IJICTR
International Journal of Information & Communication Technology Research

Volume 15- Number 3 – 2023 (53 -65)

NMF-based Improvement of DNN and LSTM Pre-Training for Speech Enhancement

Razieh Safari Dehnavi
Department of Electrical Engineering
Amirkabir University of Technology
(Tehran Polytechnic)
Tehran, Iran
r.safari@aut.ac.ir

Sanaz Seyedin*
Department of Electrical Engineering
Amirkabir University of Technology
(Tehran Polytechnic)
Tehran, Iran
sseyedin@aut.ac.ir

Received: 1 June 2023 – Revised: 19 August 2023 - Accepted: 1 September 2023

Abstract—**A novel pre-training method is proposed to improve deep-neural-networks (DNN) and long-short-term-memory (LSTM) performance, and reduce the local minimum problem for speech enhancement. We propose initializing the last layer weights of DNN and LSTM by Non-Negative-Matrix-Factorization (NMF) basis transposed values instead of random weights. Due to its ability to extract speech features even in presence of non-stationary noises, NMF is faster and more successful than previous pre-training methods for network convergence. Using NMF basis matrix in the first layer along with another pre-training method is also proposed. To achieve better results, we further propose training individual models for each noise type based on a noise classification strategy. The evaluation of the proposed method on TIMIT data shows that it outperforms the baselines significantly in terms of perceptual-evaluation-of-speech-quality (PESQ) and other objective measures. Our method outperforms the baselines in terms of PESQ up to 0.17, with an improvement percentage of 3.4%.**

Keywords: pre-training, deep neural networks (DNN), long short-term memory (LSTM), non-negative matrix factorization (NMF), speech enhancement, basis matrix, noise classification

Article type: Research Article

Publisher: ICT Research Institute

[DOI: 10.61186/itrc.15.3.53]

I. INTRODUCTION

Speech signal enhancement is a widely used practical block in many applications such as speech recognition (ASR), and mobile speech communication [1]. The main purpose of speech enhancement is to remove static or non-static noise from the noisy signal. Speech enhancement techniques may be categorized into three divisions of statistical methods, machine learning-based procedures (and more specifically deep learning ones), and association of statistical and sparse models with machine learning as the brand-new approach [2, 3, 4, 5]. Spectral subtraction methods are among the classic ones in the category of statistical procedures. There is a possibility of music noise in spectral subtraction methods [1]. Statistical methods are implemented based on specific statistical assumptions [4]. Wiener filter [6], minimum mean square error (MMSE) [7, 8], and Non-Negative Matrix Factorization (NMF) are some of the statistical

* Corresponding Author

techniques used in this field [9]. Deep learning is one of the methods that has recently been considered for noise removal [10]. Due to the high performance of neural networks in eliminating various noises, this method has been one of the successful methods in improving speech signal [11].

In [12], SNR-based progressive learning method was proposed for DNN. This method achieved good results in speech quality while the number of required parameters for network training was also decreased. In [13], to find a comprehensive model against different types of noise, noise classification has been used in neural network training. In [14], LSTM was used as a mask for speech noise suppression. This structure can extract temporal information about speech and noise. A combination of LSTM and CNN layers was used for speech enhancement in [15]. With this structure, contextual information on speech signals was extracted. In [16], a method for speech enhancement was proposed with simple recurrent units (SRU). In this method, an SRU network with some layers was used to estimate clean speech signal.

In [17], NMF basis matrix was estimated using a ratio mask in the DNN structure. Clean speech was then estimated using NMF and the basis matrix. In a different approach, two autoencoders were employed to extract clean speech and noise NMF parameters in [5]. After that, the outputs of the encoder parts were used as input features in a DNN structure to estimate clean speech signals. In summary, DNNs have been suggested in many ways to improve speech signal [18],[19],[20]. Despite the good performance of neural networks in many areas, including speech improvement, local minimum is the main problem in the learning stage. Increasing the number of network layers leads to the increase of nonlinearity in feature extraction. However, the probability of facing a local minimum issue will increase as the number of layers and network parameters increases. Using pre-training methods is one of the useful techniques to overcome the local minimum problem [21]. One pre-training method is the use of restricted Boltzman machines (RBM) [22]. In this method, the weights of the DNN layers are trained using RBM and the resultant values are used as the initial weights of the DNN. Supervised and unsupervised pre-trainings for deep belief networks (DBN) have been evaluated in [23]. In this approach, bidirectional pre-training was proposed to calculate the initial values of DNN weights for image classification. The results show an improvement in the accuracy and speed of learning in the proposed method. In [24], greedy layer-wised pretraining and fine tuning was used. In this method, autoencoder (AE) networks and deep denoising autoencoder (DDAE) have been used for pre-training and fine tuning in speech enhancement, respectively. Also, noisy and clean speech signals have been used as input and output of AEs. Hence, we refer to this method as a supervised pre-training (SUP) structure.

NMF is one of the useful linear methods in improving the speech noisy signal and extracting the relationships between speech signal and noise in an appropriate way [9]. Therefore, instead of finding the initial values of the network weights using traditional pre-training, NMF can be used as a useful method in extracting speech signal information. In this paper, we propose a novel method based on NMF for both DNN and LSTM pre-training in speech enhancement to overcome such training problems of deep networks as the local minima. The proposal of employing the NMF basis matrices as the initial weights in deep networks in this paper is due to the fact that NMF is known to be an appropriate sparse model for extracting speech features [25]. In fact, NMF is trained by clean speech signals to decompose clean speech into two matrices of *basis* and *coefficients* [9]. Thus, the basis matrix works as an appropriate data-driven filter capable of finding proper speech features in the coefficients matrix. In our proposed method, we obtain the initial weights of the last layer of the network using the transpose values of NMF basis matrix. Because NMF is trained by clean speech, it is able to adjust the weights appropriately while mapping the features of the last layer to the target output (enhanced speech). We also propose the use of basis matrix of the suggested NMF pretraining in both the first and last layers of the network. Insertion of the proposed NMF pre-training approach in the supervised AE pre-training structure has also been introduced. The results show that this method is superior to previous AE supervised pre-training. In order to improve the speech enhancement network quality in noisy environments, we use a noise classification method with individual networks used for each noise type. Fig. 1 shows a simple block diagram of the proposed approach. We find the basis matrix of the NMF algorithm from clean speech signal X, and use the basis matrix for the pre-train block in Fig. 1. A DNN is trained with input noisy speech signal Y for noise classification. According to Fig. 1, we acquire the enhanced signal either from path 1 or 2. In path 1, we use individual models for matched noises (those seen in the training phase). For the mismatched noise types that have not been seen in the training step, we suggest using a general model trained with all noises to improve the generalization of our approach in different noisy environments. Please note that the noise classification parts in Fig. 1 are optional. The basic proposed block of our system which is the NMF pre-training approach is painted in green to signify its importance. Also, we have used the system structure of Fig. 1 only for the proposed model and this structure has not been used in other reference methods. Other reference methods use a general model that is trained with all noises without any NMF pre-training which is the main novelty of the current paper.

To the best of authors' knowledge, no pretraining has been used for LSTM structures for speech enhancement so far. However, according to the results obtained in our proposed approach, it is beneficial for LSTM networks as well.

Therefore, the main contributions of this paper are as follows:

[DOI: 10.61186/itrc.15.3.53]

- Proposing a novel method based on NMF for both DNN and LSTM pre-training in speech enhancement: we propose NMF pre-training in three approaches of initializing the weights of the last layer, both the first and last layer of the network based on NMF basis matrix. We also suggest inserting the proposed NMF pre-training approach in the supervised AE pre-training structure. Specifically, since NMF is trained by clean speech, it could adjust the weights appropriately while mapping the features of the last layer to the target output (enhanced speech).
- Suggesting a noise classification method with individual networks used for each noise type: we use individual models pre-trained by the proposed NMF approach for matched noises that have been seen in the training phase. For the mismatched noise types that have not been seen in the training step, we suggest using a general model which has also been pre-trained by NMF and tuned with all noises to improve the generalization of our approach in different noisy environments.

This paper is organized as follows. Section II explains the NMF structure. Section III defines the AE pre-training. In section IV, the proposed approach is presented. The experimental setting is mentioned in Section V. Discussion and Conclusion are presented in Sections VI and VII, respectively.

II. NMF DESCRIPTION

The NMF algorithm linearly decomposes X as a non-negative matrix into W and H matrices. W is called the basis matrix, while H refers to coefficients. In (1), $X \in R_{\geq 0}^{M \times N}$ is an input matrix, $W \in R_{\geq 0}^{M \times K}$ and $H \in R_{\geq 0}^{K \times N}$ are the two non-negative factors. Dot in (1) means the product of W and H. M, N, and K determine the sizes of the matrices. The relationship between these parameters is shown in (2) [18]:

$$X \approx W.H \quad (1)$$

$$K \leq \min(M, N) \quad (2)$$

N is equal to the number of frames in X. W and H matrices are updated frequently, until the objective function $D(X \mid WH)$ is minimized and the best approximation is obtained from the input matrix. Since the NMF basis and coefficients matrices are found by a training strategy on the input matrix, it is among data-driven statistical approaches. The distance between X and $W.H$ could be minimized with Frobenius, itakura-saito, or kullback-leibler method [18].

III. AE PRE-TRAINING

In pre-training with AE networks (Fig. 2), the deep network is first decomposed into a number of networks with single hidden layers. If we denote the input and output of the network as Y and X, respectively, first a single-layer network is learned

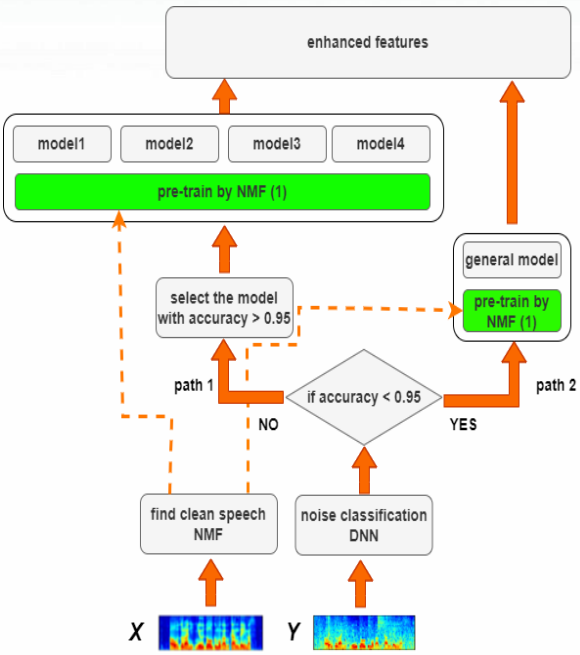

Figure 1. The simple block diagram of the proposed approach based on NMF pre-training using the noise classification strategy. X and Y refer to clean and noisy speech, respectively. Path 1 is for matched noises, while we suggest path 2 for mismatched ones. "modeli" refers to the individual models trained with noise type i for matched noises. The general model is trained with all noises to improve the results in mismatched conditions. The noise classification parts are optional and the basic blocks of our system are painted in green to signify their importance.

With input Y and target X. In the next step, the hidden layer of the trained network is considered as the input of the next network, and so on, until all single-layer networks are trained. Using the weight matrices calculated by the AE networks as shown in Fig. 2, the initial values of the DNN weights will be obtained [24]. In the fine-tuning phase, the DNN will be trained using the calculated weight matrices, and the values of the network weights in the DNN will be adjusted [24].

IV. PROPOSED PRE-TRAINING

NMF is useful in decomposing clean speech signal into two matrices of basis W, and coefficients H. In fact, W works as a data-driven filter which helps find proper speech features such as formants and harmonics reflected in H [25]. Therefore, the output of a one-layer neural network could represent these suitable features when the input is clean speech, and W has a sparse distribution that improves the generality of the network for different speech signals and noises. Hence, the proposed structure includes the use of NMF in DNN and LSTM pre-training. Also, in [26], it is shown that with fix random weight values in all layers, except the last layer which is computed analytically, we can achieve acceptable results with lower computing complexity. This research shows the importance of utilizing analytical values in the last layer and random values for other layers. Therefore, we propose a pretraining method for the last layer of the network. For further evaluation, we examine the proposed initialization for the first layer as well.

[DOI: 10.61186/itrc.15.3.53]

Initially, we extract the basis matrix of the NMF from the clean speech signal X as the target of the DNN, using (1). In the next step, we use the transpose of basis matrix as the initial weights of the last layer of the DNN. The dimensions of the basis matrix in NMF are set equal to the number of nodes in the last layer times the number of nodes in the hidden layer before it in this case when X is the target layer instead of the input as in (1). Thus, we should transpose its values to be used as the initial weights in the network. The proposed method is illustrated in Fig. 3. Once again, Y and X are equivalent to noisy and clean speech signals, respectively, and W_n denotes the basis matrix calculated by NMF in the nth layer. We call this method as PNMF1.

It is also possible to use the proposed NMF weight initialization in the supervised pre-training scenario (SUP) [24]. We use noisy and clean speech signals as input and output target values in pre-training. Fig. 4 depicts the use of the proposed NMF pretraining in supervised pre-training structure (PNMF1_SUP). The proposed method in Fig. 4 is similar to the supervised pre-training method, except that instead of the initial weights of the last layer of the neural network, the NMF basis matrix is used. In this structure, the initial layers of the neural network are displayed in light blue, and a dark blue layer represents the layers at greater depths. As seen, all the weights of the network, except the weights of the last layer, have been calculated by supervised pre-training method. In other words, we initialize the weights of the last layer by the NMF algorithm. After calculating the initial values of the network weights, the weights of all network layers are readjusted to reach an acceptable value of the network error using fine-tuning method.

We also propose using the NMF basis matrix (W) as the initial values of the first network layer for improving the pretraining method. Here, according to (1), we assume having clean speech X as the input, and W represents the weight matrix mapping X into coefficients H. In this structure, we also suggest using the transpose of the NMF basis matrix as the initial weights of the last layer of the DNN network. We call this method as PNMF2. A diagram of the PNMF2 approach is shown in Fig. 5.

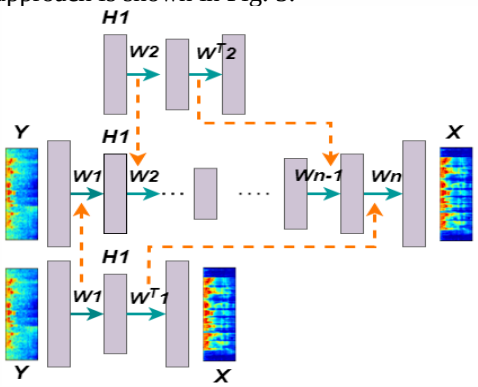

Figure 2. The pre-training method with AE networks [24]

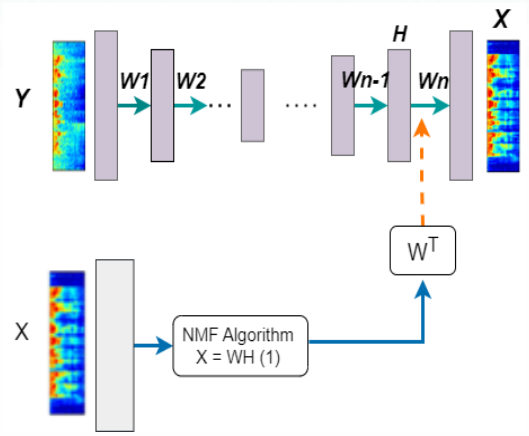

Figure 3. The proposed pre-training method (PNMF1) using NMF in the DNN structure. Wn is initialized with transpose of NMF basis matrix (W).

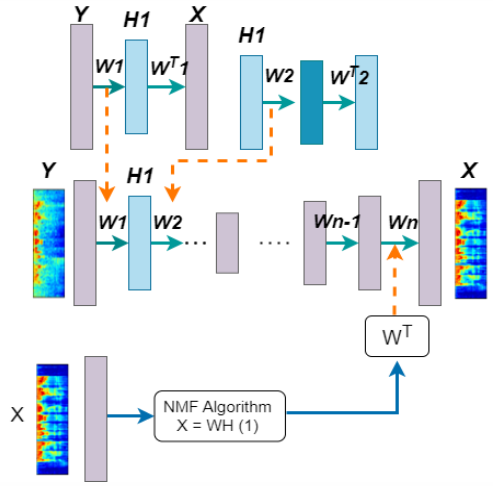

Figure 4. Inserting the proposed NMF pre-training approach in the supervised pre-training structure (PNMF1_SUP).

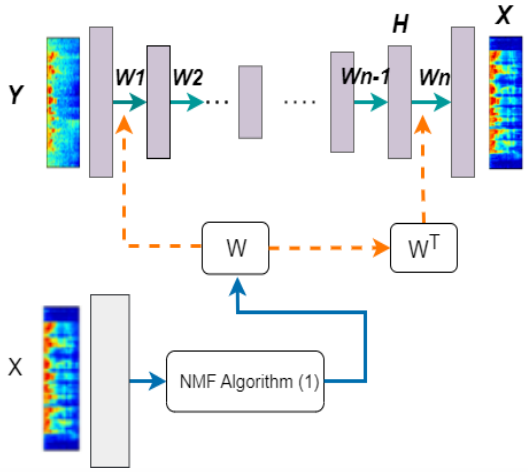

Figure 5. The proposed PNMF2 pre-training method using NMF in the DNN structure. Wn is initialized with transpose of the NMF basis matrix (WT) and W1 is initialized with W matrix.

In addition, similar to our proposed DNN pretraining method for PNMF1, we suggest using the transpose of basis matrix in the NMF method as the initial weights for the last layer of the LSTM network which we suggest tobe a linear layer. Due to the linearity of the last layer, we expect that LSTM be trained more efficiently with the proposed strategy.

[DOI: 10.61186/itrc.15.3.53]

In Fig. 6, W_f, W_i, W_o, and W_c are the weights of the forgetting gate, input gate, output gate, and new candidate one, respectively, in the LSTM network [14]. σ is the sigmoid activation function and tanh is the hyperbolic activation function. In this proposed pre-training method called PNMF2, we use the NMF algorithm basis matrix to initialize the W_i, W_f, W_c, and W_o matrices in training the LSTM network. According to the obtained experimental results shown later, this proposed pre-training approach works better than using random values as expected. One layer of the LSTM method is shown in Fig. 6. According to Fig. 6, W matrix of the NMF algorithm is used as initial weights for W_i, W_f, W_c, W_o. Also, W^T matrix is used as initial weights of the last layer.

We also propose employing the SUP pre-training method, similar to what used for the case of DNNs, in the LSTM network and evaluate the earned results of speech enhancement with our proposed approach.

In the proposed speech enhancement structure, we use different networks for different matched noise types. Thus, we initially use a noise classification procedure based on a DNN to determine the nearest noise type in each noisy mixture. The clean signal will be extracted with the specified network according to Fig. 1. The use of a DNN contributes to the accuracy of classification. If the noise classification accuracy is higher than a specified threshold, we mark it as a matched noise and thus, use one of the networks trained with a matched noise, namely either of model1, model2, model3, and model4 in Fig. 1. If noise classification accuracy is lower than the threshold, it is categorized as a mismatched noise type. Hence, we use a general network trained with all matched noises (general model in Fig. 1). This suggested strategy not only leads to better results for the matched noise types, but also improves the generalization of the network in mismatched conditions.

V. EXPERIMENTS AND EVALUATIONS

A. Experimental Setting

The data set used in this paper is the TIMIT corpus [27]. It contains 6300 sentences, of which 1334 sentences are considered as test data. Also, the total number of speakers is equal to 630 from which 168 speakers are used in the test set [27]. In this paper, we randomly select 700 sentences from TIMIT to train any individual network related to each noise type. The number of test sentences is equal to 120 for each type of noise. The data used for the test have been selected from the TIMIT test dataset and have no overlap with the training data. Also, we use the IEEE sentence database for our mismatched speech signals [28].

To evaluate the proposed system, Babble, Factory1 and F-16 noises from NATO RSG-10 dataset [29] and Car noise from AURORA-2 database [30] have been used to train the DNN and LSTM. Also, we use the Restaurant noise from AURORA-2 database, Pink noise from NATO RSG-10 dataset [28], and Piano noise from [31] for the mismatched noises. We generate noisy signals by adding clean speech signals with noise signals according to ITU-T P.56 standard [32] in SNRs of -5, 0, 5, 10, 15, and 20 dB.

The approach in [16] is one of the new methods we compare our results with. In [16], the signals are resampled to 8 kHz. Similarly, as we would like to use the proposed method for the telephony band and for fair comparison with [16], we have down-sampled the signals to 8 kHz. The features extracted from the speech signals are the magnitude of fast Fourier transform (FFT) with a frame length of 32 ms with 16 ms frame shift. Thus, the length of the feature vector used is equal to 129 (the first half of FFT).

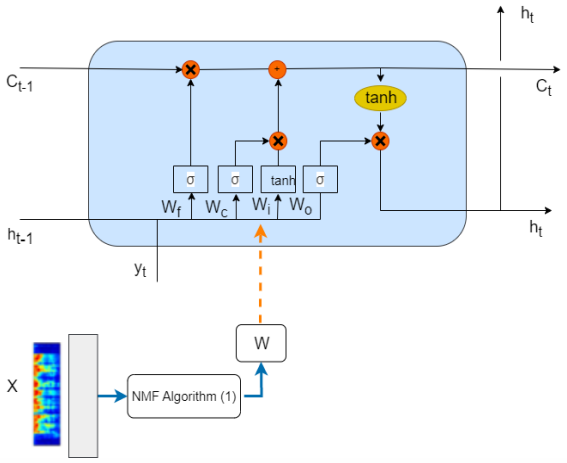

Figure 6. The proposed pre-training method (PNMF2) using NMF in the LSTM structure. Wf, Wc,Wi, Wo are initialized with the NMF basis matrix (W), while Wn (the last layer's weights not shown in the figure) is initialized with WT.

The enhancement is carried out in the time-frequency domain after the application of the short time Fourier transform (STFT). The inputs and outputs, for both DNN and LSTM, are noisy and clean speech features, respectively. We combine the noisy phase of the input signal with the enhanced features at the network output to reconstruct the enhanced speech signal.

Since the context of the speech signal is very important in speech enhancement, we suggest using both DNN and LSTM networks with context. Therefore, the DNN and LSTM models have 5 frames in the input and output of the networks with 550 nodes in hidden layers. The context features for the input and output of the network consist of the central, two past, and two future frames, so that the input and output sizes of the network are equal to 129*5 nodes. Also, the central frames of the network output are considered as the enhanced frames. The DNN trained for speech enhancement is a network with three hidden layers. The activation functions used in this network are Leaky-ReLU in all layers, except in the last layer which has a linear function. ReLU activation function has better results than sigmoidal and hyperbolic activation functions without using any unsupervised pre-training [33] and Leaky-ReLU maintains all feature information in a certain range in addition to having nonlinear properties and low computational

[DOI: 10.61186/itrc.15.3.53]

time of ReLU [34]. The LSTM model for speech enhancement has two hidden layers. These values have been found experimentally. In hidden layers, the LSTM layers are used, and the last layer is the linear layer. We use Adam for algorithm optimization. The batch size used in the DNN and LSTM is set to 1024. The negative slope in leaky-relu layers is 0.01. In DNN, the learning rate is set to 1e-4. In LSTM, the learning rate is set to 1e-3. Also, the period of learning rate decay is set to 30 for epochs greater than 5 and the multiplicative factor of learning rate decay is set to 0.5.

For fair comparison between methods, other initializations are the same for all networks and methods, and the initialization procedure is set as the default one in the Pytorch package.

We use the NMF models with rank of 550 (equal to the number of hidden layer nodes), and iteration of 100 for convergence. The initial values for NMF algorithm are randomly set. We use coordinate descent solver (cd) for this algorithm and the distance between X and $W.H$ is minimized with Frobenius method. The noise classifier network is a 5-layer DNN. The first three layers have the Leaky-Relu activation function, the fourth layer has the linear function, and the last layer is Softmax. The hidden layers of this network have 800 neurons. The data set for training the noise classifier network is the first 10 frames of each noisy utterance assuming silence at the beginning. The specified threshold in the classification method to categorize the noise type as either matched or mismatched one is set to 95 percent experimentally.

TABLE I. THE PARAMETER SETTINGS FOR THE PROPOSED APPROACH

Parameter	Value
frame length	32ms
frame shift	16ms
sampling frequency	8 kHz
network input and output size	129*5 = 645
hidden layers sizes	550
algorithm optimaization	adam
batch size	1024
negative slope in leaky-relu layers	0.01
learning rate for DNN	1.00E-04
learning rate for LSTM	1.00E-03
learning rate decay for epochs greater than 5 in LSTM	30
multiplicative factor of learning rate decay In LSTM	0.5
NMF matrice size	550*645
NMF iteration	100
solver for the NMF	coordinate descent (cd)
NMF minimization method	Frobenius

In Table 1, all parameter settings for the proposed methods are shown. All these parameters were based on the best PESQ scores in the training time as well as previous related papers.

All methods are trained using an NVIDIA GeForce GTX 1080 GPU system. The training times of DNN, LSTM, PNMF1 for DNN, and PNMF1 for LSTM networks are equal to 5287, 6265, 5849, and 7129 seconds, respectively, including the NMF training time for our proposed methods. The NMF training time for all proposed methods is 508 seconds

B. Objective Speech Quality Measures of matched noises

To evaluate the proposed methods, the widely used objective evaluations of PESQ [35], COVL [36], and fwsegSNR [37] have been used. The range of PESQ is from -0.5 to 4.5 and the range of COVL is from 1 to 5. COVL is a linear combination of the perceptual evaluation of speech quality, log-likelihood ratio (LLR), and weighted slope spectral (WSS) measures. In fwsegSNR, higher values refer to better results.

In this paper, we have proposed and tested four new methods for pretraining neural networks. These four methods are summarized as follows.

- PNMF1: The weights of the last layer of the network are initialized with the transpose of the NMF basis matrix.
- PNMF2: The weights of the first and last layers of the network are initialized with the NMF basis matrix and its transpose, respectively.
- PNMF1_SUP: The incorporation of the PNMF1 method within the SUP framework.
- PNMF2_SUP: The incorporation of the PNMF2 method within the SUP framework.

The results of the proposed methods are compared with those of the SRU [16], SUP [24], DNN [11], and LSTM [14] networks as the baseline methods. The difference between the proposed and baseline methods is in the approach taken to initialize the matrix weights in the networks described in the paper. Please note that some of the baselines such as DNN, LSTM, and SRU do not have any pre-training. Also, we have a noise classification block for the proposed methods. Table 2 briefly describe these differences.

The average results of PESQ and COVL over 120 test signals for car and F-16 noises in different SNRs are shown in Table 3 for DNN. Table 4 illustrates the average results over noisy signals contaminated with all four noises of Babble, Factory1, F-16 and Car noises for DNN in various SNRs. Table 5 shows the average results of PESQ and COVL over 120 test signals for Car and F-16 noises for LSTM in different SNRs. We have also shown the average results over noisy signals contaminated with all four noises of Babble, Factory1, F-16 and Car noises for LSTM in different SNRs in Table 6. We have achieved similar improvements for Babble and Factory1 noises for both DNN and LSTM networks too, but the results have not been reported in this paper to save space. To solidify the obtained results, the experiments of the proposed methods are repeated for five different initialization values of the weights.

For further comparison of the models, we have used the frequency-weighted segmental SNR

[DOI: 10.61186/itrc.15.3.53]

(fwsegSNR) as the evaluation method. The average results of fwsegSNR over Babble, Factory1, F-16 and Car noises for LSTM in various SNRs are shown in Table 7.

The network learning curve for Babble noise for some baselines and the best proposed method is shown in Fig. 7. In Fig. 7, the amount of network error during the learning process is plotted in terms of the number of epochs. DNN and SUP methods are used as baselines. PNMF1 has the best results and is the best-proposed method.

For more evaluation, we compare the best proposed method with SRU model [16] as the most recent baseline. The SRU network was trained with Babble, F-16, Factory1, and Car noises (Fig. 8).

In Fig. 9, the spectrograms of the noisy, clean, PNMF1 output, and SUP output are shown. Fig. 10 depicts the NMF basis values, a matrix with random weight, the last layer matrix weight of the trained LSTM model, and the last layer matrix weight of the trained PNMF1 model for LSTM.

For evaluating the effect of noise classification in the proposed method, in Fig. 11, the PESQ results of the best-proposed method (PNMF1) while using the suggested noise classification strategy, and without noise classification, as well as the SRU model are shown. The experiments emphasize that a similar trend is followed for PESQ results in Fig. 11 with that of the fwsegSNR, and the proposed method has the best results with and without noise classification. We showed the results of fwsegSNR in Table 7.

C. Objective Speech Quality Measures of mismatched noise and mismatched database

In Table 8, the results of Restaurant noise as a mismatched noise in LSTM are shown. As seen, the PNMF1 method has the best result of all methods.

In Table 9, the average results of Pink and Piano noises as two periodic and mismatched noise signals in LSTM are shown. The results illustrate the best performance for the proposed methods. Fig. 12 illustrates the average PESQ results of the mismatched speech signal (IEEE sentence database) over F-16 noise for SRU, LSTM, and PNMF1 methods.

Table 10 illustrates the average results over noisy signals contaminated with all 7 matched and mismatched noises (Babble, Factory1, F-16, Car, Restaurant, Pink, Piano) for LSTM in various SNRs.

To evaluate the statistical improvement of the results caused by the proposed method, we use the Friedman test with the Holm's post hoc test [38],[39]. To find the significance of the results, the modified statistical value (F_f) of the Friedman test with (J-1) and (I-1)*(J-1) degrees of freedom and the critical F value (F) are calculated. J is defined as the number of methods, and I is defined as the number of conditions. In this experiment, J is equal to 7 and I is equal to 48 (6 different SNRs for 7 noise types and one mismatched speech condition). The null hypothesis is rejected if the F_f value is larger than the F value, and we can compare the results with the Holm's post hoc test. In this test, if the p-value is smaller than the Holm values, the null hypothesis is rejected, and that model has significant results. In this test, the F_f value is equal to 55.59 and the F value is equal to 2.130. Thus, the F_f value is larger than the F value and we use the Holm's post hoc test to find the significant methods. The results of Holm's post hoc test are illustrated in Table 11. As seen, the proposed methods have larger Holm's values than p-values. Thus, our proposed methods pass the significance test.

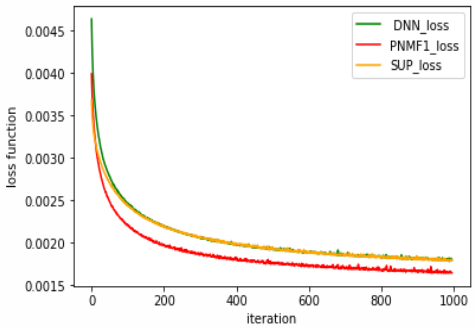

Figure 7. The amount of network error in the learning process for some baselines and the best proposed method.

TABLE II. DIFFERENT BASELINE AND PROPOSED METHODS.

method	type	explanation
DNN [11]	Reference method	A deep-neural-network is used for speech enhancement. No pre-training is used.
LSTM [14]	Reference method	A long-short-term-memory network is used for speech enhancement. No pre-training is used.
SUP [24]	Reference method	Greedy layer-wised pretraining and fine tuning was used. This method has a supervised pre-training (SUP) structure.
SRU [16]	Reference method	A method for speech enhancement with simple recurrent units (SRU). No pre-training is used.
PNMF1	Proposed method	The weights of the last layer of the network are initialized with the transpose of the NMF basis matrix.
PNMF2	Proposed method	The weights of the first and last layers of the network are initialized with the NMF basis matrix and its transpose, respectively.
PNMF1_SUP	Proposed method	The incorporation of the PNMF1 method within the SUP framework.
PNMF2_SUP	Proposed method	The incorporation of the PNMF2 method within the SUP framework.

[DOI: 10.61186/itrc.15.3.53]

TABLE III. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS IN DNN FOR CAR AND F-16 NOISES IN DIFFERENT SNRS.

PESQ - Car noise							COVL - Car noise						
SNR	-5	0	5	10	15	20	SNR	-5	0	5	10	15	20
DNN [11]	1.93	2.28	2.57	2.78	2.96	3.15	DNN [11]	2.34	2.79	3.16	3.42	3.63	3.84
SUP [24]	1.95	2.34	2.64	2.82	2.98	3.14	SUP [24]	2.38	2.85	3.22	3.44	3.63	3.81
SRU [16]	1.89	2.25	2.59	2.81	3.00	3.16	SRU [16]	2.24	2.69	3.12	3.41	3.63	3.83
PNMF1	**2.08**	**2.43**	**2.72**	**2.95**	**3.12**	**3.28**	PNMF1	**2.47**	**2.92**	**3.29**	**3.56**	**3.76**	3.95
PNMF2	2.05	2.37	2.64	2.85	2.99	3.14	PNMF2	2.44	2.86	3.22	3.47	3.65	3.81
PNMF1_SUP	2.04	2.38	2.67	2.90	3.07	3.27	PNMF1_SUP	2.45	2.90	3.27	3.55	3.75	**3.97**
PNMF2_SUP	2.01	2.34	2.64	2.84	3.00	3.15	PNMF2_SUP	2.38	2.83	3.21	3.46	3.65	3.82
PESQ - F-16 noise							COVL - F-16 noise						
SNR	-5	0	5	10	15	20	SNR	-5	0	5	10	15	20
DNN [11]	1.88	2.29	2.55	2.78	2.98	3.14	DNN [11]	2.30	2.80	3.12	3.40	3.65	3.82
SUP [24]	1.83	2.33	2.60	2.82	2.98	3.13	SUP [24]	2.26	2.84	3.17	3.43	3.63	3.79
SRU [16]	1.93	2.28	2.55	2.79	3.00	3.16	SRU [16]	2.27	2.72	3.06	3.37	3.63	3.82
PNMF1	1.95	**2.42**	**2.71**	**2.95**	3.13	3.27	PNMF1	**2.38**	**2.95**	**3.31**	**3.59**	3.80	3.95
PNMF2	**1.97**	2.38	2.64	2.87	3.02	3.15	PNMF2	2.36	2.88	3.21	3.49	3.67	3.82
PNMF1_SUP	1.95	2.38	2.68	2.95	**3.14**	**3.29**	PNMF1_SUP	2.36	2.91	3.27	3.58	**3.81**	**3.97**
PNMF2_SUP	**1.97**	2.37	2.64	2.88	3.05	3.17	PNMF2_SUP	2.33	2.85	3.19	3.49	3.69	3.84

TABLE IV. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS IN DNN OVER CAR, F-16, FACTORY1 AND BABBLE NOISES IN DIFFERENT SNRS.

PESQ							**COVL**						
SNR	-5	0	5	10	15	20	**SNR**	-5	0	5	10	15	20
DNN [11]	1.77	2.18	2.50	2.72	2.95	3.12	DNN [11]	2.17	2.68	3.07	3.35	3.62	3.81
SUP [24]	1.75	2.22	2.53	2.76	2.96	3.12	SUP [24]	2.16	2.72	3.10	3.37	3.61	3.78
SRU [16]	1.83	2.20	2.51	2.75	2.98	3.13	SRU [16]	2.14	2.62	3.02	3.33	3.61	3.79
PNMF1	1.84	**2.29**	**2.62**	**2.87**	**3.07**	**3.22**	PNMF1	**2.23**	**2.79**	**3.20**	**3.49**	**3.72**	3.89
PNMF2	1.87	2.27	2.56	2.78	2.95	3.07	PNMF2	2.22	2.73	3.11	3.38	3.59	3.73
PNMF1_SUP	1.82	2.27	2.59	2.83	3.06	**3.22**	PNMF1_SUP	2.20	2.76	3.16	3.46	**3.72**	**3.90**
PNMF2_SUP	**1.86**	2.25	2.54	2.77	2.96	3.09	PNMF2_SUP	2.20	2.70	3.08	3.37	3.60	3.75

TABLE V. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS IN LSTM FOR CAR AND F-16 NOISES IN DIFFERENT SNRS.

PESQ - Car noise							COVL - Car noise						
SNR	-5	0	5	10	15	20	SNR	-5	0	5	10	15	20
LSTM [14]	1.98	2.32	2.61	2.81	2.98	3.13	LSTM [14]	2.38	2.82	3.18	3.43	3.62	3.79
SUP	1.99	2.32	2.61	2.80	2.97	3.12	SUP	2.40	2.82	3.19	3.43	3.62	3.79
SRU [16]	1.89	2.25	2.59	2.81	3.00	3.16	SRU [16]	2.24	2.69	3.12	3.41	3.63	3.83
PNMF1	**2.10**	**2.43**	**2.72**	**2.95**	**3.11**	**3.27**	PNMF1	**2.51**	**2.95**	**3.31**	**3.57**	**3.76**	**3.93**
PNMF2	2.00	2.31	2.57	2.76	2.90	3.02	PNMF2	2.35	2.77	3.11	3.35	3.52	3.66
PNMF1_SUP	2.10	2.43	2.72	2.94	3.09	3.23	PNMF1_SUP	2.50	2.94	3.30	**3.57**	3.74	3.90
PNMF2_SUP	2.00	2.32	2.59	2.79	2.93	3.08	PNMF2_SUP	2.35	2.78	3.14	3.39	3.56	3.72
PESQ - F-16 noise							COVL - F-16 noise						
SNR	-5	0	5	10	15	20	SNR	-5	0	5	10	15	20
LSTM [14]	1.95	2.37	2.60	2.83	3.00	3.13	LSTM [14]	2.36	2.88	3.16	3.43	3.64	3.79
SUP	1.96	2.36	2.59	2.82	2.99	3.12	SUP	2.38	2.87	3.16	3.43	3.63	3.79
SRU [16]	1.93	2.28	2.55	2.79	3.00	3.16	SRU [16]	2.27	2.72	3.06	3.37	3.63	3.82
PNMF1	**2.06**	**2.47**	**2.74**	**2.98**	**3.15**	**3.29**	PNMF1	**2.47**	**2.99**	**3.31**	**3.60**	**3.80**	**3.95**
PNMF2	2.00	2.35	2.59	2.79	2.94	3.06	PNMF2	2.35	2.82	3.13	3.37	3.57	3.70
PNMF1_SUP	2.06	2.46	2.73	2.97	3.14	3.27	PNMF1_SUP	**2.47**	2.98	**3.31**	3.58	3.79	3.94
PNMF2_SUP	2.03	2.40	2.64	2.86	3.02	3.11	PNMF2_SUP	2.38	2.86	3.17	3.44	3.63	3.75

[DOI: 10.61186/itrc.15.3.53]

TABLE VI. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS IN LSTM OVER CAR, F-16, FACTORY1 AND BABBLE NOISES IN DIFFERENT SNRS.

	PESQ							COVL						
SNR	-5	0	5	10	15	20	**SNR**	-5	0	5	10	15	20
LSTM [14]	1.83	2.25	2.53	2.77	2.97	3.11	LSTM [14]	2.21	2.73	3.09	3.37	3.60	3.77
SUP	1.86	2.24	2.54	2.76	2.96	3.10	SUP	2.24	2.73	3.09	3.37	3.60	3.76
SRU [16]	1.83	2.20	2.51	2.75	2.98	3.13	SRU [16]	2.14	2.62	3.02	3.33	3.61	3.79
PNMF1	**1.94**	**2.34**	**2.64**	**2.88**	**3.08**	**3.22**	PNMF1	**2.31**	**2.83**	**3.21**	**3.49**	**3.73**	**3.88**
PNMF2	1.89	2.24	2.51	2.72	2.90	3.02	PNMF2	2.21	2.68	3.03	3.30	3.51	3.65
PNMF1_SUP	1.93	2.33	**2.64**	**2.88**	3.07	3.21	PNMF1_SUP	2.30	2.82	3.20	**3.49**	3.72	3.87
PNMF2_SUP	1.91	2.28	2.55	2.76	2.95	3.07	PNMF2_SUP	2.23	2.71	3.07	3.34	3.56	3.70

TABLE VII. AVERAGE RESULTS OF FWSEGSNR MEASUREMENTS IN LSTM OVER CAR, F-16, FACTORY1 AND BABBLE NOISES IN DIFFERENT SNRS.

	fwsegSNR					
SNR	-5	0	5	10	15	20
LSTM [14]	6.3	8.1	9.7	11.2	12.5	13.5
SUP	6.3	8.1	9.7	11.2	12.6	13.5
SRU [16]	6.0	7.7	9.5	11.3	**13.0**	**14.1**
PNMF1	**6.5**	**8.4**	**10.1**	**11.7**	**13.0**	13.9
PNMF2	6.1	7.8	9.4	10.9	12.1	12.9
PNMF1_SUP	6.5	8.3	**10.1**	11.6	**13.0**	13.9
PNMF2_SUP	6.1	7.9	9.5	11.0	12.2	13.0

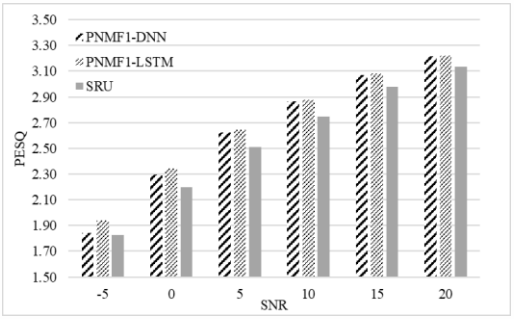

Figure 8. Average PESQ results over Babble, Factory1, F-16 and Car noises for the best proposed methods and SRU.

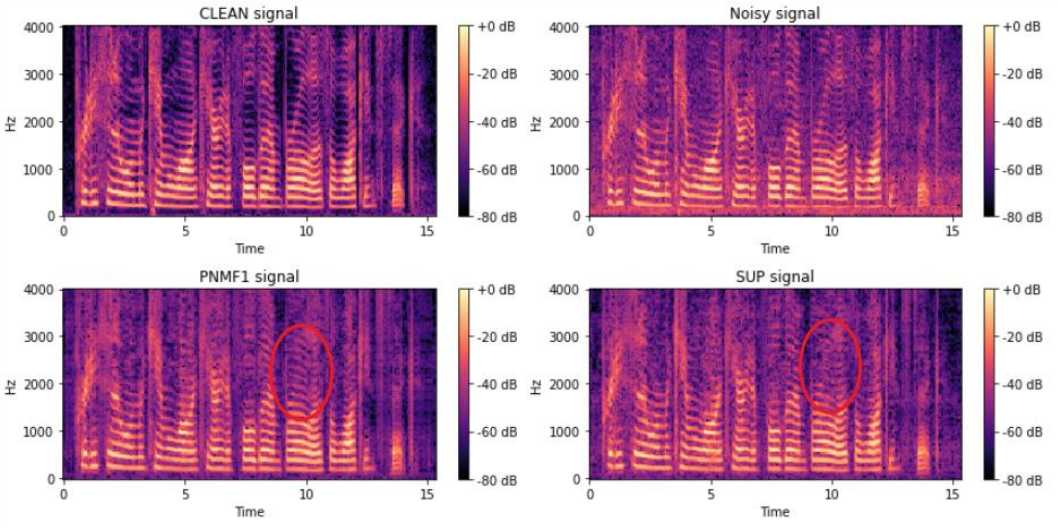

Fig. 9. The spectrograms of clean, noisy, PNMF1, and SUP methods.

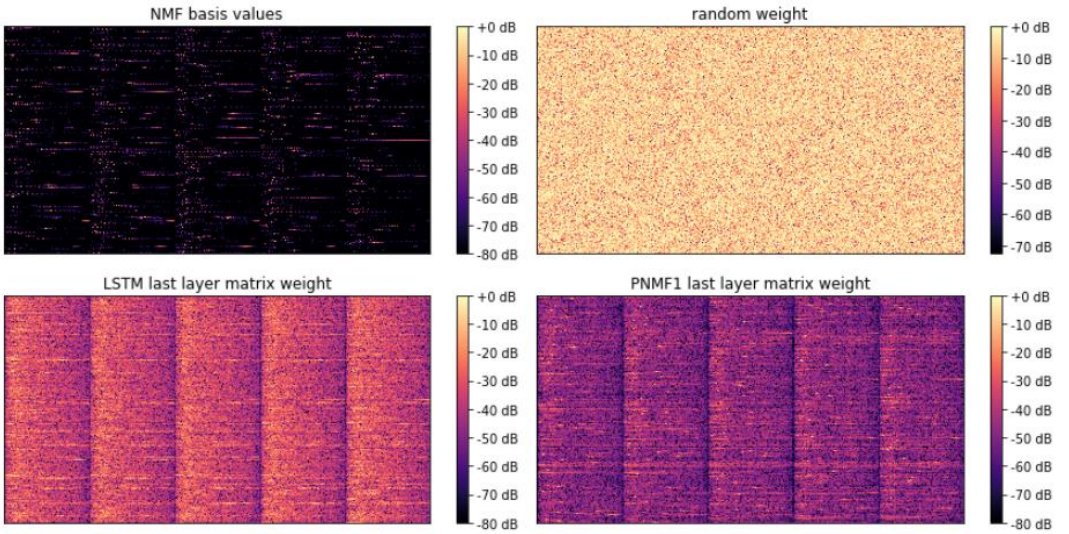

Fig. 10. The matrix weights of NMF basis values, random weight, LSTM last layer, and PNMF1 last layer.

[DOI: 10.61186/itrc.15.3.53]

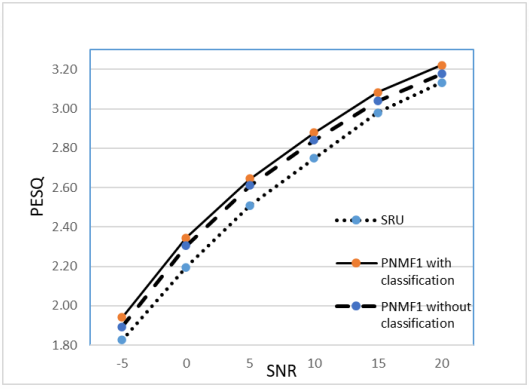

Fig. 11. Average PESQ results over Babble, Factory1, F-16 and Car noises for SRU method and the proposed PNMF1 method with and without noise classification.

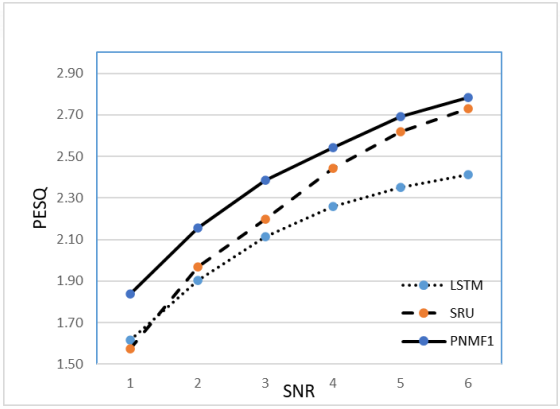

Fig. 12. Average PESQ results over F-16 noise for SRU, LSTM, and the proposed PNMF1 method for unseen speech signals (IEEE sentence dataset).

TABLE VIII. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS OVER RESTAURANT NOISE IN LSTM.

PESQ - Restaurant noise						
SNR	-5	0	5	10	15	20
LSTM [14]	1.68	2.02	2.32	2.62	2.88	3.07
SUP	1.69	2.00	2.30	2.62	2.88	3.06
SRU [16]	1.68	1.99	2.29	2.60	2.87	3.06
PNMF1	**1.70**	2.03	**2.36**	**2.70**	**2.97**	**3.15**
PNMF2	1.64	2.03	2.32	2.63	2.90	3.08
PNMF1_SUP	1.69	**2.04**	**2.36**	2.69	2.95	3.14
PNMF2_SUP	1.73	2.02	2.32	2.62	2.87	3.04
COVL - Restaurant noise						
SNR	-5	0	5	10	15	20
LSTM [14]	1.90	2.36	2.80	3.18	3.49	3.71
SUP	1.92	2.35	2.78	3.18	3.49	3.71
SRU [16]	1.91	2.33	2.77	3.17	3.50	3.72
PNMF1	1.91	**2.37**	**2.84**	**3.27**	**3.58**	**3.79**
PNMF2	1.87	2.36	2.80	3.20	3.51	3.73
PNMF1_SUP	1.92	**2.37**	**2.84**	3.26	3.56	**3.79**
PNMF2_SUP	**1.94**	2.36	2.79	3.17	3.47	3.67

TABLE IX. AVERAGE RESULTS OF THE SPEECH QUALITY MEASUREMENTS OVER PINK AND PIANO NOISES IN LSTM.

PESQ						
SNR	-5	0	5	10	15	20
LSTM [14]	1.77	2.13	2.44	2.70	2.90	3.09
SUP	1.73	2.14	2.45	2.71	2.91	3.08
SRU [16]	1.67	2.07	2.41	2.69	2.89	3.10
PNMF1	1.75	**2.19**	**2.53**	2.80	2.99	3.16
PNMF2	1.76	2.14	2.45	2.68	2.84	3.00
PNMF1_SUP	1.75	2.16	**2.53**	**2.81**	**3.01**	**3.18**
PNMF2_SUP	**1.78**	2.15	2.46	2.70	2.89	3.05
COVL						
SNR	-5	0	5	10	15	20
LSTM [14]	**2.02**	2.52	2.90	3.26	3.52	3.73
SUP	1.98	2.54	2.92	3.28	3.53	3.74
SRU [16]	1.84	2.41	2.85	3.24	3.51	3.75
PNMF1	1.97	**2.58**	3.00	3.37	3.60	3.80
PNMF2	1.99	2.51	2.91	3.23	3.44	3.62
PNMF1_SUP	1.99	2.57	**3.02**	**3.39**	**3.63**	**3.83**
PNMF2_SUP	1.99	2.50	2.90	3.24	3.48	3.68

TABLE X. AVERAGE RESULTS OF SPEECH QUALITY MEASUREMENTS IN LSTM OVER ALL 7 MATCHED AND MISMATCHED NOISES IN DIFFERENT SNRS.

PESQ						
SNR	-5	0	5	10	15	20
LSTM [14]	1.79	2.18	2.47	2.72	2.93	3.09
SUP	1.79	2.17	2.48	2.72	2.93	3.09
SRU [16]	1.76	2.13	2.45	2.71	2.94	3.11
PNMF1	**1.85**	2.23	**2.57**	**2.83**	**3.04**	**3.19**
PNMF2	1.82	2.18	2.46	2.70	2.88	3.02
PNMF1_SUP	1.85	**2.24**	**2.57**	**2.83**	3.03	**3.19**
PNMF2_SUP	1.84	2.21	2.49	2.72	2.92	3.06
COVL						
SNR	-5	0	5	10	15	20
LSTM [14]	2.11	2.61	2.99	3.31	3.56	3.75
SUP	2.12	2.62	2.99	3.32	3.56	3.74
SRU [16]	2.02	2.52	2.93	3.28	3.56	3.77
PNMF1	**2.16**	**2.69**	**3.09**	**3.42**	**3.67**	**3.84**
PNMF2	2.08	2.58	2.96	3.26	3.49	3.65
PNMF1_SUP	2.15	2.68	**3.09**	**3.42**	**3.67**	**3.84**
PNMF2_SUP	2.12	2.66	2.98	3.28	3.52	3.69

TABLE XI. THE FRIEDMAN TEST RESULTS WITH HOLM'S POST HOC **TEST**

	PESQ	
model	p-value	Holm
PNMF1	0.0000	0.0083
PNMF1_SUP	0.0000	0.0100
PNMF2_SUP	0.0015	0.0125
LSTM [14]	0.1850	0.0166
SUP	0.5700	0.0250
SRU[16]	0.6366	0.0500
PNMF2	-	-

[DOI: 10.61186/itrc.15.3.53]

VI. DISCUSSION

According to the results reported in the tables, it is observed that all proposed methods outperform the baselines. Specifically, PNMF1 method in which we propose using the transpose of NMF basis matrix in the last layer has generally the best performance in both PESQ, COVL, and fwsegSNR criteria of all other methods. As Tables 4, 6, and 7 show, the improvement of PNMF1 over other methods is significant especially in LSTM network. In the NMF method, the components of the clean speech signal are decomposed into two matrices of the basis and coefficients. By adjusting the weights of the last layer of the network to the transpose of the basis matrix of NMF, the last hidden layer of the network will estimate the values of speech features obtained from speech FFT magnitude. In fact, this is the main advantage of using NMF for decomposing clean speech signal into two matrices of W and H. W works as an appropriate data-driven filter which is able to find proper speech features when applied on clean speech signal X. Estimation of speech features leads to the extraction of useful information from clean signals, and thus better noise reduction of the noisy speech. Hence, in the fine-tuning phase for the DNN and LSTM networks, the local minimum problem is also reduced. As the NMF algorithm extracts the basis matrix from clean features and is independent from noisy features, the proposed method has a higher generalization over different noises according to the results of Tables 8, 9, 10, and also has a higher generalization for different utterances according to the results of mismatched database of Fig. 12. Thus, the proposed method is not dependent on a specific dataset. The results obtained in Table 10 also proves the claim that our proposed methods especially PNMF1 outperform other baselines and have a higher generalization over different noises for speech enhancement. Due to the linear nature of NMF method, proposing to use it for initializing the weights of the last layer, which is a linear layer, is more effective in improving the performance of the DNN and LSTM. Moreover, we have found the NMF basis matrix from clean speech signal, and not noisy one. Therefore, it works more efficiently when the input NMF matrix is very close to clean, i.e. the enhanced speech at the output layer. Hence, PNMF1 method leads to the best results in both DNN and LSTM networks according to all tables. This is due to the fact that it can map the features of the last hidden layer to the target output more appropriately, and would also solve the local minima problem. For the same reason, PNMF2 has weaker results compared to PNMF1. Since the calculation of the NMF basis matrix is on the clean input signal, it leads to lower results when the real input of the networks is the noisy speech in the first layer for PNMF2 structure. Also, in [26], the importance of the last layer values is described. Using appropriate initializing for this layer, the network will be trained more efficiently and random initialization for other weights will lead to more generalization for the neural network.

As illustrated in Fig. 8, the PNMF1 for the LSTM network outperforms the PNMF1 for DNN. This is in line with the time dependence capabilities of LSTM networks. Also, the SRU network (as the most recent baseline) has weaker results than our proposed methods. In addition, as indicated in Fig. 9 by the red ovals, PNMF1 as our best proposed method, better extracts the speech signal details in comparison with SUP (which also has a pre-training strategy) in the obtained speech spectrograms. We owe this performance to the NMF properties in extracting clean speech features.

Fig. 7 shows the process of reducing network error by increasing the number of iterations. As seen, the PNMF1 method has less error during the learning process and converges more quickly. Less error and faster network convergence in the PNMF1 method could be attributed to the selection of appropriate initial weights and the reduction of the local minimum problem.

Fig. 10 illustrates that the matrix weights of the NMF basis values, the last layer of LSTM, and the last layer of PNMF1 have the same structure. Also, the LSTM and PNMF1 matrix weights are mostly the same, but PNMF1 has more sparse structure than the LSTM model. This sparse structure of the PNMF1 method will have more generalization over different noises and speech signals. Also, this sparse structure will remove the background noise signal more efficiently. It is clear that this matrix weight of the PNMF1 method has a structure like a filter bank, which helps with the denoising of the noisy speech signal.

Moreover, the proposed noise classification strategy is useful in improving the results according to Fig. 11. The individual models trained for each noise type are more compatible with the properties of each specific noise and could lead to better results especially for matched noises. For mismatched noises, a general model trained with all models has been proposed to achieve better results. Not surprisingly, increasing the number of training noise types and using a much larger noise dataset could lead to higher results.

We also evaluate the statistical significance of the proposed methods. The statistical Friedman test in Table 11 shows that the proposed PNMF1 method is significantly better than the baselines.

VII. CONCLUSION

In this paper, we proposed a novel method for improving the pre-training of DNNs and LSTMs in speech enhancement. Since NMF is known to be an appropriate sparse model for extracting speech features, we suggested using NMF basis matrices as the initial weights in deep networks to make use of the advantages of both NMF and deep learning methods. In addition, NMF pre-training could address the local-minima problem of deep learning algorithms. In this paper, we proposed using the transpose of NMF basis matrix as the initial weights of the last layer of DNN and LSTM. The use of the proposed NMF pre-training

[DOI: 10.61186/itrc.15.3.53]

method in supervised pre-training, and the NMF basis matrix as the initial values of the first layer were also suggested. Practical observations indicated that pre-training of the last layer with the NMF method leads to better network performance and better results. This happens due to the fact that the NMF model linearly maps clean speech features to two matrices with a data-driven approach. Thus, it is able to extract appropriate clean speech features in different cases leading to an improved network output. Therefore, using this method to find the initial weights in the last layer of the network, which has the enhanced speech features as the output, was more compatible with the NMF structure, and could reconstruct clean speech features better. Moreover, we suggested a noise classification strategy in this paper by training individual models for each noise type. Using these specific models led to more improvement in the enhancement procedure due to their compatibility with noise. Furthermore, to extend the generalization of the suggested approach to unseen noises, we introduced a general model trained with all noises in mismatched conditions. The experiments showed that the proposed method could improve the PESQ and COVL of the enhanced speech signal significantly compared to previous baselines.

REFERENCES

[1] T. Kawase, M. Okamoto, T. Fukutomi, and Y. Takahashi, "Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 125-133, 2020.

[2] P. C. Loizou, Speech enhancement: Theory and Practice. CRC press, 2013.

[3] A. Pandey, and D. Wang, "A new framework for supervised speech enhancement in the time domain," *Interspeech*, pp. 1136-1140, 2018.

[4] S. K. Roy, A Nicolson, and K. K. Paliwal, "Deep learning with augmented Kalman filter for single-channel speech enhancement," *In IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1-5.

[5] M. M. Mirjalili, S. Seyedin, "Speech enhancement using NMF based on hierarchical deep neural networks with joint learning," *28th Iranian Conference on Electrical Engineering (ICEE)*, 2020.

[6] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218-1234, 2006.

[7] B. Chen, and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech communication*, vol. 49, no. 2, pp. 134-143, 2007.

[8] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1404-1415, 2020.

[9] K. Kumar, and S. Cruces, "An iterative posterior NMF method for speech enhancement in the presence of additive Gaussian noise," *Neurocomputing*, vol. 230, pp. 312-315, 2017.

[10] Z. Wang, T. Zhang, Y. Shao, B. Ding, "LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement," *Applied Acoustics*, Vol. 172, pp. 107647, 2021.

[11] Y. Xu, J. Du, L. Dai, and 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.

[12] T. Gao, J. Du, Li-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep deural network for speech enhancement," *INTERSPEECH*, pp. 3713-3717. 2016.

[13] R. Li, Y. Liu, Y. Shi, L. Dong, and W. Cui, "ILMSAF based speech enhancement with DNN and noise classification," *Speech Communication*, vol. 85, pp. 53-70, 2016.

[14] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 239-243, 2019.

[15] Z. Wang, T. Zhang, Y. Shao, and B. Ding, "LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement," *Applied Acoustics*, vol. 172, 2021.

[16] X. Cui, Z. Chen, and F. Yin, "Speech enhancement based on simple recurrent unit network," *Applied Acoustics*, vol. 157, 2020.

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

[18] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement incorporating deep neural network." *15th Annual Conference of the International Speech Communication Association*, 2014.

[19] C. Yarra, S. Nagesh, O. D. Deshmukh, and P. K. Ghosh. "Noise robust speech rate estimation using signal-to-noise ratio dependent sub-band selection and peak detection strategy," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. 1615-1628, 2019.

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

[21] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?," *13th International Conference on Artificial Intelligence and Statistics*, pp. 201-208, 2010.

[22] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, pp. 153-160, 2007.

[23] S. Z. Seyyedsalehi, and S. A. Seyyedsalehi, "A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural networks." *Neurocomputing*, vol. 168, pp. 669-680, 2015.

[24] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *Interspeech*, vol. 2013, pp. 436-440, 2013.

[25] M. Lashkari, S. Seyedin. "NMF-based cepstral features for speech emotion recognition." *4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 189-193, 2018.

[26] W. Cao, X. Wang, Z. Ming, and J. Gao, "A review on neural networks with random weights," *Neurocomputing*, Vol. 275, pp. 278-287, 2018.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." *STIN* 93, pp. 27403, 1993.

[28] E. H. Rothauser. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17, pp. 225-246, 1969.

[29] H. J. M. Steeneken, and F. W. M Geurtsen, "Description of the RSG-10 noise database," *report IZF* 3 (1988): 1988.

[30] H.-G. Hirsch, and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," In *ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*, 2000.

[31] http://www.pianosociety.com/.

[32] ITU-T, P. "Objective measurement of active speech level," *ITU-T Recommendation*, 1993.

[33] H.-W. Tseng, M. Hong, and Z.-Q. Luo, "Combining sparse NMF with deep neural network: A new classification-based approach for speech enhancement," *IEEE International*

[DOI: 10.61186/itrc.15.3.53]

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2145-2149, 2015.

[34] X. Zhang, Y. Zou, and W. Shi. "Dilated convolution neural network with LeakyReLU for environmental sound classification," *22nd International Conference on Digital Signal Processing (DSP)*, pp. 1-5, 2017.

[35] K. Kondo, Subjective quality measurement of speech: its evaluation, estimation and applications, Springer Science & Business Media, 2012.

[36] Y. Hu, and P. C. Loizou. "Evaluation of objective measures for speech enhancement," *9th International Conference on Spoken Language Processing*, 2006.

[37] Hu, Y. and Loizou, P.C. "Evaluation of objective quality measures for speech enhancement". *IEEE Transactions on audio, speech, and language processing*, Vol. 16, Issue 1, pp. 229-238, 2007.

[38] H. Damirchi, S. Seyedin, S. M. Ahadi, "Improving the loss function efficiency for speaker extraction using psychoacoustic effects," *Applied Acoustics*, Vol. 183, pp. 108301-108307, 2021.

[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets,", *The Journal of Machine learning research*, Vol. 7, pp. 1-30, 2006.

Razieh Safari Dehnavi received the B.Sc. degree in Electronic Engineering from the Yazd University, Yazd, Iran, in 2013, the M.S.c. degree in Electronic Engineering from the Amirkabir University of Technology, Tehran, Iran, in 2017. She is currently pursuing the Ph.D. degree in the Electronic Engineering, Amirkabir University of Technology, Tehran, Iran. Her research interests include AI and Machine Learning, Signal Processing and Speech Enhancement.

Sanaz Seyedin received the B.Sc. degree from Amirkabir University of Technology, Tehran, Iran, and the M.Sc. degree from Iran University of Science and Technology, Tehran, Iran, both in Electronics Engineering, in 2001, and 2005, respectively. She received the Ph.D. degree from Amirkabir University of Technology in 2010 focusing on the field of Speech Recognition. She is a Senior Member (SM) of IEEE. She is currently an Assistant Professor at the Electrical Engineering Department, Amirkabir University of Technology, teaching both undergraduate and graduate courses as well as conducting research in different areas of Machine Learning and AI, Signal Processing (audio, speech, image, biological signals), Compressive Sensing and Sparse Coding and Source Separation.

[DOI: 10.61186/itrc.15.3.53]