Journal of Integrated
SCIENCE & TECHNOLOGY

Discriminative Autoencoder architecture for Acoustic Signal enhancement

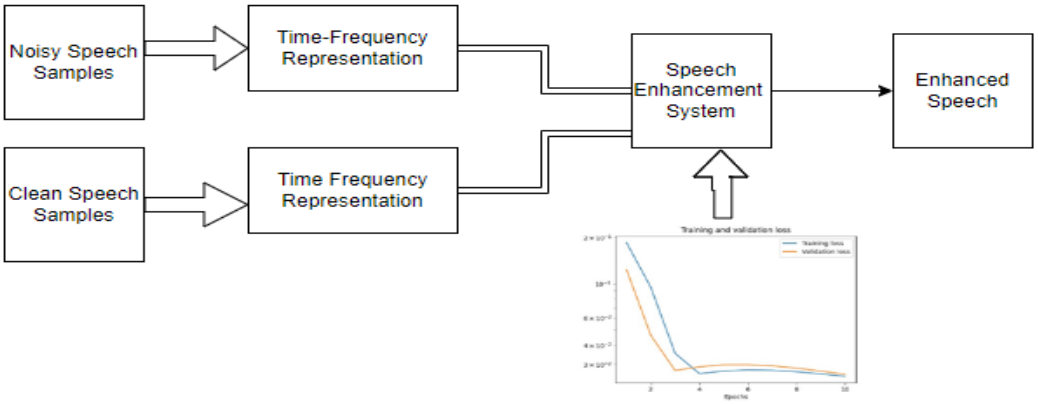
Shibani Kar,* Vishwajeet Mukherjee

Department of Electronics and Communication Engineering, Sambalpur University Institute of Information Technology, Jyoti Vihar, Burla, Odisha, India

Received on: 27-Nov-2023, Accepted and Published on: 14-Feb-2024

Article

ABSTRACT



The speech signal is an important acoustic signal. The quality of speech signal is dependent upon the surroundings of the speaker and listener of speech, sound and audio. The additive noises such as white noise and babble noise severely degrade the performance of the sound-based applications. The conventional methods for noise reduction introduce musical noises in the enhanced speech signal. The discriminative networks map the noisy speech to the clean target speech signal. In this process, the discriminative networks add unpleasant distortions to the signal. Hence, two auto encoder based discriminative approaches: Discriminative UNET model (DUNET) and Discriminative De-noising Auto encoder model (DDAE) are designed and tested with noisy speech samples available from NOIZEUS dataset. The performance of the method is compared with four baseline methods: UNET, Variational Auto encoder, Convolutional auto encoder and Pixel CNN architecture. Five evaluation indexes, PESQ, STOI, SDR, improvement in SNR, and Segmental SNR are used for the comparison of performance. The architecture provides better intelligibility and less signal distortion ratio as compared to given baseline methods.

Keywords: Discriminative network, Acoustic Signal, Auto encoder, Speech De-noising, UNET

INTRODUCTION

The speech signal is susceptible to nearby interferences. These interferences are usually additive in nature, such as babble noise, white noise. This affects the quality and intelligibility of the signal. The conventional methods used for the denoising and enhancement of signals are Minimum Mean Square Error(MMSE),[1] Wiener filtering,[2] spectral subtarction.[3] These methods introduce musical noise in the denoised speech signal and have limited performance for non-stationary noises.[4] The deep neural network is used for the classification, recognition and enhancement of image signals. The networks can learn complex features of the image signal and use these features for image classification. Nowadays, the deep neural network architecture is used for speech recognition, speech denoising and enhancement. Hence, in this work, the deep neural network based discriminative architectures are applied for the denoising of the network. The deep discriminative method maps the noisy speech to a clean speech target. In this process, unpleasant signal distortions are introduced in the enhanced speech signal.[5] Hence, a discriminative autoencoder based approach is designed for the denoising of speech signals. The first designed approach, Discriminative Denoising Autoencoder(DDAE), is the combination of a discriminator with an autoencoder and the second approach, discriminative UNET(DUNET), combines a discriminator with UNET based architecture. The comparative performance of the two methods is evaluated using five metrics: Perceptual Evaluation of Speech Quality(PESQ), Signal to Distortion Ratio (SDR), Short Time Objective Intelligibility Score(STOI), Segmental SNR(SegSNR), improvement in

*Corresponding Author: Shibani Kar
Email: shibanikar@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(5), 812.*
URN:NBN:sciencein.jist.2024.v12.812

SNR(SNR). The higher the scores, the better is the performance of the enhancement system. The segmental SNR and the improvement in SNR are time doman metrics for the evaluation of performance of enhanced systems. They work upon the enhanced speech time domain waveform. In the present work, the short time fourier transform is applied on the speech signal to extract time frequency features of the speech signal. These time frequency features are represented as two dimensional spectrograms of speech signals. Clean speech is represented as clean spectrograms and noisy speech as noisy spectrograms.The two designed discriminative approaches proposed in the paper map the noisy speech spectrograms to the clean speech spectrograms. The performance of the architectures is compared with the four baseline methods: UNET, Variational autoencoders, convolutional autoencoders and PixelCNN architecture. The objective of the paper is to apply DUNET and DDAE architecture to the speech corrupted with the babble noise at different signal to noise ratios(0 dB to 15dB). The performances are reported in terms of the evaluation metrics stated above.

Related Works

The machine learning approaches are used to solve the problem of speech denoising and enhancement.[6] The most popular network architecture is convolutional neural networks in encoder decoder form that takes noisy speech as input and provides denoised speech at the output.[7] The encoder decoder network is known as an autoencoder framework. The skip connection or residual connections provides enhanced performance in an autoencoder framework due to sharing of spatial information in between corroesponding encoder decoder layers.[8] The attention[9] concept is used with the convolutional networks and UNET networks for speech enhancement.[9–11] The time series networks used for sequential problems are applied for the denoising of speech signals. These networks comprise LSTM and recurrent neural network architectures for the enhancement of signals.[12,13] The use of encoder decoder networks work upon time domain signals and waveforms for speech enhancement is termed as WaveUNET architecture for signal enhancement.[14] The generative networks are used for mapping input signals to a known prior distribution. The decoder part of the generator then reconstructs the signal from this prior distribution. Here, the encoder decoder network act as a generator that generates a clean signal when noisy input is given to the encoder. The encoder act as an inference engine and the decoder regenerates the signal. These networks are termed as deep generative networks, such as variational autoencoder combined with non-negative matrix factorization method, generative adversarial networks, PixelCNN architecture, diffusion based generative networks for enhancement of speech.[5,15–17] Discriminative networks that are convolutional in nature are proposed for speech enhancement. They map the noisy speech to a clean speech target.[18] The performance of these enhancement methods is measured using PESQ, STOI, Segmental SNR, improvement in SNR, signal to distortion ratio.[19] The discriminative networks give unpleasant distortion in enhanced speech[5]. Hence, the paper proposes a combination of a discriminative approach with an autoencoder network. The discriminator differentiates between noisy and clean speech and aids in improving the performance of the autoencoder architecture. Two approaches are proposed: Discriminative UNET architecture(DUNET) that uses an encoder decoder architecture based upon UNET architecture.[10,20] The output of UNET architecture is given as an input to the discriminator for differentiating between the clean target. The discriminator aids the UNET network to improve the performance of DUNET to produce enhanced speech with less distortion. The second approach is Discriminative Denoising Autoencoder (DDAE) that integrates a denoising autoencoder with a discriminator to provide enhanced speech with less distortion. The network's objective is to minimize the mean square error between the denoised speech and clean target signal for the improvement of the network performance.

Description of Designed Network

The enhancement system consists of an autoencoder and a discriminator network. The noisy signal is given as an input to the network for the mapping of noisy samples to an enhanced form. The autoencoder acts as a generator. Both the generator and the discriminator are trained in a supervised manner. The autoencoder act as a generator and a discriminator is used to discriminate between the clean and noisy samples. The enhancement system uses the time frequency samples of the speech signal(i.e. spectrogram) as the feature input for the autoencoder and the discriminator. The discriminator network is a fully connected network trained with both clean and noisy speech samples. In figure 1., the spectrogram of clean and noisy speech samples are obtained and the autoencoder and discriminator are trained with these speech features. The autoencoder consists of encoder decoder networks. The encoder converts the input signal to a small dimension representation called hidden vector or latent form. The decoder takes these latent vector inputs and reconstructs the signal at the output. The autoencoder objective is to minimize the gap between the reconstructed output and clean target using a loss function "mean square error" and to update the gradients of the encoder and decoder using backpropoagation method. In this process the encoder learns to map accurately the high dmensional input to a low dimensional latent vector and the decoder uses this low dimensional input to reconstruct the signal. The discriminator is trained with the same set of inputs to map the noisy speech to clean ones. The speech enhancement system model is created by applying the generator's output as an input to the discriminator. The output of discriminator is treated as model's output. The model is trained with both noisy and clean speech samples available from NOIZEUS dataset.[17,21]

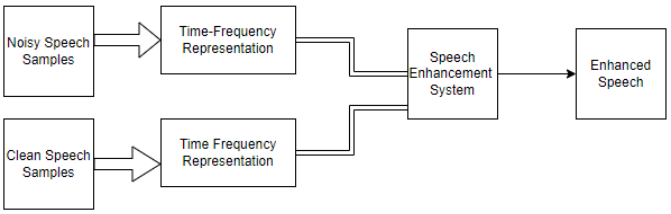


Figure 1. Discriminative Speech Enhancement System

The paper proposes two different architectures for autoencoder. In the DUNET architecture, the UNET[20] architecture is implemented as autoencoder that encodes the noisy input speech and reconstructs the enhanced speech. The DDAE architecture uses the denoising autoencoder architecture for the encoding of noisy input speech signals and for the reconstruction of speech signals. The two models are combined with a discriminator to implement the speech enhancement model.

EXPERIMENTAL SET UP

The discriminator network consist of three dense layers having units 512,256 and 1 respectively. The network is a fully connected network. The two dense layers having units 512 and 256 use "LeakyReLU" as an activation function. The last layer use "sigmoid" as an activation function for the discriminative action of the network to classify between the valid and fake samples i.e. between clean and noisy samples. The shape of the input applied to the discriminator is selected as (128,128,1). The model parameters are optimized using "ADAM" optimizer and "Mean Square Error" as the loss function. The model is trained for 10 epochs. The autoencoder act as a generative network. The discriminator model is similar to the two architectures. The DUNET architecture consist of an autoencoder and a discriminator network. The autoencoder consist of a UNET architecture[20] consisting of encoder and decoder networks where the layers of encoder are connected to decoder using the skip connections for the sharing of spatial information. The UNET based autoencoder architecture is a fully connected convolutional architecture. The encoder network uses conv2D layer, downsamplers and Leaky ReLU activation function. The decoder layer uses upsamplers , conv2D layers and 'LeakyReLU' activation function. The last layer is a fully connected layer that uses 'tanh' as an activation function. The DDAE architecture consist of an autoencoder. The encoder model consist of three dense layers having units 512,512 and 2 respectively. The "LeakyReLU" is used as an activation layer. The decoder layer consist of three dense layers having units 512,512 and 2 respectively. The "LeakyReLU" activation layer is used for first two dense layers output and "tanh" activation layer is used for the third dense layer. Both the encoder and decoder layer are trained with "adam" optimizer and "mean square error" loss function. The input size for encoder model is (128,128,1). The encoder layer compresses the input to latent dimension(value=2). The decoder input size is similar to the latent dimension(value=2). The decoder reconstructs the signal from the latent dimension and reshapes as an output having size(128,128,1). The autoencoder is trained in supervised manner with both clean and noisy samples for 10 epochs. To construct the discriminative autoencoder architecture as enhancement model, the autoencoder(i.e. generator) output is applied as an input to the discriminator and the discriminator output is selected as the output of DUNET and DDAE methods. The DUNET and DDAE architecture are trained for 80 epochs. The input size for both the architectures is (128,128,1). The loss function graph for the autoencoder, discriminator, DUNET, DDAE architecture is given below:

Figure 2. Training Loss for Discriminator (DUNET)

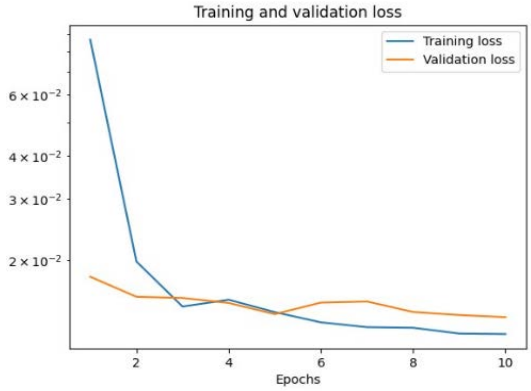


Figure 3: Training Loss for Generator (DUNET)

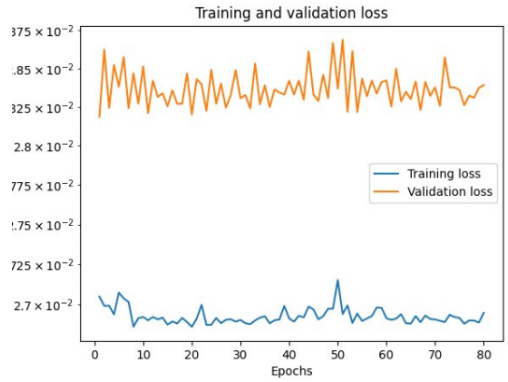


Figure 4: Training Loss for DUNET architecture

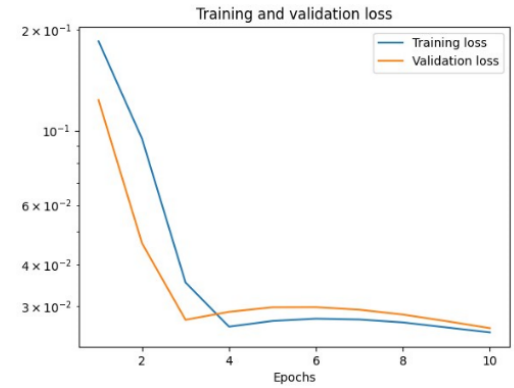


Figure 5: Training Loss for Discriminator(DDAE)

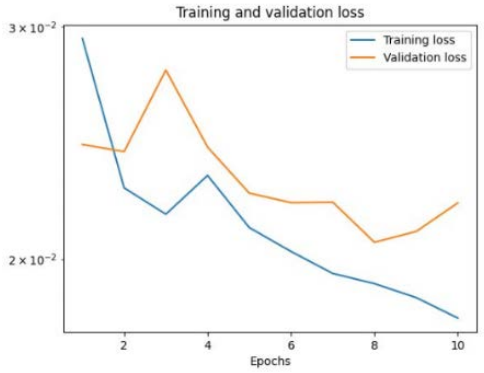


Figure 6: Training Loss for Generator(DDAE)

Figure 7: Training Loss for DDAE architecture

RESULTS

The performance of the two designed enhancement models: DUNET and DDAE for the de noising and enhancement of signal corrupted with babble noise is evaluated in terms of five objective metrics: (1)Perceptual Evaluation of Speech Quality (PESQ), (2) Short Time Objective Intelligibility (STOI), (3) Segmental SNR (Seg SNR), (4) Improvement in Signal to Noise Ratio, (5) Signal to Distortion Ratio (SDR)[21]. The performance is compared with four baseline methods (Table 1 to Table 5): UNET, VAE, Auto encoder, and Pixel CNN. The performance of six models was compared with the performance of the models proposed by Alamdari et. al.,2021 [22], Liu et.al,2020[23], Bhat et.al.,2019 [24] Ephraim et. al.[1] in Table 6, 7-8, 9 and 10 respectively. The author Alamdari et.al. has compared the performances of the three models Wiener filter, Supervised Speech de-noising (SSD) and Hybrid Speech De-noising (HSD) for the removal of babble noise from speech signal. The comparison is performed with three parameters: PESQ, STOI, Segmental SNR. The comparative performance is shown in the form of comparative chart of all nine models from figure 13 to 15. The PESQ performances are compared with models proposed by Liu et. al, Bhat et. al. and Ephraim et. al. and represented in charts from figure 16 to 18.

The PESQ performance of the designed models are given in Table1. The observations represents the perceptual evaluation of speech quality for the signal corrupted with babble noise at SNR conditions: 0dB to 15dB. The average PESQ measure for DUNET is 1.45 and DDAE is 1.43 as given in Table 6. The signal perceptual quality performance of DUNET and DDAE is satisfactory at given SNR conditions. The PESQ performance of the models is compared with the models proposed by Alamdari et. al,2021[22], Bhat et. al.,2019[24] Liu et. al., 2020[23] and Ephraim et. al.[1]. The PESQ performance of DUNET, DDAE, UNET, AAE, VAE and Pixel CNN is satisfactory as compared to the models proposed by Alamdari et. al.[22] and Bhat et. al.[24]. The PESQ performance is compared with single channel models proposed by Liu et. al[23] and found to be satisfactory.

Table 1: PESQ Performance

PESQ						
SNR Level	DUNET	DDAE	UNET	VAE	AAE	Pixel CNN
0db	1.30	1.28	1.31	1.28	1.37	1.31
5db	1.38	1.40	1.45	1.35	1.49	1.37
10db	1.51	1.48	1.58	1.45	1.70	1.52
15db	1.62	1.59	1.77	1.61	1.81	1.77
Avg	1.45	1.43	1.52	1.42	1.59	1.49

The short time intelligibility score for the DUNET and DDAE model is compared with the baselines in Table 2 for the speech signal corrupted with babble noise at SNR conditions(0dB to 15dB). The DUNET and DDAE architecture enhanced speech signal gives 0.64 intelligibility scores at 0db and 0.91 score at 15dB. This shows the model provides better intelligibility scores at low SNR. In table 6, the DUNET and DDAE model gives better STOI scores as compared to Wiener, SSD and HSD models proposed by Alamdari et. al.,2021[22]. DUNET and PixelCNN gives the highest intelligibility as compared to all models and the models proposed by Alamdari et. al.,2021[22]. The models give better intelligibility performance compared to IEM(L), IEM(R), FCN-251 and DDAE models propsoed by Liu et. al.,2020.[23]

Table 2: STOI Performance

STOI						
SNR level	DUNET	DDAE	UNET	VAE	AAE	Pixel CNN
0dB	0.64	0.61	0.62	0.57	0.62	0.64
5dB	0.75	0.73	0.72	0.71	0.76	0.74
10dB	0.85	0.81	0.82	0.80	0.85	0.84
15dB	0.91	0.90	0.88	0.86	0.88	0.91
Avg	0.78	0.76	0.76	0.73	0.77	0.78

The segmental SNR score for the DUNET and DDAE architecture is compared with baseline methods in Table 3. The UNET method provides better segmental SNR scores at all conditions. In Table 6, UNET gives better SSNR scores as compared to models: Wiener, SSD, HSD proposed by Alamdari et. al.,2021 [22] and logmmse method proposed by Ephraim et. al.

The improvement in SNR performance is given in Table 4. The DUNET and DDAE gives SNR improvement score: 2.42 and 2.41 at 0dB, whereas DUNET gives 0.29 score at 15dB.

The signal to distortion ratio score is given in Table 5. The DUNET architecture gives 1.16 SDR score at low SNRs.

Table 3: Segmental SNR Performance

SSNR						
SNR Level	DUNET	DDAE	UNET	VAE	AAE	Pixel CNN
0dB	-4.47	-3.57	0.06	-4.16	-1.73	-4.49
5dB	-2.75	-2.12	1.38	-2.37	0.22	-2.89
10dB	-1.15	-0.82	2.40	-1.02	1.64	-1.26
15dB	0.29	0.58	3.11	0.22	2.61	0.26
Avg	-2.02	-1.48	1.73	-1.83	0.68	-2.09

Table 4: Improvement in SNR Performance

SNR						
Level	**DUNET**	**DDAE**	**UNET**	**VAE**	**AAE**	**Pixel CNN**
0dB	2.42	2.41	3.56	2.24	3.05	2.09
5dB	-1.91	-1.52	-0.95	-2.03	-1.33	-0.77
10dB	-6.69	-5.88	-6.40	-6.87	-6.21	-4.43
15dB	0.29	-9.86	-12.84	-11.77	-12.03	-8.77
Avg	-1.47	-3.71	-4.15	-4.60	-4.13	-2.97

Table 5: Signal to Distortion Ratio Performance

SDR						
Level	**DUNET**	**DDAE**	**UNET**	**VAE**	**AAE**	**Pixel CNN**
0dB	1.16	0.15	-5.29	0.64	-0.33	1.18
5dB	2.05	1.49	4.32	1.48	0.51	1.79
10dB	3.00	2.21	-6.40	2.22	1.87	2.43
15dB	3.99	3.50	-1.31	2.69	2.15	4.04
Avg	2.55	1.83	-2.17	1.75	1.05	2.36

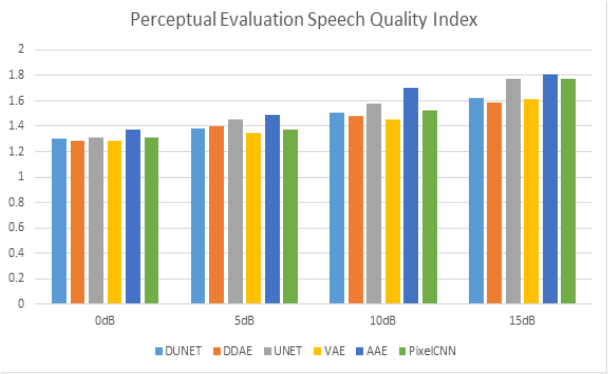


Figure 8 : Chart representing PESQ scores

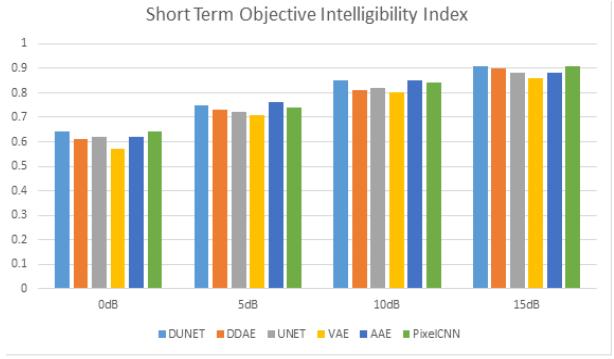


Figure 9 : Chart representing STOI scores

The DUNET and DDAE SDR score is satisfacory at higher SNR compared to other baseline methods and logmmse method proposed by Ephraim et.al..

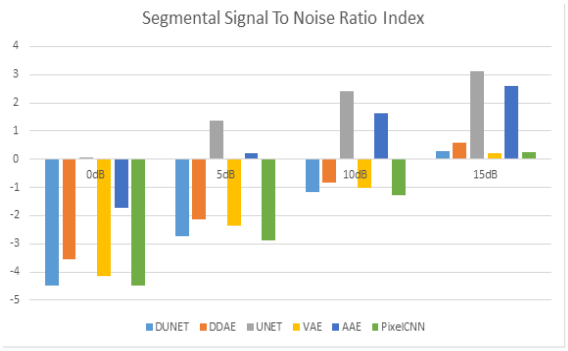


Figure 10: Chart representing SegSNR Score

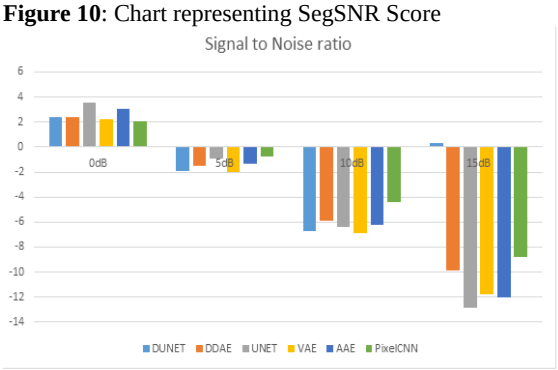


Figure 11.Chart representing SNR score

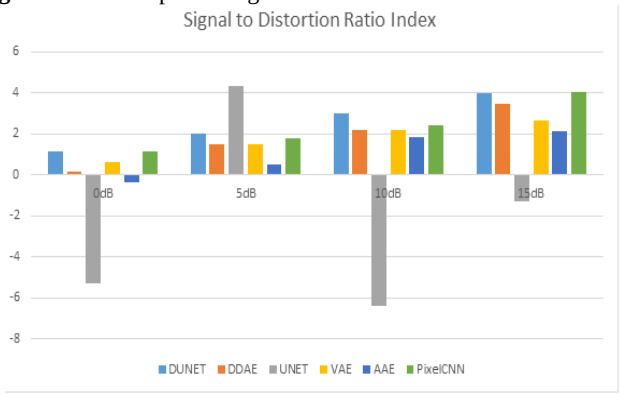


Figure 12. Chart representing Signal to Distortion Ratio

Table 6. Comparative Performance with models proposed by Alamdari et. al.[22]

Models	**PESQ**	**STOI**	**SegSNR**
DUNET	1.45	0.78	-2.02
DDAE	1.43	0.76	-1.48
UNET	1.52	0.76	1.73
VAE	1.42	0.73	-1.83
AAE	1.59	0.77	0.68
PixelCNN	1.49	0.78	-2.09
Wiener[a,22]	1.39	0.5	-2.88
SSD[a,22]	1.43	0.62	-0.96
HSD[a,22]	1.48	0.67	0.18

a: models proposed by Alamdari et.al.[22]

Table 7: Comparative Performance with models by Liu et. al.[23]

Methods	PESQ	STOI
DUNET	1.45	0.78
DDAE	1.43	0.76
UNET	1.52	0.76
VAE	1.42	0.73
AAE	1.59	0.77
PixelCNN	1.49	0.78
IEM(L)[23,b]	1.14	0.69
IEM(R)[23b]	1.10	0.69
FCN-55[23,b]	1.31	0.80
DCN-54[23,b]	1.36	0.81
FCN-251[23b]	1.17	0.72
Sinc FCN[23b]	1.47	0.84

b: models proposed by Liu et. al.[23]

Table 8: Comparative Performance with models by Liu et. al.[23]

Methods	PESQ	STOI
DUNET	1.45	0.78
DDAE	1.43	0.76
UNET	1.52	0.76
VAE	1.42	0.73
AAE	1.59	0.77
Pixel CNN	1.49	0.78
SDFCN(L)[23b]	1.63	0.86
SDFCN(R)[23b]	1.59	0.82
SDFCN[23b]	1.64	0.88
DFCN[23b]	1.56	0.86
FCN[23b]	1.44	0.83
DDAE[23b]	1.93	0.77
rSDFCN[23b]	1.98	0.89

b: models proposed by Liu et. al.[23]

Table 9: Comparative Performance with models proposed by Bhat et. al.[24].

Models	PESQ
DUNET	1.45
DDAE	1.43
UNET	1.52
VAE	1.42
AAE	1.59
Pixel CNN	1.49
Proposed CNN[24c]	1.26

c: models proposed in Bhat et. al. 2019[24]

Table 10: Comparative Performance with LOGMMSE method proposed by Ephraim et. al.,1985[1]

Models	PESQ	STOI	SegSNR	Improvement in SNR	SDR
DUNET	1.45	0.78	-2.02	-1.47	2.55
DDAE	1.43	0.76	-1.48	-3.71	1.83
UNET	1.52	0.76	1.73	-4.15	-2.17
VAE	1.42	0.73	-1.83	-4.60	1.75
AAE	1.59	0.77	0.68	-4.13	1.05
PixelCNN	1.49	0.78	-2.09	-2.97	2.36
LogMMSE[1]	2.05	0.77	1.71	2.79	2.06

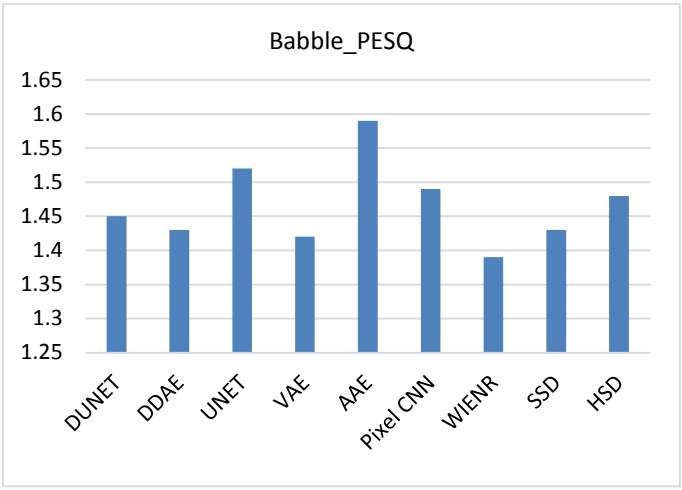


Figure 13: Comparative Performance for PESQ for Babble Noise with models proposed by Alamdari et. al.,2021[22]

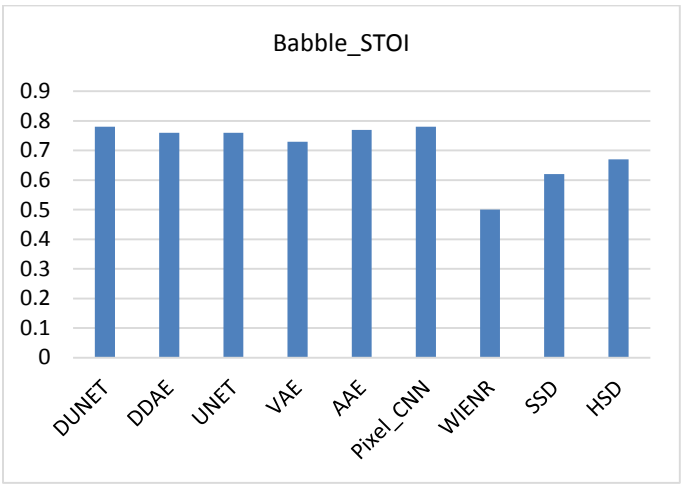


Figure 14: Comparative Performance for STOI for Babble Noise with models proposed by Alamdari et. al.,2021[22]

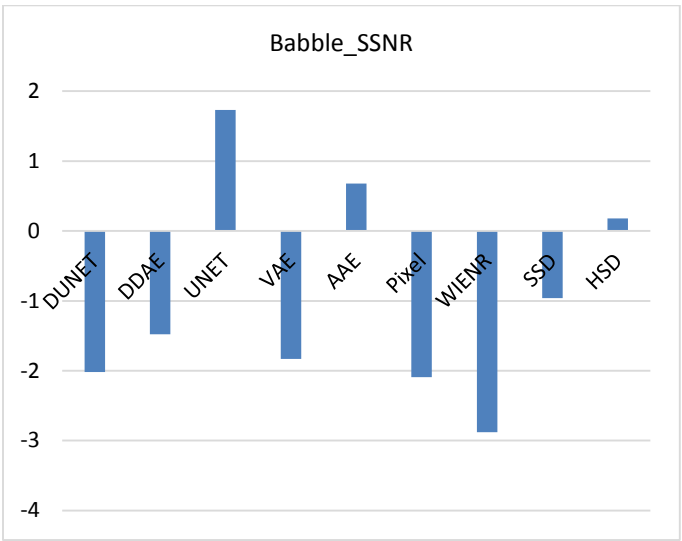


Figure 15: Comparative Performance for Segmental SNR for Babble Noise with models proposed by Alamdari et.al.,2021[22]

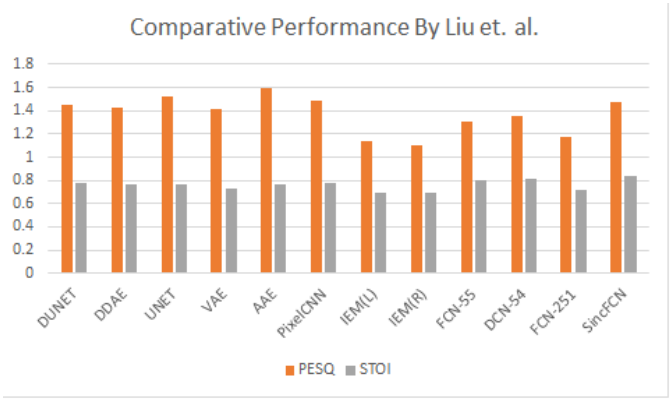


Figure 16: Comparative Performance with models proposed by Liu et. al,2020.[23]

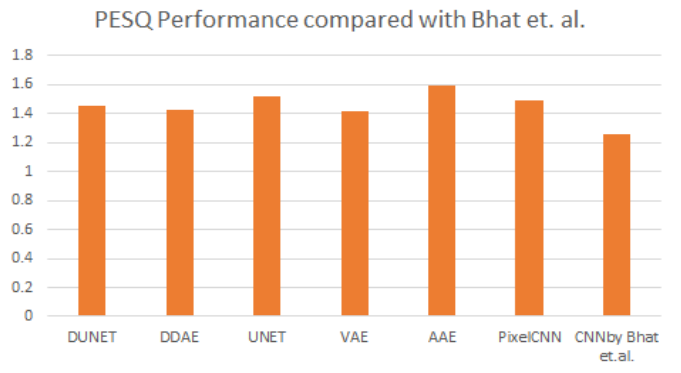


Figure 17: Comparative Performance with models proposed by Bhat et. al.,2019[24]

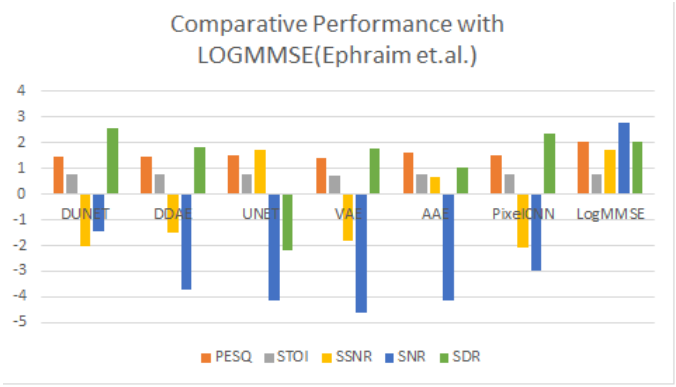


Figure 18: Comparative Performance with LOGMMSE proposed by Ephraim et. al. , 1985[1]

Conclusion

We have designed two discriminative models: DUNET and DDAE for the enhancement of speech signal corrupted with babble noises at SNRs 0dB, 5dB, 10dB and 15dB. The performance of the two methods is compared with four baseline methods: UNET, VAE, Auto encoder, Pixel CNN architecture. The baseline methods are implemented, trained and tested on the same speech samples. The performances are compared w.r.t five objective metrics: PESQ, STOI, and Improvement in SNR, Segmental SNR, and Signal to Distortion Ratio.

The PESQ measure for DUNET is 1.45 and DDAE is 1.43. The PESQ performance of the designed models are better than the Wiener Filter proposed by Alamdari et. al.[22]. The PESQ performance of the two models is similar to SSD and HSD models proposed by Alamdari et.al.[22]. The PESQ performance of models is compared with single channel models proposed by Liu et. al.[23] and Bhat et. al.[24] The PESQ performance is better than the models proposed by Liu et. al. and Bhat et. al. The STOI score for the DUNET and DDAE architecture is 60 percent at 0dB and 90 percent at high SNR. This represents the improvement in intelligibility of the signals after enhancement. The STOI scores are compared with model proposed by Alamdari et. al. [22] and the intelligibility performance of DUNET and DDAE is better than Wiener, SSD and HSD models. The intelligibility performance is better than IEM model proposed by Liu et. al. and similar to the model proposed by Ephraim et. al.

The segmental SNR score for the two architectures improves at 15dB SNR. The UNET based model gives better segmental SNR score in comparison with Wiener, SSD and HSD models proposed by Alamdari et. al. and LogMMSE method proposed by Ephraim et. al.. The improvement in SNR score is good at 0dB SNR and the DUNET gives better performance at 15dB. The DUNET and DDAE represents better signal to distortion ratio. The SDR performance of DUNET(2.55) and PixelCNN(2.36) is better than LogMMSE method(2.06) proposed by Ephraim et. al. This represents less distortion after signal enhancement. The performance analysis demonstrates the efficiency of the DUNET, DDAE, UNET, AAE and PixelCNN architectures for the de-noising and enhancement of signal corrupted by babble noise.

Conflict of Interest Statement

Authors declare that there is no financial or academic conflict of interest.

References and Notes

1. Y. Ephraim, D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, 33 (2), 443–445.
2. D. Ribas, A. Miguel, A. Ortega, E. Lleida. Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement. *Appl. Sci.* **2022**, 12 (18), 9000.
3. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, 27 (2), 113–120.
4. Y. Hu, P.C. Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *J. Acoust. Soc. Am.* **2007**, 122 (3), 1777–1786.
5. J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Gerkmann. Speech Enhancement and Dereverberation With Diffusion-Based Generative Models. *IEEEACM Trans. Audio Speech Lang. Process.* **2023**, 31, 2351–2364.
6. D. Hepsiba, J. Justin. Role of Deep Neural Network in Speech Enhancement: A Review. In *Artificial Intelligence*; Hemanth, J., Silva, T., Karunananda, A., Eds.; Communications in Computer and Information Science; Springer Singapore, Singapore, **2019**; Vol. 890, pp 103–112.
7. S.R. Park, J. Lee. A Fully Convolutional Neural Network for Speech Enhancement. arXiv September 22, 2016.
8. A. Cohen-Hadria, A. Roebel, G. Peeters. Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation. In

2019 27th European Signal Processing Conference (EUSIPCO); IEEE, A Coruna, Spain, **2019**; pp 1–5.
9. Z. Zhang, S. Xu, S. Zhang, T. Qiao, S. Cao. Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing* **2021**, 453, 896–903.
10. S. Xu, Z. Zhang, M. Wang. Channel and temporal-frequency attention UNet for monaural speech enhancement. *EURASIP J. Audio Speech Music Process.* **2023**, 2023 (1), 30.
11. K.Md. Nayem, D.S. Williamson. Attention-Based Speech Enhancement Using Human Quality Perception Modeling. *IEEEACM Trans. Audio Speech Lang. Process.* **2024**, 32, 250–260.
12. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt. Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. *EURASIP J. Adv. Signal Process.* **2020**, 2020 (1), 49.
13. J. Abdulbaqi, Y. Gu, S. Chen, I. Marsic. Residual Recurrent Neural Network for Speech Enhancement. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE, Barcelona, Spain, **2020**; pp 6659–6663.
14. D. Rethage, J. Pons, X. Serra. A Wavenet for Speech Denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE Press, Calgary, AB, Canada, **2018**; pp 5069–5073.
15. Mixture of Inference Networks for VAE-Based Audio-Visual Speech Enhancement | IEEE Journals & Magazine | IEEE Xplore https://ieeexplore.ieee.org/document/9380713 (accessed Sep 22, 2023).
16. S. Pascual, J. Serrà, A. Bonafonte. Time-domain speech enhancement using generative adversarial networks. *Speech Commun.* **2019**, 114, 10–21.
17. S. Kar. Acoustic signal enhancement using autoregressive PixelCNN architecture. *J. Integr. Sci. Technol.* **2024**, 12 (3), 770–770.
18. H. Chung, E. Plourde, B. Champagne. Discriminative Training of NMF Model Based on Class Probabilities for Speech Enhancement. *IEEE Signal Process. Lett.* **2016**, 23 (4), 502–506.
19. Y. Hu, P.C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2008**, 16 (1), 229–238.
20. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F., Eds.; Lecture Notes in Computer Science; Springer International Publishing, Cham, **2015**; pp 234–241.
21. S. Kar, V. Mukherjee. Convolutional Neural Network for Removal of Environmental Noises from Acoustic Signal. In *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*; **2023**; pp 640–645.
22. N. Alamdari, A. Azarang, N. Kehtarnavaz. Improving deep speech denoising by Noisy2Noisy signal mapping. *Appl. Acoust.* **2021**, 172, 107631.
23. C.-L. Liu, S.-W. Fu, Y.-J. Li, et al. Multichannel Speech Enhancement by Raw Waveform-Mapping Using Fully Convolutional Networks. *IEEEACM Trans. Audio Speech Lang. Process.* **2020**, 28, 1888–1900.
24. G.S. Bhat, N. Shankar, C.K.A. Reddy, I.M.S. Panahi. A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. *IEEE Access* **2019**, 7, 78421–78433.
25. N. Saleem, M. Irfan, X. Chen, M. Ali. Deep Neural Network based Supervised Speech Enhancement in Speech-Babble Noise. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*; IEEE, Singapore, **2018**; pp 871–874.