Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Ideal Ratio Mask Estimation using Supervised DNN Approach for Target Speech Signal Enhancement

Poovarasan Selvaraj (poovarasan.cs@buc.edu.in)
Bharathiar University

Chandra Eswaran
Bharathiar University

Research Article

Keywords: Speech enhancement, SWEMD-VVMDH, DNN, Ideal ratio mask, Speech quality, Speech intelligibility, Generalizability

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-474390/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at Journal of Intelligent & Fuzzy Systems on October 31st, 2021. See the published version at https://doi.org/10.3233/JIFS-211236.

Ideal Ratio Mask Estimation using Supervised DNN Approach for Target Speech Signal Enhancement

Poovarasan Selvaraj, Dr.E.Chandra
Research Scholar, Professor
Department of Computer Science
Bharathiar University
Coimbatore-641046
poovarasan.cs@buc.edu.in, crcspeech@gmail.com

ABSTRACT

The most challenging process in recent Speech Enhancement (SE) systems is to exclude the non-stationary noises and additive white Gaussian noise in real-time applications. Several SE techniques suggested were not successful in real-time scenarios to eliminate noises in the speech signals due to the high utilization of resources. So, a Sliding Window Empirical Mode Decomposition including a Variant of Variational Model Decomposition and Hurst (SWEMD-VVMDH) technique was developed for minimizing the difficulty in real-time applications. But, this is the statistical framework that takes a long time for computations. Hence in this article, this SWEMD-VVMDH technique is extended using Deep Neural Network (DNN) that learns the decomposed speech signals via SWEMD-VVMDH efficiently to achieve SE. At first, the noisy speech signals are decomposed into Intrinsic Mode Functions (IMFs) by the SWEMD Hurst (SWEMDH) technique. Then, the Time-Delay Estimation (TDE)-based VVMD was performed on the IMFs to elect the most relevant IMFs according to the Hurst exponent and lessen the low- as well as high-frequency noise elements in the speech signal. For each signal frame, the target features are chosen and fed to the DNN that learns these features to estimate the Ideal Ratio Mask (IRM) in a supervised manner. The abilities of DNN are enhanced for the categories of background noise, and the Signal-to-Noise Ratio (SNR) of the speech signals. Also, the noise category dimension and the SNR dimension are chosen for training and testing manifold DNNs since these are dimensions often taken into account for the SE systems. Further, the IRM in each frequency channel for all noisy signal samples is concatenated to reconstruct the noiseless speech signal. At last, the experimental outcomes exhibit considerable improvement in SE under different categories of noises.

Keywords—Speech enhancement, SWEMD-VVMDH, DNN, Ideal ratio mask, Speech quality, Speech intelligibility, Generalizability

I. INTRODUCTION

In the globalized era, the most essential for SE is rejecting the microphone interferences due to the noises in the speech utterances. Several researchers have been designed many SE techniques for diminishing different categories of noises from the speech signals [1]. The compromised accurate assessment of the noise information is the key challenge in the SE systems. The standard assessments are primarily focused on Voice Activity Detectors (VAD) [2-3]. Later, the power distribution of noise characteristics was chosen as a smooth modification of its prior ranges for the speech duration. This approach suggests fair reliability for stationary noises, but the time-varying spectrum was not precisely assessed. For long speech segments and poor SNR, the difficulty of recognizing non-stationary noises emerges towards more intuitive [4]. These criteria were tackled by different power spectrum-based strategies [5-6].

The SE systems based on Time-Frequency (TF) [7] were suggested in preceding decades by the EMD method [8]. The speech signals were decomposed into a set of oscillatory IMF and a residual feature [9]. This method does not involve the set of essential components to analyze the target signal effectively. As well, it should not limit the stationary signals. So, a new SE technique based on EMD was suggested to resolve issues in non-stationary noisy scenarios [10]. Every IMF's noise characteristics were defined and chosen via its Hurst exponent data. Also, IMF selection and speech recovery were conducted in frame-by-frame via accounting target prediction for characteristics and reliability. Regardless, this technique has a high consumption of time and energy. Therefore, a significant improvement may not have been successful under Babble noise scenarios.

So, the SWEMDH technique [11] was recommended according to the EMD estimation in a fairly limited window that is sliding with the time axis. The window size was dependent on the speech signal frequency spectrum. The specific IMF distortions between windows were eliminated through the number of modes and the filtering iterations. For every element, the number filtering iterations can be adjusted based on the sampling rate, assessed signal, its consistency, and frequency band. This was done via decomposing the signal windows according to the usual method and calculating the actual number of filtering iterations. Also, the window sliding process was achieved by extracting the features related to the window

after obtaining the mode in the current iteration. These features were accumulated in the IMFs array with a suitable time index. It may happen that the window dimension was very small for spline interpolation so various window dimensions were needed for every mode. In this system, the default window dimension was used. The time indexes of the initial windows in the current iteration were stored. If the mode was extracted efficiently, then this index was maximized and the window for this mode was sliding forward in the subsequent iteration. If there were not adequate extrema, then the sliding task was terminated and the index of the initial windows for the current mode was not shifted. So in the subsequent iteration, the window was increased 2 times. On the contrary, the SWEMDH was not successful in backgrounds with white noise. As a result, a SWEMD-VVMDH technique was developed for minimizing the difficulty in real-time applications [12]. First, the noisy speech signals were decomposed into IMFs using the SWEMDH technique. Then, the TDE-based VVMD was performed on the IMFs to elect the most relevant IMFs according to the Hurst exponent and lessen the low- as well as high-frequency noise elements in the speech signal. Accordingly, the SE was achieved under various noise scenarios include additive white Gaussian noise, street noise, Babble noise, airport noise, and so on. Nonetheless, these are statistical frameworks that take a long time for computations. Hence, it needs to introduce advanced deep learning techniques to reduce the computational difficulty and improve the speech signal quality efficiently.

Therefore in this paper, the SE technique is developed using the DNN that learns the decomposed speech signals via SWEMD-VVMDH for enhancing the Speech Quality (SQ) efficiently. The main goal of SWEMD-VVMDH-DNN is to execute a systematic analysis of the generalizability abilities of the SE technique based on the estimated SQ and Speech Intelligibility (SI). For each signal sequence/frame, the target features are chosen and fed to the DNN that learns these features to estimate the IRM in a supervised manner. The abilities of DNN are enhanced for the categories of background noise, and the SNR of the target speech signals. Also, the DNN framework is differed and measured as a dimension. Besides, the noise category dimension and the SNR dimension are chosen for training and testing manifold DNNs since these are dimensions often taken into account for the SE systems. Further, the IRM in each frequency channel for all noisy signal samples is concatenated to reconstruct the noiseless speech signal.

The remaining article is prepared as follows: Section II presents the previous studies related to the SE systems using deep learning methods. Section III explains the SWEMD-VVMDH-DNN technique and Section IV displays its performance. Section V summarizes the work.

II. LITERATURE SURVEY

Xu et al. [13] designed a huge training set and fed it to the DNN for classifying the speech signals. Also, global variance equalization was used to avoid the over-smoothing issue of the regression framework. Further, the dropout and noise-aware training methods were employed for enhancing the generalization ability of DNN to invisible noise criteria. However, only a small number of data was used that cannot help to achieve better coverage of various acoustic situations like the speaker and morphological inconsistencies.

Williamson et al. [14] presented a scheme for enhancing the perceptual quality of speech segregated from background noise at low SNRs. First, an IRM was determined which segregates the speech from noise with reasonable sound quality. After, the DNN was used for approximating the clean speech via determining the activation weights from the ratio-masked speech where the weights linearly pool components from a Nonnegative Matrix factorization (NMF) speech framework. But, the performance was not effective.

Lee et al. [15] defined a new Spectro-Temporal Detection-Reconstruction (STDR) scheme in which the speech was extracted from background noise via learning a spectro-temporal feature space continuously. Also, a static nonlinearity was applied for projecting the noisy speech. Then, time-frequency gains were decoded together to adjust the noisy speech for creating clean speech estimation. But, the performance was not effective while considering the small amount of data.

Huang et al. [16] developed a common monaural source segregation scheme for mutually modeling all sources within a combination as targets to a Deep Recurrent Neural Network (DRNN). In this scheme, the limits between the actual combination and output predictions were used via time-frequency mask functions and mutually optimizing the time-frequency functions through DRNN. In contrast, the vanishing gradient properties of DRNN were not discarded that may affect the performance.

Chen et al. [17-18] suggested a DNN-based supervised speech segregation model with large-scale training for enhancing speech intelligibility. The trained DNN was employed for segregating the speech from noises such as multi-talker babble and cafeteria noise. But, it was

not suitable for real-time uses since it considers only magnitude spectra. Also, the generalizability dimensions were predetermined i.e., the dimensions of feed-forward DNN were fixed.

Amodei et al. [19] developed RNN with one or more convolutional layers followed by many recurrent layers, one fully connected layer, and a softmax layer. The entire network was trained end-to-end via the Connectionist Temporal Classification (CTC) loss function which facilitates direct recognition of the sequences of characters from the input speech signal. Conversely, the training was difficult while increasing the network's size and depth i.e., using more recurrent layers.

Zhang et al. [20] proposed multi-context networks for analyzing the monaural speech segregation. The initial multi-context network was used for averaging the outputs of many DNNs whose inputs use various window lengths. The second multi-context network was a stacked multiple DNNs where every DNN yields the concatenation of actual acoustic features and expanses the soft output of the lower unit as its input to identify the IRM of the target speaker. But, gradient descent-based training may cause vanishing gradient problems and degrade the efficiency.

Vidya et al. [21] preferred Hilbert Huang transform which consists of EMD for producing the IMFs and Hilbert spectral testing for detecting the regional properties of the speech signal. Also, Hilbert amplitude spectrum and phase spectrum were investigated to prevent the direct amplitude and phase of a speech signal. Moreover, the marginal and normalized Hilbert spectrum was determined. But, it needs to develop the analysis for further processing.

Mukherjee et al. [22] suggested a VAD method that utilizes the Line Spectral Frequency-based Statistical attributes called LSF-S merged with an extreme learning-based classifier. First, the audio signals were pre-processed and the attributes of multifarious sizes were extracted. These attributes were given to the extreme learning-based classifier to detect the voice and non-voice signals. But, it does not handle the live audio signals and needs to analyze the robustness under different noise scenarios.

Karjol et al. [23] presented a variant of a multiple DNN-based SE scheme which directly estimates the clean speech spectrum as a weighted mean of outputs from manifold DNNs. Initially, the weights were obtained via a gating network. After, the manifold DNNs and the gating network were trained together. Moreover, the objective function was assigned as the

mean square logarithmic error between the target clean spectrum and the estimated spectrum. In contrast, it requires optimization of DNN parameters for further performance improvement.

Zhao et al. [24] suggested a short-time objective intelligibility factor in the loss along with the Mean Squared Error (MSE) for improving the SE system. The perceptually guided loss was optimized using DNN according to the use of reinforcement learning with the previously defined time-varying rewards and a group of mask templates for further improving the speech intelligibility. However, the efficiency was degraded due to the fixed hyper-parameter during training.

Saleem et al. [25] suggested a Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) for extracting the acoustic features at the frame level. Also, an Auto-Regressive Moving Average (ARMA) filter was used for smoothing the temporal curves of extracted features. Then, the less aggressive Wiener filter was employed as an extra layer on the top of a DNN for generating an improved magnitude spectrum. At last, the noisy speech phase was used for restoring the enhanced speech. But, the parameters used in DNN were fixed i.e., fixed-size DNN cannot able to effectively improve the performance.

Wang et al. [26] developed the DNN framework for every microphone to improve the recorded noisy speech signal and all the noisy recordings were fused into the huge feature structure. A channel-dependent DNN was applied for enhancing the respective noisy input and all the channel-wise improved outcomes were fed to the DNN fusion framework for creating the almost clean signal. On the other hand, the recording channels were not accurately chosen rather than using all the channels as input.

Khaleelur Rahiman et al. [27] decomposed the noisy speech signal into frames and these were fed to the deep Convolutional Neural Network (CNN) for estimating the frequency channel. Then, the speech-dominated cochlear implant channels were considered for generating the electrical stimulation according to the estimated frequency channels. But, the parameters like window function and stride used in CNNs were fixed.

Llombart et al. [28] developed the progressive SE using convolutional and residual neural network structures. In this system, 2 different conditions were used for optimizing the loss factor such as weighted and homogeneous progressive. But, it needs more signals to provide better performance.

III. PROPOSED METHODOLOGY

In this section, the proposed SWEMD-VVMDH-DNN technique for achieving SE is explained briefly. The major tasks in the SWEMD-VVMDH-DNN technique are depicted in Figure 1.

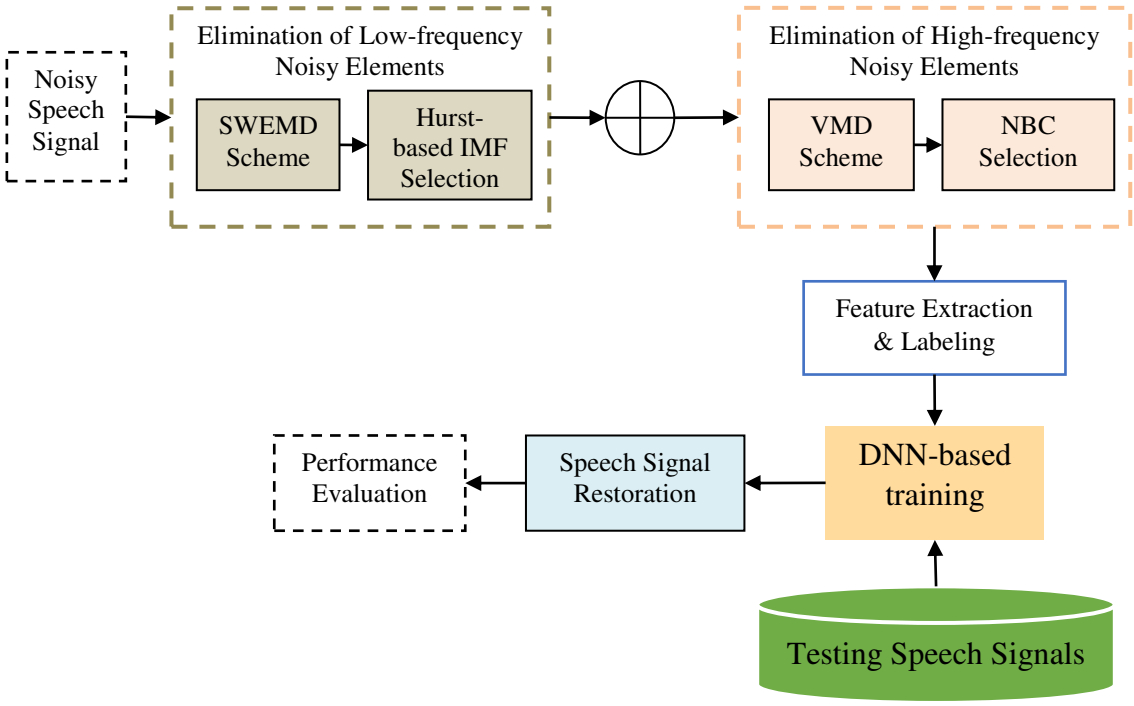

Figure.1 Block Diagram for SWEMD-VVMDH-DNN Technique

3.1 Speech Signals Decomposition using SWEMD-VVMDH Technique

First, the noisy speech signals are decomposed into the group of IMFs using the SWEMDH scheme [11]. Let N be the amount of IMFs extracted. For every IMF, the Hurst exponent H is calculated and the IMFs having $H \leq \frac{1}{2}$ are decided. So, $H > \frac{1}{2}$ is vital for eliminating the low-frequency noise elements from the input speech signals. Let K be the decided IMFs. Then, the sum of all K IMFs including input parameters is fed to the VVMD method to decompose into the group of Narrow-Band Components (NBCs) [12]. For every NBC, the mean $(\mu_{\bar{f}_C})$ and their respective standard variances $(\sigma_{\bar{f}_C})$ are determined. After, the NBCs having $\mu_{\bar{f}_C} + \sigma_{\bar{f}_C} \leq 3.6KHz$ are decided. So, $\mu_{\bar{f}_C} + \sigma_{\bar{f}_C} > 3.6KHz$ is essential for eliminating the high-frequency noise elements including DC elements from the input speech signals. Let P be the decided NBCs. Thus, both low- and high-frequency noise elements are

successfully eliminated from the input speech signals. Then, all the residual elements are considered for the DNN learning and reconstructing the speech signals effectively.

3.2 Feature Extraction and Labeling

The selection of training targets such as IRMs is desirable. So, the DNN is trained in a supervised manner for estimating the IRM from a feature representation of a noisy speech signal. The TF representation used for making the IRM is according to the gammatone filter bank with 64 filters linearly spaced on a Mel-frequency scale from 50Hz-8kHz and with a bandwidth identical to one Equivalent Rectangular Bandwidth (ERB). The output of the filter bank is split into 20ms frames with 10ms overlap and a sampling frequency of 16kHz, every TF unit denotes a vector of 320 samples. These frames and sampling frequency are highly satisfactory i.e., optimal to achieve a better spectral resolution of speech signals and mostly preferred to reconstruct the speech signal from its short-time magnitude spectrum

Consider $x(n,\omega)$ indicate the TF unit of the noiseless speech signal at frame n and frequency channel ω. Also, consider $d(n,\omega)$ be the respective TF unit of the noise signal. After, the IRM is determined as:

$$IRM(n,\omega) = \left(\frac{\|x(n,\omega)\|^2}{\|x(n,\omega)\|^2 + \|d(n,\omega)\|^2}\right)^{\beta} \quad (1)$$

In Eq. (1), $\|x(n,\omega)\|^2$ denotes the squared 2-norm i.e., the noiseless speech power of TF unit n in frequency channel ω. Similarly, $\|d(n,\omega)\|^2$ denotes the noise power of a TF unit n in frequency channel ω. The variable β is a tunable parameter and set to 0.5.

Every frame is converted into the 1845-dimensional feature vector for separating the discriminative and noise-robust features. The features considered are 31 Mel-Frequency Cepstrum Coefficients (MFCC), 15 Amplitude Modulation Spectrogram (AMS), 13 RASTA-PLP, and 64 Gammatone Filter bank Energies (GFE). Also, delta and double delta features are determined and a context of 2 past and 2 future frames are used and so obtaining the 1845-dimensional feature vector. Each feature vector is normalized to zero mean and unit difference.

3.3 DNN Architecture and Training

The DNNs follow a feed-forward framework with an 1845-dimensional input layer and three hidden layers and 64 output units. Each hidden layer has 1024 hidden units. The activation functions for the hidden units are Rectified Linear Units (ReLUs) and the sigmoid function is used for the output units. The DNN structure for SE is portrayed in Figure 2 wherein $x_1, \dots, x_n$ are the input units i.e., speech signals and y is the estimated $\widehat{IRM}(n, \omega)$.

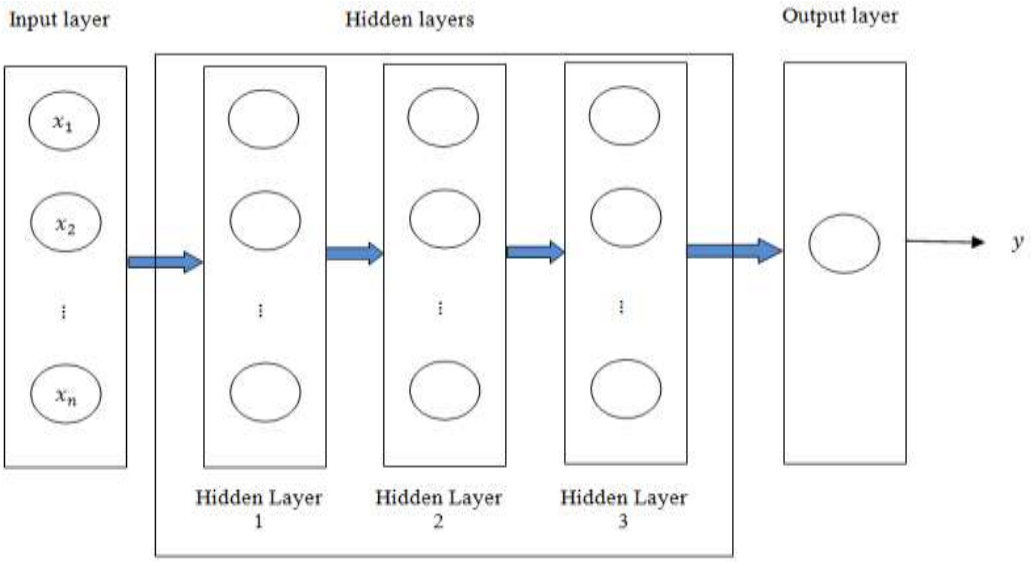

Figure 2. Architecture of DNN for SE

The hidden layers are initialized via the GlorotUniform scheme [29]. Moreover, the DNN has nearly 4M tunable parameters such as weights and biases. The values of these parameters are determined by the Stochastic Gradient Descent (SGD) following the AdaGrad method [30]. The gradients are calculated by the backpropagation according to the MSE function using a batch size of 1024. Also, 20% dropout is applied to each hidden layer during training for decreasing the overfitting. To further diminish the problem of overfitting, an early-stopping training mechanism is used which terminates the training while the MSE of the validation set is not reduced with greater than 1% for more than 20 epochs.

3.4 Speech Signal Restoration

Once DNN training is completed, the IRM is determined for a given test speech signal sequence via forward propagating its feature representation for each frame via the DNN. The outcome of the DNN is the expected $\widehat{IRM}(n, \omega)$ for the given frame. Then, the expected IRM is applied to the TF representation of the noisy speech signal by multiplying the given

entry of the mask to all noisy signal samples of a TF unit. Each TF unit in every frequency channel is concatenated and all overlapping elements are summed. After, the frequency channels are synthesized into the time domain signal via initial compensating for various group delays in the different channels and adding the frequency channels. The group delay compensation is executed via time reversing the signals, transferring them through the gammatone filter bank, and time reversing the signals again.

Algorithm:

Input: Training speech signal samples with their feature vectors $X = \{x_1, \dots, x_n\}$

Output: Expected $\widehat{IRM}(n, \omega)$

Initialize all weights and biases in the network;

$\boldsymbol{while}(termination\ condition\ is\ not\ satisfied)$

{

$\boldsymbol{for}(each\ training\ signal\ sample\ X)$

{

$\boldsymbol{for}(each\ input\ layer\ node\ j)$

$O_j = X_j$; //X_j & O_j: Input & Output of node j in the input layer

$\boldsymbol{for}(each\ hidden\ layer\ node\ j)$

$H_j = \frac{1}{1+e^{-j}}$; //H_j: Sigmoid function of node j in the hidden layer

$O_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right)$; //$O_j$: Output of $node\ j$ in the hidden layer, $f(\cdot)$: Transfer function, w_{ij}: Weight values, b_j: Bias values & x_i: Input feature vectors

$\boldsymbol{for}(each\ node\ j\ in\ output\ layer)$

$E_j = \frac{1}{2}\left(t_j^p - o_j^p\right)^2$; //$E_j$: Error of $node\ j$ in the output layer, t_j^p: Desirable output for the p-th observation in $node\ j$ & o_j^p: Real output for the p-th observation in $node\ j$.

Update the weight and bias values based on the E_j;

}

}

Return the expected $\widehat{IRM}(n, \omega)$;

end

IV. EXPERIMENTAL RESULTS

In this section, the SWEMD-VVMDH-DNN technique is implemented for evaluating its efficiency as compared to the SWEMD-VVMDH using MATLAB 2014a which has many audio tools and deep learning functions for execution. In this experiment, the TIMIT database [31] is taken into consideration for gathering the speech signals of American English in .wav format, which are directly digitized at a sampling rate of 20kHz by the Digital Sound Corporation (DSC) 200 along the anti-aliasing filter at 10kHz. After, the speech is digitally filtered, debiased, and downsampled to 16kHz. From this database, a group of 12 speakers i.e., a total of 120 speech signal frames encompassing 7 men and 5 women is randomly picked. Also, 8 are combined from every 10 signals per speaker and applied for training whereas the remaining 2 are divided for testing. Every 12×2=24 test signals are degraded by various categories of noises include additive white Gaussian noise, airport, babble, car, exhibition, restaurant, station, street, and train with different SNR of 0dB, 5dB, 10dB, and 15dB. These categories of noises are collected from the NOISEX-92 database [32]. Therefore, a total of 36×24=864 testing are carried out (9 categories of noises×4 categories of SNR=36). Figure 3 represents the magnitude spectrogram of a sample speech signal.

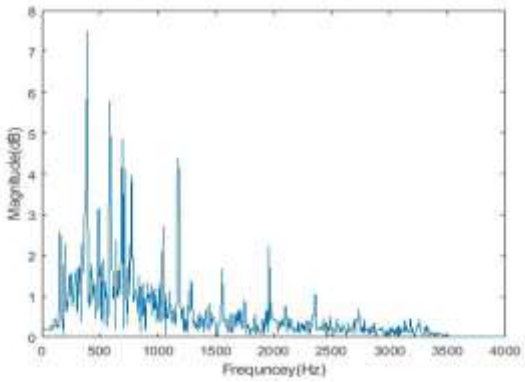

Figure 3. Magnitude Spectrogram of Sample Speech Signal

The performance metrics used for evaluation are:

- **SNR:** The rate of a mean power of speech signal (P_{signal}) to the mean power of noise (P_{noise}) is called the SNR.

$$SNR(dB) = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \quad (2)$$

It is rewritten as:

$$SNR(dB) = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \quad (3)$$

In Eq. (3), A_{signal} and A_{noise} are the Root Mean Square (RMS) amplitude of signal and noise, accordingly.

- **MSE:** The cumulative squared error between the reconstructed/enhanced speech signal and real speech signal frame is called the MSE.

$$MSE = \frac{1}{l}\sum_{i=1}^{n} e_i^2 \text{ where } e = \hat{x}(t) - x(t) \quad (4)$$

In Eq. (4), l denotes the signal length and e denotes the error between the real signal $x(t)$ and reconstructed/enhanced signal $\hat{x}(t)$.

- **Peak Signal-to-Noise Ratio (PSNR):** The rate of the maximum signal power to the noise power is called PSNR.

$$PSNR(dB) = 10\log_{10}\frac{255^2}{MSE} \quad (5)$$

- **Mean Absolute Error (MAE):** The overall uncertainty between the enhanced or reconstructed speech signal and a real signal is called the MAE.

$$MAE = \frac{1}{l}\sum_{i=1}^{n} e_i \quad (6)$$

- **Perceptual Evaluation of Speech Quality (PESQ):** It is used to analyze an end-to-end value for distinguishing the listening quality as apparent by users.

$$PESQ = \alpha_0 - \alpha_1 . D - \alpha_2 . A \quad (7)$$

Where $\alpha_0 = 0.1, \alpha_1 = 0.1, and\ \alpha_2 = 0.0309$

- **Weighted Spectral Slop Measure (WSSM):** It is the interval averaged WSSM where only the best frames are averaged. It computes the weighted variances of spectral slope over 25 critical frequency bands between any 2 related signal frames.

$$WSSM = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{\sum_{f=1}^{24} W(f)[\Delta E_S(f) - \Delta E_{\bar{S}}(f)]^2}{\sum_{f=1}^{24} W(f)}\right] \quad (8)$$

In Eq. (8), $\Delta E_s(f)$ and $\Delta E_{\bar{s}}(f)$ denote the spectral slope at frequency band S and $\bar{S}$, $W(f)$ refers to the weight in every band and N stands for the amount of best frames.

Table 1 gives the MSE outcomes for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission.

Table 1. MSE Comparison

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	0.0058	0.0034	0.0040	0.0037	0.0018	0.0010	0.0013	0.0012	0.0016	0.0008	0.0011	0.0010	0.0014	0.0010	0.0009	0.0008
Babble	0.0094	0.0054	0.0041	0.0042	0.0033	0.0018	0.0013	0.0022	0.0031	0.0016	0.0011	0.0022	0.0029	0.0014	0.0008	0.0022
Car	0.0070	0.0048	0.0027	0.0024	0.0028	0.0017	0.0008	0.0007	0.0026	0.0015	0.0006	0.0005	0.0024	0.0013	0.0009	0.0003
Exhibition	0.0074	0.0047	0.0038	0.0036	0.0033	0.0018	0.0013	0.0012	0.0031	0.0016	0.0011	0.0010	0.0029	0.0014	0.0009	0.0008
Restaurant	0.0096	0.0053	0.0037	0.0036	0.0034	0.0018	0.0013	0.0012	0.0032	0.0015	0.0011	0.0010	0.0030	0.0014	0.0009	0.0008
Station	0.0058	0.0047	0.0037	0.0043	0.0027	0.0016	0.0013	0.0022	0.0025	0.0014	0.0011	0.0022	0.0023	0.0012	0.0009	0.0022
Street	0.0073	0.0055	0.0048	0.0041	0.0032	0.0021	0.0021	0.0022	0.0030	0.0019	0.0019	0.0020	0.0028	0.0017	0.0017	0.0018
Train	0.0081	0.0052	0.0036	0.0034	0.0041	0.0023	0.0013	0.0012	0.0039	0.0021	0.0011	0.0010	0.0037	0.0019	0.0009	0.0008

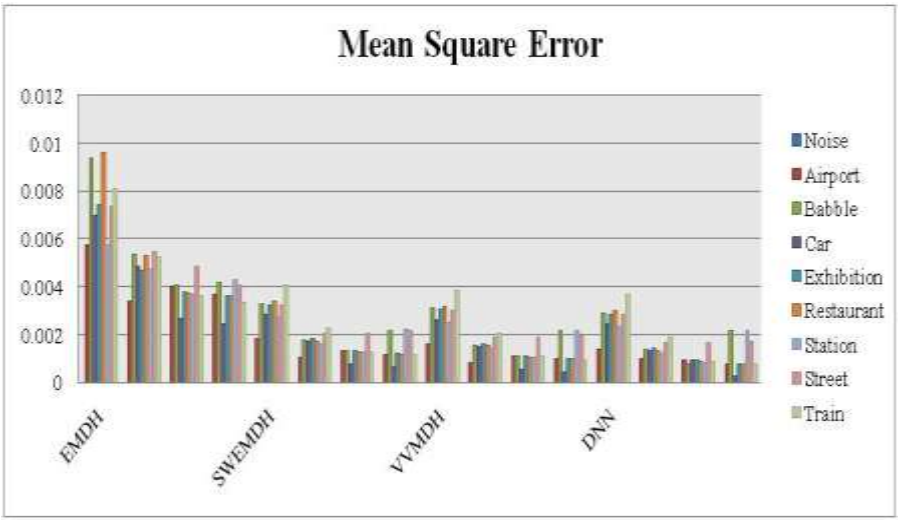

Figure 4. MSE Comparison

Figure 4 shows the graphical representation of MSE for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises. The x-axis denotes the different techniques and the y-axis denotes the MSE ranges. It recognizes that the DNN algorithm lessens the MSE than the other methods. E.g., if airport noise circumstance is considered with the SNR of 10dB, the MSE of DNN is 74.7%, 30.1%, and 17.7% minimized than the EMDH, SWEMDH, and VVMDH, respectively. This is attained because of the backward error correction and ability to accurately deal with different kinds of noise.

Table 2 gives the MAE results for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission.

Table 2. MAE Comparison

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	0.0581	0.0416	0.0412	0.0364	0.0326	0.0225	0.0225	0.0192	0.0290	0.0189	0.0189	0.0156	0.0254	0.0155	0.0153	0.0120
Babble	0.0716	0.0522	0.0410	0.0383	0.0419	0.0290	0.0222	0.0252	0.0383	0.0254	0.0186	0.0216	0.0347	0.0218	0.0132	0.0180
Car	0.0611	0.0471	0.0320	0.0285	0.0378	0.0273	0.0167	0.0139	0.0342	0.0237	0.0131	0.0103	0.0306	0.0201	0.0159	0.0067
Exhibition	0.0641	0.0466	0.0379	0.0360	0.0423	0.0287	0.0214	0.0196	0.0387	0.0251	0.0178	0.0160	0.0351	0.0215	0.0142	0.0124
Restaurant	0.0727	0.0516	0.0362	0.0362	0.0427	0.0284	0.0201	0.0189	0.0391	0.0248	0.0165	0.0153	0.0355	0.0212	0.0129	0.0117
Station	0.0544	0.0458	0.0362	0.0381	0.0382	0.0262	0.0200	0.0255	0.0346	0.0226	0.0164	0.0219	0.0310	0.0190	0.0128	0.0183
Street	0.0621	0.0463	0.0427	0.0368	0.0411	0.0271	0.0266	0.0252	0.0375	0.0235	0.0230	0.0216	0.0339	0.0199	0.0194	0.0180
Train	0.0663	0.0487	0.0335	0.0297	0.0470	0.0315	0.0198	0.0167	0.0434	0.0279	0.0162	0.0131	0.0398	0.0243	0.0126	0.0095

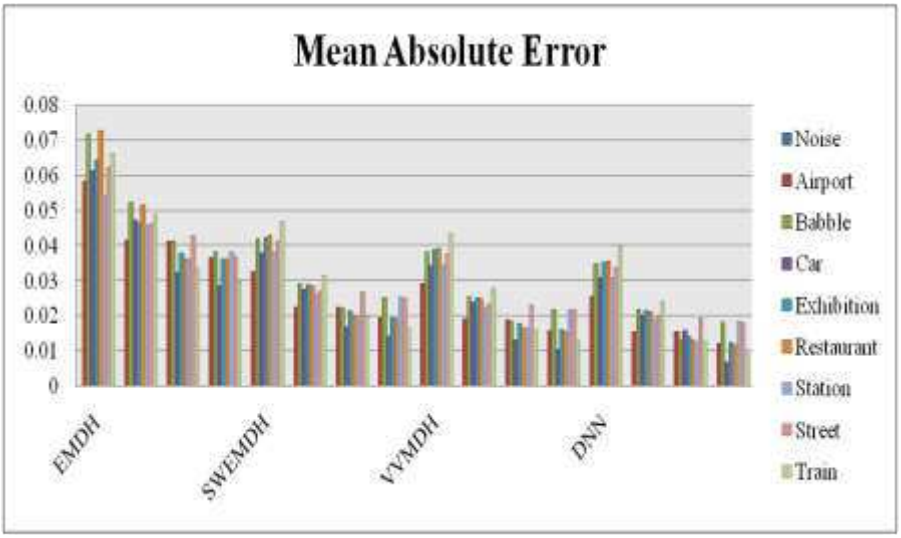

Figure 5. MAE Comparison

The graphical representation of MAE values for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises is portrayed in Figure 5. The x-axis denotes the different techniques and the y-axis denotes the MAE ranges. This analysis indicates that the DNN can

reduce the MAE compared to other methods. If the airport noise circumstance is taken into account with the SNR of 10dB, the MAE of DNN is 62.88%, 31.98%, and 19.04% reduced compared to the EMDH, SWEMDH, and VVMDH. This is because of an accurate tradeoff for mistakes between the significant bands and their respective modes which are concurrently determined via DNN.

Table 3 lists the SNR outcomes for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission.

Table 3. SNR Comparison (in dB)

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	2.6596	2.3340	1.8918	1.8312	5.7596	5.4340	1.8945	2.0288	8.7596	8.4340	4.8945	5.0288	11.1096	12.2796	7.2445	7.3788
Babble	2.6591	2.2435	1.9243	1.8478	6.8591	6.4435	3.4243	3.3478	9.8591	9.4435	6.4243	6.3478	12.2091	11.7935	11.1722	8.6978
Car	2.2961	2.5118	2.0742	2.0187	5.3962	6.7118	2.3071	2.6175	8.3962	9.7118	5.3071	5.6175	10.7462	12.0618	8.7820	7.9675
Exhibition	1.8031	1.7409	1.6864	1.7704	4.9031	5.5409	1.8448	2.0199	7.9031	8.5409	4.8448	5.0199	10.2531	10.8909	7.1948	7.3699
Restaurant	2.7790	2.2223	2.3393	1.7957	6.1790	4.9223	6.5393	2.0447	9.1790	7.9223	9.5393	5.0447	11.5290	10.2723	11.8893	7.3947
Station	1.4322	2.4314	2.3274	1.9001	1.8295	6.2314	3.8274	5.0001	4.8295	9.2314	6.8274	8.0001	7.1795	11.5814	9.1774	10.3501
Street	1.6990	2.5477	2.3765	1.7535	1.8802	6.7477	5.7765	4.0535	4.8802	9.7477	8.7765	8.9535	10.7005	12.0977	11.1265	11.3035
Train	1.6888	2.0173	2.1394	2.2150	4.7888	5.1173	3.6394	6.0150	7.7888	8.1173	6.6394	9.0150	10.1388	10.4673	8.9894	11.3650

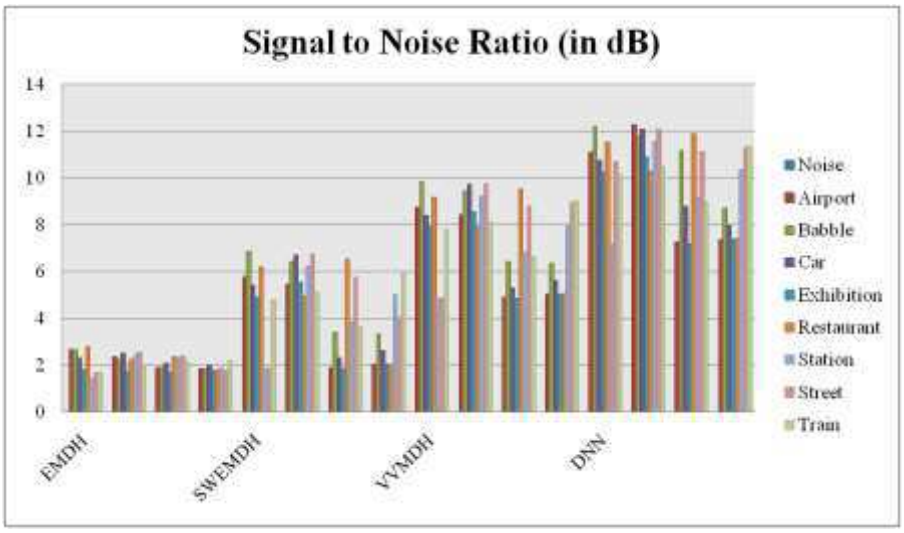

Figure 6. SNR Comparison

Figure 6 portrays the graphical representation of SNR values for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises. The x-axis denotes the different techniques and the y-axis denotes the SNR ranges (in dB). From this analysis, it is identified that the DNN algorithm increases the SNR compared to the EMDH, SWEMDH, and VVMDH methods. For the case of airport noise with the SNR of

10dB, the SNR of DNN is 7.2445 dB which is greater than EMDH, SWEMDH, and VVMDH methods whose SNR values are 1.8918 dB, 1.8945 dB, and 4.8945 dB, respectively. This is attained due to the narrow-band properties related to the current measure of the mode's center-frequency to the signal estimation residual of all other modes.

Table 4 gives the PSNR results for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission.

Table 4. PSNR Comparison (in dB)

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	12.6832	15.5304	15.1277	14.7907	17.6900	20.6958	19.9140	19.6507	18.1946	21.6208	20.6226	20.4431	18.7655	23.2590	21.4698	21.4131
Babble	12.6488	13.9453	14.2013	14.0989	17.1735	18.7410	19.0280	16.9647	17.4437	19.2596	19.7307	16.9665	17.7318	19.8488	22.3158	16.9665
Car	15.9196	17.2740	16.2434	24.7646	16.6770	21.7680	21.6248	22.3132	20.1690	22.3052	22.9169	23.8678	20.5126	22.9183	20.6891	26.3121
Exhibition	13.6671	13.5521	15.4931	14.9107	17.2422	17.6192	20.0242	19.7009	17.5173	18.1203	20.7253	20.4912	17.8111	18.6870	21.5617	21.4579
Restaurant	11.2770	13.3684	17.3195	14.8587	15.7940	18.1592	21.9581	19.6991	16.0577	18.6853	22.6945	20.5034	16.3385	19.2841	23.5818	21.4912
Station	14.4142	16.2747	17.3226	13.8600	17.6759	20.8709	21.9918	16.7419	18.0071	21.4358	22.7410	16.7777	18.3657	22.0853	23.6470	16.7777
Street	13.2661	15.3365	15.4882	13.9844	16.8453	19.5611	19.1347	16.7649	17.1238	20.0006	19.5715	17.1882	17.4213	20.4896	20.0573	17.6573
Train	14.0786	15.4111	17.3774	17.9914	17.0534	18.9970	21.8842	22.6148	17.2723	19.3946	22.6167	23.4362	17.5028	19.8322	23.4982	24.4502

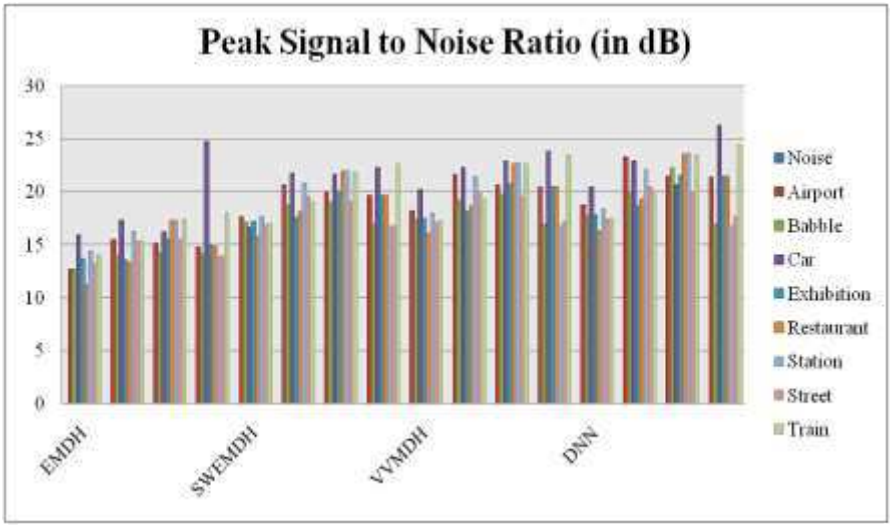

Figure 7. PSNR Comparison

The graphical representation of PSNR values for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises is depicted in Figure 7. The x-axis denotes the different techniques and the y-axis denotes the PSNR ranges (in dB). This analysis observes that the DNN can attain the maximum PSNR compared to other methods. E.g., if airport noise with

the SNR of 10dB is considered, the PSNR of DNN is 41.92%, 7.81%, and 4.11% maximized compared to the EMDH, SWEMDH, and VVMDH.

Table 5 gives the PESQ outcomes for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission. Similarly, Figure 8 displays its graphical representation. Here, the x-axis denotes the different techniques and the y-axis denotes the PESQ ranges. This scrutiny identifies that the DNN can improve the PESQ compared to other methods. E.g., consider the airport noise with the SNR of 10dB, the PESQ of DNN is 15.24%, 10.3%, and 5.47% maximized than EMDH, SWEMDH, and VVMDH. This is realized due to the proper adjustment of the center-frequency of low- and high-frequency harmonics which are identified at the tolerable quality and restored without faults.

Table 5. PESQ Comparison

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	3.5467	3.6631	3.3460	3.5071	3.7567	3.8731	3.4960	3.6571	3.9167	4.0331	3.6560	3.7571	4.0667	4.0478	3.8560	3.9371
Babble	3.6508	3.5891	3.6181	3.6256	3.8108	3.7491	3.7681	3.7756	3.9708	3.9291	3.8981	3.9856	4.0708	4.1391	4.1171	4.1456
Car	3.5010	3.7469	3.7168	3.5983	3.7110	3.7377	3.8168	3.6983	3.8710	3.9377	3.9968	3.8583	4.0210	4.0877	3.7754	4.0183
Exhibition	3.4376	3.4665	3.5156	3.6021	3.6476	3.6665	3.6956	3.7021	3.8076	3.8165	3.8556	3.8621	3.9576	3.9165	4.0156	4.0221
Restaurant	3.6158	3.3615	3.7524	3.3722	2.6942	3.4915	3.9124	3.5822	2.9042	3.6515	4.1124	3.7422	3.0542	3.8615	4.2624	3.9422
Station	2.9603	3.7972	3.8534	3.7621	3.1703	3.9972	3.6706	3.6889	3.3803	4.0972	3.7706	3.8889	3.5103	4.2972	3.9206	4.0189
Street	3.2797	3.5666	3.6083	3.7785	3.4397	3.7266	3.7883	3.8785	3.5897	3.8766	3.9483	4.0985	3.7497	4.0366	4.1083	4.2785
Train	3.2468	3.4448	3.5683	3.7313	3.4568	3.6548	3.7183	3.9313	3.5174	3.8148	3.9283	4.1413	3.5174	4.0148	4.0883	4.3513

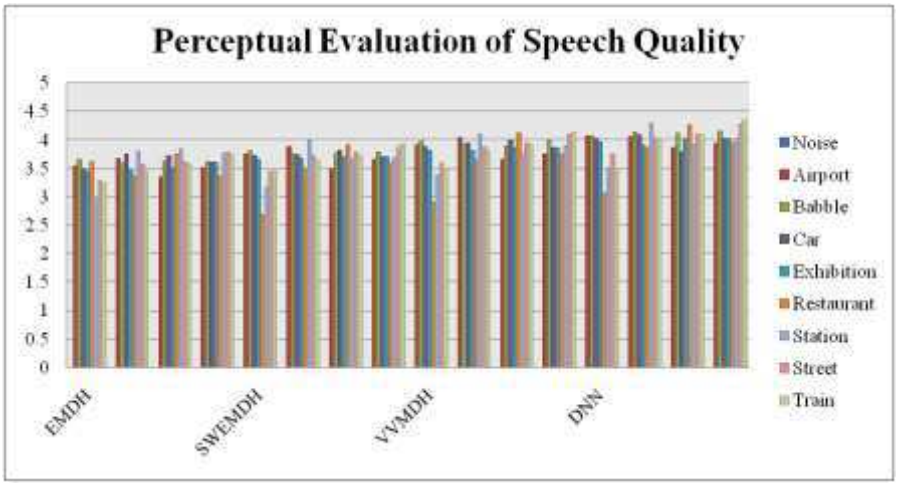

Figure 8. PESQ Comparison

Table 6 gives the WSSM results for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises that corrupt the speech signal during transmission.

Table 6. Weighted Spectral Slop Measure (WSSM) Comparison

Noise	EMDH				SWEMDH				VVMDH				DNN			
	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
Airport	0.3323	0.3383	0.3777	0.3981	0.2458	0.5118	0.5266	0.7243	0.7863	1.1131	0.9810	1.3386	0.8363	0.9102	1.0310	1.3886
Babble	0.3602	0.3409	0.4407	0.3782	0.4552	0.4693	0.7117	0.7137	0.9298	1.0638	0.9055	0.8804	0.9798	1.1138	1.1089	0.9304
Car	0.3227	0.3616	0.3623	0.3590	0.2077	0.2413	0.8262	0.5339	0.5667	0.7617	0.9135	1.0397	0.6167	0.8117	0.9112	1.0897
Exhibition	0.3419	0.3630	0.3819	0.4097	0.3780	0.6689	0.7466	0.8834	0.9475	1.0195	1.2339	1.1959	0.9975	1.0695	1.2839	1.2459
Restaurant	0.3205	0.3518	0.3902	0.4015	0.2335	0.3848	1.1889	0.6998	1.2673	1.0031	1.3001	1.5547	1.3173	1.0531	1.3501	1.6047
Station	0.3717	0.3220	0.4429	0.3621	0.5565	0.2621	0.8359	1.0695	0.8975	0.6001	1.0651	1.1348	0.9475	0.6501	1.1151	1.1848
Street	0.3045	0.3539	0.3671	0.3950	0.2291	0.3011	0.6315	0.9128	1.4370	0.6862	0.8957	1.4180	1.4870	0.7362	0.9457	1.4680
Train	0.3279	0.3761	0.4253	0.3799	0.1756	0.6352	1.3000	0.7405	0.7267	0.8769	1.1541	1.1703	0.7767	0.9269	1.2041	1.2203

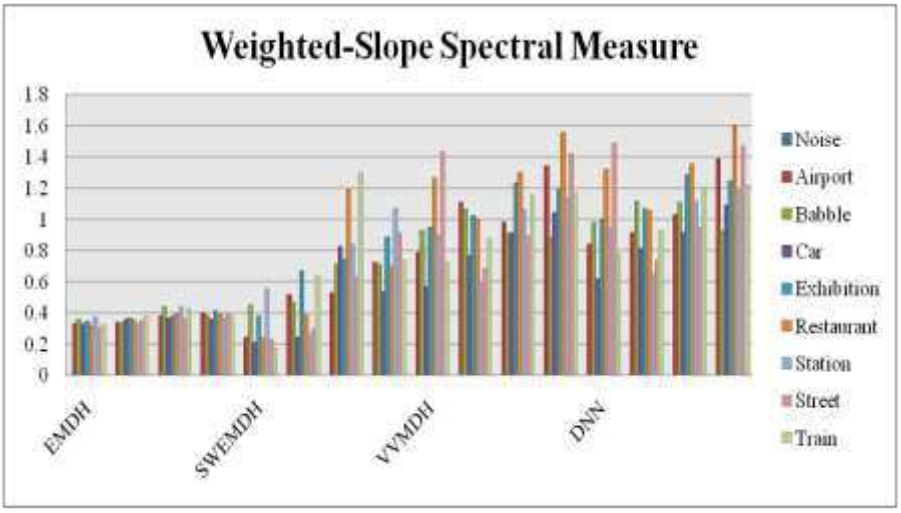

Figure 9. WSSM Comparison

Figure 9 portrays the graphical representation of WSSM for EMDH, SWEMDH, VVMDH, and DNN using different acoustic noises. The x-axis denotes the different techniques and the y-axis denotes the WSSM ranges. It observes that the DNN increases the WSSM compared to other methods, e.g., for airport noise with the SNR of 10dB, the WSSM of DNN is 1.0310 which is larger compared to the EMDH, SWEMDH, and DNN whose WSSM are 0.3777, 0.5266, and 0.9810, accordingly. Therefore, it is verified that the DNN attains enhanced efficiency in terms of greater PSNR, SNR, PESQ, WSSM and lesser MSE, MAE than the other methods.

Additionally, the computational complexity of DNN for N features is $O(N) + O(N)$ complex multiplications and additions; whereas, the other techniques need the computational complexity of $O\left(\frac{N}{2}\log_2 N\right) + O(N\log_2 N)$ complex multiplications and additions.

V. CONCLUSION

In this article, the SWEMD-VVMDH-DNN technique is proposed for learning the speech signal features to reconstruct the noiseless speech signals. The key objective of this SWEMD-VVMDH-DNN is to perform a systematic analysis of the generalizability abilities of the SE technique according to the estimated SQ and SI. Initially, the noisy speech signals are decomposed into IMFs by the SWEMDH technique. Then, the TDE-based VVMD was performed on the IMFs to elect the most relevant IMFs according to the Hurst exponent and lessen the low- as well as high-frequency noise elements in the speech signal. Also, the discriminative and noise-robust features from the speech signals are extracted via estimating the IRMs. Moreover, these are learned by DNN for the categories of background noise, and the SNR of speech signals for reconstructing the noise-free speech signals. To end, the experimental outcomes proved that the SWEMD-VVMDH-DNN technique has better efficiency than the SWEMD-VVMDH for SE under different categories of noises. The SWEMD-VVMDH-DNN achieves better performance in terms of MSE, MAE, WSSM, SNR, PSNR, and PESQ measures which are related to the SQ and SI as compared to the SWEMD-VVMDH techniques for SE.

REFERENCES

[1] Gulati, S. (2020). Comprehensive review of various speech enhancement techniques. In *International Conference on Computational Vision and Bio Inspired Computing*, Springer, Cham, pp. 536-540.

[2] Dov, D., Talmon, R., & Cohen, I. (2015). Audio-visual voice activity detection using diffusion maps. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(4), 732-745.

[3] Zhang, Y., Tang, Z. M., Li, Y. P., & Luo, Y. (2014). A hierarchical framework approach for voice activity detection and speech enhancement. *The Scientific World Journal*, *2014*.

[4] Mai, V. K., Pastor, D., Aïssa-El-Bey, A., & Le-Bidan, R. (2015). Robust estimation of non-stationary noise power spectrum for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(4), 670-682.

[5] Ji, Y., Baek, Y., & Park, Y. C. (2017). Robust noise power spectral density estimation for binaural speech enhancement in time-varying diffuse noise field. *EURASIP Journal on Audio, Speech, and Music Processing*, *2017*(1), 25.

[6] Jin, Y. G., Shin, J. W., & Kim, N. S. (2017). Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement. *The Journal of the Acoustical Society of America*, *141*(3), EL228-EL233.

[7] Samui, S., Chakrabarti, I., & Ghosh, S. K. (2019). Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network. *Applied Soft Computing*, *74*, 583-602.

[8] Khaldi, K., Boudraa, A. O., & Komaty, A. (2014). Speech enhancement using empirical mode decomposition and the Teager–Kaiser energy operator. *The Journal of the Acoustical Society of America*, *135*(1), 451-459.

[9] Lv, Y., Yuan, R., & Song, G. (2016). Multivariate empirical mode decomposition and its application to fault diagnosis of rolling bearing. *Mechanical Systems and Signal Processing*, *81*, 219-234.

[10] Zao, L., Coelho, R., & Flandrin, P. (2014). Speech enhancement with emd and hurst-based mode selection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(5), 899-911.

[11] Poovarasan, S., & Chandra, E. (2019). Speech enhancement using sliding window empirical mode decomposition and hurst-based technique. *Archives of Acoustics*, *44*(3), 429-437.

[12] Poovarasan, S., & Chandra, E. (2020). A variant of SWEMDH technique based on variational mode decomposition for speech enhancement.

[13] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (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.

[14] Williamson, D. S., Wang, Y., & Wang, D. (2015). Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality. *The Journal of the Acoustical Society of America*, *138*(3), 1399-1407.

[15] Lee, T., & Theunissen, F. (2015). A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features. *Proceedings*

of the Royal Society A: Mathematical, Physical and Engineering Sciences, *471*(2184), 20150309.

[16] Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2015). Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(12), 2136-2147.

[17] Chen, J., Wang, Y., Yoho, S. E., Wang, D., & Healy, E. W. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *The Journal of the Acoustical Society of America*, *139*(5), 2604-2612.

[18] Chen, J., Wang, Y., & Wang, D. (2016). Noise perturbation for supervised speech separation. *Speech Communication*, *78*, 1-10.

[19] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, J. (2016). Deep speech 2: end-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pp. 173-182.

[20] Zhang, X. L., & Wang, D. (2016). A deep ensemble learning method for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(5), 967-977.

[21] Vidya, S. S., Mariselvam, A. K., Samiappan, D., Subramanian, S., & Latha, S. (2017). Processes incorporated in the extraction of IMF, EMD and speech signal analysis using Hilbert huang transform. In *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering*, pp. 1195-1201.

[22] Mukherjee, H., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, *21*(4), 753-760.

[23] Karjol, P., Kumar, M. A., & Ghosh, P. K. (2018). Speech enhancement using multiple deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5049-5052.

[24] Zhao, Y., Xu, B., Giri, R., & Zhang, T. (2018). Perceptually guided speech enhancement using deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5074-5078.

[25] Saleem, N., Irfan Khattak, M., Ali, M. Y., & Shafi, M. (2019). Deep neural network for supervised single-channel speech enhancement. *Archives of Acoustics*, *44*.

[26] Wang, S. S., Liang, Y. Y., Hung, J. W., Tsao, Y., Wang, H. M., & Fang, S. H. (2020). Distributed microphone speech enhancement based on deep learning. *axXiv preprint arXiv:1911.08153*.

[27] Khaleelur Rahiman, P. F., Jayanthi, V. S., & Jayanthi, A. N. (2020). Speech enhancement method using deep learning approach for hearing-impaired listeners. *Health Informatics Journal*, 1-19.

[28] Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., & Lleida, E. (2021). Progressive loss functions for speech enhancement with deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing*, *2021*(1), 1-16.

[29] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249-256.

[30] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(7), 2121-2159.

[31] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, *93*.

[32] Varga, A., & Steeneken, H. J. (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), 247-251.

Declarations

- The speech enhancement technique is improved by advanced machine learning techniques instead of using conventional statistical models.
- In this proposed technique, Deep Neural Network (DNN) is used for learning the SWEMD-VVMDH features efficiently.
- The main aim of this approach is conducting a systematic evaluation of the generalizability capabilities of a state-of-the-art DNN based speaker enhancement system in terms of estimated Speech Quality (SQ) and Speech Intelligibility (SI).
- The capabilities of state-of-the-art DNN is discovered with respect to the types of background noise, the gender of the target speaker and the SNR.
- the architecture of DNN may be varied and taken as a dimension. For this proposed approach, the noise type dimension and the SNR dimension are selected to train and test all DNN based speech enhancement systems.

Funding: Not applicable

Conflicts of interest/Competing interests: No Conflicts of Interest

Availability of data and material: Not applicable

Code availability: Not applicable

Authors' contributions

A DNN based approach does have potential to improve Speech Enhancement in a broader range of usage Situations and it support that matching the noise type is critical in acquiring good performance for DNN based SE algorithms.

Ethics approval Not Applicable

Consent to participate Not Applicable

Consent for publication: Nil

Figures

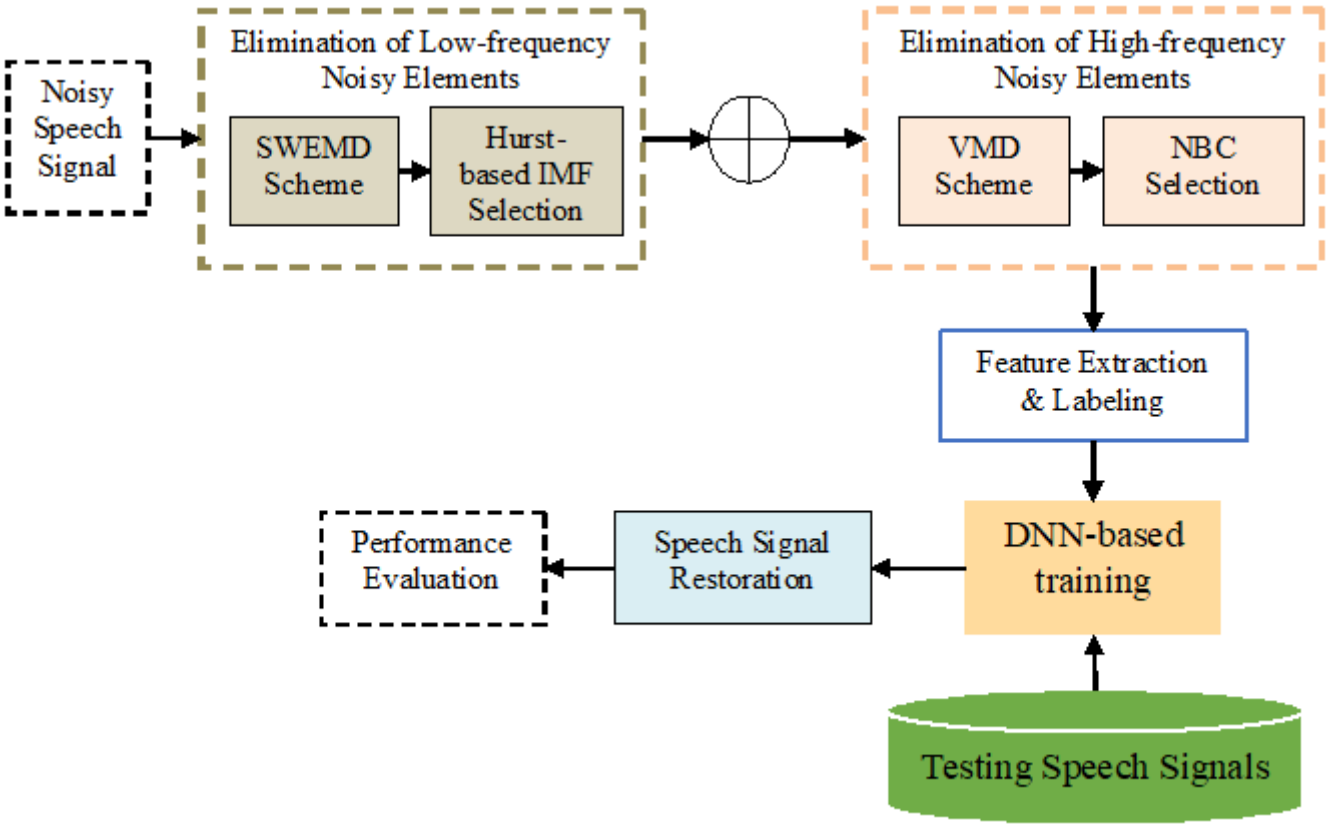

Figure 1

Block Diagram for SWEMD-VVMDH-DNN Technique

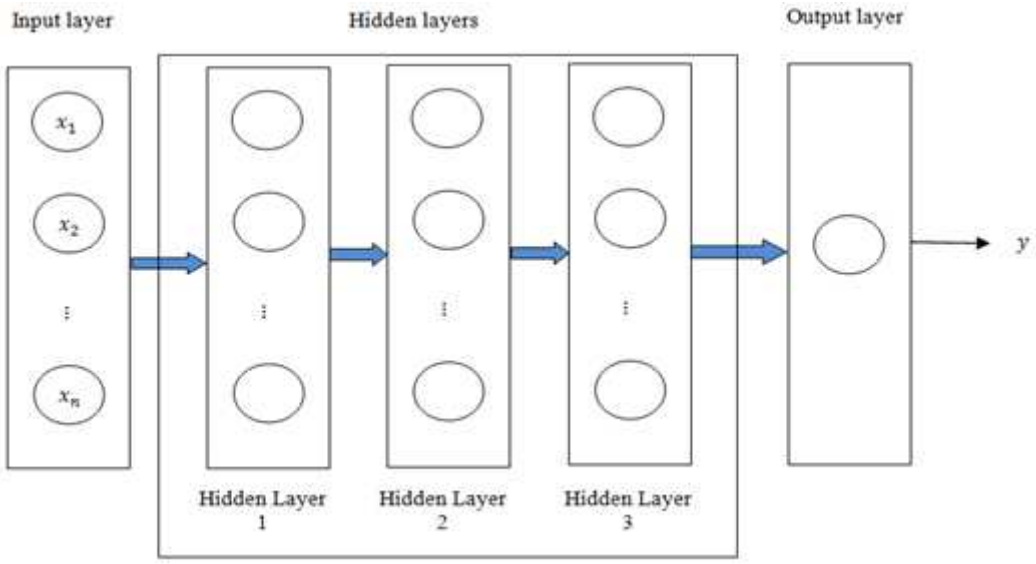

Figure 2

Architecture of DNN for SE

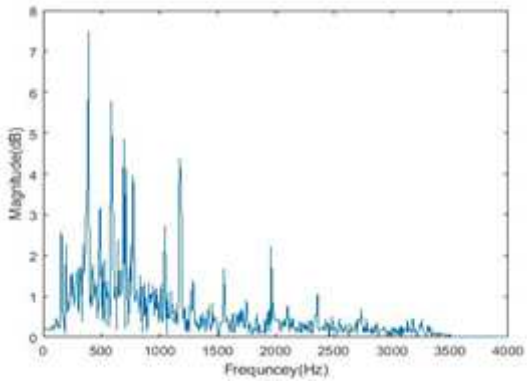

Figure 3

Magnitude Spectrogram of Sample Speech Signal

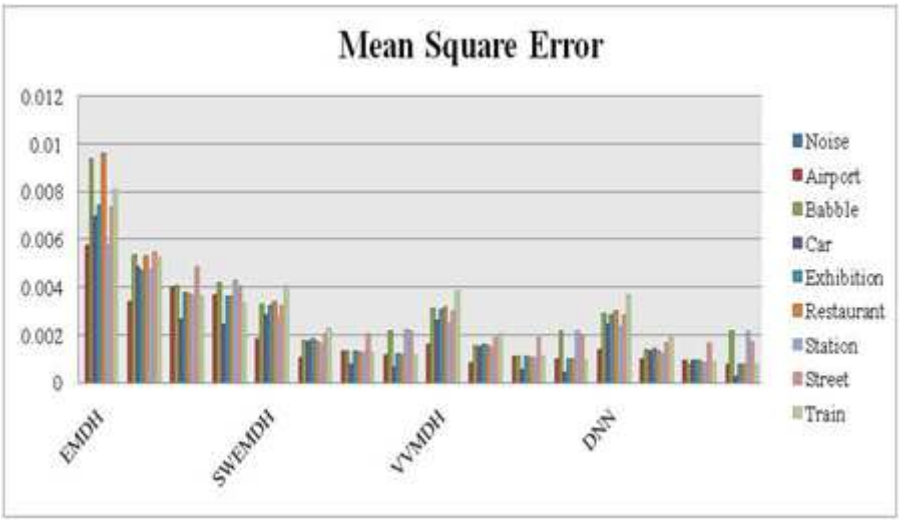

Figure 4

MSE Comparison

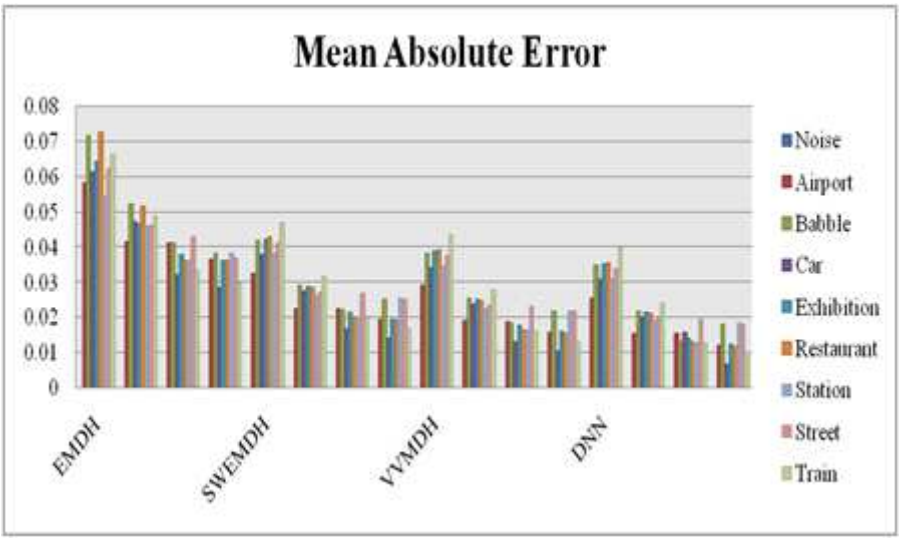

Figure 5

MAE Comparison

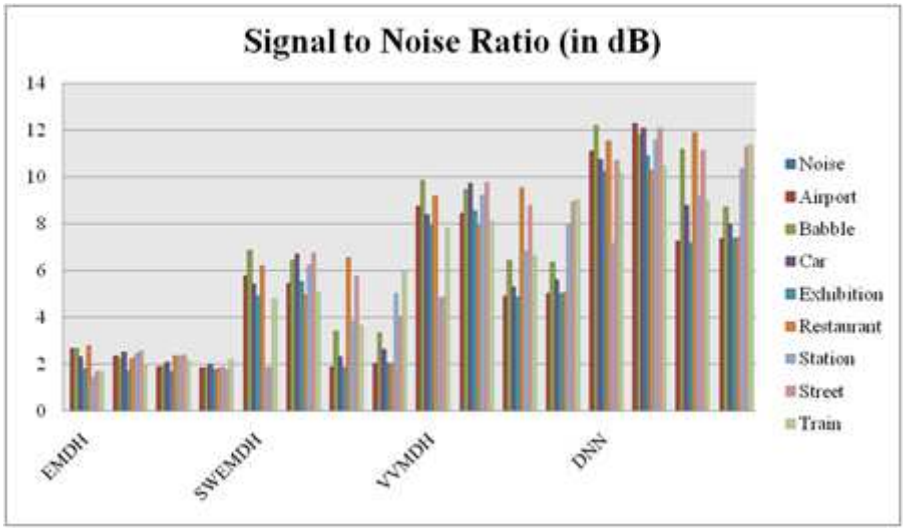

Figure 6

SNR Comparison

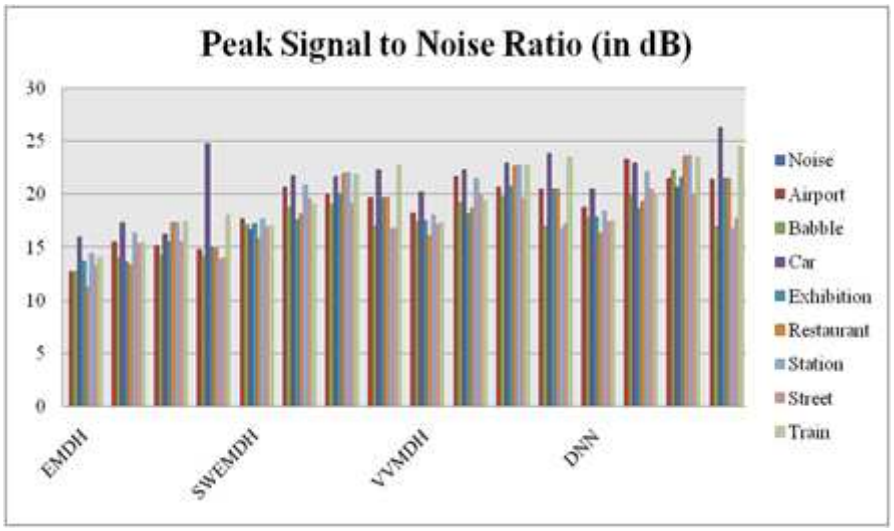

Figure 7

PSNR Comparison

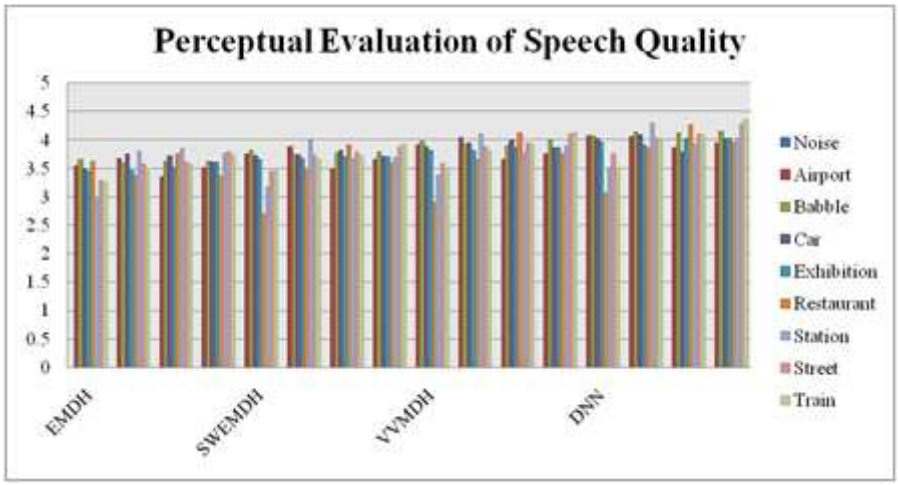

Figure 8

PESQ Comparison

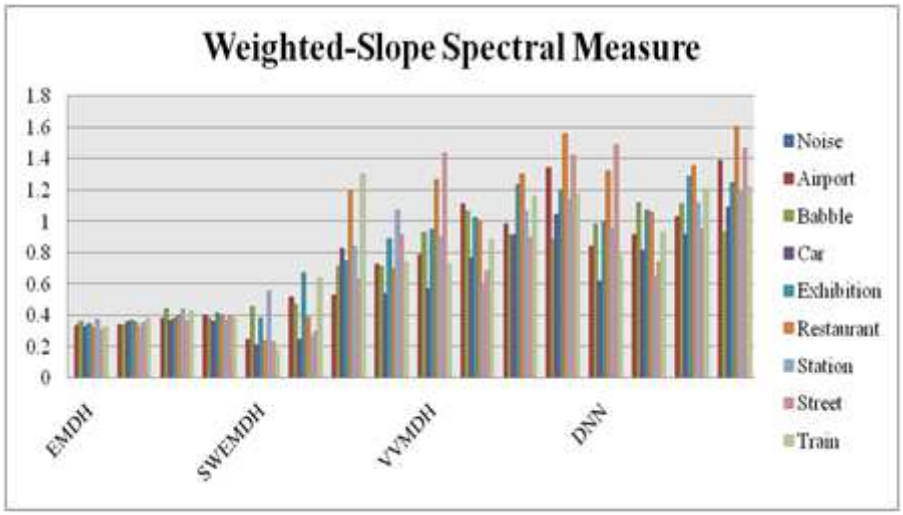

Figure 9

WSSM Comparison