A Fully Connected Deep Neural Network approach with multiple sub-frame consideration and phase recompense for noise suppression

Rohun Nisa, Asifa Mehraj Baba

Abstract—In the speech communication process, the desirable speech needs to be addressed under the influence of noise encountered in diverse environments that degrade the speech quality and intelligibility. In opposition to the unfavorable scenario particularly lowered signal-to-noise-ratio, the progress of traditional noise suppressive algorithms is hindered, introducing further distortion in speech, making them non-applicable for real-time applications. In order to reduce the complicacies of current algorithms, a hybrid approach for upgrading the quality together with intelligibility of speech is proposed for dealing with real-world hearing scenario. For improving the intelligibility of speech of interest, multiple sub-frame analysis using over-spectral subtractive factor with phase recompense approach is implemented on the multi-channel noise corrupted speech, yielding approximated speech spectrum, that constitutes the pre-processing stage. The approximated speech spectrum and clean speech spectrum forming the training set are further fed to Fully Connected Layered Deep Neural Network to reduce the mean square error with the incorporation of regression network resulting in improved quality for speech. The proposed hybrid network results in upgraded intelligibility and quality in speech signal with improved SNR measured in terms of Short-Time-Objective-Intelligibility (STOI) score, Perceptual-Evaluation-of-Speech-Quality (PESQ) score, Segmental SNR level, and Mean Square Error (MSE) in contrast to prior noise suppressive algorithms together with less complexity of the hybrid network.

Index Terms— Adam Optimization, Deep Neural Networks, Fourier Transform, Fully Connected Layers, Noise Suppression, Regression Network, Speech Processing

I. INTRODUCTION

The objective of noise cancelling algorithms includes the construction of speech without noise from corrupted signals, while enhancing the observed quality pertaining to speech segment together with intelligibility improvement. The noise removal technique can be employed in varied scenarios where background noise is encountered during the course of communication. During recent periods, numerous algorithms were considered for upgrading the modern transmission

outcome particularly in presence of noise and unwanted interference. Moreover, the knowledge regarding the performance of such algorithms continue to remain uncertain with regard to real-world hearing conditions due to unpredictable background noise characteristics. Thus, the major aspect for noise suppression algorithms to be reliable is robustness to unpredictable environmental noise. In addition, the research nowadays considers robustness related to changes involving Lombard effect, dialect, accent, stress, and emotion. The advancement regarding quality in speech signal is profoundly beneficial for the reduction in auditor's fatigue involving the scenario of high-level noise exposure for longer durations (e.g., manufacturing). The hard-of-hearing audiences using kits or cochlear implant aids generally experience complications during the process of communication in noise affected environment, as such noise suppression methods can be employed to preprocess and clean noise corrupted audio ahead of amplification.

The ground breaking work related to noise suppressive techniques was initiated by Schroeder [1][2] in the beginning. The author proposed a novel contribution in form of two patents explaining the applicability of analog method with regard to spectral-magnitude subtractive algorithm. In [3], Boll in the form of his descriptive method presented the digital aspect for spectral-subtractive algorithm. Berouti [4] explained the variant of [3] in the form of spectral over-subtractive method. Lim and Oppenheim [5] with reference to their milestone work portrayed the noise-suppressive challenge with the consideration of previous algorithms in the form of comparative analysis for upgrading the quality together with intelligibility of noise degraded speech to some extent. Ephraim and Malah [6] reported *Minimum-Mean-Square-Error-Short-Time-Spectral-Amplitude* (MMSE-STSA) criterion that involved the estimation of spectral gain with the consideration of independent Gaussian nature of noise and speech spectra together with Rayleigh and Uniform nature of amplitude and phase spectra respectively. Since then, many improved and more sophisticated algorithms based on the same approach have been proposed in [7],[8].

However, the traditional algorithms bring up certain artifacts referred as musical artifacts arising mainly due to incorrect noise spectral approximation [8]. Another approach to Spectral subtraction was employed referred as multi-band spectral

subtraction that takes into account the effect of real-world noise, the colored noise on speech signal for suppression to better enhance the quality and to overcome musical or residual noise resulting from traditional spectral subtraction method to some extent. This method relies on the selection of a particular subtraction factor for removing the required amount of residual noise from frequency bands without distorting the speech signal of interest thereby maintaining speech quality [9]. Certain variants of noise usually influence the low-frequency part of spectrum comparable to high-frequency part. Thus, there is a need to incorporate frequency-related subtractive element to take into account impact of noise variants on speech.

In recent times, machine learning algorithms have been found to be beneficial for suppressing the noise from corrupted speech. Wan and Nelson [10] carried out the preliminary work on Shallow Neural Networks (SNNs) regarding speech enhancement. However, the performance of the algorithm was limited due to constraints related to training parameter size and computation power that resulted in smaller network realization. Deep Neural Networks (DNNs) possessing multiple non-linear hidden layers have revealed greater capability in capturing the intricate relation among clean and noisy speeches for varied speakers, distinguishing noise characteristics and levels. Hu and Loizou [11] perceived noise-approximated procedures to be incorporated in static environment and demonstrated much improvement regarding the intelligibility of speech in babble noise presence followed by insignificant progress particularly in car surroundings. Yong Xu *et al.* [12] have evaluated the complex learning function that involves mapping from noise-corrupted to clean speech with the incorporation of regression based DNN. The work was carried out on large collective trained data of different parameters for noise degraded speech including SNR levels, talkers and noise variants, thereby enhancing the speech.

Yong Xu *et al.* [13] proposed a regression approach for noise suppression employing Deep Neural Network (DNN) in which a mapping function as linear predictive framework was computed among the noise corrupted speech and clean speech, resulting in subjective and objective quality improvement in comparison to traditional *MMSE* algorithm. Li *et al.* [14] investigated a novel approach as *Improved Least-Mean-Square Adaptive-Filtering (ILMSAF)* for noise suppression based on DNN with significant achievement regarding subjective and objective quality in speech in comparison to traditional Wiener filter approach by incorporating Deep Belief Networks (DBN) for computing adaptive filter parameters. Kolbæk *et al.* [15] examined DNN dependent noise suppression process regarding specific noise variants together with speakers and Signal-to-Noise-Ratio (SNR) providing an upgradation related to quality and intelligibility for desirable speech but accompanied by reduction in performance for other scenarios. Nicolson and Paliwal [16] proposed noise suppression approach that was carried out using DNN framework combined with *Minimum-Mean-Square-Error (MMSE)* method, following the objective regarding achievement of enhanced speech with improved intelligibility together with high quality, by incorporating *Residual-Long-Short-Term-Memory (ResLSTM)* structure.

Kolbæk *et al.* [17] have proposed DNN architecture for monaural noise suppression, maximizing the Short-Time-Objective-Intelligibility (STOI) parameter. With the incorporation of approximated STOI cost-function together with gradient-based optimization method for training phase, performance is shown to enhance as compared to conventional *Short-Time-Spectral-Amplitude (STSA)* DNN architecture for suppressing the noise. Sun *et al.* [18] have implemented a supervised approach for enhancement by incorporating Recurrent Neural Network (RNN) architecture for upgrading both the intelligibility and quality at low SNR situations. The architecture is designed for reduced computation burden applicable for real-time scenarios particularly for smartphone-focused binaural listening kits. Dash and Solanki [19] have put forward a hybrid methodology for upgrading the quality and intelligibility of speech. the author incorporates Adaptive approach to multiband spectral subtraction together with phase details for upgrading the intelligibility. The speech quality is further improved with the inclusion of Deep Neural Network and Nelder Mead optimization method that brings in enhanced performance.

Nossier *et al.* [20] have demonstrated a comparative analysis on the basis of three classes including the initially proposed *Deep Multi-layer Perceptron (MLP)*, *Convolutional Neural Networks (CNN)*, and *Denoising Autoencoder (DAE)*. The work carried out investigates the impact of network hyperparameter changes and data arrangement on the performance together with the Lombard effect. The authors contribution provides a detailed understanding regarding varied DNN architectures for noise suppression task, emphasizing the robustness and shortcomings for particular architecture, and providing necessary suggestions for attaining enhanced results.

Preferably, noise suppressive techniques are aimed to enhance quality and/or intelligibility for intended speech. The impact of noise on desirable speech can be minimized to a considerable extent but with the introduction of speech distortion, followed by speech intelligibility impairment. Therefore, the primary requirement in devising efficient noise suppressive technique is to minimize the influence of noise in the absence of any further distortion in the speech of interest. Until now, most of the suppressive algorithms are known to enhance quality of speech alone. Thus, the main aim of noise suppression algorithms becomes to upgrade quality together with intelligibility that will prove beneficial to hearing-kit implanted users during the course of day-to-day communication.

In this work, a Fully Connected Deep Neural Network with multiple sub-frame approach to spectral over-subtraction for noise suppression with magnitude and phase recompense is proposed. The multiple sub-frame work forms the pre-processing stage for multi-channel noise corrupted speech that will acquire rapid spectral transformations within a specified speech frame, yielding approximated speech with improved intelligibility. The clean spectrum of speech and approximated spectrum are fed as training pairs to Fully Connected Layered DNN that reduces the mean square error with the incorporation of regression network resulting in improved quality for speech,

together with the addition of phase recompense.

The paper is arranged as follows. The work proposed is illustrated in Section II with background mathematical knowledge of the proposed network described in Section III. The detailed results of the proposed work obtained are illustrated in Section IV together with the comparative investigation of state-of-art algorithms.

II. PROPOSED NETWORK

A. Multiple sub-frame analysis with spectral Over-subtractive factor as Pre-Processing Stage

The proposed method for suppressing the noise involving three stages of processing can be explained with the help of block diagram depicted in figure 1.

Fig. 1. Block Diagram of Proposed Network

The main idea depends on the generalized fact involving non-uniform impact of noise on speech signal spectrum. Considering the spectral properties of noise, part of the frequencies gets affected and corrupted due to disturbance adversely in comparison to remaining portion of spectrum. For the processing, speech signal having non-stationary characteristics is usually viewed as quasi-stationary for brief timespan. The analysis of desirable speech is thus carried out with the incorporation of short segment of windows, commonly termed as Frames. For each corresponding segment, mathematical transformation is applied by employing Short-Time-Fourier-Transformation that yields Spectrum form of speech consisting of respective magnitude and phase spectra.

The corresponding over-lapped frames are divided into N non-intersecting sub-frames followed by the application of over-spectral subtractive algorithm on respective sub-frames independently with a suitable factor. This eliminates the necessary noise estimate from noise corrupted speech for each sub-frame to yield approximated clean speech spectrum after reconstructing the corresponding frame. This constitutes the pre-processing stage before feeding the raw information of speech to network as mostly Deep Neural Network models do not function and learn right away from unprocessed form of speech signals, but are fed after transforming speech signal into its approximated spectrum domain.

For noise suppression, usually dual processes are involved for segmenting the audio signal into multiple frames, carried out as time-domain analysis incorporating band-pass filters or frequency-domain analysis employing suitable window functions. For the proposed algorithm to be computationally economical with standard implementation, latter method is

incorporated here with the utilization of hamming-window function in the beginning and hanning-window function at the end during processing of speech respectively, as represented in figure 2.

Fig. 2. Hamming-window and Hanning-window for speech processing

B. Fully Connected Deep Neural Network Stage

Further, the proposed approach is bifurcated into two modes: the training-mode and the testing-mode. Initially, for training-mode, the magnitude spectra of clean speech computed from STFT and approximated speech from pre-processing stage are fed as an input in form of vectors to Fully Connected Layered Deep Neural Network. The regression network takes the approximated speech spectrum to minimize the mean-squared-error among the clean speech spectrum and network outcome resulting in enhanced speech spectrum. The corrupted speech phase spectrum is not taken up for training mode and thus is assessed based on the SNR and estimated noise pertaining to corresponding sub-frame during pre-processing stage. After the combination of all frames, inverse Short-Time-Fourier-Transformation is employed to recover the speech signal in time format from spectrum domain with the addition of enhanced phase spectrum. The pre-processing stage involving multiple sub-frame analysis results in the improvement regarding intelligibility of speech signal together with phase information compensation, thereby dealing with multi-channel noise corrupted speech. In addition, achievement pertaining to SNR and enhanced quality in speech is attained by the incorporation of DNN with fully connected layers.

C. Phase Recompense Stage

Most of the traditional noise suppression algorithms ignore the phase approximation of clean speech spectrum as evident from preceding literature stating that noise corrupted speech phase forms the favorable alternative substituted for clean speech phase. The study is based on the matter that human-auditory recognition is unresponsive to any changes pertaining to phase knowledge regarding audio. However, latter studies reveal the importance regarding phase estimation and approximation for enhancing the speech corrupted due to noise variants by Paliwal *et al.* [21]. Later, Yukoh *et al.* [22] presented a phase reconstruction method for single-channel noise suppression with the incorporation of fundamental frequency and phase distortion features. These features were used to derive the relationship among harmonic segments together with temporal characteristics of the signal to determine the phase spectrum. The proposed method resulted in authentic

phase spectrum reconstruction over voiced intervals together with a degree of control on phase distortion with the use of a smoothing parameter and reduction in musical noise on the enhanced speech with upgraded PESQ and Segmental SNR.

III. BACKGROUND ANALYSIS OF PROPOSED NETWORK

A. Pre-Processing Stage

Considering the noise degraded speech as input that constitutes the clean speech sample from corpus considered and recorded data, distorted due to multi-channel background noise variants, the resulting sample is illustrated as

$$z[n] = x[n] + y[n] \quad (1)$$

where $z[n]$, $x[n]$, and $y[n]$ depict the sampled noise degraded speech, clear speech, and multiple channel noise, under the supposition of zero average time-domain value for noise, with n being the discrete time-index [5]. With the diverse time characteristics of speech, the speech processing is accomplished with the incorporation of Short-Time-Fourier-Transformation (STFT) on frame-by-frame analysis of degraded speech as interpreted by (1) and illustrated as

$$Z(\omega) = X(\omega) + Y(\omega) \quad (2)$$

Expressing in the polar representation with magnitude spectrum $|Z(\omega)|$ and phase spectrum $e^{j\phi_z(\omega)}$, the noise corrupted signal is given as

$$Z(\omega) = |Z(\omega)| \cdot e^{j\phi_z(\omega)} \quad (3)$$

Also, the noise spectrum $Y(\omega)$ with magnitude and phase spectra can be represented as $|Y(\omega)| \cdot e^{j\phi_y(\omega)}$. The magnitude noise spectrum $|Y(\omega)|$ is substituted by its average or estimated value $|\bar{Y}(\omega)|$ obtained during speech pauses, with the noise phase spectrum $\phi_y(\omega)$ substituted by phase of noise corrupted speech $\phi_z(\omega)$ based on the study that phase leaves lesser impact on speech intelligibility [23]. Thus, an approximate of enhanced speech spectrum is obtained as

$$|\bar{X}(\omega)| = (|Z(\omega)| - |\bar{Y}(\omega)|) \cdot e^{j\phi_z(\omega)} \quad (4)$$

The improved speech signal in time-format is thus achieved by employing reverse short time Fourier Transformation of $\bar{X}(\omega)$. This forms the basic principle of standard spectral-subtractive algorithm in magnitude representation. In terms of short-time power spectra domain, the algorithm is referred as Power Spectrum Subtraction algorithm and expressed as

$$|\bar{X}(\omega)|^2 = |Z(\omega)|^2 - |\bar{Y}(\omega)|^2 \quad (5)$$

This algorithm involves the elimination of an overestimate related to noise power spectrum, at the same time setting a specific value for resultant spectral factor rather than going below the already defined minimal value referred Spectral

Floor value, expressed in the form as

$$|\bar{X}(\omega)|^2 = \begin{cases} |Z(\omega)|^2 - \alpha|\bar{Y}(\omega)|^2 & \text{if } |Z(\omega)|^2 > (\alpha + \beta)|\bar{Y}(\omega)|^2 \\ \beta|\bar{Y}(\omega)|^2 & \text{elsewise} \end{cases} \quad (6)$$

where α refers as Over-subtractive factor ($\alpha > 1$) and β as Spectral Floor parameter ($0 < \beta \ll 1$). The main drawback encountered in traditional *Spectral-Subtraction* algorithm is the musical artifact appearance after the enhancement procedure. The two components α and β defined offer flexibility to a considerable extent for spectral-subtractive method. The component α controls the speech spectral distortion impact resulting from the subtraction factor and β regulates the extent of residual noise and the musical artifacts level left. The frame-to-frame variation in α [4] is given as

$$\alpha = \alpha_0 - \frac{3}{20} SNR, \quad -5 \text{ dB} \leq SNR \leq 20 \text{ dB} \quad (7)$$

Here, α_0 is set at accepted value of 0dB SNR with SNR computed and approximated in corresponding frame. This SNR is referred as *a posteriori* SNR estimated from the ratio of noise corrupted speech power spectrum to estimated noise power spectrum at frequency ω , given as

$$a_SNR(\omega) = \frac{|Z(\omega)|^2}{|\bar{Y}(\omega)|^2} \quad (8)$$

For the consideration and implementation of proposed multiple sub-frame approach to speech noise suppression, the corresponding frame is divided into $\mathcal{N}$ non-intersecting sub-frames. Additionally, the approximate of clean signal spectrum related to suppose j^{th} sub-frame is thus expressed as

$$|\bar{X}_j(\omega_k)|^2 = |Z_j(\omega_k)|^2 - \alpha_j \cdot \epsilon_j \cdot |\bar{Y}_j(\omega_k)|^2, \quad \beta_j \leq \omega_k \leq a_j \quad (9)$$

where $\omega_k = 2\pi k/N$ ($k = 0, 1, 2, \dots, N-1$) designate discrete frequencies, N as length of particular frame, $Z_j(\omega_k)$ as refined noise corrupted speech spectrum computed during initial periods, $\bar{Y}_j(\omega_k)$ as approximated noise spectrum, β_j and a_j as initial and final bins pertaining to j^{th} sub-frame, α_j as over-subtractive component corresponding to j^{th} sub-frame, and ϵ_j as added subtractive factor determined independently with regard to particular sub-frame.

The improvement in the SNR related to specific sub-frame referred as Segmental SNR_j is computed as

$$SegSNR_j(\text{dB}) = 10 \log_{10} \sum_{\omega_k=\beta_j}^{a_j} \left(\frac{|Z_j(\omega_k)|^2}{|\bar{Y}_j(\omega_k)|^2} \right) \quad (10)$$

The over-subtraction factor α_j forms the objective related to segmental SNR_j pertaining to j^{th} sub-frame, represented as

$$\dot{\alpha}_j = \begin{cases} 4.75 & SNR_j < -5 \\ 4 - \frac{3}{20}SNR_j & -5 \leq SNR_j \leq 20 \\ 1 & SNR_j > 20 \end{cases} \quad (11)$$

The incorporation of $\dot{\epsilon}_j$ magnitude brings in further command with regard to noise subtractive factor related to specific sub-frame, together with over-subtractive component $\dot{\alpha}_j$ and various bands of frequency. The value for $\dot{\epsilon}_j$ is analytically computed as

$$\dot{\epsilon}_j = \begin{cases} 1 & f_j \leq 1kHz \\ 2.5 & 1kHz < f_j \leq \frac{F_s}{2} - 2kHz \\ 1.5 & f_j > \frac{F_s}{2} - 2kHz \end{cases} \quad (12)$$

where f_j represent higher frequency related to j th sub-frame with F_s as sampling frequency in Hertz (Hz).

B. DNN Model Training

The main task in devising noise suppression system based on Fully Connected DNN is to respond the likely disparities and mismatching among training and testing requirements, due to varied SNR scales, talker variabilities and noise characteristics. After the pre-processing stage, the approximated clean speech spectrum $|\dot{X}_j(\omega_k)|$ from multiple sub-frame analysis, where $\dot{X}_j = \{\dot{x}_1, \dot{x}_2, \dots, \dot{x}_{N-1}, \dot{x}_N\}$ and clean speech spectrum $|\dot{Y}_j(\omega_k)|$, where $\dot{Y}_j = \{\dot{y}_1, \dot{y}_2, \dots, \dot{y}_{N-1}, \dot{y}_N\}$, forming training pairs are fed as input to Fully Connected Layered Deep Neural Network that results in the final estimated output $|\hat{E}_j(\omega_k)|$ in enhanced form without undesirable artifacts and distorted output. Initially two fully connected hidden layers with each layer characterized by 1024 neurons, accompanied by two Rectified Linear Unit (ReLU) layers constituting the activation task function and two Batch Normalization Layers that helps in normalizing the Mean and Standard Deviation for final outcome are incorporated. Another fully hidden connected layer is included followed by a Regression Layer that accounts for optimization and utilizes approximated clean spectrum to reduce the Mean-Square-Error (MSE) among the final estimated speech outcome and its clean spectrum. Considering each approximated and clean speech spectrum training set parameter $(\dot{X}_j, \dot{Y}_j)$, the parameters in reference to proposed network are optimized with the minimization of mean-square-error optimization function, computed as

$$\ddot{\chi}_{\dot{X}_j, \dot{Y}_j} = \frac{1}{N \times U} \sum_{n=1, f=1}^{n=N, f=U} \left[\left| \hat{E}_{j(n, f)}(\omega_k) \right| - \left| \dot{Y}_{j(n, f)}(\omega_k) \right| \right]^2 \quad (13)$$

where f represents the frequency index that varies depending upon the frequency axis dimension chosen, referred as spectral vector size denoted as U , n is the discrete index of time defined initially and N constitutes count of time frames formed during training mode. As the model is trained considering the magnitude spectra features of approximated and clean speech computed from STFT for estimating the enhanced outcome of

speech, the samples related to negative frequencies are initially discarded and frame size is halved depending upon the spectral symmetry features during the process.

C. Phase Compensation

For the compensation of phase, we propose an algorithm to recompense the phase related to noise corrupted speech spectrum, making it nearby to the phase spectrum of enhanced signal. The phase spectrum is compensated taking the Segmental SNR together with noise estimation achieved in existent frame. The proposed method significantly reduces the musical artifacts together with enhanced achievement way more than previous algorithms, particularly at lower SNR scale. To compensate for the phase, the phase spectrum is computed as follows. As noise corrupted speech constitutes the real-framed signal, its Fourier Transform forms conjugate symmetry [24]. The proposed algorithm takes into account the extent of reinforcement and cancellation of such conjugates by the adjustment of respective phases together with incorporation of phase spectrum recompense function impertinent to time-frequency relation. The necessary function is depicted as

$$u_j(\omega_k) = \tau e^{-SegSNR_j} \cdot \zeta_k |\dot{Y}_j(\omega_k)| \quad (14)$$

where τ is constant, ζ_k is time-frequency impertinent anti-symmetric function represented as

$$\zeta_k = \begin{cases} +1 & 0 \leq k < N/2 \\ -1 & \frac{N}{2} \leq k \leq N-1 \end{cases} \quad (15)$$

Thus, it is evident from (15) that lesser the $SegSNR_j$ level, the higher phase-spectrum recompense is achievable for the noise corrupted speech spectrum. As approximated magnitude spectrum $|\dot{Y}_j(\omega_k)|$ for noise is symmetric with antisymmetric phase spectrum as ζ_k , their product yields anti-symmetry function $u_j(\omega_k)$. The compensation regarding complex noise corrupted speech spectrum is achieved with the application of function illustrated as

$$Y(\omega_k) = Y_j(\omega_k) + u_j(\omega_k) \quad (16)$$

Together with the compensation of phase spectrum given as

$$\angle Y(\omega_k) = \arg[Y(\omega_k)] \quad (17)$$

Thus, enhanced clean spectrum of speech is acquired by merging together estimated speech signal spectrum from Deep network output and estimated phase spectrum from recompensated stage depicted as $|\hat{E}(\omega)| \angle Y(\omega_k)$, followed by the application of inverse transformation to provide time-domain version of speech signal.

IV. RESULTS AND DISCUSSIONS

The work is carried with the formation of multi-channel noise corrupted speech corpus constituting the clean samples of

speech together with diverse noise signals added through unseen distinct SNR scales to address broad spectrum of background noise encountered in actual communication scenario. The clean speech is extracted with the incorporation of ITU-T speech dataset [25]. Speech uttered with the inclusion of eight male and female speakers forming training and testing pairs, originally designed at a sampling rate of 48kHz were down-sampled to 8kHz resulting in minimized initial computational burden on the network. The clear speech is further contaminated by multi-channel real-world noise instances also down-sampled to 8kHz obtained from DEMAND dataset [26] [27], including noise of varied categories such as Domestic, Nature, Office, Public, Street, Transportation at unseen SNRs -10 dB, -5dB, 0dB, 5dB, 10dB of 8kHz sample rate, forming the noise corrupted speech corpus. The DEMAND dataset incorporated allows the training of proposed algorithm in an extensive series of adjustments for noise encountered in diverse environments represented by the spectrograms considered during pre-processing and training stages depicted in figure 3.

Fig. 3. Spectrogram analysis of noise signals at 0 dB SNR for noise variants extracted from dataset and incorporated during pre-processing stage followed by training mode of DNN (Courtesy: DEMAND Dataset [26], [27])

The implementation is carried out using MATLAB 2019b platform on Intel Xeon E3-1240v6, 32GB RAM, 4 cores, NVIDIA NVS 510 2 GB Graphics, Windows 10 Pro OS HP Z238 Workstation. Further, three unseen noise signals are recorded and added to clean speech for testing the performance of proposed algorithm. The duration of window N is chosen to be 32ms or window length of 256 samples with shift percentage or overlap of 75%, as window length of 15–35 ms is considered as favorable case for restoring the desirable speech magnitude spectrum in short-time domain [28]. For pre-processing stage, as noise is estimated during non-speech periods, the initial estimate of noise statistics is made from averaging over a few frames of silence. In this process phase, three frames of 250ms of silence period were added to corrupted speech signal, with estimate factor of the noise updated during speech pauses. Further, the 32ms frame duration is bifurcated into $N=8$ non-intersecting sub-frames of 4ms each with the application of spectral over-subtractive factor. The smoothing factor α for noise updating is set at value 9 and spectral floor parameter β of 0.03 that brings in further command with regard to noise subtractive factor related to specific sub-frame, with added subtractive factor ϵ_j of 1.5. The value of the constant τ is chosen to be 0.5 and the compensation function gets updated together with the analysis

of frame. For training mode of DNN, the samples pertaining to negative frequencies are discarded and the spectral vector size U are reduced to 129 from 256. The approximated speech spectrum thereby comprises of 8 successive vectors, as such single sample from clean spectrum and all approximated speech spectrum training vectors are fed to network, forming approximated speech vector size of 129-by-8 and clean speech vector size of 129-by-1 respectively. Initially, two hidden fully connected layers are defined with each layer comprising of 1024 neurons together with the incorporation of Rectified Linear Unit (ReLU) function followed by two Batch Normalization Layers. A third fully hidden connected layer is incorporated followed by a Regression Layer that accounts for optimization and utilizes approximated clean spectrum to reduce the mean-squared-error among the final enhanced speech outcome and clean spectrum. For the particular layer, initial learning rate is chosen to be 0.0067 with learning rate drop factor set at 0.9 for each epoch passing in the network and minibatch size of 128 with the epoch size set at 3. For normalization of two main parameters to zero mean with unit standard deviation, batch normalization layer is employed. For validation mode of the network, mean square is estimated for each epoch by setting the validation frequency taking 1% during training mode to account for overfitting of the network with the incorporation of Adaptive Moment Estimation (Adam) algorithm. This forms the training mode of the network which is further followed by the phase recompense procedure.

To analyze the performance of algorithm, segmental SNR, PESQ score, STOI score, and mean-squared-error (MSE) of estimated speech spectrum have been evaluated. The results depicted below shows the analysis of the proposed technique for the enhanced speech together with the comparative analysis of other noise suppression algorithms. The results regarding the improvement in segmental SNR achievable from Fully Connected DNN is depicted in figure 4 for the multi-channel noise variants added during the training phase and figure 5 depict the segmental SNR improvement for recorded noise during testing phase. The enhancement pertaining to quality and intelligibility for corrupted speech are illustrated in figures 6, 7, 8 and 9 in the form of PESQ score and STOI score respectively. The mean square computed during training and testing phases are pictured in figures 10 and 11 respectively.

Fig. 4. Segmental SNR improvement for Proposed Method for suppression of noise variants [26] [27] during pre-processing and training phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 5. Segmental SNR improvement for Proposed Method for suppression of recorded noise variants including Market Crowd, District Hospital Crowd and University Restaurant noises during testing phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 6. PESQ score improvement for Proposed Method for suppression of noise variants [26], [27] during pre-processing and training phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 7. PESQ score improvement for Proposed Method for suppression of recorded noise variants including Market Crowd, District Hospital Crowd and University Restaurant noises during testing phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 8. STOI score improvement for Proposed Method for suppression of noise variants [26], [27] during multi sub-frame pre-processing and training phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 9. STOI score improvement for Proposed Method for suppression of recorded noise variants including Market Crowd, District Hospital Crowd and University Restaurant noises during testing phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 10. Mean Square Error for Proposed Method for suppression of noise variants [26], [27] during pre-processing and training phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

Fig. 11. Mean Square Error for Proposed Method for suppression of recorded noise variants including Market Crowd, District Hospital Crowd and University Restaurant noises during testing phase together with comparison of methods (i) AMBSS [19] (ii) PDA [22] (iii) RDNN [12] (iv) MBSS [9], and (v) SOS [4]

The proposed approach greatly suppresses the significant level of non-stationary real-world background noise encountered. The fully connected layered network incorporated provides the advantage of predicting enhanced frame from the equivalent clean frame and approximated frame effectively, providing enhanced quality of speech. The network results in improved segmental SNR, quality and intelligibility for 15 multi-channel noise variants involving Kitchen, Living Room, Washing, Sports Field, City Park, River, Hallway, Meeting, Office, Cafeteria, Subway Station, Traffic, Bus, Car and Metro types during pre-processing and training modes. As the proposed model involves the training of multi-channel noise variants, it also turns out to be an efficient noise suppression algorithm in addressing recorded noise together with its impact on clean speech including Market Crowd, District Hospital Crowd and University Restaurant noise. The model greatly reduces the undesirable impact of musical artifacts resulting in enhanced speech in comparison to the state-of-art algorithms [19], [22], [12], [9], and [4] respectively without noise distortion as evident from results. In addition, both the multi-channel noise corrupted speech corpus and the recorded noise corrupted signals for real acoustic conditions are effectively validated, thereby providing a high-end noise suppression model. However, from the Segmental SNR graphs, there is less improvement for Market Crowd Noise in comparison to other two noise variants during testing phase due to high impact of noise at all SNRs levels.

V. CONCLUSION

In this work, a combined approach for suppressing the noise encountered in real-world scenario is proposed. For the pre-processing stage, the multiple sub-frame analysis is performed on the noise corrupted speech spectrum with precise over-subtractive factor together with the enhancement of phase spectrum resulting in approximated speech spectrum and enhanced intelligibility of speech. The clean speech spectrum and approximated spectrum form the training pairs and are fed to Fully Connected Layered Deep Neural Network that minimizes the mean square error among them yielding the upgraded speech signal with improved Segmental SNR and Quality score for each layer. The network is designed to minimize its initial computational burden during the entire training-testing process, thereby proving beneficial for real-time communication scenario. The comparative analysis of the proposed algorithm with the previous stated suppressive algorithms illustrates its enhanced performance in terms SNR, PESQ and STOI scores respectively.

REFERENCES

- [1] M. R. Schroeder: U.S. Patent No. 3180936, filed Dec. 1, 1960, issued Apr. 27, 1965
- [2] M. R. Schroeder: U.S. Patent No. 3403224, filed May 28, 1965, issued Sept. 24, 1968
- [3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, April 1979.
- [4] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, USA, pp. 208-211, 1979.
- [5] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in Proceedings of the IEEE, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.
- [6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109-1121, December 1984.
- [7] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 2, pp. 137-145, April 1980.
- [8] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, vol. 2, pp. 629-632, 1996.
- [9] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, pp. IV-4164-IV-4164, 2002.
- [10] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," Handbook of neural networks for speech processing, Artech House, Boston, USA, 1999.
- [11] G. Kim and P. C. Loizou, "Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 8, pp. 2080-2090, Nov. 2010.
- [12] Y. Xu, J. Du, L. Dai and C. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," in IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65-68, Jan. 2014.
- [13] Y. Xu, J. Du, L. Dai and C. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, Jan. 2015.
- [14] L. Ruwei, L. Yanan, S. Yongqiang, Shi, D. Liang, C. Weili, "ILMSAF based speech enhancement with DNN and noise classification," Speech Communication, vol. 85, pp. 53-70, 2016.
- [15] M. Kolbæk, Z. Tan and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 153-167, Jan. 2017..
- [16] A. Nicolson, K.K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," Speech Communication, vol. 111, pp. 44-5, Aug 2019.
- [17] M. Kolbæk, Z. Tan and J. Jensen, "Monaural Speech Enhancement Using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5059-5063
- [18] Z. Sun, Y. Li, H. Jiang, F. Chen, X. Xie and Z. Wang, "A Supervised Speech Enhancement Method for Smartphone-Based Binaural Hearing Aids," in IEEE Transactions on Biomedical Circuits and Systems, vol. 14, no. 5, pp. 951-960, Oct. 2020.
- [19] T. K. Dash, S. S. Solanki, "Speech Intelligibility Based Enhancement System Using Modified Deep Neural Network and Adaptive Multi-band Spectral Subtraction," Wireless Personal Communications, vol. 111, pp. 1073-1087, 2020.
- [20] S. A. Nossier, J. Wall, M. Moniri, C. Glackin and N. Cannings, "An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement," Electronics, vol. 10, no. 1, 2021.
- [21] K. Paliwal, K. Wójcicki, B. Shannon, "The importance of phase in speech enhancement," Speech Communication, vol. 53, no. 4, pp. 465-494, April 2011.
- [22] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, "Single-Channel Speech Enhancement With Phase Reconstruction Based on Phase Distortion Averaging," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1559-1569, Sept. 2018.
- [23] K. Paliwal, L. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," Speech Communication, vol. 45, no. 2, pp. 153-170, 2005.
- [24] K. Wójcicki, M. Milacic, A. Stark, J. Lyons and K. Paliwal, "Exploiting Conjugate Symmetry of the Short-Time Fourier Spectrum for Speech Enhancement," in IEEE Signal Processing Letters, vol. 15, pp. 461-464, 2008.
- [25] ITU-T Test Signals for Telecommunication Systems. Retrieved from: <http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm>
- [26] Diverse Environments Multichannel Acoustic Noise Database (DEMAND). Retrieved from: https://zenodo.org/record/1227121#_YMHCX6gzY2w
- [27] J. Thiemann, N. Ito, E. Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," in Proc Meetings on Acoustics, Montreal, QC, Canada, 2-7 June 2013.
- [28] K. Paliwal, K. Wójcicki, B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Communication, vol. 52, pp. 450-475, 2010.

Rohun Nisa received the B.E. degree in Electronics and Communication Engineering from University of Kashmir, Kashmir, India, in 2014 and M. Tech. degree in Information Technology from Central University of Kashmir, Kashmir, India, in 2018. From 2014 to 2016, she was working as an Assistant Professor in SSM College of Engineering and Technology, Kashmir. She is currently working towards Ph.D. degree at Islamic University of Science and Technology Awantipora, Kashmir, India. Her research interests include speech enhancement, speech recognition, speaker recognition, artificial intelligence.

Dr. Asifa Mehraj Baba born in Kashmir, has received B. Sc and M. Sc. degree in Electronics from University of Kashmir, Kashmir, India. She has received Ph.D. degree in Electronics and Communication Engineering from University of Kashmir, Kashmir, India, in 2009. She was working as Lecturer in University of Kashmir till September 2009. She joined as Assistant Professor in the Department of Electronics and Communication

Engineering, Islamic University of Science and Technology Awantipora, Kashmir, India in 2009. She is currently the Senior Assistant Professor in the same department. Her research interests include Artificial Intelligence, Deep Learning, Digital Image Processing and Speech Recognition. She has 30 publications in various reputed national and international journals. She has attended several national and international conferences within and outside India since 2015. She has in her credit two major research projects sponsored by MHRD, Government of India.