# ERBM-SE: Extended Restricted Boltzmann Machine for Multi-Objective Single-Channel Speech Enhancement

Muhammad Irfan Khattak[1], Nasir Saleem[2]*, Aamir Nawaz[2], Aftab Ahmed Almani[3], Farhana Umer[4], Elena Verdú[5]

[1] Department of Electrical Engineering, University of Engineering & Technology, Peshawar (Pakistan)
[2] Department of Electrical Engineering, FET, Gomal University, Dera Ismail Khan (Pakistan)
[3] School of Electrical Engineering, Shandong University, Jinan (China)
[4] Department of Electrical Engineering, Islamia University, Bahawalpur (Pakistan)
[5] Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja (Spain)

Received 9 March 2021 | Accepted 3 July 2021 | Published 11 March 2022

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

Machine learning-based supervised single-channel speech enhancement has achieved considerable research interest over conventional approaches. In this paper, an extended Restricted Boltzmann Machine (RBM) is proposed for the spectral masking-based noisy speech enhancement. In conventional RBM, the acoustic features for the speech enhancement task are layerwise extracted and the feature compression may result in loss of vital information during the network training. In order to exploit the important information in the raw data, an extended RBM is proposed for the acoustic feature representation and speech enhancement. In the proposed RBM, the acoustic features are progressively extracted by multiple-stacked RBMs during the pre-training phase. The hidden acoustic features from the previous RBM are combined with the raw input data that serve as the new inputs to the present RBM. By adding the raw data to RBMs, the layer-wise features related to the raw data are progressively extracted, that is helpful to mine valuable information in the raw data. The results using the TIMIT database showed that the proposed method successfully attenuated the noise and gained improvements in the speech quality and intelligibility. The STOI, PESQ and SDR are improved by 16.86%, 25.01% and 3.84dB over the unprocessed noisy speech.

## Keywords

Restricted Boltzmann Machine, Spectral Masking, Speech Enhancement, Speech Intelligibility, Speech Quality, Supervised Machine Learning.

DOI: 10.9781/ijimai.2022.03.002

## I. Introduction

THE aim of speech enhancement (SE) is to attenuate/suppress the background noise and recover the clean speech from the noise contaminated speech with better intelligibility and speech quality. The speech enhancement is mainly used in a speech communication system to improve the voice quality, recorded multimedia contents, to boost the automatic speech recognition (ASR) accuracy and for robust hearing aids. Many signal processing-based speech enhancement methods are proposed in literature to improve the performance of aforesaid applications which include spectral subtraction [1] and variants [2]-[4], Wiener filtering [5] and variants [6]-[7], minimum mean square error (MMSE) estimator [8] and the variants [9]-[10]. These methods are apt in various real-time speech applications because of less computational complexity. But, they show poor performance in many non-stationary acoustic conditions. To overcome this problem, supervised learning-based speech enhancement methods are opted [11]-[12]. Learning approaches, such as the regression, spectral-mapping, and spectral-masking [13]-[19], Gaussian mixture models-based SE (GMM) [20]-[21], support vector machines-based SE (SVM) [22] and non-negative matrix factorization (NMF) [23]-[24] have been developed and examined for the speech enhancement. In the past few years, speech enhancement is considered as a supervised learning problem, motivated from the time-frequency (T-F) masking in Computational Auditory Scene Analysis (CASA). In such methods, a trained learning machine directly estimates the clean speech or estimates a T-F mask such as ideal binary mask (IBM) and ideal ratio mask (IRM) which are then applied to the T-F representation of the contaminated speech to reconstruct clean speech [25]-[26]. Perhaps, paradigms of data-driven methods present a convenient explanation to grasp the complex mechanism of the acoustic speech distortion. Recently, a number of deep neural network (DNN) frameworks are developed with encouraging results. Starting from the autoencoders to feed-forward DNN, many frameworks have been designed for speech enhancement [27]-[33]. DNN-based methods deal with three attributes: complementary acoustic features, learning algorithm, and training-target. Pursuant to the above explanation, DNN-based supervised speech enhancement methods are categorized into the masking-based and mapping-based enhancement methods. However, we are dealing with masking-based method in this paper.

* Corresponding author.
E-mail address: nasirsaleem@gu.edu.pk

## II. Related Literature

In the recent past, the supervised learning methods for speech enhancement have achieved enormous performance gain and outperformed the conventional signal processing-based speech

TABLE I. Gap Analysis of Literature

| Reference | Neural Network | Pre-Training | Phase Estimate |
|---|---|---|---|
| [32] | DBN with multiple Mask Estimation | Networks are pre-trained with RBM without raw data | No Phase Estimation |
| [38] | DBN with Bayesian Estimators | Networks are pre-trained with RBM without raw data | No Phase Estimation |
| [37] | MCMC and SGD DBN pre-training with RBM | Networks are pre-trained with RBM without raw data | No Phase Estimation |
| [39] | Recurrent RBM for Pre-training | Networks are pre-trained with RBM without raw data | No Phase Estimation |
| [30] | Feed Forward DNN with Mask Estimation | Networks are randomly initialized without raw data | Phase Estimated |
| [26] | Feed Forward DNN with Mask Estimation | Networks are randomly initialized without raw data | No Phase Estimation |
| Proposed | Feed Forward DBN with Mask Estimation | Networks are pre-trained with RBM with raw data | Phase Estimated |

enhancement. The masking-based SE methods outperformed the mapping-based SE methods; but, large performance deterioration can happen as a result of the mismatch conditions. A large performance gain can be achieved if DNNs are layer-wise pre-trained by stacked-multiple RBM (Restricted Boltzmann Machine) [34]. DNN is proposed for the binary classification and feed-forward DNNs and RBM pre-training are used for subband classification for IBM estimation [27]. DNNs are pre-trained with Fuzzy RBM [32], [35] instead of the regular RBM and achieved significant performance by estimating various T-F masks [36]. A unified method based on Monte Carlo Markov Chain (MCMC) and Stochastic Gradient Descent (SGD) for RBM pre-training is proposed [37]. Bayesian estimators are designed for RBM pre-training [38]. Other variants of RBM such as the recurrent-temporal RBM [39], Gaussian RBM, cardinality RBM, pointwise gated RBM, and conditional RBM have been formulated by modifying regular RBM. Recurrent neural network-based speech enhancement method is formulated which exploited recurrent-temporal RBM to explore temporal-correlation between speech frames [35]. The idea is extended to the features of input and output signals into elemental feature-spaces. The network was fine-tuned by jointly optimized RNN with additional masking layer with a reconstruction constraint. A detailed review of the RBMs and their deep structures can be studied in [40]. Many recent studies on deep learning can also be found in [41]-[47]. The gap analysis is given in Table I. It can be observed that in literature either various networks have initialized the parameters with RBM without phase estimation or randomly initialized the parameters with phase estimation. But, the proposed method initialized the network parameters with a more robust way and also the phase is estimated.

In this paper, we examined the supervised learning algorithms in order to train DBN for time-frequency mask (T-F) estimation. Deep learning in speech enhancement is the arrangement of many hidden layers such that the network learns from the input features. Different from shallow neural networks, DNN should not be trained directly by using standard backpropagation algorithm. Since errors propagate through the network and the gradient becomes infinitesimally small that can affect the weights updating in the previous layers. Gradient-vanishing is one of the core challenges in the deep learning. To address the vanishing problem, a multi-layered framework is used, known as Deep Belief Network, a pre-trained DNN with multiple-stacked RBMs. Following the pre-training, the standard backpropagation algorithm is employed. The acoustic features are progressively extracted by multiple-stacked RBMs during the pre-training phase. The hidden acoustic features from the previous RBM are combined with the raw input data to serve as the new inputs to the current RBM. By adding the raw data to each RBM, layer-wise features related to raw data can be progressively extracted, which is helpful to mine valuable information in the raw data [48]. The aim of this work is not to design a state-of-the-art, but rather to examine the proposed pre-training method and compare the performance with DNN using the typical RBM-based pre-training for speech enhancement. The contributions of this paper are summarized and discussed as. (**i**): A novel pre-training method is proposed by stacking RBMs. The acoustic features are gradually extracted by multiple-stacked RBMs during pre-training phase. The hidden acoustic features from a previous RBM are combined with the raw input data to serve as the new inputs to the current RBM. The network parameters are initialized in the unsupervised fashion using RBM. The parameters are further fine-tuned via adaptive gradient descent and backpropagation algorithm. It is observed that the proposed pre-training method outperformed the DNNs which are initialized randomly or pre-trained with typical RBM. (**ii**): Less computational complexity and fast convergence is achieved by the proposed method as compared to the conventional DNN and DBN frameworks. With similar number of the hidden layers and quantity of hidden neurons, the proposed method achieved better speech quality and intelligibility. The reason for the quick network convergence (less MSE errors) is the adaptation of new pre-training method.

The reminder of this paper is organized as follows. Section III recapitulates the RBM and DBN frameworks. The proposed RBM for speech enhancement is presented in Section IV. Experiments are given in Section V. Results are explained in Section VI. The conclusions are given in Section VII.

## III. Restricted Boltzmann Machine (RBM) and Deep Belief Network (DBN)

Restricted Boltzmann Machine [34] is an elemental part of DBN framework and is mainly composed of visible and hidden layers used for many applications including speech enhancement. Unlike Boltzmann machine (BM), a RBM confines the interconnections of peer neurons in order to guarantee the mutual independence. The typical RBM structure is shown in Fig. 1. RBM gives probabilistic models and their parameters are consisting of the weights and biases. Let a RBM be represented by $v$ visible layer and $h$ hidden layer, respectively. The joint probability density of $v$ and $h$ is given as:

$$p(v,h) = \frac{e^{(-E(v,h))}}{\iint_{v,h} e^{(-E(v,h))}} \tag{1}$$

Where, E ($v$, $h$) indicates the energy function where type of function is determined by the nature of variables in the visible layer. Two common variables in the visible layer are the binary and Gaussian. The binary-binary and Gaussian-Gaussian energy functions are given by equations as:

$$E(v,h) = -\sum_{j\in vis} \alpha_j v_j - \sum_{i\in hid} \beta_i h_i - \sum_{j\in vis, i\in hid} v_j h_i w_{ji} \tag{2}$$

$$E(v,h) = \sum_{j\in vis} \frac{(v_j-\alpha_j)^2}{2\sigma_j^2} + \sum_{i\in hid} \frac{(h_i-\beta_i)^2}{2\sigma_i^2} - \sum_{j\in vis, i\in hid} \frac{v_j}{\sigma_j}\frac{h_i}{\sigma_i} w_{ji} \tag{3}$$

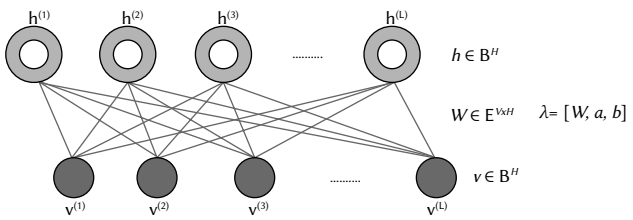



Fig. 1. RBM Network Structure.

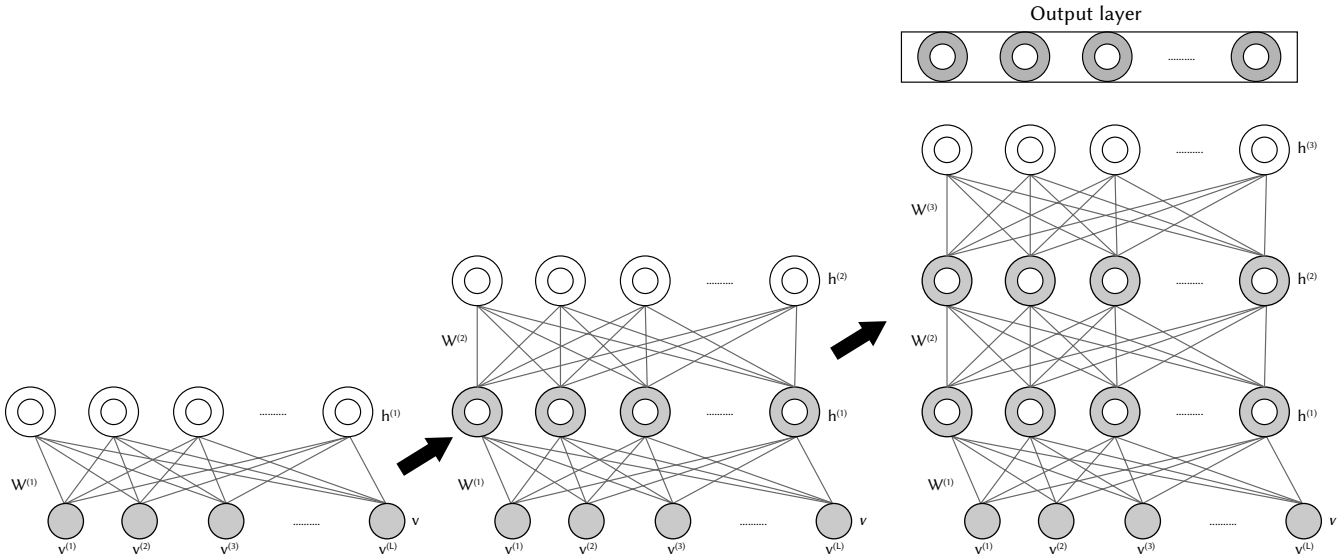



Fig. 2. The DBN framework with Three Hidden Layers.

where $v_j$ and $h_i$ indicate the activation states of the hidden layer neuron $j$ and visible layer neuron $i$, respectively; $\alpha_j$ and $\beta_i$ are bias terms whereas $w_{ji}$ indicates the weights used to connect the $v_j$ and $h_i$. $\sigma_j$ and $\sigma_i$ indicate the standard deviation terms. With the joint and marginal probabilistic distributions, i. e, $p(v, h)$, $p(v)$ and $p(h)$, the conditional probabilistic distributions $p(h|v)$ and $p(v|h)$ can be achieved by the Bayesian presumption as:

$$p(h|v) = \frac{p(v,h)}{p(v)} = \frac{e^{(-E(v,h))}}{\int_h e^{(-E(v,h))}} \tag{4}$$

$$p(v|h) = \frac{p(v,h)}{p(h)} = \frac{e^{(-E(v,h))}}{\int_v e^{(-E(v,h))}} \tag{5}$$

For Gaussian neurons, the conditional distributions follow the normal distributions. Deep belief networks consist of multiple stacked RBMs and an output layer added over the final RBM, as shown in Fig. 2. The training process of DBN includes a layer-wise unsupervised pre-training and fine-tuning. During the pre-training step, layer-wise greedy scheme is used for RBM training. Once a RBM is trained, its hidden layer is served as a visible layer to the next RBM. Thus, all RBMs in the network are trained in this fashion by maximizing input data probabilities. Contrastive divergence (CD) method [49]-[50] is applied for parameters updating. After pre-training, a T-F masking layer is appended to the final hidden layer. The entire DBN network is further fine-tuned by reducing the errors between estimated and preset masks. The backpropagation is employed to gradually pass the errors from the final to base input layer. In this way, the entire network parameters are continuously updated.

## IV. Proposed RBM-based Speech Enhancement Method

Though DBN framework effectively extracts features and achieves quick convergence by executing pre-training and fine-tuning, yet there can be a room to improve the learning performance. In deep learning, by increasing the number of hidden layers and with layer-wise compression process, important information in the raw data is usually lost in higher layers. To reduce this problem, we extended conventional DBN to amply detain the important information in the raw data by multiple stacked-RBMs. By using the raw data as supplementary inputs to the visible layers to pre-train every RBM, the input raw data participates in entire compression process. As a result, the extracted acoustic features are greatly related to input raw data and the potential important information is copiously kept. Unlike conventional DBN, the proposed extended version of DBN framework can repetitively extract the important information from input raw data, thereby provides deep compressed representations which are in correlation with the input raw data. Fig. 3 illustrates the proposed DBN framework which consists of the pre-training and fine-tuning procedures, respectively. In the pre-training process, the input raw data is appended to visible layers of all RBMs. The weight matrices are composed of $w_I$ and $w_H$, where $w_I$ connects the hidden layer with the input raw data, and $w_H$ connects the hidden features of the preceding RBM with upper hidden layer. After that, contrastive divergence and the maximum likelihood rules are used to update the RBM parameters. By doing so, we can improve the network learning potential, and can accurately initialize the network parameters for fine-tuning process. During the fine-tuning process, an output layer is added for mask estimation. Finally, the backpropagation is performed iteratively in order to update parameters of network by minimizing the MSE loss function between estimated and preset mask.

### A. Pre-Training

The pre-training process of the proposed DBN is to train all RBMs individually. For the first RBM, there is no need to extend input raw data. Every RBM updated its weights and biases which are based on the $k$-step CD learning (CD-$k$) method and maximum likelihood rule. In general, maximum likelihood rule is applied to achieve suitable parameters of the network ($\delta = \{w, \alpha, \beta\}$) that excellently fit the input data distribution. By determining logarithmic partial derivatives of data $p(\mathbf{v}=\mathbf{v}_{\text{Data}})$, the gradient updating for network parameters is given as:

$$\frac{\partial \log p\,(v = v_{Data})}{\partial \delta} = -\int_h p(h|v_{Data}) \frac{\partial E(v_{Data},h)}{\partial \delta} + \int_v p(v) \int_h p(h|v) \frac{\partial E(v,h)}{\partial \delta} \tag{6}$$

Since it is challenging to determine the second terms in Eq. (6) precisely, CD-$k$ method is used to achieve the estimated solution. It is intended to transfer the data among hidden and visible layers for $k$ times, such that the network states can characterize the model distribution after k iterations. The input to RBM $\mathbf{v}_{\text{Data}}$ can be articulated as $\mathbf{v}^{(0)}$ as shown in Fig. 2. By determining the sampling of the probability $p(\mathbf{h}|\mathbf{v}^{(0)})$, the state of the hidden neurons can be achieved as $\mathbf{h}^{(0)}$. Also, $\mathbf{v}^{(1)}$ can be achieved by sampling of $p(\mathbf{v}|\mathbf{h}^{(0)})$. The learning procedure, from $\mathbf{v}^{(0)}$ to $\mathbf{v}^{(1)}$, is known as one-step Gibbs sampling. After performing $k$-step Gibbs sampling ($k \rightarrow \infty$), a stationary distribution is achieved, whereas $\mathbf{v}^{(k)}$ reflects the model distribution. By predicting expectations over $p(\mathbf{v})$, equation (6) can be expressed as:

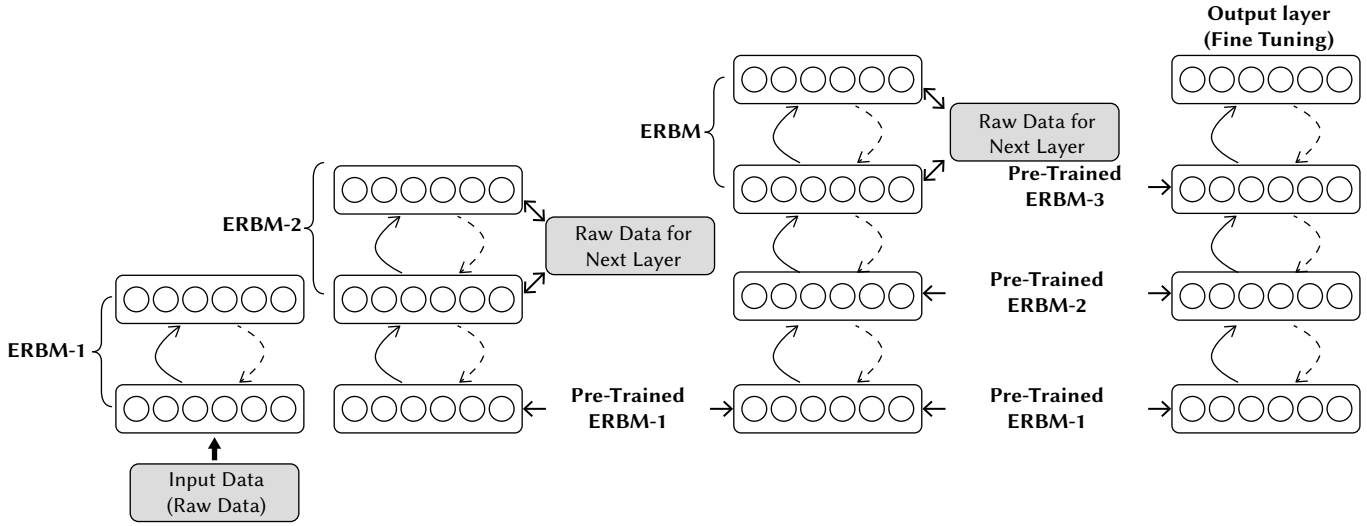



Fig. 3. The Proposed DBN framework structure with three hidden layers during pre-training and fine-tuning.

$$CD_k(\delta, v^{(0)}) = -\frac{\partial E(v^{(0)}-h^{(0)})}{\partial \delta} + \frac{\partial E(v^{(k)}-h^{(k)})}{\partial \delta} \tag{7}$$

The first term is called as negative-term whereas the second term is known as the positive-term, respectively. The negative-term reflects the distribution of raw data and the positive-term reflects distribution of the model. During training, the $k$-step CD takes $k = 1$, that meets the required of calculation for the accuracy. If we substitute Eq. (3) into Eq. (7), equations (Eq. (8)-Eq. (10)) can be obtained. The RBM parameters ($\delta_{PRE} = W_{PRE}, \alpha_{PRE}, \beta_{PRE}$) can be updated via following equations:

$$\Delta w_{ji} = v_j^{(0)} h_i^{(0)} - v_j^{(k)} h_i^{(k)} \tag{8}$$

$$\Delta \alpha_j = v_j^{(0)} - v_j^{(k)} \tag{9}$$

$$\Delta \beta_i = h_i^{(0)} - h_i^{(k)} \tag{10}$$

$$\delta_{PRE}^{(Epoch+1)} = \delta_{PRE}^{(Epoch)} + \mu \Delta \delta_{PRE}^{(Epoch-1)} + \upsilon \Delta \delta_{PRE}^{(Epoch)} \tag{11}$$

Where, the parameters $\mu$ and $\upsilon$ indicate momentum and learning rate, respectively, whereas weighting matrix $w_{PRE}$ is composed of $w_I$ and $w_H$ such that $w_{PRE} = [w_I, w_H]$. Algorithm-A and Algorithm-B explains the $k$-step CD rule and the pre-training of proposed DBN, respectively.

**Algorithm-A:** $k$-step CD Process

Input: RBM ($v$, $h$), Training Batch B

Output: Estimated Gradient $\Delta w$, $\Delta \alpha$, $\Delta \beta$

1: init $\Delta w_{ji} = \Delta \alpha_j = \Delta \beta_i = 0$, for $j = 1, ..., m$; $i = 1, ..., n$

2: for all samples $\in B$ *do*

3: $v^{(0)} \leftarrow$ sample

4: *for* $t = 0, ..., k$-1 *do*

5: *for* $i = 1, ..., n$ *do* sample $h_i^{(t)}$ from $p(h_i \mid v^{(t)})$

6: *for* $j = 1, ..., n$ *do* sample $v_i^{(t+1)}$ from $p(v_j \mid h^{(t)})$

7: sample $h_i^{(k)}$ from $p(h_i \mid v^{(k)})$ *for* $i = 1, ..., n$

8: *for* $j = 1, ..., m$, $j = 1, ..., n$ *do*

9: $\Delta w_{ji} \leftarrow \Delta w_{ji} + v_j^{(0)}.h_i^{(0)} - v_j^{(k)}.h_i^{(k)}$

10: $\Delta \alpha_j \leftarrow \Delta \alpha_j + v_j^{(0)} - v_j^{(k)}$

11: $\Delta \beta_i \leftarrow \Delta \beta_i + h_i^{(0)} - h_i^{(k)}$

**Algorithm-B:** DBN layer-by-layer Pre-Training

Input: Training Set Y

Output: Pre-Trained DBN Framework

1: for all *RBM* in *DBN* framework

2: init Network Parameters; $w$, $\alpha$, $\beta$

3: if training model is *RBM* then input $\leftarrow Y$

4: else *input* $\leftarrow$ combine $H$ and $Y$

5: for *epoch* $= 1, ...., e$ do

6: for $k$=1, ..., *floor* $(\frac{N_{samples}}{N_{Batchsize}})$ do

7: $B \leftarrow$ take batch from *input*

8: $\Delta w, \Delta \alpha, \Delta \beta \leftarrow$ Algorithm-1: $k$-CD

9: $w \leftarrow w + \mu.\Delta w$

10: $\alpha \leftarrow \alpha + \mu.\Delta \alpha$

11: $\beta \leftarrow \beta + \mu.\Delta \beta$

12: $H \leftarrow$ Input $\times w + \beta$

## B. Fine-Tuning

In fine-tuning of the proposed DBN, the additional layer for output is appended at final hidden layer in order to get probabilities of the samples. The parameters in the proposed DBN $\{(w_{ft}^{(i)}, \beta_{ft}^{(i)})\}_{i=1,2,3,...,m}$ are initialized by pre-trained parameters $\{(w_h^{(i)}, \beta_{PRE}^{(i)})\}_{i=1,2,3,...,m}$ as:

$$\begin{cases} w_{ft}^{(i)} = w_h^{(i)} \\ \beta_{ft}^{(i)} = \beta_{PRE}^{(i)} \end{cases}, i = 1,2,3,4,\ldots,m \tag{12}$$

Where, $m$ indicates the hidden layer's number, the raw data neurons and subsequent parameters that are dropped after pre-training. Random values are used for parameters $(w_{ft}^{(o)}, \beta_{ft}^{(o)})$ of output layer. Thus, the parameters $\delta_{ft}^{(i)}=\{(w_{ft}^{(i)}, \beta_{ft}^{(i)}, w_{ft}^{(o)}, \beta_{ft}^{(o)})\}_{i=1,2,3,...,m}$. By using the standard forward-propagation, the loss errors MSE can be computed between estimated and preset mask. Finally, based on the adaptive moment estimation (Adam), parameters of the proposed DBN are further tuned by following equations:

$$\upsilon^{(Epoch+1)} = \upsilon^{(0)} \sqrt{\frac{1-\nu_2^{Epoch}}{1-\nu_1^{Epoch}}} \tag{13}$$

$$\mu_1^{(Epoch+1)} = \nu_1 \mu_1^{(Epoch)} + (1-\nu_1) \Delta \delta_{ft}^{(Epoch)} \tag{14}$$

$$\mu_2^{(Epoch+1)} = \nu_2 \mu_2^{(Epoch)} + (1-\nu_2)(\Delta\delta_{ft}^{(Epoch)})^2 \quad (15)$$

$$\delta_{ft}^{(Epoch+1)} = \delta_{ft}^{(Epoch)} - \upsilon^{(Epoch+1)}\left(\frac{\mu_1^{(Epoch+1)}}{\sqrt{\mu_2^{(Epoch+1)}+\rho}}\right) \quad (16)$$

Where $\upsilon^{(Epoch+1)}$ indicates learning rate with initial rate of $\upsilon^{(0)}$ set according to the requirement. The terms $\mu_1^{(Epoch)}$ and $\mu_2^{(Epoch)}$ show the first momentum estimate and second raw momentum estimate. The $\nu_1$, $\nu_1$, $\rho$ are Adam parameters.

## V. Experiments

### A. Dataset

In experiments, we selected clean speech utterances from TIMIT database [51]. TIMIT corpus includes time-aligned and phonetically balanced 16-bit, 16 kHz speech waveform files. The clean utterances are used for speech enhancement and speech recognition. It contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. In order to evaluate the performance of the proposed method in different noisy backgrounds, 10 different noise sources are selected from the Aurora-4 [52] database, given in Table II. The spectrograms of the noise sources are demonstrated in Fig. 4. To produce the noisy speech, we used four signal-to-noise (SNR) levels, -4dB to 2dB with a 2dB step. To train the proposed DBN framework, we have used 2000 speech utterances from different speakers of both genders. For all SNRs, the input training utterances are mixed with 10 noise sources (2000 x 4 =8000 speech utterances). To test the proposed method, 1000 speech utterances from different speakers are used. The experimental results are averaged over 10 noise sources.

TABLE II. Background Noise Sources (N1-N10)

| |
|---|
| N1: Airport Noise, N2: Babble Noise, N3: Buccaneer, N4: Car Noise, N5: Café Shop Noise, N6: Destroyerengine Noise, N7: Destroyerops Noise, N8: Factory Noise, N9: Hall Noise, N10: Street Noise |

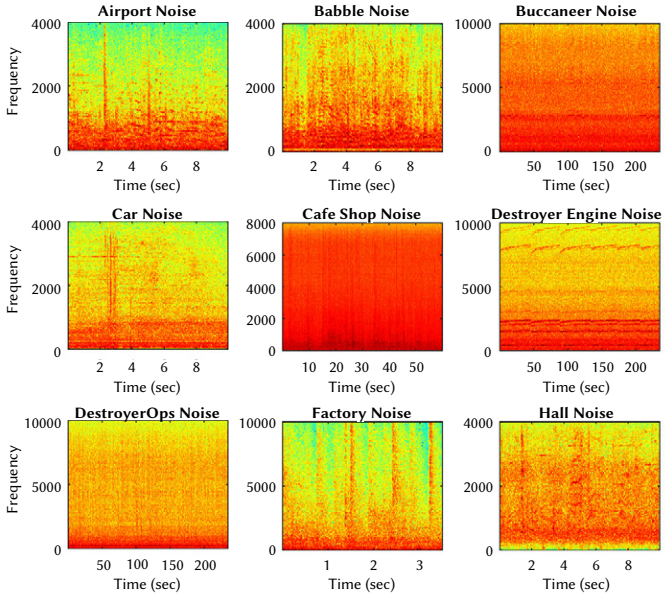



Fig. 4. Spectrograms of Background Noise Sources.

### B. Acoustic Features

The acoustic features are extracted from the input speech frames. The frame length and shift in the proposed method are fixed to 20 msec and 10 msec, respectively. The acoustic features set is composed of 13-*d* relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), 31-*d* Mel-frequency cepstral coefficients (MFCC), 64-*d* gammatone filter-bank energies (GFE), 15-*d* amplitude modulation spectrogram (AMS), where *d* represents feature dimensions. The GFE features are extracted from the T-F representation, known as Cochleagram usually employed in computational auditory scene analysis. Cochleagram expresses the mechanism of the human auditory system. We used a 64-channel gammatone filterbank to extract the GFE features. Furthermore, delta features are computed and affixed to the acoustic feature sets. RASTAMAT toolbox is utilized to extract all the acoustic features. We have used second order auto-regressive moving average filter (ARMA) to obtain the flat temporal trajectories of the acoustic features. The mathematical expression for ARMA is given as:

$$\bar{A}(t) = \frac{\bar{A}(t-k) + \cdots + A(t) + \cdots + A(t+k)}{2k+1} \quad (17)$$

where $A(t)$ shows the feature vectors at time frame $t$, $\bar{A}(t)$ corresponds to the filtered feature vectors and $k$ is the order of filter. In order to add the temporal information, a context window of five frames is used in the proposed method. As a result, we attained 1230-*d* feature vectors. All feature vectors are normalized to zero mean and unit variance before fed to the deep neural networks.

### C. Network Architecture

In this paper, a DBN network with a novel pre-training method is employed to learn the magnitude+phase aware spectral-mask. The network architecture is described in this section. DBNs are learning machines and have shown to perform better in speech enhancement. The DBN architecture in this study consists of five layers; an input layer, three hidden layers, and an output layer. The size of the input layer is 1230 neurons, that is, 246*5=1230, including 246-*d* acoustic features and features window composed of 5 frames. Each hidden layer consists of 1024 hidden neurons and the output layer contains 517 visible neurons. From the input to output layer, architecture of the proposed DBN has [1230, 1024, 1024, 1024, 517] neurons. Backpropagation and dropout regularization are used during fine-tuning. Adaptive gradient descent algorithm with a momentum parameter $\mu$ is used to optimize DBN. 512 samples batch size is used. The scaling factor for adaptive gradient descent is set to 0.0010 and the learning rate $\upsilon$ is reduced linearly from 0.06 to 0.002. 100 epochs are used during the process. For the first few epochs, the $\mu$ is fixed at 0.5 and the rate is increased to 0.8 for remaining epochs. The MSE loss function based on the mask approximation is considered. In supervised spectral masking-based SE, the loss functions are usually formulated to estimate the masking parameters that efficiently restore the clean speech by attenuating undesired noise components in T-F units. The time-domain enhanced speech signals are finally recovered by applying inverse STFT (*i*STFT) using the noisy phase or estimated phase. In this study, the enhanced speech is recovered by using the estimated phase. Spectral-masking methods are found to be successful as T-F masks are dynamically bounded; therefore, achieves quick convergence. In deep learning-based SE, many approaches are opted to estimate a T-F mask and depend on the training-target or the optimization-domain. In mask-approximation (MA) domain, the T-F masks are estimated such that mean square error (MSE) with preset T-F mask is minimized [53], and is given by equation as:

$$MSE_{MA} = \frac{1}{2L}\sum_{k=1}^{K-1}\left[(M_S(t,f) - \widehat{M}_S(t,f))^2\right] \quad (18)$$

Where $\widehat{M}_S(t,f)$ and $M_S(t,f)$ indicates the estimated and preset T-F masks. The rectified linear unit (ReLU) activation converts a weighted sum of the inputs to the model neuron's output. Recent studies show that deep MLPs with ReLU function can successfully be trained by using large training data. Thus, ReLU is used as activation function in hidden layers and sigmoid activation function is used in output layer. The reason for selecting the sigmoid as an output activation function is its dynamic range [0 1]. It is used for models that predict the output probabilities, since probability exists between 0 and 1. Also, the dynamic range of IRM mask exists between 0 and 1. The activation functions are:

$$f(\kappa) = max(0, \kappa); \quad f(\kappa) = \frac{1}{1+e^{\kappa}} \tag{19}$$

### D. Evaluation Metrics and Parameters

We extensively evaluated the proposed method by using four objective measures. Perceptual evaluation of speech quality (PESQ) [54] and signal-to-distortion ratio (SDR) are used to quantify speech quality. PESQ, an ITU-T P.862 recommendation calculates the speech quality of enhanced speech with an output value ranging from 0.5 to 4.5. A high PESQ value implies better quality. SDR also measures the quality. Short-time objective intelligibility (STOI) and extended STOI (ESTOI) are used to quantify the intelligibility. STOI [55] and ESTOI [56] measure the intelligibility of the enhanced speech with an output value ranging from 0 to 1. A high STOI and ESTOI value implies better intelligibility. The STOI and ESTOI values are obtained by correlation between clean and enhanced speech signals in short-time overlapped segments. Segmental SNR (SSNR) and output SNR ($SNR_O$) are used to quantify the residual noise in enhanced speech.

## VI. Results

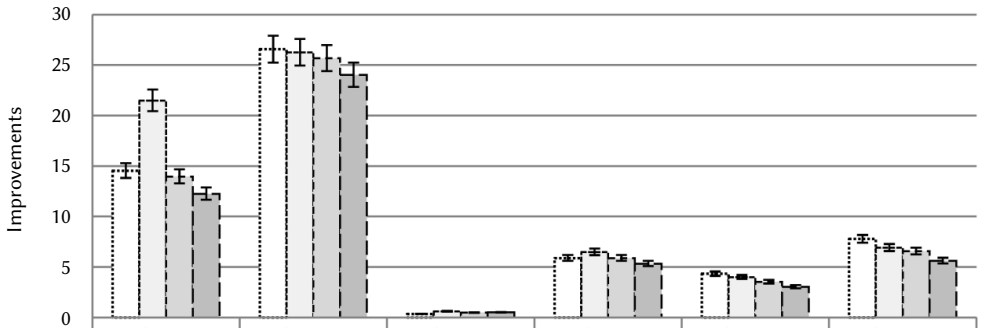
In this section, we provide the major findings of this study. We objectively evaluated the proposed SE method and compared the proposed method with baseline DBN. We additionally compared the proposed method with other related speech enhancement methods from various classes.

TABLE III. Performance Evaluation in Terms of STOI and ESTOI in Four Input SNRs Using TIMIT Corpus and Three Background Noises. $DBN_P$: Proposed DBN and $DBN_B$: Baseline DBN

| Noise Type → | Airport Noise | | Babble Noise | | Factory Noise | |
|---|---|---|---|---|---|---|
| SNR -4dB | | | | | | |
| Methods | STOI | ESTOI | STOI | ESTOI | STOI | ESTOI |
| Noisy | 62.82 | 30.61 | 57.31 | 23.46 | 56.58 | 23.53 |
| $DBN_P$ | 78.47 | 49.48 | 66.27 | 37.84 | 78.22 | 42.80 |
| $DBN_B$ | 76.94 | 47.82 | 65.74 | 37.18 | 75.69 | 39.14 |
| SNR -2dB | | | | | | |
| Noisy | 67.17 | 36.05 | 61.18 | 28.68 | 60.74 | 27.71 |
| $DBN_P$ | 81.03 | 53.11 | 71.53 | 43.22 | 69.10 | 37.38 |
| $DBN_B$ | 80.59 | 52.64 | 71.62 | 42.84 | 68.42 | 35.22 |
| SNR 0dB | | | | | | |
| Noisy | 71.83 | 41.63 | 65.43 | 34.01 | 65.24 | 33.24 |
| $DBN_P$ | 85.05 | 64.87 | 76.33 | 51.63 | 74.97 | 48.00 |
| $DBN_B$ | 85.02 | 64.63 | 76.36 | 51.64 | 74.44 | 47.81 |
| SNR 2dB | | | | | | |
| Noisy | 76.14 | 47.70 | 70.79 | 40.14 | 69.87 | 39.21 |
| $DBN_P$ | 86.95 | 67.73 | 80.66 | 57.32 | 79.11 | 55.98 |
| $DBN_B$ | 86.96 | 67.36 | 80.31 | 57.24 | 78.88 | 55.51 |

TABLE IV. Performance Evaluation in Terms of SDR and PESQ of DBNs in Four Input SNRs Using TIMIT Corpus and Three Background Noises. $DBN_P$: Proposed DBN and $DBN_B$: Baseline DBN

| Noise Type → | Airport Noise | | Babble Noise | | Factory Noise | |
|---|---|---|---|---|---|---|
| SNR -4dB | | | | | | |
| Methods | SDR | PESQ | SDR | PESQ | SDR | PESQ |
| Noisy | -3.79 | 1.58 | -4.88 | 1.44 | -3.71 | 1.34 |
| $DBN_P$ | 3.31 | 1.83 | 1.52 | 1.64 | -0.17 | 1.61 |
| $DBN_B$ | 2.94 | 1.70 | 1.16 | 1.41 | -0.53 | 1.48 |
| SNR -2dB | | | | | | |
| Noisy | -1.83 | 1.72 | -1.80 | 1.60 | -1.81 | 1.44 |
| $DBN_P$ | 4.79 | 1.96 | 2.98 | 1.78 | 4.51 | 1.89 |
| $DBN_B$ | 4.34 | 1.90 | 2.91 | 1.63 | 4.47 | 1.87 |
| SNR 0dB | | | | | | |
| Noisy | 0.12 | 1.82 | 0.13 | 1.73 | 0.15 | 1.56 |
| $DBN_P$ | 6.09 | 2.23 | 4.56 | 1.94 | 6.31 | 2.11 |
| $DBN_B$ | 5.75 | 2.01 | 4.55 | 1.77 | 5.97 | 1.98 |
| SNR 2dB | | | | | | |
| Noisy | 2.10 | 1.95 | 2.12 | 1.85 | 2.12 | 1.69 |
| $DBN_P$ | 7.38 | 2.35 | 6.26 | 2.15 | 7.44 | 2.29 |
| $DBN_B$ | 7.32 | 2.17 | 6.27 | 2.07 | 7.32 | 2.09 |




| | ΔSTOI | ΔESTOI | ΔPESQ | ΔSDR | ΔSSNR | ΔSNRO |
|---|---|---|---|---|---|---|
| -4 | 14.53 | 26.56 | 0.34 | 5.88 | 4.32 | 7.77 |
| -2 | 21.49 | 26.26 | 0.6 | 6.48 | 3.99 | 6.91 |
| 0dB | 13.96 | 25.68 | 0.48 | 5.88 | 3.53 | 6.57 |
| 2dB | 12.25 | 24.02 | 0.5 | 5.33 | 3.03 | 5.62 |

Fig. 5. The average PESQ, SDR, STOI, and ESTOI improvements in all noise sources.

TABLE V. Average Comparison Performance of $DBN_P$ and $DBN_B$ in All Noise Sources at Four Input SNRs Using TIMIT Corpus

| Input SNRs | $DBN_P$ | | | | | | $DBN_B$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | STOI | ESTOI | SDR | PESQ | SSNR | SNRo | STOI | ESTOI | SDR | PESQ | SSNR | SNRo |
| -4dB | 73.43 | 45.08 | 1.77 | 1.82 | 2.48 | 3.55 | 71.59 | 43.16 | 1.09 | 1.63 | 2.13 | 3.13 |
| -2dB | 77.49 | 51.62 | 4.67 | 2.08 | 3.43 | 4.57 | 75.99 | 49.89 | 4.14 | 1.89 | 3.01 | 4.34 |
| 0dB | 81.46 | 57.35 | 6.04 | 2.21 | 3.80 | 6.43 | 80.13 | 55.90 | 5.72 | 2.06 | 3.09 | 5.98 |
| 2dB | 84.51 | 63.01 | 7.44 | 2.33 | 4.95 | 7.38 | 83.31 | 61.78 | 7.26 | 2.23 | 4.39 | 7.23 |
| Avg. | 79.22 | 54.27 | 4.98 | 2.11 | 3.66 | 5.48 | 77.75 | 52.68 | 4.55 | 1.95 | 3.15 | 5.17 |

### A. Objective Evaluation

We reported the in depth evaluation for three noise sources, for example, using TIMIT database in Table III and Table IV, where we have used MA-based MSE loss function for training networks. We examined DBN with the proposed pre-training scheme and compared to the DBN with regular pre-training scheme. The T-F mask with the proposed pre-training scheme significantly improved the quality and intelligibility of the noisy speech. Clearly, DBNP outscored the conventional baseline DBNB. For example, at -4dB airport noise, DBNP improved the STOI and ESTOI by 9.65% and 18.87% over the unprocessed noisy speech. Similarly, the STOI and ESTOI at -4dB airport noise are improved by 1.54% and 1.66% over the DBNB. In addition, DBNP improved the STOI and ESTOI by 8.96% and 14.38% over noisy speech at -4dB babble noise. Equally, DBNP improved the STOI and ESTOI at -4dB factory noise by 2.53% and 3.66% over the DBNB. DBNP improved the SDR at -4dB and -2dB airport noise by 7.10dB and 6.62dB over the noisy speech. Similarly, DBNP improved the SDR at -4dB and -2dB babble noise by 0.40dB and 0.45dB over the DBNB. At 0dB factory noise, the DBNP improved the SDR by 5.96dB and 0.34dB over the unprocessed noisy speech and DBNB, respectively. Similarly, the PESQ at -4dB, -2dB and 0dB babble noise are improved by 13.88%, 11.25%, and 12.13% over the noisy speech. Also, the PESQ at -4dB, -2dB and 0dB airport noise are improved by 15.83%, 13.95%, and 10.4% over the DBNB, respectively. The PESQ, SDR, STOI, and ESTOI gains of the proposed pre-training scheme are improving in all noise sources. The average PESQ, SDR, STOI, ESTOI, SSNR and SNRO improvements are demonstrated in Fig. 5. The average PESQ, SDR, STOI, ESTOI, SSNR and SNRO scores with the DBNP and DBNB are given in Table V. The outputs of various objective measures indicate that the proposed pre-training scheme is performing better. The average outputs (STOI, ESTOI, PESQ, SDR, SSNR and $SNR_O$) are improved over $DBN_B$ by 1.47%, 1.59%, 9.45%, 8.02%, 16.20% and 6.0%, respectively. In order to examine the noise reduction potentials of the proposed method, we used Segmental SNR (SSNR) and output SNR ($SNR_O$). It is clear from Table V that the proposed method attenuated the background noise and achieved better SSNR and $SNR_O$ as compared to other neural networks with the conventional pre-training scheme. Time-varying spectral analysis graphically demonstrates the vital speech patterns over the time at different frequency bands. In order to envisage performance of the proposed SE, spectrograms of the clean, noisy and enhanced speech samples are plotted in Fig. 6. For better understanding, PESQ, STOI, SDR and SSNR are pointed out over the spectrograms. It is noticeable that $DBN_P$ successfully attenuated the background noise frequencies, and provides a better reconstructed speech compared to the $DBN_B$. In order to envisage the impacts of phase estimation in the proposed method, spectrograms of the clean, noisy speech, $DBN_P$, and $DBN_B$ outputs are plotted in Fig. 6. The proposed pre-training scheme considerably improved the speech quality and intelligibility.

### B. Comparison With Related Methods

The proposed DBN-based SE method is further judged against other related SE methods including baseline DBN ($DBN_B$), DNN, deep denoising autoencoder (DDAE) [57], and LMMSE to validate the performance. It is observed that the DBN with proposed pre-training scheme ($DBN_P$) achieved considerable improvements in terms of the PESQ, STOI, and SDR as well as outscored the related SE methods. On the other hand, the PESQ, STOI, and SDR scores of the baseline DBN underperformed as compared to the DNN and DDAE. Table VI validated that $DBN_P$ outscored the baseline DBN, DNN and DDAE, as well as LMMSE with reasonable margins. For illustration, the STOI is improved from 67.12% with $DBN_B$ at -4dB airport noise to 72.13% with $DBN_P$ and improved STOI by 5.01%. Similarly, the PESQ is improved from 1.56 with $DBN_B$ at -4dB airport noise to 1.79 with $DBN_P$ and improved PESQ by 14.74%. Also, the SDR is improved from 0.98dB with DNN, 0.73dB with DDAE and 0.23 with LMMSE to 1.57dB with

TABLE VI. Average Performance Evaluation Against Related Speech Enhancement Methods

| Processing Methods | -4dB | | | -2dB | | | 0dB | | | 2dB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | STOI | PESQ | SDR | STOI | PESQ | SDR | STOI | PESQ | SDR | STOI | PESQ | SDR |
| Noisy | 58.90 | 1.45 | -4.11 | 63.0 | 1.58 | -1.81 | 67.50 | 1.70 | 0.13 | 72.26 | 1.83 | 2.11 |
| $DBN_P$ | 72.13 | 1.79 | 1.57 | 76.45 | 2.02 | 4.33 | 80.76 | 2.18 | 5.94 | 83.11 | 2.23 | 7.23 |
| $DBN_B$ | 67.12 | 1.56 | 1.09 | 73.54 | 1.81 | 3.97 | 78.60 | 1.95 | 5.52 | 82.05 | 2.11 | 6.97 |
| DNN | 71.33 | 1.61 | 0.98 | 75.87 | 1.87 | 4.02 | 80.43 | 2.02 | 5.23 | 83.47 | 2.19 | 7.08 |
| DDAE | 70.16 | 1.54 | 0.73 | 73.77 | 1.73 | 3.21 | 77.71 | 1.89 | 4.32 | 80.82 | 2.00 | 6.42 |
| LMMSE | 65.33 | 1.48 | 0.23 | 69.31 | 1.67 | 2.18 | 71.33 | 1.78 | 2.98 | 75.11 | 1.93 | 3.87 |

TABLE VII. Output SNR and SSNR Performance at Input SNRs Against Related SE Methods

| Methods | -4dB | | | -2dB | | | 0dB | | | 2dB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $SNR_O$ | ΔSNR | SSNR | $SNR_O$ | ΔSNR | SSNR | $SNR_O$ | ΔSNR | SSNR | $SNR_O$ | ΔSNR | SSNR |
| $DBN_P$ | 3.56 | 7.67 | 2.43 | 4.75 | 6.75 | 3.31 | 6.36 | 6.36 | 3.73 | 7.21 | 5.21 | 4.94 |
| $DBN_B$ | 2.98 | 6.98 | 1.92 | 4.03 | 6.03 | 2.92 | 5.76 | 5.76 | 3.08 | 6.31 | 4.31 | 3.67 |
| DNN | 3.01 | 7.01 | 1.98 | 4.12 | 6.12 | 3.01 | 5.89 | 5.89 | 3.11 | 6.53 | 4.53 | 3.98 |
| LMMSE | 2.23 | 6.23 | 1.61 | 2.73 | 4.73 | 1.79 | 4.88 | 4.88 | 2.74 | 5.79 | 2.73 | 3.27 |

$DBN_P$ and improved SDR by 0.56dB, 0.84dB and 1.34dB, respectively. At 2dB SNR, the average STOI is improved from 82.05% with $DBN_B$ to 83.11% with $DBN_P$ and improved STOI by 1.06%. Similarly, the PESQ is improved from 2.11 with $DBN_B$ to 2.23 with $DBN_P$ and improved PESQ by 5.68%. Also, the SDR is improved from 7.08dB with DNN, 6.42dB with DDAE and 3.87dB with LMMSE to 7.23dB with $DBN_P$ and improved SDR by 0.15dB, 0.81dB and 3.36dB, respectively.

Table VII demonstrates the performance of the proposed pre-training method in terms of the $SNR_O$, ΔSNR, and SSNR, respectively. The SSNR measure is employed in order to quantify the residual noise distortion in the output speech. The proposed $DBN_P$ considerably improved the $SNR_O$ and achieved significant performance gain in terms of the $SNR_O$. The ΔSNRs for $DBN_P$ are higher that for the related SE methods. For example, the $SNR_O$ at -4dB is improved from 2.98dB with $DNN_B$, 3.01dB with DNN and 2.23dB with LMMSE to 3.56dB with $DNN_P$ and increased $SNR_O$ by 0.58dB, 0.55dB, and 1.33dB, respectively. In case of the SSNR, a consistent output score signifies that DBN-based SE with the proposed pre-training notably attenuated the background noise, confirmed by time-varying spectrograms in Fig. 6.

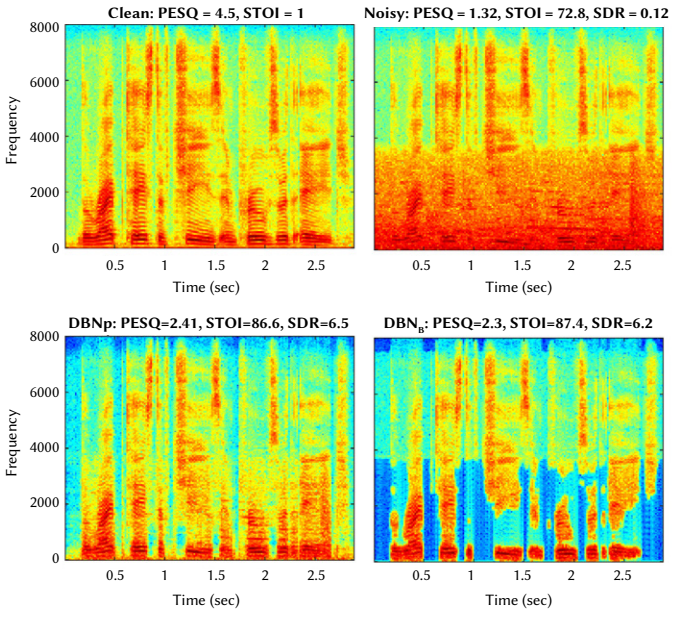



Fig. 6. Time-varying spectral analysis.

## C. Network Complexity and Convergence

The complexity of $DBN_B$/$DBN_P$ relies on the number of network parameters and the forward-backward propagation during tuning of the neural network. In the proposed DBN-based SE method, we have initialized the network parameters by a novel pre-training scheme instead of the randomly initialization. We observed that a network initialized with the proposed pre-training scheme converges quickly as compared to the randomly initialization or initialization with typical pre-training scheme. Moreover, the network complexity also relies on the quantity of hidden neurons and their weights. Greater the quantity of hidden neurons greater will be the network complexity. All DBNs have similar network architecture, quantity of hidden layers, neurons in the hidden and visible layers; however, the proposed DBN converged quickly and showed less complexity. The reason behind the quick convergence (less MSE loss) is incorporation of the novel pre-training scheme. With similar hidden neurons quantity, the proposed DBN-based SE method provided lower MSE errors, and this fact can be observed in Fig. 7. The complexity of the proposed DBN is illustrated in Table VIII, symbolized by "$O$". The forward-backpropagation relies on input features dimension: $F_D$, training data quantity: $T_D$, quantity of hidden neurons: $T_H$, quantity of output neurons $T_O$, and quantity of epochs for parameters tuning $T_E$.

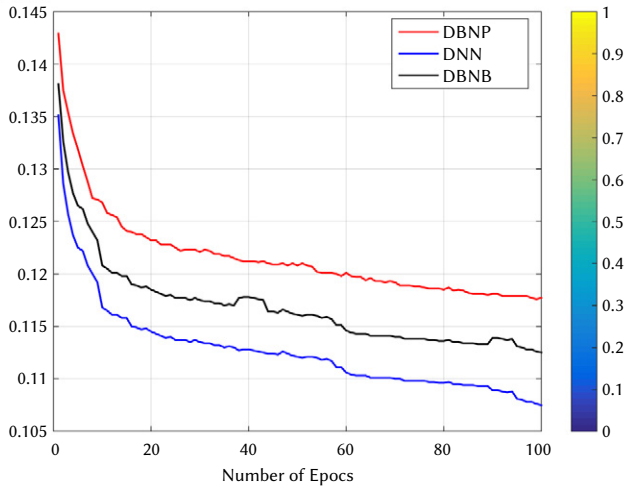



Fig. 7. MSE loss function.

TABLE VIII. Complexity of the Network

| Operations | Proposed Network |
|---|---|
| Forward-Backward Propagation | $O(T_D T_E (F_D + T_H + 2T_H^2 + T_H T_O))$ |
| Average Pre-Training Time | 3.33 Hours for 2000 Utterances |
| Average MSE at 100 Epochs | 0.08 Approximately |

## VII. Discussion and Conclusions

We have proposed and examined a supervised DBN-based speech enhancement method to reduce/attenuate the background noise in single-channel systems. We pre-trained and fine-tuned the DBNs by employing a novel pre-training scheme which incorporates important information available in the raw data to learn a T-F mask. The estimated mask is applied to the noisy speech by using the noisy phase to achieve the enhanced version of degraded speech. In the proposed DBN framework, the acoustic features are progressively extracted by multiple-stacked RBM during the pre-training. The hidden acoustic features from the preceding RBM are combined with raw input data to serve as the new inputs to in-progress RBM. By feeding the raw data to RBMs, layer-wise features related to the raw data can progressively be extracted, which showed useful to mine valuable information in the raw data. The proposed study used the estimated phase during the speech reconstruction to further improve the performance. All acoustic features are the integration of the raw acoustic features in windows, since temporal-dynamics provides important information for speech. The fundamental perception to utilize temporal-dynamics is to employ the DBN architecture, an extension of the feedforward DNN. The DBN framework grabs the long-term temporal-dynamics by using the pre-trained RBM parameters. $DBN_P$ are pre-trained to estimate the IRM, and achieved by 1.47%, 1.59%, 9.45%, 8.02%, 16.20% and 6.0% improvements over the $DBN_B$ in terms of the STOI, ESTOI, PESQ, SDR, SSNR and $SNR_O$, respectively. The achieved improvements are significant in the speech enhancement. In order to test the generalization ability of the proposed DBN, we have employed the TIMIT database which is composed of male and female speakers. The ΔSNRs and SSNR for $DBN_P$ are higher compared to the related SE methods. We achieved less computational complexity and quick convergence as compared to the baseline DBNs. The spectrogram of the $DBN_P$ indicates a better reconstructed speech signal, suggesting the benefits of the proposed pre-training scheme. To summarize, the proposed DBN-based SE is simple and performed better in terms of improving the intelligibility and quality in the background noisy environments.

### Future Work

Presently, most of the speech processing algorithms operate only with the spectral magnitude, leaving the spectral phase unexplored. With recent advancement in deep neural networks, the phase processing became more important as an innovative and emergent prospective of the DNN-based speech enhancement. In the future, the authors will develop the DBN with phase estimation to test the intelligibility and quality potentials in the complex noisy environments. The unsupervised learning algorithms with modifications can also lead to comparable performances [58] - [60].

## References


[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on acoustics, speech, and signal processing, vol. 27, no. 2, pp.113-120, 1979.

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

[3] S. Nasir, A. Sher, K. Usman, U. Farman, "Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation," Research Journal of Applied Sciences, Engineering and Technology, vol. 6, no. 6, pp.1081-1087, 2013.

[4] B.L Sim, Y.C. Tong, J.S. Chang, C.T. Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE transactions on speech and audio processing, vol. 6, no. 4, pp.328-337, 1998.

[5] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp.197-210, 1978,

[6] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (Vol. 2, pp. 629-632), IEEE, 1996.

[7] Y. Sandoval-Ibarra, V.H. Diaz-Ramirez, V. I. Kober, V.N. Karnaukhov, "Speech enhancement with adaptive spectral estimators," Journal of Communications Technology and Electronics, vol. 61, no. 6, 672-678, 2016.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 6, pp. 1109-1121, 1984.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE transactions on acoustics, speech, and signal processing, vol. 33, no. 2, pp.443-445, 1985.

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

[11] N. Mohammadiha, P. Smaragdis, A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2140-2151, 2013.

[12] I. Tashev and M. Slaney, "Data driven suppression rule for speech enhancement," in 2013 Information Theory and Applications Workshop (ITA) (pp. 1-6). IEEE, 2013.

[13] Y. Xu, J. Du, L.R. Dai, C.H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp.7-19, 2014.

[14] Y. Xu, J. Du, L.R. Dai, C.H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal processing letters, vol. 21, no. 1, pp.65-68, 2013.

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

[16] N. Saleem, M.I. Khattak, A.B. Qazi, "Supervised speech enhancement based on deep neural network," Journal of Intelligent & Fuzzy Systems, vol. 37, no. 4, pp.5187-5201, 2019.

[17] N. Saleem, M. Irfan Khattak, M.Y. Ali, M. Shafi, "Deep neural network for supervised single-channel speech enhancement," Archives of Acoustics, vol. 44, 2019.

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

[19] Y. Wang, A. Narayanan, D. Wang, "On training targets for supervised speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp.1849-1858, 2014.

[20] G. Kim, Y. Lu, Y. Hu, P.C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," The Journal of the Acoustical Society of America, vol. 126, no. 3, pp.1486-1494, 2009.

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

[22] J.H. Chang, Q.H. Jo, D.K. Kim, N.S. Kim, "Global soft decision employing support vector machine for speech enhancement," IEEE Signal Processing Letters, vol. 16, no. 1, pp.57-60, 2008.

[23] K. Kwon, J.W. Shin, N.S. Kim, "NMF-based speech enhancement using bases update," IEEE Signal Processing Letters, vol. 22, no. 4, pp.450-454, 2014.

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

[25] N. Saleem, M.I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," Applied Soft Computing, vol.

95, p.106666, 2020.
[26] N. Saleem, M.I. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp.84-90, 2020.
[27] Y. Wang, D. Wang, "Towards scaling up classification-based speech separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp.1381-1390, 2013.
[28] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa, M. Iwahashi, "Noise robust voice activity detection using joint phase and magnitude based feature enhancement," Journal of ambient intelligence and humanized computing, vol. 8, no. 6, pp.845-859, 2017.
[29] P.S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp.2136-2147, 2015.
[30] N. Saleem, M.I. Khattak, E.V. Perez, "Spectral Phase Estimation Based on Deep Neural Networks for Single Channel Speech Enhancement," Journal of Communications Technology and Electronics, vol. 64, no. 12, 1372-1382, 2019.
[31] X.L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 24, no. 5, pp.967-977, 2016.
[32] S. Samui, I. Chakrabarti, S.K. Ghosh, "Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network," Applied Soft Computing, vol. 74, pp.583-602, 2019
[33] N. Saleem, M.I. Khattak, A. Jan, "Multi-objective long-short term memory recurrent neural networks for speech enhancement," Journal of Ambient Intelligence and Humanized Computing, pp.1-16, 2020.
[34] R. Karakida, M. Okada, S.I. Amari, "Dynamical analysis of contrastive divergence learning: Restricted Boltzmann machines with Gaussian visible units," Neural Networks, vol. 79, pp.78-87, 2016.
[35] S. Samui, I. Chakrabarti, S.K. Ghosh, "Deep Recurrent Neural Network Based Monaural Speech Separation Using Recurrent Temporal Restricted Boltzmann Machines," in INTERSPEECH (pp. 3622-3626), 2017.
[36] Z. Chen, Y. Huang, J. Li, Y. Gong, "Improving Mask Learning Based Speech Enhancement System with Restoration Layers and Residual Connection,", in INTERSPEECH (pp. 3632-3636), 2017.
[37] A. Fischer, C. Igel, "An introduction to restricted Boltzmann machines," in Iberoamerican congress on pattern recognition (pp. 14-36). Springer, Berlin, Heidelberg, 2012.
[38] M. Aoyagi, "Learning coefficient in Bayesian estimation of restricted Boltzmann machine," Journal of Algebraic Statistics, vol. 4, no. 1, pp. 31-58, 2013.
[39] I. Sutskever, G.E. Hinton, G.W. Taylor, "The recurrent temporal restricted boltzmann machine," in Advances in neural information processing systems, (pp. 1601-1608), 2009.
[40] N. Zhang, S. Ding, J. Zhang, Y. Xue, "An overview on restricted Boltzmann machines," Neurocomputing, vol. 275, pp.1186-1199, 2018.
[41] S.R. Chiluveru and M. Tripathy, "Low snr speech enhancement with dnn based phase estimation," International Journal of Speech Technology, vol. 22, no. 1, pp. 283-292, 2019.
[42] S.K. Roy, A. Nicolson, K.K. Paliwal, "DeepLPC: A deep learning approach to augmented Kalman filter-based single-channel speech enhancement," IEEE Access, vol. 9, pp. 64524-64538, 2021.
[43] K. Tan, D. Wang, "Towards Model Compression for Deep Learning Based Speech Enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1785-1794, 2021.
[44] S.K. Roy, A. Nicolson, K.K. Paliwal, "DeepLPC-MHANet: Multi-Head Self-Attention for Augmented Kalman Filter-based Speech Enhancement," IEEE Access, vol. 9, pp. 70516-70530, 2021.
[45] A. Pandey, D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1270-1279, 2021.
[46] S. Abdullah, M. Zamani, 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, 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] Y. Wang, Z. Pan, X. Yuan, C. Yang, W. Gui, "A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network," ISA transactions, vol. 96, pp. 457-467, 2020.
[49] G.E. Hinton, "A practical guide to training restricted Boltzmann machines," In Neural networks: Tricks of the trade (pp. 599-619). Springer, Berlin, Heidelberg, 2012.
[50] G.E. Hinton, S. Osindero, Y.W. Teh, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp.1527-1554, 2006.
[51] V. Zue, S. Seneff, J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech communication, vol. 9, no. 4, pp. 351-356, 1990.
[52] D. Pearce and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep., 2002.
[53] Q. Wang, J. Du, L.R. Dai, C.H. Lee, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp.1185-1197, 2018.
[54] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 2, pp. 749-752). IEEE, 2001.
[55] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in 2010 IEEE international conference on acoustics, speech and signal processing (pp. 4214-4217). IEEE, 2010.
[56] J. Jensen and C.H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 11, pp. 2009-2022, 2016.
[57] H.P. Liu, Y. Tsao, C.S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," Speech Communication, vol. 104, pp.106-112, 2018.
[58] T. Lavanya, T. Nagarajan, P. Vijayalakshmi, "Multi-Level Single-Channel Speech Enhancement Using a Unified Framework for Estimating Magnitude and Phase Spectra," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1315-1327, 2020.
[59] N. Saleem, M.I. Khattak, E. Verdú, "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, 2020.
[60] N. Saleem and T.G. Tareen, "Spectral Restoration based speech enhancement for robust speaker identification," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 1, pp. 34-39, 2018.

Muhammad Irfan Khattak

Dr. Muhammad Irfan Khattak is working as an Associate Professor in the Department of Electrical Engineering in University of Engineering and Technology Peshawar. He did his B. Sc Electrical Engineering from the same University in 2004 and did his PhD from Loughborough University UK in 2010. His research interest involves Antenna Design, On-Body Communications, Machine learning, Speech processing and Speech Enhancement.

Nasir Saleem

Dr. Nasir Saleem received BS and M.S degree in Electrical Engineering from UET Peshawar in 2008 and 2012. He received Ph.D. Electrical Engineering, major in Deep Learning for speech processing from UET Peshawar, Pakistan. He is now an Assistant Professor in Department of Electrical Engineering, Gomal University, Pakistan. He published a number of research papers on the deep learning applications in speech processing. His research interests are in the area of Machine learning, digital speech processing and speech enhancement.

Aamir Nawaz

Dr. Aamir Nawaz received BS and MS degree in Electrical Engineering from UET Peshawar and UET Taxila, Pakistan in 2009 and 2014. He received Ph. D in Electrical Engineering, major in Power Systems from Shandong University China. He is serving as Lecturer in the Department of Electrical Engineering, Gomal University, Pakistan. His research interests are in the area of Power Engineering, Machine learning for Power system protection and analysis.

Aftab Ahmed Almani

Mr. Aftab Ahmed Almani received BS and MS degree in Electrical Engineering and pursuing Ph. D in Electrical Engineering, major in Power Engineering from Shandong University China. His research interests are in the area of Power Engineering, Machine learning for Power systems.

Farhana Umar

Dr. Farhana Umar received BS and MS degree in Electrical Engineering from Mehran University of Science & Technology, Jamshoro and Ph.D. Electrical Engineering, from Selçuk Üniversitesi Konya, Turkey. Her research interests are in the area of Power Engg., Machine learning for Power Engg.

Elena Verdú

Elena Verdú Pérez received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e-learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, speech and image processing, data mining and expert systems.