

An improved deep learning approach for speech enhancement

Malek Miled¹, Mohamed Anouar Ben Messaoud^{1,2}

¹Departamento de Engenharia Electrica, Instituto de Engenharia, Universidade do El Manar, Rua El. Belvedere Tunis, 1002 Tunis, Tunisia (malek.miled@yahoo.fr) ORCID [0009-0002-4456-3748](https://orcid.org/0009-0002-4456-3748)




² Departamento de Physica, Faculdade de Ciencia, Universidade do El Manar, Rua El. Belvedere Tunis, 1002 Tunis, Tunisia (anouar.benmessaoud@yahoo.fr) ORCID [0000-0002-7190-2736](https://orcid.org/0000-0002-7190-2736)

Abstract

Single-channel speech enhancement refers to the task of improving the quality and intelligibility of a speech signal in a noisy environment. Time-domain and time-frequency-domain methods are two main categories of approaches for speech enhancement. In this paper, we propose a approach based on a cross-domain framework. This framework utilizes our knowledge of the spectrogram and overcomes some of the limitations faced by time-frequency domain methods. First, we apply the intrinsic mode functions of the empirical mode decomposition and an improved version of principal component analysis. Then, we design a cross-domain learning framework to determine the correlations along the frequency and time axes. At low SNR = -5 dB, the effectiveness of our proposed approach is demonstrated by its performance based on objective and subjective measures. With average scores of -0.49, 2.47, 2.44, and 0.68 for SegSNR, PESQ, Cov, and STOI, respectively. The results highlight the success of our approach in addressing low SNR conditions.

Author Keywords. Speech Enhancement, Empirical Mode Decomposition, Principal Component Analysis, Learning Model.

Type: Research Article

 Open Access  Peer Reviewed  CC BY

1. Introduction

In many speech-processing applications, such as speech recognition, telecommunication devices and hearing aids, speech enhancement is a crucial task. There have been extensive studies on this subject in the past, and numerous effective models have been proposed. According to specific applications, such as to enhance the speech signal quality, and to increase the effectiveness of the voice communication device, speech enhancement nevertheless continues to be a difficult challenge in a single-channel real world context (Loizou 2013). Time-frequency (T-F) domain approaches and time-domain methods are the two broad categories into which speech enhancement techniques can be divided. In T-F domain methods, the speech signal is analyzed and modified in the joint time-frequency domain, typically using transforms such as the Short-Time Fourier Transform (STFT) or the Mel-Frequency Cepstral Coefficients (MFCC). These methods exploit the spectral information and temporal evolution of the speech signal to enhance its quality. They often involve techniques like spectral subtraction, Wiener filtering, or mask estimation to suppress noise and enhance speech. Furthermore, common amplitude and frequency modulations are often present in speech signals. These modulations result from various linguistic and articulatory factors and

contribute to the perception of speech intelligibility. T-F domain methods can exploit these modulations by analyzing the variations in amplitude and frequency content across different T-F bins in the spectrogram. This allows for the identification and separation of speech components from noise (Hershey 2017). However, T-F domain methods may suffer from limitations such as the trade-off between time and frequency resolution, the uncertainty principle, and the need for accurate estimation of noise statistics. In addition, the metric mismatch problem in the T-F domain for speech enhancement arises when there is a discrepancy between the objective metric used for training a deep learning system and the subjective perceptual quality that humans perceive.

On the other hand, time-domain methods operate directly on the time-domain waveform of the speech signal. They leverage signal processing techniques, statistical models, and machine learning algorithms to enhance the speech quality (Pascual 2017). These methods often utilize features such as fundamental frequency, harmonicity and temporal correlations to distinguish between speech and noise components. Time-domain methods offer advantages such as preserving the fine details of the speech signal, handling non-stationary noise, exploit the long-term temporal context of the speech signal and avoiding the limitations imposed by the uncertainty principle in T-F domain methods. However, they may face challenges related to training complexity, robustness to different noise types, and accurate modeling of speech and computationally demanding and requires a large amount of training data. It can be more challenging to design effective loss functions and evaluate the performance of the models. Additionally, these methods may be more sensitive to noise and require careful regularization techniques to avoid overfitting.

In this study, we take the advantages of approaches in the two domains. Fortunately, we have observed that the benefits of time-frequency domain methods are primarily prominent in the early stages of the network, while the advantages of time-domain methods are more pronounced in the later stages of the network. Therefore, we propose a new single-channel speech enhancement approach based on the application of the empirical mode decomposition, the optimal principal component analysis, and a learning block. Our approach is decomposed into three essential stages. First, a speech denoising strategy based on spectral intrinsic mode functions (IMFs) of the empirical mode decomposition (EMD) is applied with the advantage that the basic functions are derived from the signal itself and we obtain an adaptive analysis. To further improve denoising character of IMFs, in the second stage, we recover the low-rank matrix, the sparse matrix, from the obtained IMFs spectrogram under the perturbation of a residual matrix. In these two steps, we examine an unsupervised mode of analysis that possessed the benefits of EMD's signal extraction of the dominant mode, sparse decomposition, and low-rank matrix. These upgrades can deliver the outcomes and make the decomposition more logical in an unsupervised manner. However, we apply a learning system that consider the benefits from both time-frequency domain and time domain methods. The fundamental concept behind our work is to identify long-range correlations in the time-frequency domain that were acquired from the unsupervised part. Then, a learned decoder is applied in time-domain to enhance the speech. The suggested method can handle a variety of noise implications, including Gaussian white, babble, and factory noises.

The rest of this work is arranged as follows. The related works are presented in Section 2. Section 3 describes the detailed of our proposed approach. To test the effectiveness of the proposed approach using the TIMIT database will be carried out in section 4. In section 5, we draw our conclusions.

2. Related Work

In this section, we will review single-channel speech enhancement approaches in both the time-frequency domain and the time-domain. In the T-F domain, one commonly used approach is spectral subtraction, which estimates the noise power spectrum and subtracts it from the observed spectrum to enhance the speech components (Ben 2016). Other T-F domain methods include Wiener filtering, where a time-varying gain function is applied to the noisy speech spectrogram, and non-negative matrix factorization, which decomposes the spectrogram into a sum of non-negative basis components (Shao 2011, Islam 2015, Liu 2020). Additionally, wavelet transform methods (WT) is a time-frequency domain analysis technique (Hu 2004, Shao 2011, Islam 2015, Liu 2020). By decomposing a signal into different frequency components at different time scales, the wavelet transform provides a localized representation of the signal in both time and frequency. An alternative category is sub-space approach. It is a technique used to separate the desired speech signal from background noise or interference by exploiting the differences in their spatial characteristics. The idea is to capture the subspace or spectral properties of the speech and noise signals separately. This is achieved by analyzing the signal in the joint time-frequency domain (Martin 2005, Candès 2011). Subspace decomposition techniques, such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF), are commonly used in the time-frequency domain for speech enhancement applications (Toh 2010, Sahin 2019). Deep learning-based speech enhancement approaches in the T-F domain have gained significant attention and demonstrated promising results in improving speech quality. One widely used technique is the deep neural network (DNN)-based speech enhancement. DNNs, such as feedforward neural networks or convolutional neural networks (CNNs), are trained to learn the mapping between noisy and clean speech spectrograms. The network takes the noisy spectrogram as input and produces an enhanced spectrogram as output. A considerably DNN with dilated convolution and bi-LSTM is created in (Zhao 2018). The receptive field is expanded via dilated convolution, and bi-LSTM learns long-range correlations along the time axis. Leglaive et al. proposed a variational auto-encoders approach (Leglaive 2020). The recurrent variational is fine-tuned at test time with a Gaussian noise model based on a non-negative matrix factorization. To record harmonic correlations along the frequency axis, the network's front end employs a frequency transformation block. At the conclusion of the network, a biLSTM is utilized to capture temporal dependencies. Another study by (Tolooshams 2020) aims to enhance the quality of multichannel speech signals by leveraging a combination of dense U-Net architecture and channel attention mechanism. The architecture of the proposed model is based on the U-Net, which consists of an encoder and a decoder. The encoder captures the hierarchical representations of the input speech signals, while the decoder reconstructs the enhanced speech from these representations. Various time-frequency domain methods have been developed to leverage the rich auditory patterns present in the time-frequency spectrogram. However, a common observation in prior works is that the learning of long-range correlations along the time and frequency axes is typically carried out separately. In contrast, we believe that considering the long-range correlations along both axes is crucial, as harmonics are inherent in speech signals and noise characteristics require long-term statistical analysis.

On the other hand, time-domain methods directly model the waveform of the mixture signal using an encoder-decoder framework. Empirical mode decomposition (EMD) method has been applied to analyze non-linear and non-stationary signals such as speech, and real noise. It consists to decompose a noisy speech signal into a set of IMFs (Huang 1998, He 2011). EMD

is particularly suitable for analyzing nonlinear and non-stationary signals, such as speech, where frequency components may vary over time. Therefore, EMD operates in the time domain rather than the frequency domain (Amezquita-Sanchez 2015, Pan 2016). Time-domain methods often benefit from the ability to capture long-term context and exploit the temporal correlations in the speech signal. In the last decades, speech enhancement based on deep learning is proposed due to recent advances in deep neural networks (DNNs) (Xu 2013, Xu 2014). In (Wang 2013, Narayanan 2013), the authors have used the DNN to predict the mask. These approaches typically employ deep learning models such as convolutional neural networks (CNNs) (Zhang 2017, Wang 2018) or recurrent neural networks (RNNs) to capture the temporal dependencies and reconstruct the clean speech waveform. Among the recent approaches, we can cite the generative adversarial networks method (GANs) that is used for generating realistic samples by learning the underlying distribution of the training data (Fu 2019). In the context of speech enhancement, GANs can be employed to generate enhanced speech signals that align with certain metric scores. While time-domain methods successfully address the limitations associated with T-F domain methods, it is important to note that T-F domain representations provide clear distinctions between speech and noise patterns, which time-domain methods cannot fully exploit due to the lack of prior knowledge.

3. Proposed Approach

Figure 1 shows that our proposed approach is decomposed into three stages. In the first stage, the noisy speech frames is decomposed into its corresponding IMFs based on the empirical mode decomposition. The second stage consists to apply an improved version of principal component analysis to determine the IMFs that are less corrupted by noise signal (enhanced IMFs) and to utilize them for the reconstruction of speech. The third stage is based on a deep learning model to make the classification of each enhanced IMF.

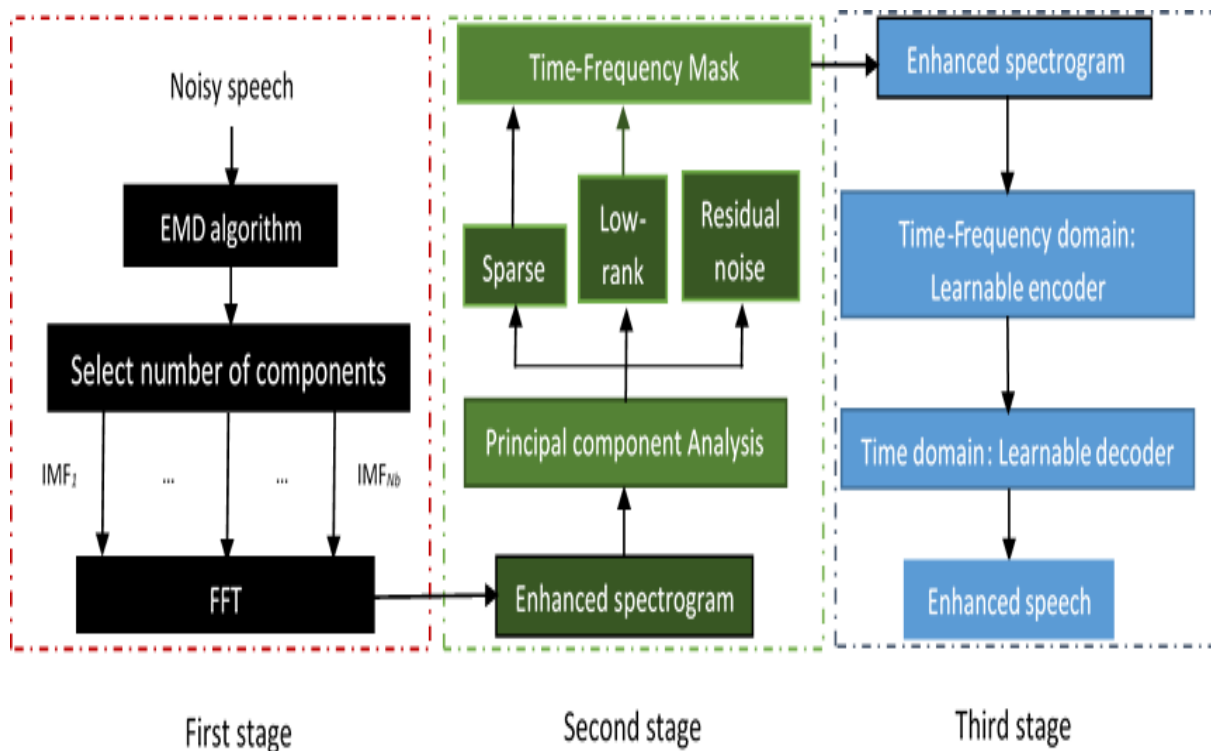


Figure 1: Block diagram of the proposed speech enhancement approach

3.1. EMD Decomposition

In the first stage, the empirical mode decomposition is applied to extract the oscillatory modes embedded in the noisy speech signal without any requirement of linearity or stationarity of the data. The time series are decomposed into components with instantaneous frequencies that have been defined. This stage allows identifying the physical time scales between successive maxima and minima intrinsic to the speech signal (He 2011). So, we obtain the Intrinsic Mode Function (IMF) that describes each characteristic oscillatory mode and replaces the speech signal details on a certain scale. Among the advantages of IMF, it has unique local frequency, it is symmetric, and it has the number of maxima and minima equal or different at most by one to the number of zero crossings to exhibit the same frequency at the same time for different IMFs. The empirical mode decomposition analysis guaranteed that we obtain a complete decomposition and the number of maxima and minima decreases when going from one residual to the next and allow separating the signal into elementary components constituting the original speech signal in order to suppress the noise.

We obtain the following equation (Equation 1):

$$x(t) = \sum_{i=1}^{Nb} f_i(t) + r_{Nb}(t) \quad (1)$$

Where $x(t)$ is the noisy speech frame, $f_i(t)$ is the i^{th} intrinsic mode functions (IMFs) with the same number of zero crossings and extrema, $r_{Nb}(t)$ is the residue and Nb is the number of selected IMFs.

The empirical mode decomposition process consists of four steps for decomposing a signal $x(k)$ using EMD. In the first step, we define the stopping criterion threshold ϵ and initialize the index of the i^{th} IMFs to 1. In the second step, we initialize the residual signal $r_{i-1}(k)$ as the signal $y(k)$. In the third step, we extract the IMF components iteratively using the sifting process. For this, we initialize the $f_{i,j-1}(k)$ to $r_{i-1}(k)$ with j is the sifting loop index. Then, we compute the upper and lower envelopes of $f_{i,j-1}(k)$ using cubic spline interpolation to fit the local maxima and minima from $f_{i,j-1}(k)$ and obtain the mean envelope the difference between the local mean that is the average of the upper and lower envelopes and the $f_{i,j-1}(k)$. Finally, we determine the number of selected IMFs automatically by computing the standard deviation criterion $\gamma(i)$ with $\gamma(i) = \frac{\sum_{n=1}^N |f_{i,j-1}(k) - f_{i,j}(k)|^2}{(f_{i,j-1}(k))^2}$ and we check if the sum of absolute differences between $r_{i-1}(k)$ and its neighboring IMF components is smaller than a threshold δ until the desired number of IMF components is reached.

In the fourth step, the Fourier transform of $x(k)$ is performed to obtain the IMFs observation matrix noted $|I(m,n)|$ of noisy speech in spectral domain by stacking every frame of the signal magnitude spectrum as column vectors over time sequences. We compute the FFT of the Nb selected IMFs components $f_i(k)$ for each frame i by using the complex exponential function to decompose the time-domain IMFs into its frequency components. We obtain the matrix I that correspond to the Nb IMFs in the spectral domain.

(Equation 2) gives the FFT of the result signal:

$$I(m,n) = \sum_{k=0}^{N-1} \sum_{i=1}^{Nb} f_i(k) w(m-k) e^{-j2\pi kn/N} \quad (2)$$

Where n is the index of the discrete frequency, m refers to the index of the time-frame, Nb is the number of selected IMFs determined in the previous step based on the sifting process, N

is the length of the frequency analysis, and $w(m)$ is an analysis window function. In the frequency domain, we accumulate all frames of the speech spectrum magnitude $|I(m, n)|$ as column vectors to obtain the matrix representation I .

In this stage, the noisy speech is divided into F frames of $N = 512$ samples with half-length overlap. We obtain a matrix $I = [I_1, I_2, \dots, I_F]$ with dimension $F \times N$. Each column in the noisy speech data matrix is decomposed into Nb intrinsic mode functions (IMFs).

The second stage will be detailed in the following sub-section.

3.2. Improved principal component analysis

In this stage, our contribution consists essentially to decompose the IMFs spectrogram of speech signal $I = [I_1, I_2, \dots, I_F]$ obtained by the first stage. Then we apply the principal component analysis to determine three subspaces (the sparse matrix Sp , low-rank matrix L , and the residual matrix R). Consequently, we impose non-negative constraints, and consider that the constraints for low-rank, and sparse, are not specified beforehand.

By applying the principal component analysis, we obtain the following equation in the spectral domain (Equation 3):

$$I = Sp + L + R \tag{3}$$

Where L , Sp and R represent respectively the low-rank matrix, the sparse matrix and the residual matrix. The goal is to separate the low-rank, sparse, and residual subspaces structures of clean speech from the noisy speech. The low-rank L decomposition is presented by a non-negative factorization GF_K , where G corresponds to the time-varying gains, and F_K is a set of basis. Therefore, the input matrix is described by the following equation (Equation 4):

$$I = Sp + R + GF_K \tag{4}$$

Then, we estimate the three subspaces described in (Equation 5):

$$\min_{Sp, F_K} \|I - Sp - GF_K\|_F^2 \quad \text{s.t.} \quad \text{card}(Sp) \leq s, F_K \in \mathcal{I}, G \in \mathcal{I} \tag{5}$$

Where $\text{card}(Sp)$ is the cardinality of Sp , $\|\cdot\|_F$ makes reference to the Frobenius norm of a matrix.

In order to make optimization, we solve the following three minimizations until convergence that is described by the (Equation 6):

$$\begin{cases} Sp = \arg \min_{\text{card}(Sp) \leq s} \|I - Sp - GF_K\|_F^2 \\ G = \arg \min_{G \geq 0} \|I - Sp - GF_K\|_F^2 \\ F_K = \arg \min_{F_K \geq 0} \|I - Sp - GF_K\|_F^2 \end{cases} \tag{6}$$

(Equation 7) solves the first and second minimizations as follow:

$$\begin{cases} G \leftarrow \frac{|F_K^T(Sp-I)| - F_K^T(Sp-I)}{2(GF_K^T F_K)} \odot G \\ F_K \leftarrow \frac{|(Sp-I)G^T| - (Sp-I)G^T}{2(GG^T F_K)} \odot F_K \end{cases} \tag{7}$$

Where \odot denotes the element-wise division and the multiplication between the matrices. The updating of Sp_i can be performed by selecting the top s largest non-zero entries of $|I - G_i F_k|$. The proof of the convergence of our method to a local minimum is given as follows:
 In the i^{th} iteration, we obtain the objective function by solving three sub-spaces, denoted as, Sub_i^1 , Sub_i^2 and Sub_i^3 respectively, is obtained (Equation 8):

$$\begin{cases} Sub_i^1 = \arg \min_{G \geq 0} \|I - Sp_{(i-1)} - G_{(i)} F_{K(i-1)}\|_F^2 \\ Sub_i^2 = \arg \min_{F_K \geq 0} \|I - Sp_{(i-1)} - G_{(i)} F_{K(i)}\|_F^2 \\ Sub_i^3 = \arg \min_{card(Sp) \leq s} \|I - Sp_{(i)} - G_{(i)} F_{K(i)}\|_F^2 \end{cases} \quad (8)$$

On the one hand, when comparing with $Sp_{(i-1)}$, the global optimality of $Sp_{(i)}$ ensure that $Sub_i^2 \geq Sub_i^3$. On the other hand, the updating strategy, which fixes $F_{K(i-1)}$ to find a more suitable $G_{(i)}$ and fixes $G_{(i)}$ to find a more suitable $F_{K(i)}$, leads to a decrease in the objective value. This deduction implies that the objective function consistently decreases throughout the iterative process. Furthermore, as the constraints are always satisfied, the objective function exhibits monotonic decrease and eventually converges to a local minimum.

To estimate the cardinality of the sparse matrix Sp , we optimize the threshold to determine the parameter s . This optimization problem is addressed using an alternate optimization method, described by (Equation 9):

$$\min_{card(Sp) \leq s} \|I - Sp - GF_k\|_F^2 + \alpha \|Sp\|_1 \quad (9)$$

Where $\alpha = \frac{1}{\sqrt{\max(m_1, m_2)}}$ is a trade-off parameter between the speech distortion and noise reduction.

After the the low-rank and sparse components are determined, residual noise R is derived as $R = I - Sp - GF_k$ when Sp , G , and F_k are determined. Then, we apply an ideal binary mask (Wang 2005).

Figure 2 shows respectively the spectrograms of clean speech, the noisy speech, and the enhanced speech signal using the improved version of the principal component analysis to the IMFs spectrogram of noisy speech signal.

In Figure 2.c), we can observe that the sparse component accords with sparsity of speech energy in the frequency domain.

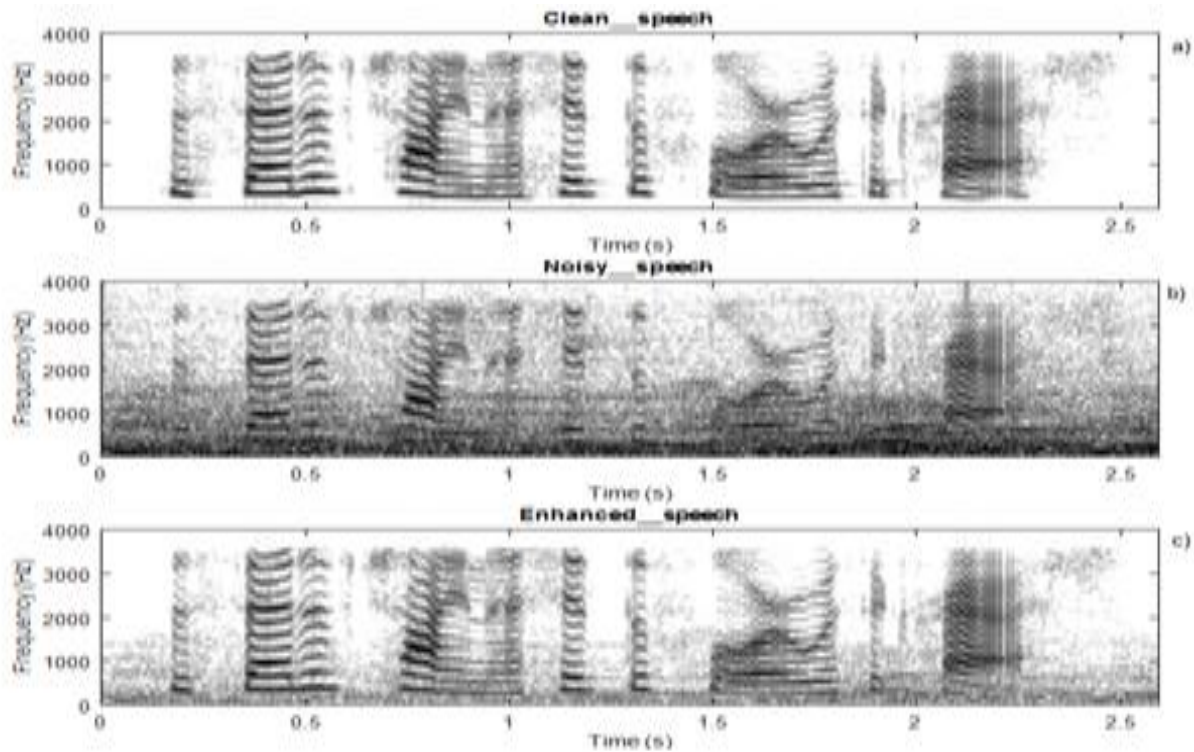


Figure 2: Spectrograms after application of two stages. a) Spectrogram of clean speech, b) Spectrogram of noisy speech with Babble noise at SNR = -5 dB, c) Spectrogram of the enhanced speech by our proposed two stages

3.3. Deep Learning block

For deep learning, we apply the framework described in Figure 3. For this, the spectrogram of enhanced IMFs obtained by the second stage is considered as the input of the feature of our deep learning model and to obtain directly the enhanced speech signal as an output time-domain. The input of the model is the spectrogram's complex value $Ie \in \mathbb{R}^{Fb \times Ts \times 2}$, which Fb corresponds to the number of frequency bands and Ts the number of time steps. Two 2D convolution layers receive the input matrix Ie as input.

As a result, the proposed model's feature, $IF \in \mathbb{R}^{Fb \times Tb \times Nbc}$, is obtained. Then it is divided into Nlb learning blocks with Nbc channels, and is considered as such. Every block produces IF_i features with i varied from 1 to Nlb hyper-parameter. Four convolution layers and a dual-path attention block make up each learning block. A batch normalization is used after each convolution layer's size of 3×3 . To find local correlations, it is used. Then, to find the long-range correlations, we use a dual-path attention block. Finding harmonic correlations along the frequency axis and global correlations along the time axis are both made possible by the dual-path attention block. The 2D spectrogram is transformed into two vectors. The first vector has a dimension of $Fb \times Nbc$ along the frequency axis. While the second vector has a dimension of $Ts \times Nbc$ along the time axis. The last block's enhanced spectrogram is supplied to the decoder layer in order to obtain the enhanced speech signal.

Figure 3 illustrates the detail of learning framework of the proposed model.

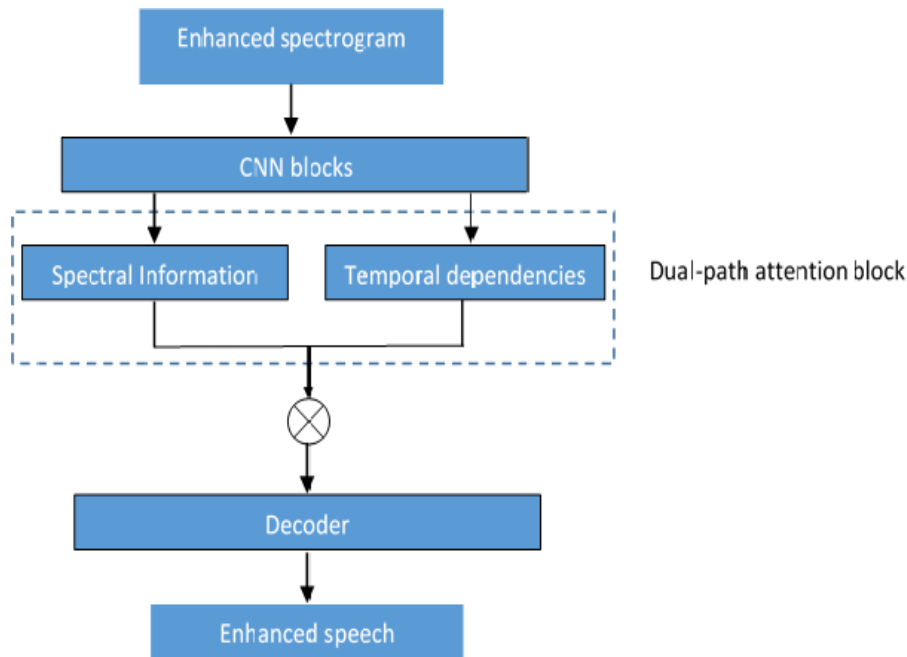


Figure 3: Learning framework for speech enhancement

4. Experiments and Results

We evaluate and compare our proposed approach for speech enhancement in this section.

4.1. Simulation conditions and dataset

Pytorch has been used to implement our model. With a sample frequency of 16 kHz, a Hann window of length 32 ms, and an FFT size of 512 points, all speech signals are calculated. Each convolution layer uses 100 channels, the rectified linear activation function was applied after each layer, and batch normalization was applied after all convolutional layers. The loss function is the log mean square error, and Adam gradient is the optimizer. On the basis of a feature window made up of ten frames, we estimate the training target.

For the speech enhancement evaluation in our simulations, we assessed the test set of the TIMIT and the NOIZEUS dataset. We make use of the NOISEX-92 database's white, factory, and babble noises (Varga 1993). The dataset is created by combining the voice signals and sounds at four different signal-to-noise ratios (SNRs) ranging from -10 to 5 dB.

Our proposed speech enhancement approach is compared to seven state-of-the-art methods to evaluate its performance. The first method is an unsupervised approach called Robust Principal Component Analysis (RPCA), which decomposes the noisy speech into low-rank and sparse components (Candès 2011). Another unsupervised method, Empirical Wavelet Transform (EWT) that combines wavelet transform with EMD to segment the frequency spectrum (Amezquita-Sanchez 2015). A supervised deep neural network (DNN) model by Narayanan et al. (Narayanan 2013) utilizes fully connected hidden layers with rectified linear units and dropout. Wang and Tan (Wang 2018) proposed a supervised convolutional neural network (CNN) model with two Long Short-Term Memory (LSTM) layers to capture long-term context. Leglaive et al. (Leglaive 2020) introduced a supervised recurrent variational auto-encoder called RNV, which fine-tunes a deep generative speech model using a Gaussian noise model based on non-negative matrix factorization. GAN model generate a realistic samples by learning the underlying distribution of the training data (Fu 2019). Finally, U-net applied an encoder and a decoder to enhance the quality of multichannel speech signals (Tolooshams 2020). Comparing our approach to these methods allows us to assess its effectiveness in enhancing speech quality and reducing noise interference.

To test the performance of our proposed approach, intelligibility tests and subjective results were applied.

4.2. Objective Results

We assess the effectiveness of the proposed speech enhancement approach by comparing it to other methods, using the average scores of Segmental SNR (SegSNR), Perceptual Evaluation of Speech Quality (PESQ), Composite measures (Cov), and short-time objective intelligibility (STOI). The evaluation and comparison of the different methods are based on the following detailed metrics. The Perceptual Evaluation of Speech Quality (PESQ) metric involves mapping the original and enhanced speech signals onto an internal representation using a perceptual model. It provides a score range of [-0.5, 4.5] and assesses the perceived quality of the enhanced speech. The Segmental Signal-to-Noise Ratio (SegSNR) measures the speech quality by averaging the frame-level Signal-to-Noise Ratio (SNR) estimation. The Combination Objective Measure (Cov) combines evaluation measures from the frequency-domain, time-domain, and perceptual field. It is calculated using (Equation 10):

$$\text{Cov} = 1.594 + 0.805 * \text{PESQ} - 0.512 * \text{LLR} - 0.007 * \text{WSS} \quad (10)$$

where LLR represents the log-likelihood ratio and WSS denotes the weighted spectral slope. The definitions of these measures can be found in (Loizou 2013). Additionally, the Short-Time Objective Intelligibility (STOI) metric is designed to predict the intelligibility of speech processed by the proposed speech enhancement approach, with a score range between 0 and 1. The average results of SegSNR, PESQ, Cov, and STOI measures for three types of noise are presented in the respective tables, enabling the evaluation and comparison of the different methods.

Table 1 illustrates the results obtained with SegSNR metric, and the overall quality of proposed approach based on average PESQ scores.

According to Table 1, the proposed approach outperforms U-net, Gan, RNV, CNN, DNN, RPCA, and EWT in terms of SegSNR for the two non-stationary noises at four SNRs. With white Gaussian noise the U-net-based model achieved the best results the at all SNRs. The primary cause is because before applying the model, U-net and RNV does a significant amount of off-line training. We can also see that for low input SNR, EWT and RPCA approaches perform less well. This is due to the unsupervised nature of the EWT and RPCA speech augmentation approaches.

Additionally, our proposed approach has the greatest PESQ scores with results that are only slightly superior to those of the examined methods at low SNR. The findings are comparable to those of the U-net, GAN, RNV and DNN-based models for all types of noise.

Noise	Method	SegSNR (dB)				PESQ			
		-10	-5	0	5	-10	-5	0	5
White	Proposed	-0.26	-0.15	2.38	2.57	1.99	2.09	2.45	2.99
	U-NET	-0.18	-0.13	2.48	2.67	1.97	2.06	2.46	2.98
	GAN	-0.21	-0.11	2.43	2.62	1.95	2.03	2.41	2.94
	RNV	-0.39	-0.11	2.47	2.65	1.87	1.92	2.43	2.95
	CNN	-0.42	-0.29	2.14	2.33	1.78	1.83	2.26	2.81
	DNN	-0.45	-0.75	2.07	2.38	1.69	1.74	2.22	2.89
	RPCA	-1.37	-1.18	1.86	2.32	1.45	1.56	2.18	2.61
	EWT	-1.42	-1.21	1.38	1.76	1.41	1.52	2.09	2.52

Noise	Method	SegSNR (dB)				PESQ			
Babble	Proposed	-1.71	-1.57	-0.40	1.21	2.52	2.58	2.94	3.26
	U-NET	-1.79	-1.59	-0.43	1.17	2.49	2.51	2.92	3.21
	GAN	-1.74	-1.61	-0.45	1.18	2.43	2.47	2.89	3.19
	RNV	-1.81	-1.63	-0.41	1.18	2.27	2.35	2.87	3.18
	CNN	-1.84	-1.86	-0.52	1.07	2.13	2.22	2.68	3.12
	DNN	-1.88	-1.78	-0.98	1.05	2.09	2.19	2.54	3.16
	RPCA	-3.59	-3.47	-1.61	-0.29	1.98	2.11	2.42	2.94
	EWT	-3.96	-3.88	-1.85	-0.86	1.81	1.97	2.27	2.76
Factory	Proposed	0.19	0.25	1.29	2.91	2.71	2.75	2.96	3.13
	U-NET	0.19	0.25	1.28	2.90	2.69	2.74	2.96	3.12
	GAN	0.19	0.21	1.27	2.91	2.63	2.67	2.92	3.11
	RNV	0.16	0.22	1.27	2.89	2.45	2.51	2.91	3.07
	CNN	-0.18	-0.13	1.16	2.68	2.27	2.33	2.83	2.89
	DNN	-0.14	0.05	1.02	2.34	2.15	2.25	2.85	2.92
	RPCA	-1.37	-1.28	0.93	1.97	1.58	1.62	1.83	2.07
	EWT	-1.69	-1.52	0.78	1.99	1.51	1.59	1.76	1.99

Table 1: Average of SegSNR value and PESQ score for different speech enhancement methods

Table 2 gives respectively the Cov, and the results of the STOI measures for the above-mentioned methods, over all noise conditions.

Noise	Method	Cov				STOI			
		-10	-5	0	5	-10	-5	0	5
White	Proposed	2.29	2.37	2.99	3.73	0.68	0.74	0.82	0.91
	U-NET	2.32	2.36	3.02	3.71	0.63	0.74	0.81	0.91
	GAN	2.29	2.31	2.91	3.57	0.59	0.72	0.78	0.89
	RNV	2.21	2.29	2.87	3.41	0.59	0.71	0.79	0.89
	CNN	1.97	2.02	2.75	3.39	0.56	0.70	0.75	0.85
	DNN	1.87	1.95	2.58	3.24	0.53	0.71	0.75	0.84
	RPCA	1.74	1.76	2.33	2.90	0.37	0.59	0.63	0.69
	EWT	1.77	1.81	2.39	2.84	0.39	0.56	0.61	0.66
Babble	Proposed	2.59	2.64	3.12	3.28	0.55	0.64	0.79	0.85
	U-NET	2.57	2.61	3.10	3.23	0.53	0.64	0.77	0.85
	GAN	2.48	2.55	3.08	3.17	0.49	0.61	0.73	0.83
	RNV	2.47	2.51	3.07	3.14	0.49	0.62	0.70	0.82
	CNN	2.32	2.38	2.95	3.05	0.47	0.63	0.72	0.83
	DNN	2.14	2.21	2.84	2.91	0.48	0.63	0.75	0.82
	RPCA	1.97	2.06	2.52	2.87	0.39	0.52	0.59	0.63
	EWT	1.80	1.84	2.56	2.84	0.37	0.51	0.58	0.63
Factory	Proposed	2.44	2.45	3.24	3.51	0.61	0.67	0.72	0.85
	U-NET	2.41	2.47	3.27	3.53	0.61	0.69	0.73	0.86
	GAN	2.39	2.41	3.19	3.47	0.58	0.65	0.71	0.84
	RNV	2.32	2.37	3.13	3.45	0.57	0.64	0.69	0.83
	CNN	1.84	1.91	2.78	2.79	0.52	0.64	0.71	0.82
	DNN	1.87	1.99	2.71	2.84	0.51	0.63	0.71	0.81
	RPCA	1.46	1.51	2.23	2.49	0.47	0.54	0.61	0.64
	EWT	1.57	1.65	2.38	2.46	0.45	0.55	0.58	0.62

Table 2: Average of Cov and STOI measures for different speech enhancement methods

Our approach exhibits a significant performance advantage over the compared speech enhancement methods. This superiority is evident through the highest Cov values recorded in [Table 1](#). These results serve as evidence for the effectiveness of combining empirical mode decomposition with our improved principal component analysis. Several factors contribute to the outperformance of our method. For the U-NET Model, we can see that our approach consistently achieves superior results with the exception of white noise scenarios. The DNN, CNN, RNV, and GAN-based Models require the specification of noise rank. If the rank is set too low, it fails to adequately address the noise, while setting it too high leads to additional noise dimensions affecting speech segments and causing distortion. For unsupervised RPCA technique, the effectiveness of the RPCA method heavily relies on the careful selection of parameters to distinguish between low-rank sub-spaces and sparse. For the EWT method, we can remark the introduction of a ringing residual noise component.

As depicted in [Table 2](#), our approach consistently outperforms the compared methods in terms of the STOI measure, closely followed by U-net, GAN, RNV, CNN, and DNN-based models under high input SNR conditions. RPCA demonstrates the efficacy of PCA technique in high input SNR scenarios. However, the EWT method exhibits poor performance, attributed to the introduction of ringing residual noise.

4.3. Subjective Results

The inclusion of subjective results in the study serves an important purpose in clarifying and providing further insight into the proposed work. While objective metrics provide quantitative measures to evaluate speech enhancement algorithms, they may not fully capture the subjective perception of speech quality by human listeners. Therefore, conducting subjective evaluations allows the authors to gather feedback and opinions from human subjects who listen to and assess the enhanced speech. By incorporating subjective evaluations, the authors aim to provide a more comprehensive assessment of the proposed method's performance and its effectiveness in improving the perceived quality of speech. The subjective results provide valuable information about the subjective listening experience and the overall preference of listeners, helping to validate and reinforce the findings obtained from the objective metrics. This multi-faceted evaluation approach enhances the understanding of the proposed work and ensures that both objective and subjective aspects of speech quality are considered.

For subjective listening tests, the mean opinion score (MOS) is conducted to evaluate our approach with compared methods. It consists to evaluate the overall quality ([Brawata 2015](#)). One hundred eighty five sentences from ten male, and seven female are randomly selected from the two databases and three background noises recorded in a white, babble, and factory was added to these sentences, at an SNR of 5 dB, and 0 dB. The resulting sequences were then enhanced using our approach and seven compared methods. The sequences are introduced to 16 expert listeners via head-phones ([Praxiling 2021](#)). [Table 3](#) presents the average listening test scores over all sentences and listeners. As can be seen in [Table 3](#), listeners considered that our approach is the most effective. Also, we can remark that the proposed approach, U-net, and GAN based-models performs almost equal when set to a SNR = 5 dB. In contrast, our approach outperforms the U-net, and GAN based-model at SNR = 0 dB. The listeners have denoted an increased speech distortion for RPCA, and EWT at low SNR. Also, they observed that some unprocessed noisy sequence obtained perceived more natural than some enhanced sequences by EWT approach.

Type of noise	SNR level	Proposed Approach	U-Net	GAN	RNV	DNN	CNN	EWT	RPCA
White	5 dB	3,28	3,26	3,27	2,82	3,14	3,02	2,49	2,49
	0 dB	2,65	2,38	2,27	2,12	2,05	1,89	2,67	1,61
Babble	5 dB	3,54	3,51	3,48	3,03	3,52	3,41	1,79	2,37
	0 dB	3,23	3,11	3,01	2,88	2,84	2,73	2,88	1,89
Factory	5 dB	3,92	3,90	3,82	3,25	3,59	3,37	2,65	2,98
	0 dB	3,46	3,19	3,06	2,98	2,84	2,61	3,04	2,03

Table 3: Subjective evaluation of proposed approach, and compared methods

5. Conclusions

In this paper, we propose a speech enhancement approach. It is decomposed into three stages. In the first stage, the empirical model decomposition is applied. The second stage consists to use an improved version of principal component analysis in speech enhancement systems. The main concept is to use low-rank matrix, sparse matrix, and residual component decomposition to noisy speech. The third stage's objective is to create a cross-domain learning framework that can take advantage of long-range frequency and temporal correlations. We perform and evaluate our proposed approach using objective measurement, and listening tests. Results show that the combination of the EMD and improved PCA technique followed by a deep learning outperforms the state-of-the-art methods. Our approach achieves the highest PESQ, SegSNR, Cov and STOI metrics among compared methods at low SNR level. The subjective results confirmed the performance of our approach.

Finally, in order to expand the suggested technique to monaural speech de-reverberation and separation, we intend to investigate the impact of loss functions and the choice of training objectives on the proposed approach.

References

- Amezquita-Sanchez, Juan P., and Hojjat Adeli. 2015. "A New Music-Empirical Wavelet Transform Methodology for Time-Frequency Analysis of Noisy Nonlinear and Non-Stationary Signals." *Digital Signal Processing* 45 (October): 55–68. <https://doi.org/10.1016/j.dsp.2015.06.013>.
- Brawata, Krzysztof, Pawel Malecki, Adam Pilch, and Tadeusz Kamisinski. 2015. "Subjective Assessment of Commercial Sound Enhancement System." In *Audio Engineering Society Convention 138*. <http://www.aes.org/e-lib/browse.cfm?elib=17730>.
- Cai, Jian-Feng, Emmanuel J. Candès, and Zuowei Shen. 2008. "A Singular Value Thresholding Algorithm for Matrix Completion." <https://doi.org/10.48550/ARXIV.0810.3286>.
- Candès, Emmanuel J., Xiaodong Li, Yi Ma, and John Wright. 2011. "Robust Principal Component Analysis?," *Journal of the ACM*, 58: 11–37. <https://doi.org/10.48550/ARXIV.0912.3599>.
- Fu, Szu-Wei, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. 2019. "MetricGAN: Generative Adversarial Networks Based Black-Box Metric Scores Optimization for Speech Enhancement." In *Proceedings of the 36th International Conference on Machine Learning*, 2031–41. <https://doi.org/10.48550/ARXIV.1905.04874>.
- Gaumin, Pierre-Olivier. 2021. "Praxiling UMR 5267 CNRS." *Praxiling UMR 5267 CNRS - Université Paul Valéry Montpellier* 3. October 7, 2021. <https://praxiling.cnrs.fr/search/2021/>.
- Gilles, Jérôme. 2013. "Empirical Wavelet Transform." *IEEE Transactions on Signal Processing* 61 (16): 3999–4010. <https://doi.org/10.1109/TSP.2013.2265222>.

- He, Ling, Margaret Lech, Namunu C. Maddage, and Nicholas B. Allen. 2011. "Study of Empirical Mode Decomposition and Spectral Analysis for Stress and Emotion Classification in Natural Speech." *Biomedical Signal Processing and Control* 6 (2): 139–46. <https://doi.org/10.1016/j.bspc.2010.11.001>.
- Hu, Y., and P.C. Loizou. 2004. "Speech Enhancement Based OnWavelet Thresholding the Multitaper Spectrum." *IEEE Transactions on Speech and Audio Processing* 12 (1): 59–67. <https://doi.org/10.1109/TSA.2003.819949>.
- Huang, Norden E., Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. 1998. "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis." *Proceedings: Mathematical, Physical and Engineering Sciences* 454 (1971): 903–95. <http://www.jstor.org/stable/53161>.
- Islam, Md Tauhidul, Celia Shahnaz, Wei-Ping Zhu, and M. Omair Ahmad. 2015. "Speech Enhancement Based on Student t Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients and a Custom Thresholding Function." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (11): 1800–1811. <https://doi.org/10.1109/TASLP.2015.2443983>.
- Leglaive, Simon, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. 2020. "A Recurrent Variational Autoencoder for Speech Enhancement." In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 371–75. Barcelona, Spain: IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053164>.
- Lin, Zhouchen, Minming Chen, and Yi Ma. 2013. "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices." *Journal of Structural Biology* 181 (2): 116–27. <https://doi.org/10.1016/j.jsb.2012.10.010>.
- Liu, H., L. Xue, J. Yang, Z. Wang, and C. Hua. 2023. "Speech Enhancement Based on Discrete Wavelet Packet Transform and Itakura-Saito Nonnegative Matrix Factorisation." *Archives of Acoustics, Archives of acoustics*, 45 (4): 565–72. <https://doi.org/10.24425/aoa.2020.134072>.
- Loizou, Philipos C. 2013. *Speech Enhancement: Theory and Practice*. 2nd ed. CRC Press. <https://doi.org/10.1201/b14529>.
- Luo, Yi, Zhuo Chen, John R. Hershey, Jonathan Le Roux, and Nima Mesgarani. 2017. "Deep Clustering and Conventional Networks for Music Separation: Stronger Together." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 61–65. New Orleans, LA: IEEE. <https://doi.org/10.1109/ICASSP.2017.7952118>.
- Martin, R. 2005. "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors." *IEEE Transactions on Speech and Audio Processing* 13 (5): 845–56. <https://doi.org/10.1109/TSA.2005.851927>.
- Messaoud, Mohamed Anouar Ben, and Aicha Bouzid. 2016. "Speech Enhancement Based on Wavelet Transform and Improved Subspace Decomposition." *Journal of the Audio Engineering Society* 63 (12): 990–1000. <https://www.aes.org/e-lib/browse.cfm?elib=18057>.
- Narayanan, Arun, and DeLiang Wang. 2013. "Ideal Ratio Mask Estimation Using Deep Neural Networks for Robust Speech Recognition." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7092–96. Vancouver, BC, Canada: IEEE. <https://doi.org/10.1109/ICASSP.2013.6639038>.

- Pan, Jun, Jinglong Chen, Yanyang Zi, Yueming Li, and Zhengjia He. 2016. "Mono-Component Feature Extraction for Mechanical Fault Diagnosis Using Modified Empirical Wavelet Transform via Data-Driven Adaptive Fourier Spectrum Segment." *Mechanical Systems and Signal Processing* 72–73 (May): 160–83. <https://doi.org/10.1016/j.ymsp.2015.10.017>.
- Park, Se Rim, and Jinwon Lee. 2016. "A Fully Convolutional Neural Network for Speech Enhancement." arXiv. <https://doi.org/10.48550/arXiv.1609.07132>.
- Pascual, S., Bonafonte, A., and J. Serra. 2017. "SEGAN: Speech Enhancement Generative Adversarial Network." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE. <https://doi.org/10.48550/arXiv.1703.09452>.
- Sahin, Mehmet Fatih, Armin Eftekhari, Ahmet Alacaoglu, Fabian Latorre, and Volkan Cevher. 2019. "An Inexact Augmented Lagrangian Framework for Nonconvex Optimization with Nonlinear Constraints." <https://doi.org/10.48550/ARXIV.1906.11357>.
- Shao, Yu, and Chip-Hong Chang. 2011. "Bayesian Separation With Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 41 (2): 284–93. <https://doi.org/10.1109/TSMCA.2010.2069094>.
- Tan, Ke, and DeLiang Wang. 2018. "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement." In *Interspeech 2018*, 3229–33. ISCA. <https://doi.org/10.21437/Interspeech.2018-1405>.
- Toh, K., and S. Yun. 2010. "An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Least Squares Problems." In *Pacific J. Optimization*, 20:615–40. <https://www.semanticscholar.org/paper/An-accelerated-proximal-gradient-algorithm-for-norm-Toh-Yun/6d389485b399ae7b60c1f426f1168f4eacaba64f>.
- Tolooshams, Bahareh, Ritwik Giri, Andrew H. Song, Umut Isik, and Arvinhd Krishnaswamy. 2020. "Channel-Attention Dense U-Net for Multichannel Speech Enhancement." In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 836–40. Barcelona, Spain: IEEE. <https://doi.org/10.1109/ICASSP40776.2020.9053989>.
- Tu, Ming, and Xianxian Zhang. 2017. "Speech Enhancement Based on Deep Neural Networks with Skip Connections." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5565–69. New Orleans, LA: IEEE. <https://doi.org/10.1109/ICASSP.2017.7953221>.
- Varga, Andrew, and Herman J.M. Steeneken. 1993. "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems." *Speech Communication* 12 (3): 247–51. [https://doi.org/10.1016/0167-6393\(93\)90095-3](https://doi.org/10.1016/0167-6393(93)90095-3).
- Wang, DeLiang. 2005. "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis." In *Speech Separation by Humans and Machines*, edited by Pierre Divenyi, 181–97. Boston: Kluwer Academic Publishers. https://doi.org/10.1007/0-387-22794-6_12.
- Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014. "An Experimental Study on Speech Enhancement Based on Deep Neural Networks." *IEEE Signal Processing Letters* 21 (1): 65–68. <https://doi.org/10.1109/LSP.2013.2291240>.
- Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2015. "A Regression Approach to Speech Enhancement Based on Deep Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (1): 7–19. <https://doi.org/10.1109/TASLP.2014.2364452>.

- Yuxuan Wang and DeLiang Wang. 2013. "Towards Scaling Up Classification-Based Speech Separation." *IEEE Transactions on Audio, Speech, and Language Processing* 21 (7): 1381–90. <https://doi.org/10.1109/TASL.2013.2250961>.
- Zhao, Han, Shuayb Zarar, Ivan Tashev, and Chin-Hui Lee. 2018. "Convolutional-Recurrent Neural Networks for Speech Enhancement." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2401–5. Calgary, AB: IEEE. <https://doi.org/10.1109/ICASSP.2018.8462155>.