Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments

Nasir Saleem*, Muhammad Irfan Khattak

Department of Electrical Engineering, Gomal University, D.I.Khan (Pakistan)
Department of Electrical Engineering, University of Engineering & Technology, Peshawar, Kohat Campus (Pakistan)

Received 9 February 2019 | Accepted 31 May 2019 | Published 18 June 2019

Abstract

In this paper, we considered the problem of the speech enhancement similar to the real-world environments where several complex noise sources simultaneously degrade the quality and intelligibility of a target speech. The existing literature on the speech enhancement principally focuses on the presence of one noise source in mixture signals. However, in real-world situations, we generally face and attempt to improve the quality and intelligibility of speech where various complex stationary and nonstationary noise sources are simultaneously mixed with the target speech. Here, we have used deep learning for speech enhancement in complex-noisy environments and used ideal binary mask (IBM) as a binary classification function by using deep neural networks (DNNs). IBM is used as a target function during training and the trained DNNs are used to estimate IBM during enhancement stage. The estimated target function is then applied to the complex-noisy mixtures to obtain the target speech. The mean square error (MSE) is used as an objective cost function at various epochs. The experimental results at different input signal-to-noise ratio (SNR) showed that DNN-based complex-noisy speech enhancement outperformed the competing methods in terms of speech quality by using perceptual evaluation of speech quality (PESQ), segmental signal-to-noise ratio (SNRSeg), log-likelihood ratio (LLR), weighted spectral slope (WSS). Moreover, short-time objective intelligibility (STOI) reinforced the better speech intelligibility.

Keywords

Deep Neural Networks, Intelligibility, Deep Learning, Speech Enhancement, Time-Frequency Masking, Ideal Binary Mask.

DOI: 10.9781/ijimai.2019.06.001

I. Introduction

SPEECH enhancement is a vital research problem in many audio and speech signal processing applications. The aim of the speech enhancement is to improve the quality and intelligibility of a noisy speech signal. In applications such as hearing aids, automatic speech recognition (ASR), mobile communication etc., speech enhancement has been an active research area and countless approaches have been proposed in the recent past to solve this problem [1]-[7]. One of the simplest methods to eliminate the additive background noise was spectral subtraction proposed by Boll [8]. The wiener filtering [9] based method was proposed to estimate the noise in means square error (MSE) manner. Another important method is MMSE [10], which performs nonlinear estimation of the short-time spectral amplitude (STSA) of the speech signals. An excellent adaptation of the MMSE estimation, acknowledged as Log-MMSE attempts to minimize the MSE in the log-spectral domain [11]. Additional approaches include the signal-subspace [12], sparse coding [13], and empirical mode decomposition (EMD) [14] based methods, which are frequently used to perform the task of speech enhancement.

Recently deep neural networks (DNNs) based deep learning architectures have been found to be exceptionally successful in the automatic speech recognition (ASR) [15]-[16]. This achievement of DNNs in ASR directed to investigate the DNNs for noise elimination for speech enhancement [3], [17]-[19]. Fig. 1 shows the DNN based speech enhancement framework. The key idea behind using a deep neural network for speech enhancement is that, the degradation of the speech by noise signal is a difficult process and a complex nonlinear architecture like deep neural network is suitable to model it. A very few in-depth studies based on DNNs for speech enhancement are available in the literature; however, DNNs have shown remarkable outcomes and outperformed many classical speech enhancement methods. A general feature of such studies [18] [20] is evaluated in matching noise conditions. Matching noise conditions implies that the testing noise source is similar to training noise source. Mismatched noise a condition means to the situations when a DNN model has not been aware of testing noise sources during training. Xu in [18] provided a prominent study related to speech enhancement in mismatch conditions using DNNs. During this study, DNN was trained based on the variety of noise sources and showed that large improvements are achievable in mismatched conditions by exposing DNNs to a large number of noise sources. Mismatched noise conditions are relatively difficult scenarios compared to matched conditions. In real-world environments, we expect the DNN not to only execute well on large noise sources but also on nonstationary noises. Generally, speech signals are degraded by multiple noise sources in the real world situations and therefore elimination of single noise source in previous works is limiting. In environments around us, multiple noise sources simultaneously mix with the target speech and this multiple noise types situations are obviously much difficult to eliminate/suppress. To examine the speech

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

enhancement in such complex nonstationary situations, we suggest moving to an environment explicit prototype.

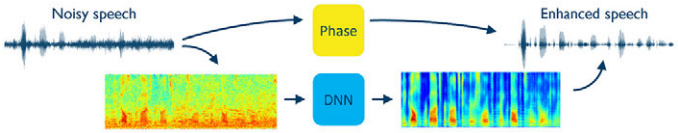

Fig. 1. DNN based Speech Enhancement.

In this paper we focus on a situation where complex noises are combined, for example, people talking in the street while vehicles are moving and the construction works are in progress. We synthesized complex noisy stimuli by adding multi-talker babble, airport and street noise simultaneously with clean utterances at different input SNRs. We have considered the deep learning based approach for speech enhancement in complex-noisy environments and an ideal binary mask (IBM) is considered as a binary classification function by using deep neural network (DNN). IBM is used as a target function during training and testing; the trained DNNs are used to estimate the IBM. The mean square error (MSE) is used as an objective cost function. The estimated target function is then applied to multi-noise mixtures to obtain the target speech.

The remaining paper is organized as: section II describes the basic problem and training DNNs for speech enhancement in complex-noisy environments, section III provides an explanation of the experiments and results and finally we concluded in section IV.

II. DNN Based Speech Enhancement

Our objective is to enhance a noisy speech in the complex-noisy conditions; where a number of possible different noise sources simultaneously degrades the quality of target speech utterances. The complex-noisy environments contain both stationary as well as nonstationary complex noise sources of completely different acoustic characteristics and are very close to real-world environments. Fig. 2 shows the time-domain waveforms and power spectral densities of various noise sources. Speech degradation under such conditions is a difficult and complex process compared to the single noise source, consequently enhancement of noisy speech becomes a more difficult task. Deep Neural Networks have high non-linear modeling capabilities and are presented in this paper for speech enhancement in complex multi-noise conditions. Prior to the actual DNN depiction, it is imperative to specify target function for DNN processing. Ideal ratio mask (IRM), ideal binary mask (IBM), short-time Fourier transform (STFT) magnitude and its mask, Mel-frequency spectrum and log-power spectra are potential target functions. We have selected IBM as target function [21]. During training, DNNs are trained and features are extracted from the noisy as well as clean speech utterances. A combined version of MFCC and RASTA-PLP features [22] are used in this paper. The extracted features are coupled with delta features to obtain Δ+DNN models. The time-frequency (T-F) representation utilized to create IBM which used a gammatone filter bank having 64 linearly spaced filters on a MEL frequency scale 50 Hz to 8 kHz and a bandwidth is equal to one Equivalent Rectangular Bandwidth (ERB) [23]. The output of the filter bank is divided into 20 ms frames with 10 ms overlap and with sampling frequency of 16 kHz. Let the noisy speech given as:

$$y(t) = s(t) + d(t) \tag{1}$$

Where $s(t)$ and $d(t)$ denote the clean speech and noise signals, respectively. The frequency-domain depiction of $y(t)$ is obtained as:

$$Y(k,\omega) = S(k,\omega) + D(k,\omega) \tag{2}$$

Where, ω and k denote frequency bin and time frame. During enhancement/testing, trained DNN is supplied with the features of the noisy speech to predict the coefficients of time-frequency mask. We have computed the coefficients of IBM, given as:

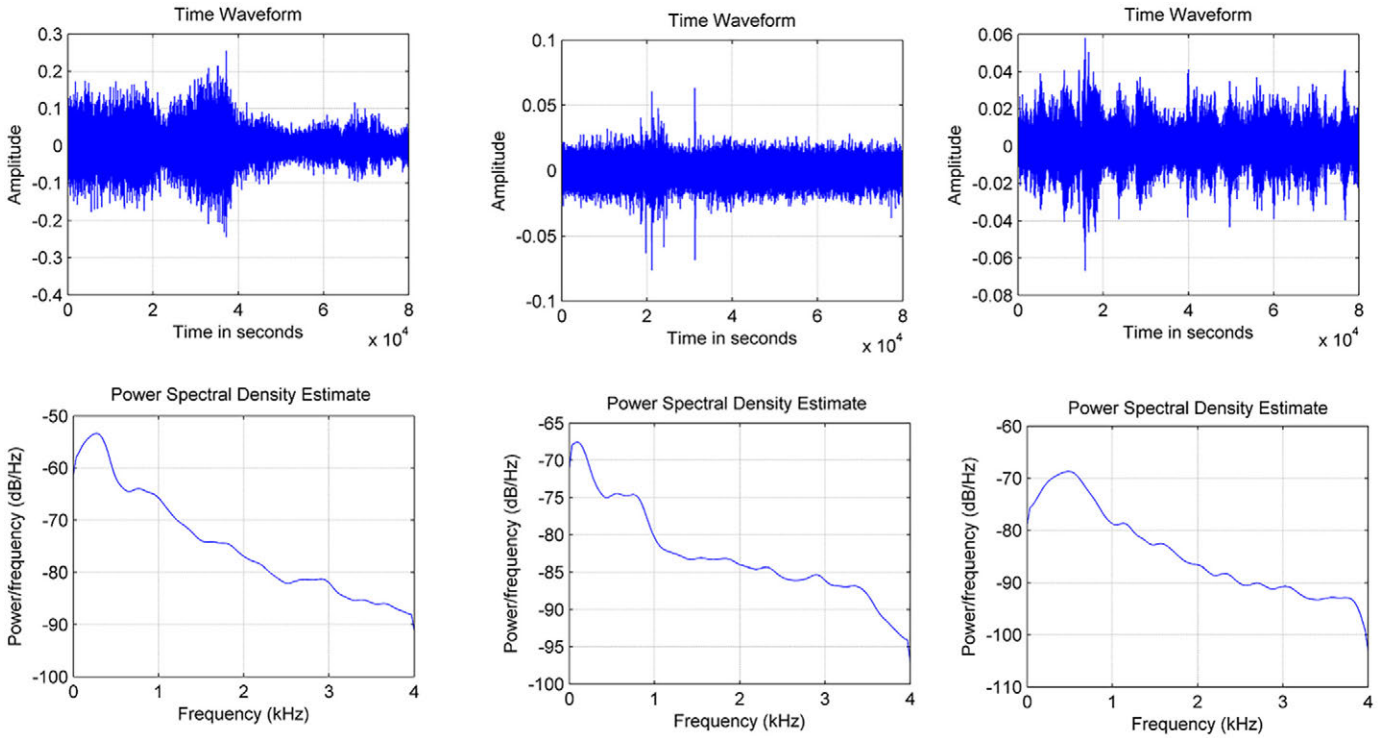

Fig. 2. Time-domain waveforms and Power Spectral Densities of Noise sources:Street, Exhibition hall and 32-multitakler babble noise (From left to Right).

$$IBM(k,\omega)=\begin{cases}1 & if\ SNR(k,\omega)>LC\\ 0 & if\ SNR(k,\omega)<LC\end{cases} \quad (3)$$

Where $S(k,\omega)$ and $D(k,\omega)$ denote T-F units of the speech and noise energies, whereas LC is local criterion. We found that $LC=0dB$ to be the finest choice for the mask estimation. The estimated magnitude of the clean speech is achieved by multiplying the estimated IBM mask with the noisy speech magnitudes. We have extracted the phase directly from the noisy speech because human auditory system remains unresponsive to small phase errors. Finally, inverse filtering is applied to reconstruct time-domain speech.

Speech enhancement formulates noisy speech signals to enhanced signals with better perceptual quality and intelligibility and usually is considered as estimate of clean speech. Supervised speech enhancement maps this process as a supervised learning problem so that mapping is determined absolutely from the input data. The proposed method contains four modules: feature extraction, training, decoding of DNN and waveform reconstruction. In training stage, DNN model is trained by using features of the noisy and underlying clean speech signals. The acoustic feature sets include the PLP, RASTA-PLP, MFCC, GFCC and AMS. We have selected the combination of RASTA-PLP MFCC and AMS acoustic features. The features are coupled with related delta features. Auto-regressive moving average (ARMA) filter is applied to smooth temporal curves of extracted features to improve speaker identification rates:

$$\hat{F}(t)=\frac{\hat{F}(t-k)+.....+F(t)+.....+F(t+k)}{2k+1} \quad (4)$$

Where $F(t)$ shows the feature vector at time frame t, $\hat{F}(t)$ is filtered feature vector and k is the order of filter. A second order ($k=2$) ARMA filter is used.

A. Network Architecture and Training

The DNN follows the feed-forward structure with five hidden layers, every layer contains 1024 hidden units and 64 output units. Rectified Linear Unit (ReLU) [24] activation functions are used in the hidden layers and also used in output units. ReLU is non-liner in the nature; hence more suitable for speech signals. Additionally, if considering the sparsity of the activation functions, sigmoid or Tanh processes all the neurons, hence make the network dense and costly. On the other hand, ReLU do not activate all the neurons and thereby makes the activations sparse and more efficient. Moreover, ReLU is less computationally expensive than sigmoid and Tanh since it involves simpler mathematical operations. Generally, almost all DNN based speech enhancement use either RBM or autoencoder based pretraining for learning. Yet, for sufficiently large and varied datasets, the pretraining stage can be eliminated and in this paper we use random initialization to initialize DNNs. Additionally, 20% dropout rate is applied to five hidden layers at training stage to decrease the overfitting phenomenon. The adaptive gradient descent (AGD) [25] is coupled with a momentum term κ to optimize the DNNs. For the first initial few epochs, κ rate is fixed at 0.5 but κ rate is increased and fixed at 0.8. The network is trained with mean squared error (MSE), as cost function, for error-correction. The Dropout regularization [26] is used to manage the mismatch conditions. The DNN framework used in this paper is shown in Fig. 3, where H1, H2… are a number of hidden layers. Each hidden layer contains 1024 neurons. Therefore, the total number of neurons in all layers is 5120 which shows a deep neural network.

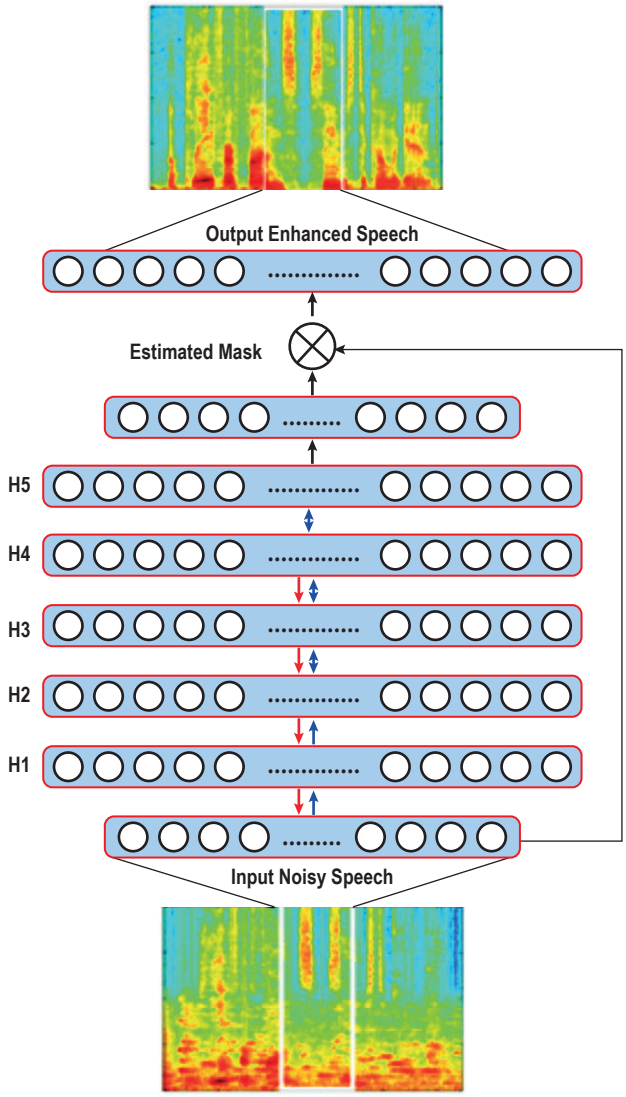

Fig. 3. DNN Training Framework.

III. Experimental Setup

A set of 720 IEEE [27] speech utterances is used during DNN training. The testing set contains 300 speech utterances from unknown speakers of both genders. We have used a real-world environment where multi-noise sources are degrading the target speech utterance. A restaurant like environment is considered in the experiments. The aim of selecting such noisy environment is two-folds: (a) this kind of noisy environment is complex to handle by a speech enhancement algorithm, thus appropriate to test the proposed algorithm; (b) this kind of noise contains diverse noise sources with different power spectral densities shown in Fig. 2. Such noise environment is more practical as individuals are exposed frequently to this noise type. The restaurant noise in our experiments is the mixture of 32-talker multi-babble noise, fan noise and noise originated from striking of the spoons, is supposed to be a real-world environment. The noise source contains both nonstationary (talking people and striking spoons) and stationary (fan noise) conditions. The duration of noise source is approximately 10 minutes.

To build the training set, the first-half of noise source is mixed with training utterances at -5dB, 0dB, 5dB and 10dB SNRs, respectively. The testing mixtures are built by mixing the last half of noise source.

We used five DNN models based on number of the hidden layers, represented by Δ+DNN$_1$, Δ+DNN$_2$, Δ+DNN$_3$, Δ+DNN$_4$ and Δ+DNN$_5$ models. For objective speech quality evaluation, we used Perceptual Evaluation of Speech Quality (PESQ) [28] whereas to evaluate the noise suppression, Segmental SNR (SNRSeg) [29] is used. Short-Time Objective Intelligibility (STOI) [30] is used to predict speech intelligibility. STOI refers to correlation between clean and enhanced signals and has been demonstrated to show high correlation to human speech intelligibility. To examine the DNNs, two distance measures, LLR and WSS, are used. The smaller values of distance measures indicate better result whereas the high values of PESQ, SNRseg and STOI indicate better performance. In our experiments, we have selected Wiener filtering (WF) and non-negative dynamical system (NNDS) as competing methods. The Wiener filtering is an unsupervised approach where NNDS is a supervised method. The DNNs represent a class of supervised methods for speech enhancement. Hence, we have compared the DNN approach with both supervised and unsupervised methods.

IV. Results and Discussions

As declared earlier, the goal of this study is to enhance a noisy speech using DNN in conditions similar to real-world environments. We have selected restaurant-environment for this study. We measured the quality and the intelligibility of the reconstructed enhanced speech in terms of the PESQ and STOI. Table I shows PESQ analysis for noisy, WF, NNDS and DNN with different layers at four input SNRs. The high PESQ scores of DNN show better performance. It is evident that speech quality achieved by DNN with four hidden layers (Δ+DNN$_4$) are higher than the noisy speech, two competing methods and DNN models with three and five hidden layers, that suggests improved speech quality of Δ+DNN$_4$. Table II presents the values of SNRSeg to indicate the suppression capabilities of DNN and other competing methods. Again Δ+DNN$_4$ performed better as compare to Δ+DNN$_3$, Δ+DNN$_5$ and competing methods. The noise is effectively reduced by DNN. Tables III-IV show performance analysis in terms of LLR and WSS. Clearly Tables III-IV indicate that distance between clean and reconstructed speech utterances is less for Δ+DNN$_4$. From Tables I-IV, it is clear that DNN with four hidden layers performed better as compared to DNN with other hidden layers. Therefore, DNN with four hidden layers is suggested to improve the quality and intelligibility of speech degraded by this particular real-time like noise source. The improvements in terms of the PESQ, SNRSeg, LLR and WSS are evident in Tables I-IV.

TABLE I. PESQ Analysis

SNR	Noisy	WF	NNDS	DNN$_3$	DNN$_4$	DNN$_5$
-5	1.57	1.61	1.66	1.68	1.72	1.67
0	1.77	1.94	2.08	2.19	2.22	2.21
5	2.08	2.21	2.39	2.42	2.46	2.42
10	2.46	2.63	2.59	2.65	2.68	2.66
Avg.	1.97	2.10	2.18	2.24	**2.27**	2.24

TABLE II. SNRSeg Analysis

SNR	Noisy	WF	NNDS	DNN$_3$	DNN$_4$	DNN$_5$
-5	0.50	1.24	1.54	1.73	1.78	1.70
0	1.08	2.18	2.98	3.14	3.21	3.20
5	2.30	3.31	4.54	4.93	4.98	4.92
10	4.40	5.43	6.29	6.92	7.02	6.93
Avg.	2.07	3.04	3.83	4.18	**4.24**	4.18

TABLE III. LLR Analysis

SNR	Noisy	WF	NNDS	DNN$_3$	DNN$_4$	DNN$_5$
-5	2.07	1.27	1.01	0.77	0.75	0.76
0	1.61	1.04	0.91	0.86	0.60	0.71
5	1.14	0.78	0.62	0.43	0.33	0.45
10	0.82	0.60	0.59	0.52	0.44	0.40
Avg.	1.41	0.92	0.78	0.64	**0.53**	0.58

TABLE IV. WSS Analysis

SNR	Noisy	WF	NNDS	DNN$_3$	DNN$_4$	DNN$_5$
-5	70.72	57.40	56.01	50.21	48.37	49.07
0	61.02	50.88	48.53	44.52	43.35	44.15
5	53.53	49.31	47.00	42.02	41.22	43.60
10	43.03	37.71	35.54	33.24	31.97	32.60
Avg.	57.07	48.82	46.77	42.49	**41.22**	42.35

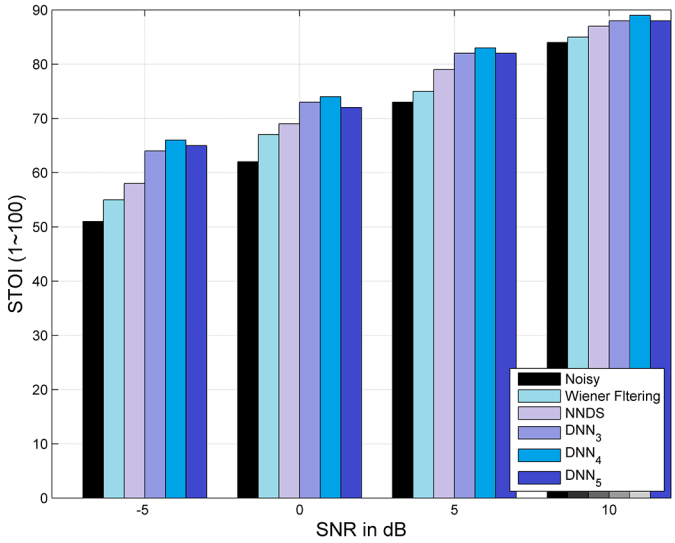

Fig. 4. Intelligibility Analysis.

Speech enhancement in order to improve the speech intelligibility using STOI measures is presented in Fig. 4. The achieved STOI scores with the WF, NNDS and DNN approaches in real-time like condition at four SNRs shows that a better performance is observed with DNN approach than with the two competing approaches and noisy unprocessed speech. Again STOI is computed for three DNN models i.e., Δ+DNN$_3$, Δ+DNN$_4$ and Δ+DNN$_5$ and two competing approaches. It is obvious from Fig. 4 that Δ+DNN$_4$ achieved the highest STOI score as compare to other models and competing approaches. The average STOI score is increased from 50% with noisy speech to 66.2% with Δ+DNN$_4$ at -5dB SNR. Similarly, the average STOI score is increased from 67% with WF to 78% with Δ+DNN$_4$ at 0dB SNR. Moreover, the average STOI score is increased from 79% with NNDS to 84% with Δ+DNN$_4$ at 5dB SNR. By observing the STOI scores, it is evident that DNN with four hidden layers (Δ+DNN$_4$) performed well in improving speech intelligibility. Fig. 5 shows the MSE cost-function values for representing errors at 20 epochs at all input SNRs. We have considered one noise source: 32-talkers babble, for example, and shown that lowest MSE is achieved at 20th epoch. At low SNRs (-5dB and 0dB) a considerable MSE is achieved at epochs greater than 16. For higher SNRs (5dB and 10dB), the MSE is greater as compared to low SNRs.

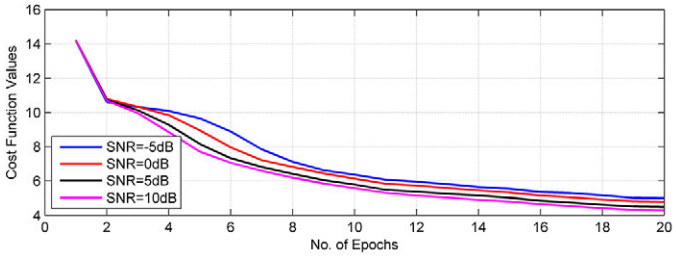

Fig. 5. Objective Cost Function at epochs.

Recently in [31], it has been shown that HIT–FA measure correlates well with human intelligibility. The term HIT indicates the percentage of correctly classified target-dominant T-F units and FA indicates the false alarm or the percentage of wrongly classified interference-dominant T-F units. A fine estimate of IBM ought to have high HIT rate and low FA rate respectively, which guides to high HIT–FA rates. We have used HIT–FA rates in our study to indicate the classification and estimation errors. Fig. 6 shows the HIT-FA rates for the estimated masks for different DNN models. It is evident that Δ+DNN$_4$ has achieved high HIT-FA rates.

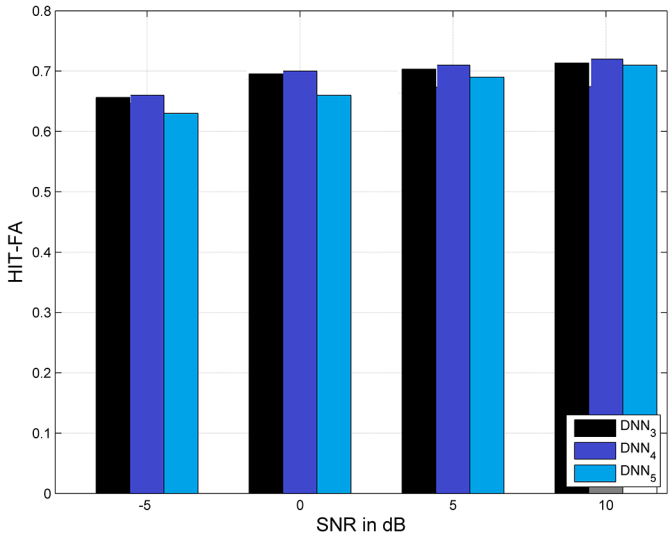

Fig. 6. HIT-FA analysis.

In speech enhancement, distortion is a vital parameter that indicates the ability of understanding a spoken enhanced speech utterance (intelligibility). The distorted utterance would lose vital speech contents, which results in loss of the speech intelligibility. Consequently, it is vital to perform enhancement of noisy speech in such a way that noise is reduced but not at the cost of intelligibility. To examine the speech distortion and residual noise, we have conducted spectrogram analysis. The spectrograms of the enhanced speech obtained with all processing methods are depicted in Fig. 7(A). The spectrograms of WF and NNDS have lost some important speech contents, hence provided less speech intelligibility as compared to DNN model Δ+DNN$_4$ which is evident in the Fig. 7(A). If we note the spectrogram of DNN, we obtained a close replica of clean speech spectrogram and important speech contents are effectively preserved. Also a low residual noise is observed in the spectrogram of DNN output speech. The time-domain waveforms of the enhanced speech utterances obtained with all the processing methods are depicted in Fig. 7(B). The waveforms of WF and NNDS have some residual noise, hence provided less segmental SNR (quality) as compared to DNN model Δ+DNN$_4$ which is evident in the Fig. 7(B). If we observe, the waveform of DNN is a close replica of clean speech waveform and important speech contents are effectively preserved. Also a low residual noise is observed in the waveform of DNN output speech.

V. Summary and Conclusions

This paper considered the restaurant noise problem for speech enhancement which is identical to real-world environments and many noise sources that concurrently degrade quality and intelligibility of a target speech. The existing studies on the speech enhancement principally focus on the presence of one noise source. However, in real-world situations, attempts are made to improve the speech quality and intelligibility of speech where many stationary and nonstationary noise sources are simultaneously mixed with target speech. To address such problem, we have used Deep Neural Networks approach and used ideal binary mask (IBM) as a binary classification method and target function during training. The mean square error (MSE) objective cost function is used during training to reduce errors. The experimental results at different input SNRs have confirmed the superiority of DNN-based multi-noise speech enhancement in terms of PESQ, SNRSeg, LLR, WSS and STOI. Our experimental results in particular noisy situations have demonstrated an average 7% improvement in speech quality as compared to noisy speech. Similarly, an average 6.5% improvement in speech intelligibility is noted during experiments. Moreover, a large improvement in terms of the SNRSeg, LLR and WSS is recorded during experiments, shown in Tables II-IV for reference. At low SNRs (-5dB) the DNN based speech enhancement in this particular noise source performed exceptionally and attained large improvements. The time-varying spectral analysis confirmed that the DNN with four hidden layers has the capacity to reduce considerable noise and the speech contents are preserved to an acceptable level of understanding. The overall analysis of DNN architecture has validated that Δ+DNN$_4$ has a great potential to deal this noise type as compared to other two competing unsupervised and supervised methods.

VI. Future Work

The greater part of the speech processing algorithms operate only with the spectral magnitude, leaving spectral phase unstructured and unexplored. With recent advancement in deep neural networks, the phase processing became more important as an innovative and emergent prospective of the DNN based speech enhancement. The authors will develop the DNN with phase estimation in future to test the speech intelligibility and quality potentials in the complex noisy environments.

References

[1] Rehr, R., & Gerkmann, T. (2018). On the importance of super-Gaussian speech priors for machine-learning based speech enhancement. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 26(2), 357-366.

[2] Saleem, N., Khattak, M. I., & Shafi, M. (2018). Unsupervised speech enhancement in low SNR environments via sparseness and temporal gradient regularization. Applied Acoustics, 141, 333-347.

[3] Saleem, N., Irfan, M., Chen, X., & Ali, M. (2018). Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 871-874). IEEE.

[4] Saleem, N., Shafi, M., Mustafa, E., & Nawaz, A. (2015). A novel binary mask estimation based on spectral subtraction gain-induced distortions for improved speech intelligibility and quality. University of Engineering and Technology Taxila. Technical Journal, 20(4), 36.

[5] Saleem, N., & Irfan, M. (2018). Noise reduction based on soft masks by incorporating SNR uncertainty in frequency domain. Circuits, Systems,

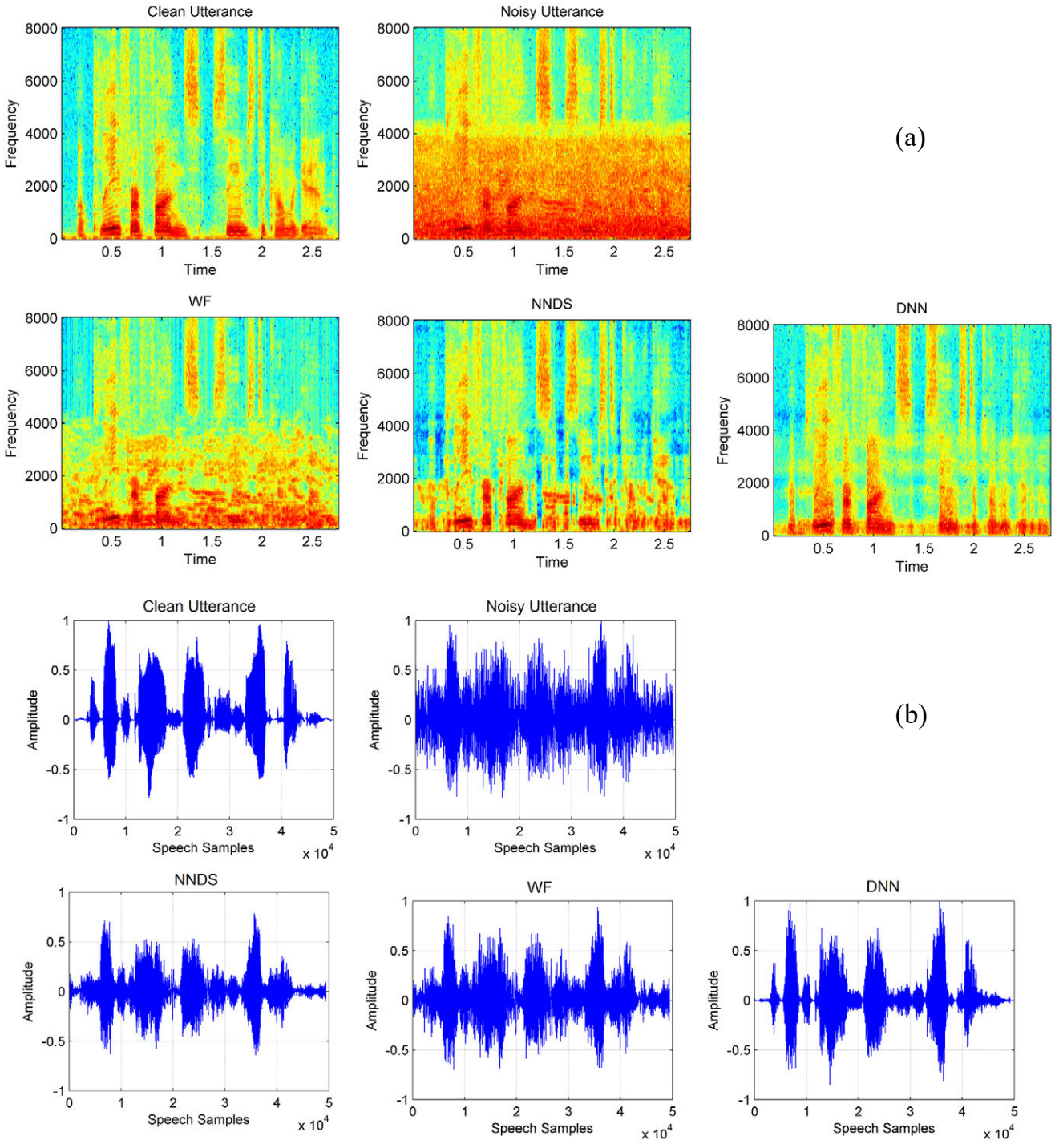

Fig. 7. Spectrogram and Waveform Analysis.

and Signal Processing, 37(6), 2591-2612.

[6] Saleem, N., & Khattak, M. I. (2018). Regularized sparse decomposition model for speech enhancement via convex distortion measure. Modern Physics Letters B, 32(22), 1850262.

[7] Saleem, N., & Tareen, T. G. (2018). Spectral Restoration Based Speech Enhancement for Robust Speaker Identification. International Journal of Interactive Multimedia and Artificial Intelligence, 5(1), 34-39.

[8] Boll, S. (1979, April). A spectral subtraction algorithm for suppression of acoustic noise in speech. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. (Vol. 4, pp. 200-203). IEEE.

[9] Scalart, P. (1996, May). Speech enhancement based on a priori signal to noise estimation. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on (Vol. 2, pp. 629-632). IEEE.

[10] Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing, 32(6), 1109-1121.

[11] Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE transactions on acoustics, speech, and signal processing, 33(2), 443-445.

[12] Ephraim, Y., & Van Trees, H. L. (1995). A signal subspace approach for speech enhancement. IEEE Transactions on speech and audio processing, 3(4), 251-266.

[13] Sigg, C. D., Dikk, T., & Buhmann, J. M. (2010, March). Speech enhancement with sparse coding in learned dictionaries. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 4758-4761). IEEE.

[14] 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.

[15] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8599-8603). IEEE.

[16] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.

[17] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters, 21(1), 65-68.

[18] 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 (TASLP), 23(1), 7-19.

[19] Kolbk, M., Tan, Z. H., Jensen, J., Kolbk, M., Tan, Z. H., & Jensen, J. (2017). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(1), 153-167.

[20] Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013, August). Speech enhancement based on deep denoising autoencoder. In Interspeech (pp. 436-440).

[21] Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(12), 1849-1858.

[22] Chen, J., Wang, Y., & Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), 1993-2002.

[23] Brown, A. M., Gaskill, S. A., Carlyon, R. P., & Williams, D. M. (1993). Acoustic distortion as a measure of frequency selectivity: relation to psychophysical equivalent rectangular bandwidth. The Journal of the Acoustical Society of America, 93(6), 3291-3297.

[24] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

[25] Klein, S., Pluim, J. P., Staring, M., & Viergever, M. A. (2009). Adaptive stochastic gradient descent optimisation for image registration. International journal of computer vision, 81(3), 227.

[26] Wager, S., Wang, S., & Liang, P. S. (2013). Dropout training as adaptive regularization. In Advances in neural information processing systems (pp. 351-359).

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

[28] Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on (Vol. 2, pp. 749-752). IEEE.

[29] Loizou, P. C. (2007). Speech enhancement: theory and practice. CRC press.

[30] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 4214-4217). IEEE.

[31] Kim, G., Lu, Y., Hu, Y., & Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. The Journal of the Acoustical Society of America, 126(3), 1486-1494.

Nasir Saleem

Engr. Nasir Saleem received the B.S degree in Telecommunication Engineering from University of Engineering and Technology, Peshawar-25000, Pakistan in 2008 and M.S degree in Electrical Engineering from CECOS University, Peshawar, Pakistan in 2012. He was a senior Lecturer at the Institute of Engineering and Technology, Gomal University, D.I.Khan-29050, Pakistan. He is now an Assistant Professor in the Department of Electrical Engineering, Gomal University, Pakistan. His research interests are in the area of digital signal processing, speech processing and speech enhancement.

Muhammad Irfan Khattak

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