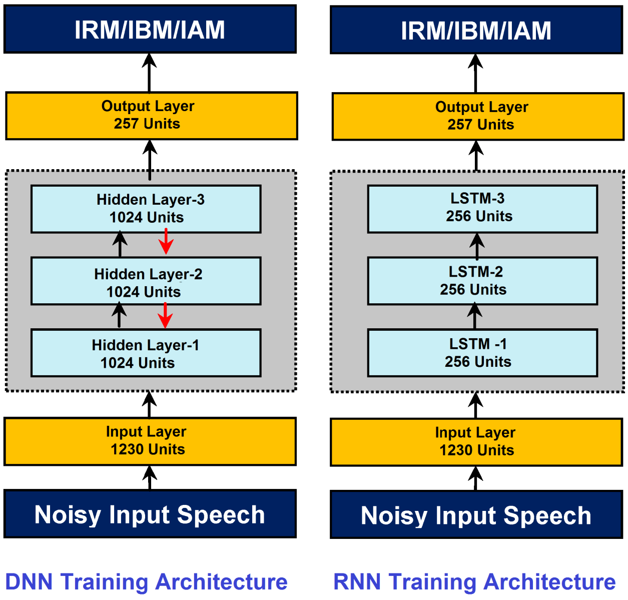

FIGURE 5. RNN and DNN training architectures.

C. INTELLIGIBILITY IMPROVEMENT FILTER BASED ON THE CRITICAL BAND IMPORTANCE FUNCTIONS

Critical band importance functions refer to the American National Standards Institute (ANSI) S3.5 standard [45]. They indicate that the higher weights of band importance functions contribute more towards improving the speech intelligibility of the noisy speech. There are 21 frequency bands, given in Table 2. The values of the band importance

TABLE 2. Critical band importance functions (CBIF) in various frequency bands.

Frequency Bands	100-200	200-300	300-400
CBIF Weights	0.010	0.026	0.041
Frequency Bands	400-510	510-630	630-770
CBIF Weights	0.057	0.057	0.057
Frequency Bands	770-920	1480-1720	1720-2000
CBIF Weights	0.057	0.057	0.057
Frequency Bands	2000-2320	2320-2700	4400-5300
CBIF Weights	0.057	0.057	0.046
Frequency Bands	5300-6400	6400-7700	7700-9500
CBIF Weights	0.034	0.023	0.011

functions for the frequency bands indicate their impacts on intelligibility. Based on critical band importance functions, an intelligibility-improvement filter (IIF) formulated and applied to the clean training waveforms. The weights are multiplied to the training data in order to further improve the intelligibility given as:

$$\bar{X}_t^F(t,f) = IIF\left(X_t^F(t,f)\right) \tag{7}$$

$$IIF\left(X_t^F(t,f)\right) = \sqrt{\frac{\sum\limits_{t=1}^{T}\sum\limits_{f}^{F}(X_t^F(t,f))^2}{\sum\limits_{t=1}^{T}\sum\limits_{f}^{F}\left(\alpha(n)X_t^F(t,f)\right)^2}}\alpha(t)X_t^F(t,f) \tag{8}$$

where, $\bar{X}_t^F(t,f)$ show filtered speech waveform, T is total number of speech frames and $\alpha(t)$ shows filter coefficients:

$$\alpha(t) = \lambda^{(M)}, \quad \text{when} \quad t \in \left[f_L^{(M)}, f_H^{(M)}\right] \tag{9}$$

where $f_L^{(M)}, f_H^{(M)}$ denotes lower and higher bounds whereas $\lambda^{(M)}$ represents the weights in M-th frequency band. The IIF filter is directly applied to the clean waveforms and the filtered waveforms are used as training data. In testing, the neural networks generate the enhanced speech waveforms by incorporating the effects of IIF filter.

D. ITERATIVE TIME-DOMAIN SPEECH RECOVERY

After generating the estimate of magnitude spectra by the neural networks, time-domain speech signals are recovered by using inverse STFT (*i*STFT). One approach to recover the time-domain signals is to apply *i*STFT using estimated magnitude of neural network and the phase of time-domain noisy speech waveforms. However, background noise also degrades the phase of the clean speech, and this degradation typically produces perceptual disturbances and has negative impacts associated to the speech quality. Moreover, Fourier transforms of overlapping speech frames are concatenated and then STFT is computed which is a redundant version of the time-domain signal. The magnitude spectrograms of the recovered time-domain speech signal perhaps different from the intended recovered signal [46], [47]. This difference is considered for neural network estimated magnitudes. To reduce mismatch between the magnitude and phase from which we prefer to recover a time-domain speech signal, an iterative procedure is adopted [46] in order to recover the time-domain speech signal, given as Algorithm 1. In this algorithm the phase is updated iteratively at every step and replaces it with phase of STFT of its *i*STFT, whereas the estimated magnitude of neural network output always remains the fixed. The algorithm acquires as input is the estimated magnitudes from DNN/RNN outputs that need to be reconstructed. The phases are not known and need to be solved for reconstructing the estimate of the original signal. The iterations identify the closest achievable magnitude spectrogram consistent with given magnitude spectrogram. GLA is an accepted phase recovery algorithm which is based on the spectrogram consistency [43]. It recovers a complex-valued spectrogram, which remains consistent and also retains given magnitude by a projection procedure. The GLA is obtained for an optimization problem [46].

Algorithm 1 Iterative Time-Domain Speech Recovering

Input: Output DNN/RNN Magnitude $\hat{X}^0$ and Noisy Phase ϕ^0 and Number of iterations K
Output: Time-domain Speech Signal $\hat{x}$

```
$\hat{X} \leftarrow \hat{X}^0, \phi \leftarrow \phi^0, k \leftarrow 1$
While $k \leq K$ do
    $\hat{x}^k \leftarrow i\mathrm{STFT}(\hat{X}, \phi)$
    $(\hat{X}^k, \phi^k) \leftarrow \mathrm{STFT}(\hat{x}^k)$
    $\hat{X} \leftarrow \hat{X}^0$
    $\phi \leftarrow \phi^0$
    $k \leftarrow k + 1$
end While
$\hat{x} \leftarrow \hat{x}^K$
```

IV. EXPERIMENTS

A. DATASET

The experiments are performed on a speech dataset that is produced from the TIMIT database [48]. In order to access the performance of processing methods in various noisy backgrounds, 15 different noise types are selected from the NOISEX-92 [49] and Aurora-4 [50] databases, as given in Table 3. To create noisy speech signals, three signal-to-noise (SNR) levels are used that ranged from -3dB to 3dB with a 3dB step size. For training, 2000 speech utterances from 100 different speakers of both genders are reproduced for each time-frequency mask (three times) for each SNR level, and mixed with the 15 noise types. Hence, a total of 18000 speech utterances (about 15 hours training data) are used during training. Moreover, 800 speech utterances from 30 different speakers are used for the testing purpose. To evaluate the processing methods, 150 speech utterances from 16 speakers of both genders are used at random. All noise sources are used

TABLE 3. Background noise types (N1-N15).

N1: Babble Noise, N2: Airport Noise, N3: Factory Noise
N4: Car Noise, N5: Destroyerengine Noise, N6: F16
N7: Buccaneer, N8: Destroyerops Noise, N9: Café Noise
N10: Street Noise, N11: Pink Noise, N12: Volvo Noise
N13: White Noise, N14: HF Channel Noise, N15: Hall

TABLE 4. Neural networks architecture settings.

Networks	T-F Mask	Loss	Hidden Units/ ReLU	Visible Units/ Sigmoid	Post Processing
RNN-IAM	IAM	L_{MSE}	1536	512	Speech Recovery And CBIF
RNN-IRM	IRM	L_{MSE}	1536	512	
RNN-IBM	IBM	L_{MSE}	1536	512	
DNN-IAM	IAM	L_{MSE}	3072	257	Speech Recovery And CBIF
DNN-IRM	IRM	L_{MSE}	3072	257	
DNN-IBM	IBM	L_{MSE}	3072	257	
RNN_B-IAM	IAM	L_{MSE}	1536	512	No Speech Recovery And CBIF
RNN_B-IRM	IRM	L_{MSE}	1536	512	
RNN_B-IBM	IBM	L_{MSE}	1536	512	
DNN_B-IAM	IAM	L_{MSE}	3072	257	No Speech Recovery And CBIF
DNN_B-IRM	IRM	L_{MSE}	3072	257	
DNN_B-IBM	IBM	L_{MSE}	3072	257	

in training and testing. The results are averaged over 15 noise types.

B. EVALUATION METRICS AND PARAMETERS

We quantitatively evaluated various spectral masking-based speech enhancement methods by four objective measures. Short-time objective intelligibility (STOI) and the extended STOI (ESTOI) are used as intelligibility indicators whereas perceptual evaluation of speech quality (PESQ) and signal-to-distortion ratio (SDR) are used as the quality indicators, respectively. PESQ [51], an ITU-T P.862 recommendation predicts the perceptual quality of the enhanced speech by giving an output value ranged from 0.5 to 4.5, where a high value implies better speech quality. SDR [52] measures the speech quality. STOI [53] predicts the intelligibility of the enhanced speech by providing an output value ranged from 0 to 1 and a high value implies better speech intelligibility. The STOI values are based on the correlation between clean and the enhanced speech signals in short-time overlapped segments. ESTOI [54] predicts intelligibility of enhanced speech by providing an output value ranged from 0 to 1.

C. SYSTEM REPRESENTATION

Based on the time-frequency masks and learning methods, various deep spectral masking-based speech enhancement methods are realized, given in Table 4. To express all the speech enhancement methods, an interpretation is followed: (< Neural Network >-< Mask Type >-< Post Processing >). A speech enhancement method "**DNN-IRM**" indicates that the feedforward DNN is used with IRM as a time-frequency mask and used *no* iterative time-domain speech recovery and IIF filter. Also, a speech enhancement method "**RNN-IBM**" indicates that recurrent neural network is used with IBM as a time-frequency mask and used *no* iterative time-domain speech recovery and IIF filter. Similarly, a speech enhancement method "**DNN-IRM-*i*SR**" indicates that the feedforward DNN is used with IRM as a time-frequency mask and used iterative time-domain speech recovery and IIF filter. Finally, a speech enhancement method "**RNN-IBM-*i*SR**" indicates that recurrent neural network is used with IBM as a time-frequency mask and used iterative time-domain speech recovery and IIF filter. The baseline deep networks are represented as $\mathbf{DNN_B}$ and $\mathbf{RNN_B}$, respectively. All neural networks are trained with same training dataset. Intel Core i7-3210M 3.2GHz processor and Nvidia GTX 950 GPU are used to conduct all the experiments.

V. RESULTS AND ANALYSIS

In this section, we discussed the main findings of this study. We first subjectively compared spectral-masking methods with time-frequency masks without iterative time-domain speech recovery algorithm and intelligibility improvement filter. Secondly, we compared the proposed RNN, DNN and related speech enhancement methods. Thirdly, we evaluated the speech recognition performance of the proposed method. We finally conducted subjective listening tests to further evaluate the proposed method in terms of the speech quality and intelligibility.

A. OBJECTIVE EVALUATION

We report the detailed comparison results for three noise types on the TIMIT database of both genders in Table 5 and Table 6 respectively, where mask approximation was used as the training loss function of all neural network models. From the Tables, we observed that the spectral masking-based methods with iterative time-domain speech recovery and the intelligibility improvement filter performed better when applied with RNN and DNN frameworks. The time-frequency masks with the speech recovery and IIF filter improved the speech quality and intelligibility over their counterparts and unprocessed noisy speech. Explicitly, RNN-based learning spectral masking outperformed DNN-based spectral masking. For example in Table 5 at −3dB babble noise, RNN-IRM-*i*SR improved the STOI by 17.6% over the noisy speech and by 1.25% over the RNN-IRM counterpart, respectively. Similarly, at −3dB factory noise, RNN-IAM-*i*SR improved the STOI by 22.38% over the noisy speech and by 1.08% over RNN-IAM. In addition, RNN-IBM-*i*SR improved the ESTOI and SDR by 1.58% and 3.85% over RNN-IBM at −3dB factory noise. In the same way, RNN-IRM-*i*SR, RNN-IBM-*i*SR and RNN-IAM-*i*SR improved the PESQ at −3dB white noise by factors 0.85, 0.81 and 0.86 over the noisy speech signal whereas improved the PESQ by 2%, 2.04% and 1.01%

TABLE 5. Performance evaluation of RNN frameworks for three SNR levels using TIMIT corpus in three example noisy backgrounds.

Methods	Babble Noise				Factory Noise				White Noise			
	-3dB				-3dB				-3dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	61.9	26.3	-2.81	1.29	61.2	25.5	-2.77	1.27	69.9	33.66	-2.73	1.19
RNN-IRM-*i*SR	72.8	45.0	4.09	1.64	75.31	46.01	5.43	1.67	82.1	57.71	10.27	2.04
RNN-IBM-*i*SR	72.3	44.8	4.05	1.61	74.81	45.71	5.52	1.64	81.2	57.00	9.63	2.00
RNN-IAM-*i*SR	73.5	46.1	4.02	1.60	74.91	45.11	5.40	1.60	82.0	57.48	10.30	2.05
$\mathrm{RNN_B}$-IRM	71.9	43.5	3.93	1.58	74.21	44.61	5.34	1.61	80.9	56.49	9.05	2.00
$\mathrm{RNN_B}$-IBM	71.8	44.1	3.90	1.57	74.01	45.21	5.44	1.58	80.0	55.78	8.41	1.96
$\mathrm{RNN_B}$-IAM	72.9	45.2	3.92	1.54	74.11	43.21	5.27	1.52	80.8	56.26	9.08	1.98
Methods	0dB				0dB				0dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	68.9	34.5	0.13	1.51	68.0	34.0	0.15	1.49	74.6	40.1	0.19	1.29
RNN-IRM-*i*SR	81.7	57.6	6.58	2.00	82.3	56.6	7.68	1.99	86.6	65.9	12.2	2.33
RNN-IBM-*i*SR	80.9	56.8	6.41	1.97	81.4	56.7	8.44	1.96	86.2	65.1	12.1	2.28
RNN-IAM-*i*SR	81.6	57.4	6.46	1.94	82.4	56.0	7.57	1.97	86.6	65.5	12.2	2.30
$\mathrm{RNN_B}$-IRM	80.5	56.4	5.36	1.94	81.1	55.3	6.46	1.91	85.4	64.7	11.0	2.29
$\mathrm{RNN_B}$-IBM	79.7	55.6	5.18	1.88	80.2	55.5	7.22	1.87	85.0	63.8	10.9	2.21
$\mathrm{RNN_B}$-IAM	80.4	56.1	5.24	1.87	81.1	54.8	6.35	1.90	85.3	64.3	10.9	2.22
Methods	3dB				3dB				3dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	75.9	43.6	3.09	1.86	78.8	43.3	3.11	1.65	79.0	46.6	3.13	1.45
RNN-IRM-*i*SR	87.9	68.1	8.94	2.31	87.7	66.5	9.91	2.23	90.4	73.13	14.2	2.58
RNN-IBM-*i*SR	87.1	66.6	8.71	2.14	87.0	66.1	9.87	2.11	90.0	71.63	14.1	2.46
RNN-IAM-*i*SR	87.7	67.4	8.89	2.32	87.7	66.3	9.9	2.20	90.3	72.83	14.1	2.51
$\mathrm{RNN_B}$-IRM	86.7	66.9	7.72	2.22	86.5	65.3	8.69	2.18	89.2	71.93	13.0	2.52
$\mathrm{RNN_B}$-IBM	85.9	65.4	7.49	2.10	85.8	64.9	8.65	2.07	88.8	70.43	12.9	2.40
$\mathrm{RNN_B}$-IAM	86.5	66.2	7.67	2.24	86.5	65.1	8.68	2.16	89.1	71.63	12.9	2.47

over RNN-IRM, RNN-IBM and RNN-IAM, respectively. The STOI, ESTOI, SDR and PESQ performance gains are higher in the nonvocal noisy backgrounds, i.e. factory, and white noise than vocal babble noise. On the other hand, DNN-based spectral masking after incorporating the IIF filter and time-domain speech recovery performed better compare to the counterparts. For example in Table 6 at 0dB babble noise, DNN-IRM-*i*SR improved the STOI, SDR and PESQ by 16.83%, 5.22dB and 19% over noisy speech utterances. Similarly, DNN-IRM-*i*SR, DNN-IBM-*i*SR and DNN-IAM-*i*SR improved STOI and PESQ by 1.51%, 1.52% and 1.51% over DNN-IRM, DNN-IBM and DNN-IAM counterparts, respectively. The average improvements in values of the STOI, ESTOI, SDR and PESQ are given in Figure 6.

We separately compared the performance gains between time-frequency masks generated by neural networks-based spectral masking methods. The results of the comparative study processed by three time-frequency mask based on RNN and DNN are given in Table 7. In terms of the STOI and ESTOI, IAM-*i*SR performed better than IRM-*i*SR and IBM-*i*SR. Similarly, in terms of the SDR and PESQ, IRM-*i*SR performed better than IAM-*i*SR and IBM-*i*SR. For example, RNN-IAM-*i*SR improved the STOI by 16.40% over noisy speech as compare to RNN-IRM-*i*SR

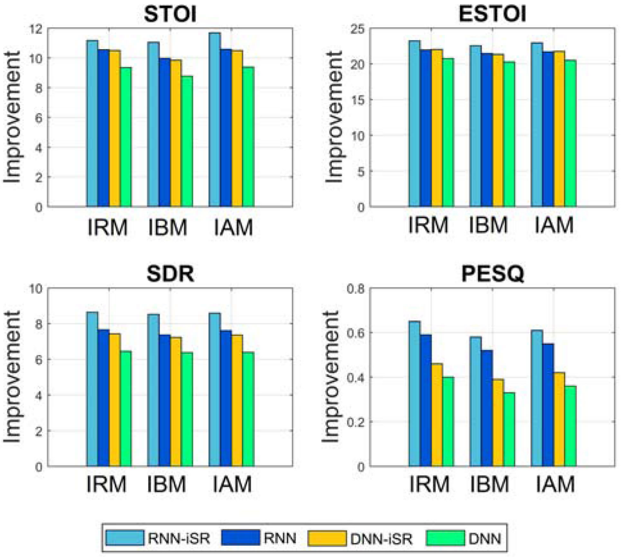

FIGURE 6. STOI, ESTOI, SDR and PESQ improvements.

TABLE 6. Performance evaluation of DNN frameworks for three SNR Levels using TIMIT corpus in three example noisy backgrounds.

Methods	Babble Noise				Factory Noise				White Noise			
	-3dB				-3dB				-3dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	61.9	26.3	-2.81	1.29	61.2	25.5	-2.77	1.27	69.9	33.66	-2.73	1.19
DNN-IRM-*i*SR	71.6	43.8	2.88	1.45	74.1	44.8	4.22	1.48	80.9	56.50	9.06	1.85
DNN-IBM-*i*SR	71.1	43.6	2.84	1.42	73.6	44.5	4.31	1.45	80.0	55.79	8.42	1.81
DNN-IAM-*i*SR	72.3	44.9	2.81	1.41	73.7	43.9	4.19	1.41	80.8	56.27	9.09	1.86
DNN$_{B}$-IRM	70.7	42.3	2.72	1.39	73.0	43.4	4.13	1.42	79.7	55.28	7.84	1.81
DNN$_{B}$-IBM	70.6	42.9	2.69	1.38	72.8	44.0	4.23	1.39	78.8	54.57	7.20	1.77
DNN$_{B}$-IAM	71.7	44.0	2.71	1.35	72.9	42.0	4.06	1.33	79.6	55.05	7.87	1.79
Methods	0dB				0dB				0dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	68.9	34.5	0.13	1.51	68.0	34.0	0.15	1.49	74.6	40.1	0.191	1.29
DNN-IRM-*i*SR	80.5	56.4	5.35	1.81	81.1	55.4	6.45	1.80	85.4	64.7	11.00	2.14
DNN-IBM-*i*SR	79.7	55.6	5.17	1.78	80.2	55.5	7.21	1.77	85.0	63.9	10.86	2.09
DNN-IAM-*i*SR	80.4	56.2	5.23	1.75	81.2	54.8	6.34	1.78	85.4	64.3	10.93	2.11
DNN$_{B}$-IRM	79.3	55.2	4.13	1.75	79.9	54.1	5.23	1.72	84.2	63.5	9.78	2.10
DNN$_{B}$-IBM	78.5	54.4	3.95	1.69	79.0	54.3	5.99	1.68	83.8	62.6	9.64	2.02
DNN$_{B}$-IAM	79.2	54.9	4.01	1.68	79.9	53.6	5.12	1.71	84.1	63.1	9.71	2.03
Methods	3dB				3dB				3dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	75.9	43.6	3.09	1.86	78.8	43.3	3.11	1.65	79.0	46.6	3.132	1.45
DNN-IRM-*i*SR	86.7	66.9	7.71	2.12	86.5	65.3	8.68	2.04	89.2	71.9	13.01	2.39
DNN-IBM-*i*SR	85.9	65.4	7.48	1.95	85.8	64.9	8.64	1.92	88.8	70.4	12.84	2.27
DNN-IAM-*i*SR	86.5	66.2	7.66	2.13	86.5	65.1	8.67	2.01	89.1	71.6	12.88	2.32
DNN$_{B}$-IRM	85.5	65.7	6.49	2.03	85.3	64.1	7.46	1.99	88.0	70.7	11.79	2.33
DNN$_{B}$-IBM	84.7	64.2	6.26	1.91	84.6	63.7	7.42	1.88	87.6	69.2	11.62	2.21
DNN$_{B}$-IAM	85.3	65.0	6.44	2.05	85.3	63.9	7.45	1.97	87.9	70.4	11.66	2.28

TABLE 7. Average comparison performance evaluation at all noise types and three SNR levels using TIMIT Corpus for both Genders.

Methods	RNN				DNN			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	71.27	36.39	0.17	1.44	71.27	36.39	0.17	1.44
IRM-*i*SR	82.95	59.61	8.81	2.09	81.77	58.41	7.60	1.90
IBM-*i*SR	82.32	58.93	7.69	2.02	81.12	57.73	7.40	1.83
IAM-*i*SR	82.96	59.34	8.76	2.05	81.76	58.14	7.53	1.86

and RNN-IBM-*i*SR that improved the STOI by 16.38% and 15.5% over noisy speech, respectively. In addition, DNN-IRM-*i*SR improved the PESQ by a factor 0.46 over noisy speech as compare to RNN-IBM-*i*SR and RNN-IAM-*i*SR structures that improved the PESQ values by the factor 0.39 and 0.42, respectively. Time-varying spectrogram graphically shows and analyzes the important speech patterns over the time at various frequency bands. To visualize and compare the performance of the speech enhancement for both RNN and DNN, spectrograms of the clean and noisy speech samples as well as for enhanced speech signal are plotted in Fig. 7. For clear understandings, STOI, SDR and PESQ values of the speech utterances are mentioned over the spectrograms. It is evident that both RNN-*i*SR and DNN-*i*SR successfully reduced the background noise components, and RNN-*i*SR provides a better recovered speech signal compare to RNN.

Also, DNN-*i*SR provides a better recovered speech signal than DNN. To visualize impacts of the phase recovery and IIF in the proposed speech enhancement, spectrograms of the clean and noisy speech samples as well as for the phase recovered-only, RNN output with phase recovery and RNN output with IIF filter effects are plotted in Fig. 8. Phase recovery and integration of IIF filter significantly improved the speech quality and indelibility.

TABLE 8. Average performance evaluation at all noise types and three SNR levels using TIMIT corpus for both genders against various competing speech enhancement methods.

Processing Methods	-3dB				0dB				3dB			
	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ	STOI	ESTOI	SDR	PESQ
Noisy	64.3	28.5	-2.7	1.25	70.5	36.2	0.16	1.43	77.9	44.5	3.11	1.65
RNN-*i*SR	76.5	49.4	6.52	1.76	83.3	59.7	8.84	2.08	88.4	74.6	10.9	2.32
DNN-*i*SR	75.3	48.2	5.31	1.57	82.1	58.3	7.61	1.89	87.2	67.5	9.73	2.13
RNN_B	75.6	48.2	6.03	1.71	82.0	58.5	7.62	2.01	87.3	67.5	9.74	2.26
DNN_B	74.4	47.1	4.82	1.52	80.9	57.3	6.40	1.82	86.0	66.3	8.51	2.07
NMF	67.0	30.6	2.76	1.39	72.5	36.8	3.93	1.48	76.6	40.5	5.82	1.56
NNDS	68.5	36.5	2.53	1.32	74.4	48.9	4.55	1.64	80.6	60.1	7.12	1.95
NRPCA	66.8	34.0	2.32	1.37	73.3	46.9	3.88	1.61	77.2	50.2	6.73	1.91
LMMSE	67.2	38.6	2.39	1.41	73.6	47.0	3.92	1.67	79.0	52.3	6.77	1.97
DDAE	73.1	45.2	4.21	1.49	78.7	56.2	6.03	1.79	83.1	63.2	7.98	2.01

TABLE 9. Output SNR, Δ SNR and SSNR performance at three input SNRs.

Processing Methods	-3dB			0dB			3dB		
	SNR_O	ΔSNR	SSNR	SNR_O	ΔSNR	SSNR	SNR_O	ΔSNR	SSNR
Noisy	-0.92	2.08	0.97	0.79	0.79	1.58	3.41	0.41	2.52
RNN-*i*SR	6.04	9.04	4.22	7.51	7.51	5.09	9.39	6.39	6.33
DNN-*i*SR	4.89	7.89	3.70	6.44	6.44	4.59	8.66	5.66	5.88
RNN_B	5.82	8.82	3.91	7.16	7.16	4.83	9.02	6.02	6.10
DNN_B	5.76	8.76	3.66	6.23	6.23	4.34	8.17	5.17	5.54
NMF	0.13	3.13	0.43	3.11	3.11	2.92	5.12	2.12	3.10
NNDS	1.74	4.74	1.48	4.32	4.32	4.07	6.51	3.51	4.77
NRPCA	1.64	4.64	1.31	4.13	4.13	3.98	6.22	3.22	4.33
LMMSE	1.13	4.13	0.98	3.89	3.89	3.44	5.76	2.76	4.01
DDAE	5.76	8.76	3.11	6.92	6.92	4.11	8.59	5.59	5.22

TABLE 10. Time complexity of DNN and RNN training.

Operation	Proposed Networks
Forward-Backward Propagation	$O\left(N_D N_E \left(U_I + N_H + 2N_H^2 + N_H N_O\right)\right)$

Network Architectures	DNN=[1230 1024 1024 1024 257]
	RNN=[1230 256 256 256 257]

Average MSE at 100 Epochs	DNN=0.0951 (Approx)
	RNN=0.0366 (Approx)

B. COMPARISON WITH RELATED METHODS

Additionally, the spectral-masking learning methods are compared to various state-of-the-art speech enhancement methods including deep neural network (DNN) [16], recurrent neural network (RNN) [31], non-negative matrix factorization (NMF) [10], non-negative dynamical system (NNDS) [55], robust principle component analysis (RPCA) [56], log-minimum mean square error (LMMSE) [7] and deep denoising autoencoder (DDAE) [57] in order to confirm the performance of the proposed speech enhancement method. It is evident that both learning methods have attained a significant improvement over the competing methods, with improved PESQ, STOI, ESTOI and SDR values. The intelligibility and quality values of NNDS are consistently greater than NMF and RPCA-based methods. The results in Table 8 demonstrated that proposed RNN-*i*SR and DNN-*i*SR outscored their counterparts, RNN and DNN, as well as other competing methods, NMF, NNDS, RPCA, LMMSE, and DDAE with reasonable margins. For example, the STOI values are improved from 67% with NMF at −3dB noise to 76.5% with RNN-*i*SR and improved STOI by 14.18%. Similarly, the PESQ values are improved from 1.64 with NNDS at 0dB noise to 1.89 with DNN-*i*SR and improved the PESQ by 15.24%. Likewise, The SDR values are improved from the 6.73 with RPCA at 0dB noise to 10.90 with RNN-*i*SR and improved the SDR by 4.17 dB.

Table 9 shows the results in terms of the output SNR (SNR_O), improvement in overall SNR (ΔSNR), and Segmental SNR (SSNR), respectively. The SSNR is used to measure the residual noise in the enhanced speech signals. The proposed speech enhancement methods, RNN-*i*SR and DNN-*i*SR significantly improved the SNR_O and achieved a significant gain in SNR_O. The overall ΔSNRs for RNN-*i*SR and DNN-*i*SR are higher than the competing state-of-the-art methods. Similarly, the consistent SSNR values indicate that proposed speech enhancement methods significantly reduced

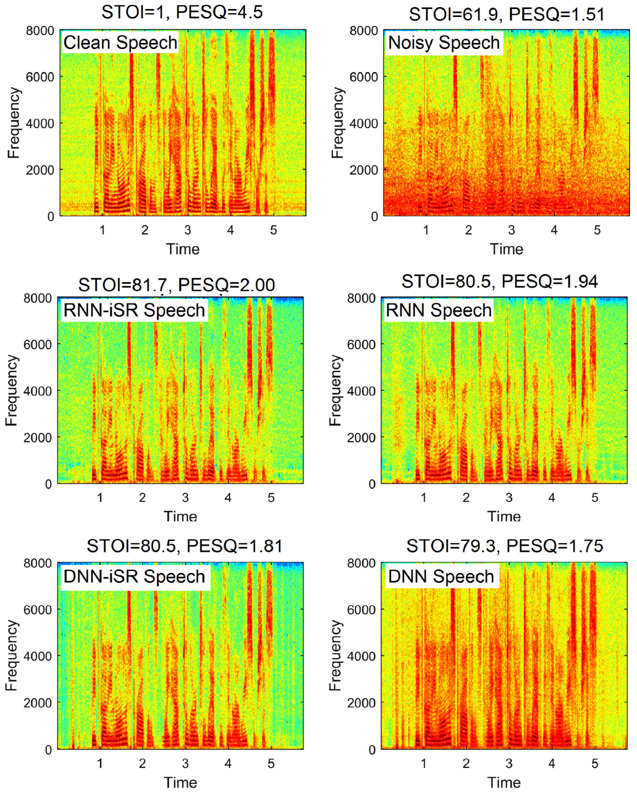

FIGURE 7. **Sample spectro-temporal analysis of the processing methods. Speech utterance degraded by babble noise at 0dB SNR. The spectrograms belong to IRM time-frequency making.**

the residual noise which is confirmed from Fig. 6 (time-varying spectrograms).

C. COMPLEXITY AND NETWORK CONVERGENCE

The complexity to train DNN/RNN depends on the network parameters and forward-backward propagation for network tuning. In the proposed speech enhancement methods, we have randomly initialized the parameters of networks. The complexity also depends upon quantity of neurons in the hidden layers and weights. Higher the number of neurons more will be the complexity of network. Observe Table 4, all DNNs/RNNs are trained with same number of layers, quantity of neurons in hidden and visible layers, LSTM units, but the proposed deep networks performed better and converged faster. The reason for fast network convergence (less loss function) is adaptation of critical band weights which are directly applied to the clean training data. The input data is pre-processed with CBIF weights which certainly improved the network performance. With equal quantity of neurons, the proposed methods provided lower values of loss functions, and this fact can be observed in Fig. 3 where all proposed methods converged at epoch ≥ 35. Based on the convergence results, we have fixed the epoch's number to 50 in the proposed speech enhancement methods. The complexity of DNN/RNN is given in Table 10, represented by "O". The forward and backpropagation propagation depends on U_I: dimension of the input acoustic features, N_D: training data point's numbers, $N_{H:}$ number of hidden neurons in layers, N_O: number of the output neurons, N_E: epochs for parameters tuning.

D. ROBUST SPEECH RECOGNITION

The above evaluations showed that DNN and RNN-based spectral masking significantly attenuated the background

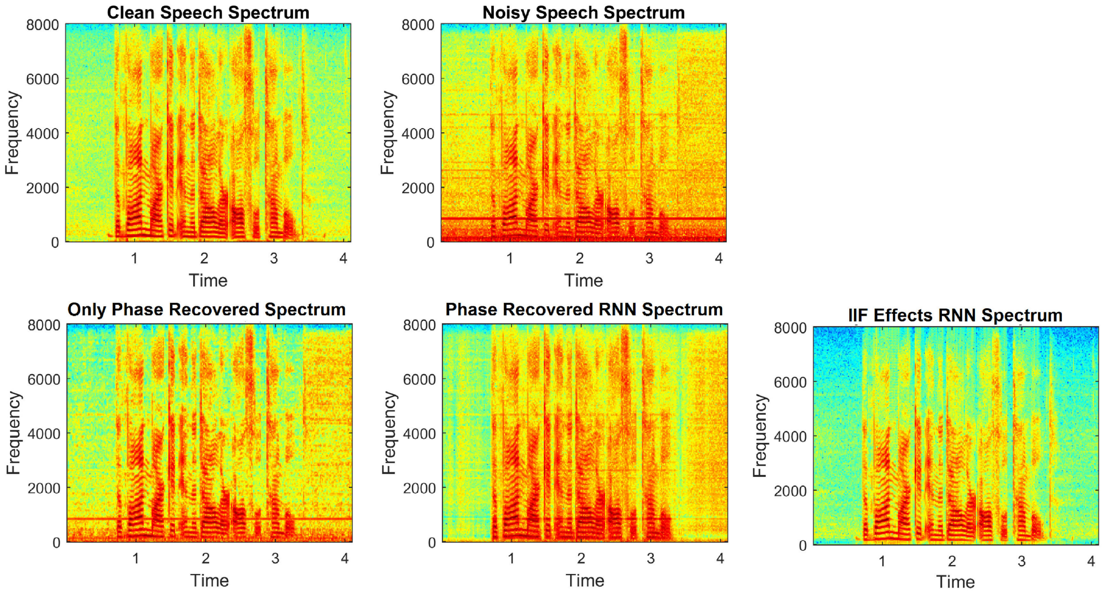

FIGURE 8. **Sample spectro-temporal analysis. Top (Clean and Noisy speech spectrograms). Bottom (Left): Phase Recovery-only, (Middle) RNN output with phase recovery and (Right): RNN output with IIF Filtering.**

TABLE 11. WERs for different processing approaches.

WERs in %	Clean Speech	Noisy Speech	RNN-*i*SR	RNN$_B$	DNN-*i*SR	DNN$_B$
	0.0%	48.12%	12.90%	16.13%	19.22%	21.58%

TABLE 12. Biographical data from the listeners tested.

Listeners	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
Age	30	28	35	38	38	40	40	42	45	45	48	50	53
Gender	M	M	F	M	F	F	M	M	M	F	M	M	M

noise and produced fine estimates of the magnitude spectrogram of the clean speech. Since automatic speech recognition methods only utilize magnitude spectrogram, one would expect DNN and RNN approaches to improve ASR performance in background noisy environments. To perform the automatic speech recognition, DNN and RNN-based spectral masking approaches are treated as front-end to enhance all speech utterances. We used Google ASR [58] in the experiments to evaluate ASR performance in terms of the word error rates (WERs). We provided average WERs results across all background noises and SNR levels. As shown in Table 11, both RNN-*i*SR and DNN-*i*SR achieved lower WERs than DNN and RNN in background noisy conditions. RNN-*i*SR and DNN-*i*SR-based speech enhancement considerably boosted the ASR performance, where the improvements are 28.29% (absolute) for the DNN-*i*SR and 35.22% for the RNN-*i*SR over noisy speech utterances. The ASR advantage gradually decreases as the SNR increases, partly because the noise becomes smaller. The ASR experiments aim to show the potential of RNNs and DNNs rather than to achieve the state-of-the-art results.

E. SUBJECTIVE EVALUATION

In addition to the objective evaluation, subjective listening tests are also performed to evaluate the perceptual quality and speech intelligibility of the enhanced speech. Speech utterances with an input SNR of −3dB, 0dB and 3dB are randomly selected from three noise sources (babble, factory, and white noise). In total 100 speech utterances are used to compare DNN-*i*SR and RNN-*i*SR. A total of 06 participants are asked to select the correctly perceived words in order to measure the speech intelligibility in terms of the word recognition rate (WRR). In experiments none of speech utterances are repeated. The tests are performed in isolated room using high quality headphones.

Figure 9 demonstrates the subjective listening results in terms of the subjective intelligibility (WRR). From Fig. 9, we can observe that RNN-*i*SR achieved better results at all input SNRs. DNN-*i*SR significantly improved the results at −3dB and 0dB. The results indicate the advantages of the iterative speech recovery and IIF filter in the proposed speech enhancement methods. In order to investigate the statistical significance, we performed analysis of variance (ANOVA) for the achieved WRR scores. The critical value of F is 3.49 at p <0.05 (95% confidence level) probability. The ANOVA results for WRR at −3dB and 0dB are [F (4, 13) = 23.98, p <0.0003] and [F (4, 13) = 17.84, p <0.0001] which indicates statistical significance. But, ANOVA results for WRR at 3dB is [F (4, 13) = 3.86, p <0.0111] which indicates slight statistical significance. The reason for the slight statistical significance is the favorable SNR.

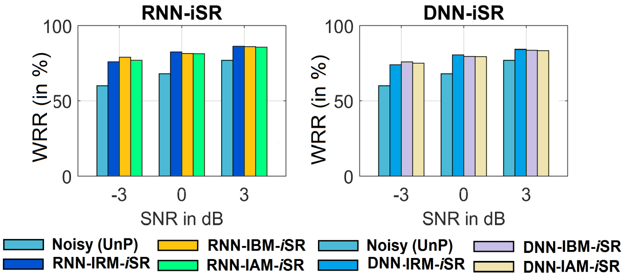

FIGURE 9. Subjective listening tests in terms of the speech intelligibility (WRR).

To measure the speech quality of the enhanced speech subjectively, a total of 13 participants are asked to select the speech utterance that they preferred in terms of the mean opinion score (MOS). The biographical data of the listeners participated in the subjective listening tests for the speech quality is given in Table 12. A total of 200 speech utterances are randomly selected which are mixed with the three noise sources (babble, factory, and white noise) at −3dB, 0dB and 3dB SNR. The processed speech utterances are used to compare the performance of proposed DNN-*i*SR and RNN-*i*SR. In experiments none of speech utterances are repeated. Training sessions are organized to disseminate the listeners about the procedure. The tests are performed in an isolated room using high quality headphones. Figure 9-10 demonstrates the subjective listening tests in terms of MOS for speech quality. Both DNN-*i*SR and RNN-*i*SR showed better performance. The average MOS scores at negative SNRs is higher than 2.75 (MOS≥2.75 at −3dB) which shows significant improvement. But at SNR≥0dB, the average MOS score surpassed 3 (MOS≥3.0 at 0dB and 3dB). The individual MOS scores for all listeners in the tests is also depicted in Fig. 9-10. The ANOVA for MOS at −3dB, 0dB and 3dB are

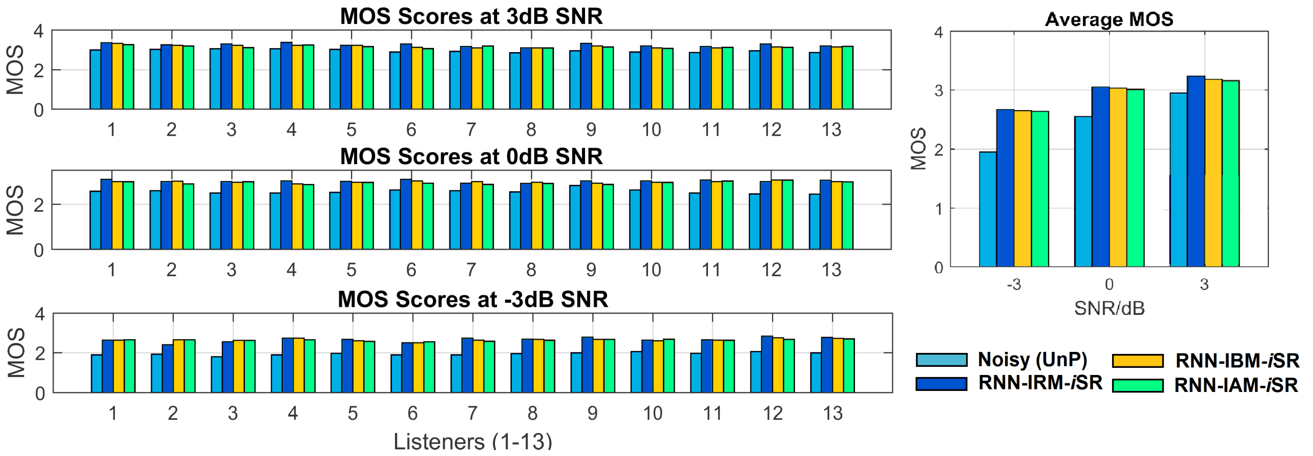

FIGURE 10. Subjective listening tests in terms of the speech Quality (MOS) for proposed RNN-iSR. MOS scores for all participants and average MOS of all participants at −3dB, 0db and 3dB SNR.

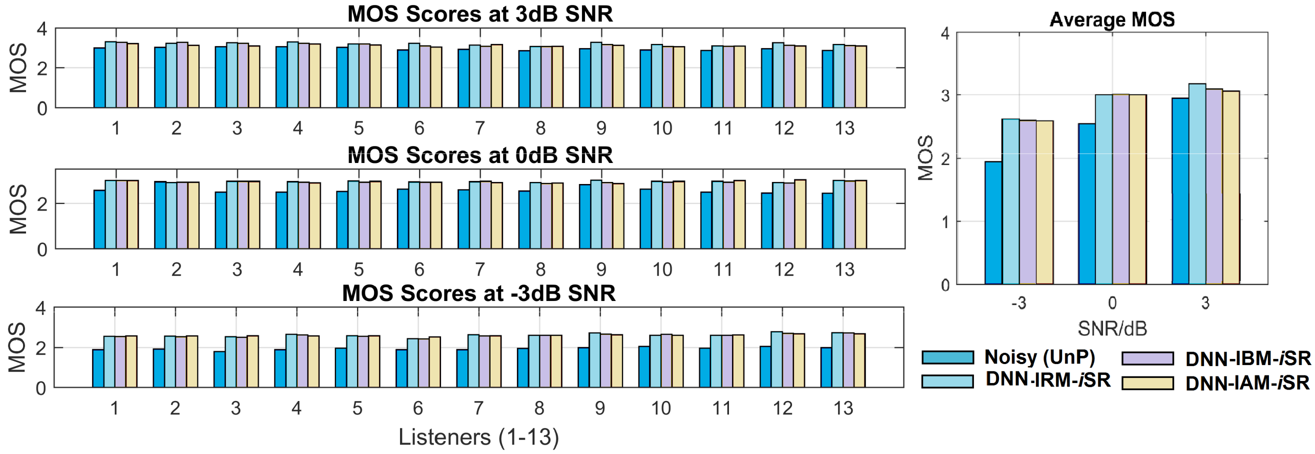

FIGURE 11. Subjective listening tests in terms of the speech Quality (MOS) for proposed DNN-iSR. MOS scores for all participants and average MOS of all participants at −3dB, 0db and 3dB SNR.

[F (4, 13) = 330.61, p < 0.0001], [F (4, 13)= 164.93, p < 0.0001] and [F (4, 13)= 89.12, p < 0.0001] which indicates statistical significance of achieved MOS scores.

VI. DISCUSSION AND CONCLUSION

We have proposed supervised spectral masking-based learning approaches to perform the single-channel speech enhancement. RNNs and DNNs are trained to learn the spectral masking between the degraded and clean speech signals. Our study trained the RNNs and DNNs without unsupervised pretraining to address single-channel speech enhancement. The presented study used the intelligibility improvement filter and an iterative reconstruction method to improve the outputs of neural networks and produced better recovered speech. In this study, all acoustic features are concatenation of the raw acoustic features in a window, since temporal dynamics gives more valuable information for the speech signals. A more elemental concept to exploit the temporal information is use of the RNN architecture, which is a fundamental extension of a feedforward network. The RNN architecture aims to grab the long-term temporal dynamics utilizing the time-delayed self-connections and is trained sequentially. We have trained RNN architectures for the spectral masking, and yielded 4.76%, 12.59%, 2.75dB and 8.67% improvements in terms of the STOI, ESTOI, SDR and PESQ. These improvements are worth significant. In our study, we also trained the DNN architectures. To test the generalization of RNNs and DNNs, we used the TIMIT database that included both male and female speakers. The overall ΔSNRs and SSNR for RNN-*i*SR and DNN-*i*SR are higher than the competing state-of-the-art methods. The listening tests indicate that RNN-*i*SR approach achieved better results at all input SNRs. DNN-*i*SR improved the results at −3dB and 0dB significantly. In addition, ASR experiments were conducted, which showed that the proposed speech enhancement method is robust to the automatic speech recognition task. It is important to mention that we first recovered the time-domain speech signals from RNNs and DNNs outputs and then performed automatic speech recognition task which is based on the processed speech signals. We achieved less computational complexity and fast convergence as compare to the baseline DNN/RNN speech enhancement methods. According to our experiments, the iterative speech recovery and IIF filter improved the predicted speech intelligibility and quality values as well as significantly improvement the ASR. Comparing Fig. 6(c) with Fig. 6(e), the spectrogram of RNN-*i*SR output is better than the spectrogram of RNN recovered signal, suggesting the benefits of iterative speech recovery and IIF filter. As RNN-*i*SR and DNN-*i*SR outputs show the better spectral representation, they yielded better ASR performance. In summary, we have proposed to use RNNs and DNNs to learn the spectral-masking from degraded speech to clean speech for single-channel speech enhancement task. The proposed supervised learning approaches are conceptually uncomplicated and have improved the performance in terms of the predicted speech intelligibility and quality, and boosted the ASR results in various noisy conditions.

REFERENCES

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

[2] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, Jun. 2008.

[3] B. Lim Sim, Y. Chow Tong, J. S. Chang, and C. Tuan Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[4] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

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

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

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

[8] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Commun.*, vol. 54, no. 2, pp. 282–305, Feb. 2012.

[9] I. Tashev and M. Slaney, "Data driven suppression rule for speech enhancement," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2013, pp. 1–6.

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

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

[12] Y. 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.

[13] W. Jiang, F. Wen, and P. Liu, "Robust beamforming for speech recognition using DNN-based time-frequency masks estimation," *IEEE Access*, vol. 6, pp. 52385–52392, 2018.

[14] N. Saleem, M. Khattakm, and A. Qazi, "Supervised speech enhancement based on deep neural network," *J. Intell. Fuzzy Syst.*, vol. 37, pp. 5187–5201, Jan. 2019.

[15] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, and W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access*, vol. 5, pp. 25542–25554, 2017.

[16] 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.

[17] 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.

[18] 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.

[19] D. Shan, X. Zhang, C. Zhang, and L. Li, "A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement," *IEEE Access*, vol. 6, pp. 62638–62644, 2018.

[20] N. Saleem, M. Khattak, M. Ali, and M. Shafi, "Deep neural network for supervised single-channel speech enhancement," *Arch. Acoust.*, vol. 44, pp. 3–12, Oct. 2019.

[21] A. Li, R. Peng, C. Zheng, and X. Li, "A supervised speech enhancement approach with residual noise control for voice communication," *Appl. Sci.*, vol. 10, no. 8, p. 2894, Apr. 2020.

[22] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, and W. A. Jassim, "Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103485–103504, 2019.

[23] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1104–1117, Mar. 2020.

[24] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Boston, MA, USA: Springer, 2005, pp. 181–197.

[25] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, Nov. 2006.

[26] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.

[27] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 625–638, May 2009.

[28] K. Han and D. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Amer.*, vol. 132, no. 5, pp. 3475–3483, Nov. 2012.

[29] May, T. and Dau, T., 2014, "Computational speech segregation based on an auditory-inspired modulation analysis," *The J. Acoust. Soc. Amer.*, vol. 136, no. 6, pp. 3350–3359.

[30] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[31] 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.

[32] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[33] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 577–581.

[34] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, R. J. Le, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.* Cham, Switzerland: Springer, Aug. 2015, pp. 91–99.

[35] 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.

[36] E. Wang, G. Brown, and C. Darwin, "Computational auditory scene analysis: Principles, algorithms and applications," *Acoust. Soc. Amer. J.*, vol. 124, p. 13, Oct. 2008.

[37] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[38] N. Saleem and M. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 1, pp. 1–7, Aug. 2019.

[39] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[40] Q. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 7, pp. 1185–1197, Jul. 2018.

[41] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network," *Appl. Soft Comput.*, vol. 74, pp. 583–602, Jan. 2019.

[42] A. Gruslys, R. Munos, I. Danihelka, M. Lanctot, and A. Graves, "Memory-efficient backpropagation through time," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4125–4133.

[43] I. Jindal, M. Nokleby, and X. Chen, "Learning deep networks from noisy labels with dropout regularization," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 967–972.

[44] Q. Song, Y. Wu, and Y. Chai Soh, "Robust adaptive gradient-descent training algorithm for recurrent neural networks in discrete time domain," *IEEE Trans. Neural Netw.*, vol. 19, no. 11, pp. 1841–1853, Nov. 2008.

[45] *American National Standard: Methods for Calculation of the Speech Intelligibility Index*, American National Standards Institute, New York, NY, USA, 1997.

[46] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 236–243.

[47] R. J. Le, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, Sep. 2008, p. 10.

[48] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA, Washington, DC, USA, Tech. Rep. LDC93S1, 1993.

[49] 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.

[50] D. Pearce and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Mississippi State Univ., Starkville, MN, USA, Tech. Rep. AU/384/02, 2002.

[51] 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.*, 2009, pp. 749–752.

[52] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Commun.*, vol. 95, pp. 28–39, Dec. 2017.

[53] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[54] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[55] C. Fevotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3158–3162.

[56] G. Min, X. Zhang, X. Zou, and M. Sun, "Mask estimate through itakura-saito nonnegative RPCA for speech enhancement," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[57] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Commun.*, vol. 104, pp. 106–112, Nov. 2018.

[58] (2017). *Cloud Speech API*. [Online]. Available: https://cloud.google.com/speech/

NASIR SALEEM received the B.Sc. degree in telecommunication engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2008, and the M.Sc. degree in electrical engineering (communication) from CECOS University, Peshawar, in 2012. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University of Engineering and Technology, Peshawar, under the supervision of Dr. Muhammad Irfan Khattak (an Associate Professor). He was a Lecturer with the Institute of Engineering and Technology, Gomal University, Dera Ismail Khan, Pakistan. He is also an Assistant Professor with the Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University. His research interests include digital signal processing, speech processing and speech enhancement, and machine learning for speech enhancement.

MUHAMMAD IRFAN KHATTAK received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Peshawar, in 2004, and the Ph.D. degree from Loughborough University, U.K., in 2010. After doing his Ph.D. degree, he was appointed as the Chairman of the Electrical Engineering Department, UET Peshawar Bannu Campus, for five years and took care of the academic and research activities at the department. Later in 2016, he was appointed as the Campus Coordinator of UET Peshawar Kohat Campus and took the administrative control of the campus. He is currently working as an Associate Professor with the Department of Electrical Engineering, University of Engineering and Technology, Peshawar. He is also heading the research group Microwave and Antenna Research Group, where he is also supervising the postgraduate students working on latest trends in antenna technology like 5G and graphene nano-antennas for terahertz, optoelectronic, and plasmonic applications etc. His research interests include antenna design, on-body communications, anechoic chamber characterization, speech processing, and speech enhancement. Besides his research activities, he is also a certified OBE Expert with the Pakistan Engineering Council for organizing OBA-based accreditation visits.

MUATH AL-HASAN (Senior Member, IEEE) received the B.A.Sc. degree in electrical engineering from the Jordan University of Science and Technology, Jordan, in 2005, the M.A.Sc. degree in wireless communications from Yarmouk University, Jordan, in 2008, and the Ph.D. degree in telecommunication engineering from the Institut National de la Recherche Scientifique (INRS), Université du Québec, Canada, in 2015. From 2013 to 2014, he was with Planets Inc., CA, USA. In May 2015, he joined Concordia University, Canada, as a Postdoctoral Fellow. He is currently an Assistant Professor with Al Ain University, United Arab Emirates. His current research interests include antenna design at millimeter-wave and terahertz, channel measurements in multiple-input and multiple-output (MIMO) systems, and machine learning and artificial intelligence in antenna design.

ABDUL BASEER QAZI (Member, IEEE) received the M.S. degree in information and communication systems from the Hamburg University of Technology, Hamburg, Germany, and the Ph.D. degree from UNU-MERIT, University of Maastricht, The Netherlands. He is currently a Senior Assistant Professor with the Department of Software Engineering, Bahria University, Islamabad. Prior to this, he was an Assistant Professor with the Capital University of Science and Technology, Islamabad, Pakistan. He also worked as an Assistant Professor with CECOS University, Peshawar, Pakistan. His industrial experience includes working for four Fortune 500 companies: IBM, Siemens, Philips Medical Systems, and NXP Semiconductors in Germany and The Netherlands. His research interests include wireless networks, antennas, technology policy, innovation, and entrepreneurship.