Journal of Electrical Engineering, Vol. 74, No. 5, 2023, pp. 365-373

sciendo

PAPERS

Performance analysis of speech enhancement using spectral gating with U-Net

Jharna Agrawal[1] , Manish Gupta[1] , Hitendra Garg[1]

Many speech processing systems' crucial frontends include speech enhancement. Single-channel speech enhancement experiences a number of technological challenges. Due to the advent of cloud-based technology and the use of deep learning systems in big data, deep neural networks in particular have recently been seen as a potent means for complex classification and regression. In this work, spectral gating noise filter is combined with deep neural network U-Net to enhance the performance of speech enhancement network. Further, for performance analysis three distinct objective functions namely, Mean Square Error, Huber Loss and Mean Absolute Error are considered as loss functions. In addition, comparison of three different optimizers Adam, Adagrad and Stochastic Gradient Descent is presented. Proposed system is tested and evaluated on LibriSpeech and NOIZEUS datasets and compared to other state-of-the-art systems. It demonstrates that, in comparison to other state-of-the-art models, the proposed network outperformed them with PESQ scores of 2.737420 for training and 2.67857 for testing, along with better generalization ability.

Keywords: speech enhancement, spectral gating, deep neural network, U-Net, optimizers

1 Introduction

Recently, in the realm of speech processing, speech enhancement has gained a lot of popularity. Generally, speech is invariably corrupted by additive noise, echo and reverberation in real-world applications, for example, background noise from different sound sources such as other speakers, distortion in voice being delivered to the far-end audience, spoken conversation over cell phones and conference calls. Furthermore, hands-free devices necessitate the amplification or isolation of the near-end speaker's voice from intrusive speakers and ambient sounds [1]. These distortions reduce speech quality and intelligibility, particularly when the signal-to-noise ratio (SNR) is low. Speech enhancement, for attenuating acoustic interference, would greatly facilitate a variety of acoustic applications, including ASR, hearing aids, acoustic-based control, speaker identification, human-machine interaction, and mobile communications [2]. Furthermore, speech enhancement and separation are also important pre-processing measures in today's personal assistants, GPS, video game consoles, and medical dictation systems for reliable comprehension [3]. Recently, deep learning (DL) has demonstrated remarkable success for a wide range of learning tasks in multiple domains. DL has shown optimal performances in many domains such as speech enhancement [4], spoofing in e-health digital twin [5], spoofing attack on the fingerprint scanners [6], ASR via wireless sensors [7], natural language processing (NLP) [8] and acoustic noise suppression [9] in recent years. Presently, there has been a lot of study into improving speech in noisy environments. Therefore, speech enhancement has been increasingly prevalent and has received a lot of attention in the realm of speech processing.

Speech enhancement's primary purpose is to minimize the noise from noisy speech in order to recreate clean speech, thus improving SNR and the intelligibility of noise-corrupted speech [10]. Single-channel speech enhancement is difficult task with one microphone whereas multi-channel voice amplification is a successful solution where many microphones are present because it takes advantage of spatial input. Moreover, the two main categories of speech enhancement are supervised learning methods and unsupervised learning approaches. The first is additionally referred to as the traditional speech enhancement technique. This kind of approach has a low computational and hardware need and does not rely on priori speech information. As a result, its real-time performance is usually good [11]. The spectral-subtraction approach [2,12], Wiener filtering [2,13], minimum mean square error (MMSE) method [14] and subspace method [2] are examples of traditional speech enhancement algorithms. Unsupervised speech enhancement performs well in contexts with high SNR and stationary noise, but poorly in environments with low SNR and non-stationary noise, according to recent

GLA University, Mathura, India
jharna.agw@gmail.com

https://doi.org/10.2478/jee-2023-0044, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X
© This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

research [15]. Recently, several efforts have been made to apply U-Net architectures to speech enhancement tasks [16], which was first introduced for biomedical image segmentation [17]. Recently, U-Net has been trained to predict a noise-reduced version of the input signal, given a noisy version of the signal as input [16]. Furthermore, noisy speech in same domain was improved using U-Net along with discrete cosine transform (DCT) without the need of adversarial training [18]. Another application of Wave-U-Net [19] was to improve speech in the time-domain without GAN. In addition to this, researchers utilized Wave-U-Net [20] to improve speech in the time-domain with a smaller number of hidden layers. In this work, spectral gating along with U-Net-based denoising is proposed to enhance the quality of mix speech signal by removing non-speech sound. Mix speech signal is obtained from one speaker blended with different noises. Moreover, performance analysis of three optimizers with different number of epochs is carried out, in this study.

2 Speech enhancement model

2.1 Mathematical notation

In this work, for real-time applications, a mathematical model for single-channel speech enhancement is described in this subsection. Considering an audio noisy signal, $\boldsymbol{x}\in\boldsymbol{R}^T$, contains a clean speech signal $\boldsymbol{y}\in\boldsymbol{R}^T$, that is corrupted by an additive background noise $\boldsymbol{n}\in\boldsymbol{R}^T$, such that $x=y+n$. Since the speaker utterances have fixed duration for samples, T has a fixed value. The aim is to determine an enhancement function f such that $f(x) \approx y$.

Here, f is estimated using the U-Net architecture, which was initially developed for the goal of segmenting biological images and subsequently adapted for the task of speech enhancement. By taking into account that y_i is the true value, x_i is the predicted value, error is computed for deep neural network (DNN).

2.2 Proposed architecture

In this work, proposed architecture consists of spectral gating to enhance the quality of noisy speech signal along with U-Net which is a DNN. The combination of spectral gating with U-Net becomes a powerful tool to remove both types of static and random noise from speech signal. But, this combination poses following challenges:

Integration of spectral gating

The primary challenge in this work is dealing with random noise present in the speech. This requires to filter speech signal through an appropriate filter before feeding to deep neural network. Combining spectral gating with a deep neural network introduces the challenge of effectively fusing these techniques.

Algorithm complexity

Deep neural networks, such as U-Net are intricated in nature and require substantial computational resources. Issues related to training time, resource allocation, and model optimization pose difficulty.

Hyperparameter tuning

Neural networks often involve a variety of hyperparameters that need careful tuning to achieve optimal performance.

Generalization and overfitting

Generally, models are overfit with a small dataset but by training the model with so many samples and iterations tasks are overfitted. Therefore, to overcome overfitting challenge a dropout layer has been implemented in the proposed network. Furthermore, to get a generalized model an adequate amount of varied dataset is used for network training.

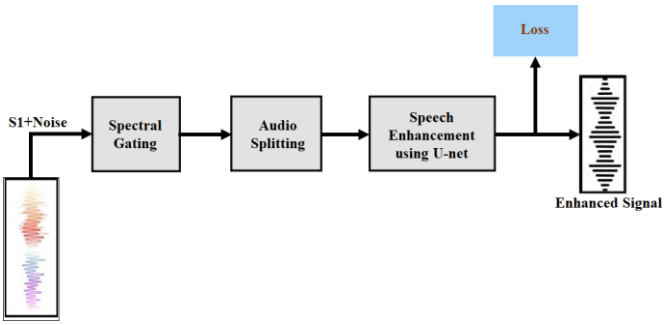


Fig. 1. Proposed architecture for speech enhancement

Figure 1 depicts that initially, a noisy speech signal containing speech and the noise is provided as input to spectral gating noise filter which is a unique filter. It removes static and some random noise added in the signal. Further splitting of audio signal is carried out to break the signal into small chunks of fixed length. Finally, to enhance the quality of speech signal DNN is used which removes rare noise. Here, U-Net is considered for this. It is evident that DNN attenuates the rare noise signals, which is denoted as a loss function. Loss function is evaluated by comparing predictions Y' from U-Net and true values Y. Based on this, a loss score is generated which is considered the objective function and serves as input to the optimizers for updating the weights for every epoch.

Spectral gating

Spectral gating is a type of noise gate. A central frequency and bandwidth parameters are used partitioning the incoming signal into its above and below frequency ranges. The threshold gates stated above are set using a noise gate by the spectral gating technique. The open gate is another term for the lower gate since the signal that crosses above the lower gate threshold will be captured. The higher gate is known as a closed gate if the signal reaches the designated upper threshold, at which point the gate closes and the signal ceases to be recorded. As a result, the open and close gate thresholds must be adjusted over time. The signal between the two gates is carefully taken into account, and the signal above or below is regarded as noise. Thus, just the necessary or noise-reduced audio will be available at the output [16].

U-Net architecture

It is a U-shaped encoder-decoder network architecture which is collectively built by joining several convolution layers followed by other necessary layers. It is comprised of an expanding and contracting path. The path of contraction follows the typical design of a CNN, for example. It consists of two 3×3 recurrent convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU). Similar to VGG and ResNet, encoder is a pre-trained classification network where convolution blocks are typically followed by maxpool downsampling. The decoder uses upsampling followed by concatenation, and at last standard convolution processes. The audio signal vector feature space has been downsampled to a fixed size and then upsampled to the same size as provided in the first layer. One important characteristic of this DNN is that it has very few parameters [21].

There are certain factors that affect the performance of the proposed model. As it has been shown in Fig. 1, the input to the model is a noisy speech signal, which is first processed by the spectral gating noise filter to remove static and some random noise. Then, in the subsequent stage, known as Audio Splitting, the audio signal is split into small chunks of fixed length. Random noise and variation in length of the chunks are the main factors that affect the performance of the proposed speech enhancement system.

3 Experimental configuration

For experimental purpose, based on PyTorch, SpeechBrain [22] toolkit is used. SpeechBrain is an open-source, all-inclusive speech toolkit, whereas, PyTorch is an open-source machine learning library. Google Colab is used to train the model.

3.1 Dataset

A substantial corpus of read English speech, the LibriSpeech dataset [23], is around 1000 hours long and is typically sampled at 16 kHz. Data are taken from audio books on LibriVox for randomly chosen speakers. The dataset used in this study for speaker 1 (S1) is 2.5 hours long and sampled at 16 kHz. Further, chunks of 2 seconds in length are created, producing 4515 files altogether. In order to train the speech enhancement network from the chunks of speaker 1, the original dataset is processed to create noisy signals. Mix signal is blended at 0 dB level with noise signals from the NOIZEUS database [24] which contains an airport, babble (crowd of people), automobile, exhibition hall and restaurant noise files. The noise files include recordings from various places to give a glimpse of different real-time noise.

3.2 U-net implementation summary

Generally, U-Net architecture has ReLU but in this work, U-Net has leaky ReLU activation layer after each convolution layer. Equation (1) gives the function for leaky ReLU as below:

$$f(x) = \max(0.01\,x, x) \tag{1}$$

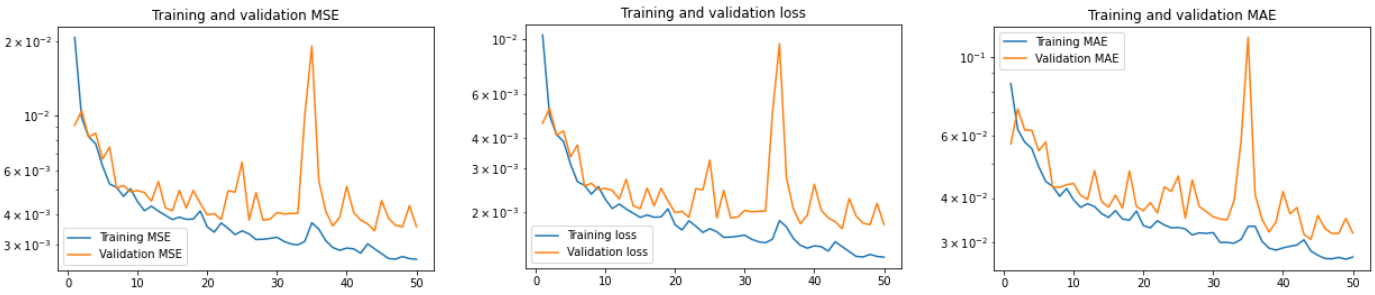
Leaky ReLU provides a very little linear component of x which is 0.01 to negative inputs in order to solve the issue of zero gradient for negative values. The ReLU function's range is widened with the help of the leak.

Further, a 2×2 maximum pooling operation with stride 2 for downsampling is considered. To avoid overfitting and enhance model performance overall, dropout layer is used. The size at the first convolution layer is taken as 128×128. The audio signal vector feature space has been downsampled to 8 and then upsampled to the same size as provided in the first layer. The number of feature channels doubles with each step of downsampling. An upsampling of the feature map, a 2×2 convolution (“up-convolution”) that divides the set of feature channels in half, a concatenation with the similarly cropped feature map from the contracting path, and two 3×3 convolutions, each accompanied by a Leaky ReLU, are all included in each step along the expanding path. Cropping is necessary since every convolution result in the loss of boundary pixels. Further, final layer uses a 1×1 convolution to transfer signal vector to the appropriate output by using tanh activation function. The final layer is settled with Tanh activation function whereas the other layers are followed by Leaky ReLU activation layer. Network is trained by taking 1000 samples, 10 epochs. Batch size is considered as 10. Network summary is given below:

- Padding type: Unpadded
- Activation function: Leaky ReLU (in hidden layers) and tanh (in output layer).
- Optimizer: The experiment uses three different optimizers, each one at a different execution period. The optimizers that have been considered are Adam, Adagrad, and Stochastic Gradient Descent (SGD), with different numbers of epochs as an experimental procedure. The execution took place with 50, 100, and 150 numbers of epochs for result formulation.
- Objective function: For training error and Testing error three objective functions MSE, Huber loss and MAE are used for continuous result generation at each iteration.
- Evaluation metrics: For the evaluation purpose the compile function has been treated with two different evaluation measures while training and testing which are perceptual evaluation of speech quality (PESQ) and STOI metrics [25].

The signal-to-noise ratio (SNR), which can be used with any signal, is the most used technique for objective evaluation for any speech enhancement model. A more specialized speech evaluation method is required to provide performance measurement for the speech signal that is more pertinent. Perceptual Evaluation of Speech Quality (PESQ) was created for this purpose by the International Telecommunication Union, Telecommunication Standardisation sector (ITU-T) in its P.862 Recommendation [26]. PESQ, one of the common metrics connected to human perception, has been shown to have a strong association with the quality ratings given by humans. A degraded audio sample's subjective opinion scores are predicted by the PESQ Algorithm. PESQ provides a score ranging from 4.5 to – 0.5. Higher scores denote higher quality.

Further, the evaluation of intelligence is necessary in speech enhancement models in order to interpret speech signals that have had their quality decreased due to additive noise and single- or multi-channel noise reduction. Short-Time Objective Intelligibility (STOI) measures objective machine-driven intelligibility and has a value range of 0 to 1. Its goal is to assess noise-reduction algorithms. STOI is a measure of intelligibility. STOI-measure is defined as a function of the clean and degraded speech signals. The higher STOI score indicates the better intelligibility.



(a) (b) (c)

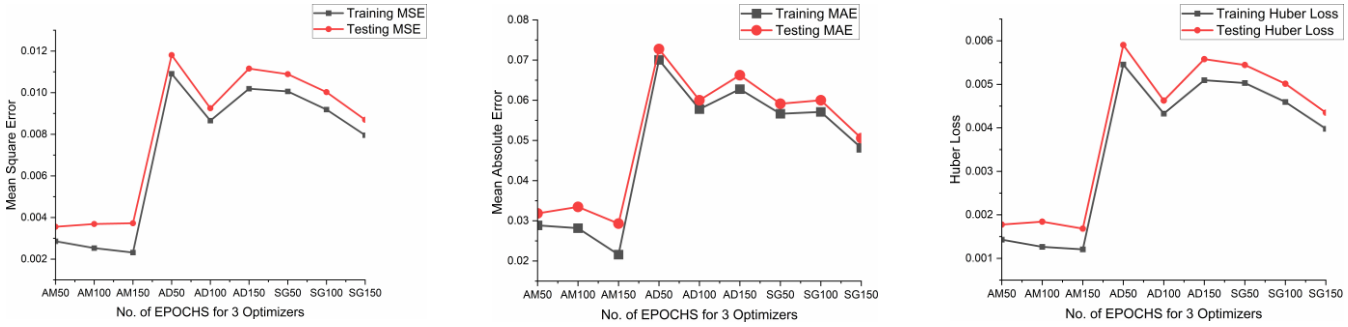
Fig. 2. Simulation results of Adam optimizer for 50 epochs using objective function: (a) MSE, (b) Huber Loss, (c) MAE

4 Result and discussion

For speech enhancement network, the experiment uses three different optimizers, each one at a different execution period. Further, comparative analysis of three objective functions for U-Net model with three optimizers is performed. Figure 2 shows simulation results for Adam optimizers at 50 epochs with three objective functions MSE, Huber Loss and MAE, in Figs. 2a, b and c, respectively.

Figure 3a compares the performance of Adam, Adagrad and SGD for training a speech enhancement model for different numbers of epochs. Increasing the number of epochs means training the model for more iterations, allowing it to potentially learn more complex relationships in the data. The evaluation metric considered is MSE. A lower Training MSE value indicates a better fit of the model to the training data. During simulation it is found that the Training MSE values decrease on increasing number of epochs from 0.002858 (for Adam with 50 epochs) to 0.002313 (for Adam with 150 epochs). Testing MSE, on the other hand, measures the average squared difference between the predicted values and the actual values of the target variable on a separate test dataset, which is not used during the training process. A lower Testing MSE value indicates a better generalization performance of the model to unseen data which is obtained at epoch 50. It is observed from simulation that Testing MSE values are increased slightly on increasing the no. of epochs which shows model is overfit for objective function MSE as the Testing MSE value is increasing while the Training MSE value is decreasing. Hence, a good model should have a low Testing MSE, while avoiding overfitting by keeping the Training MSE and Testing MSE values close to each other.

Furthermore, Fig. 3b shows the results of running different optimization algorithms (Adam, Adagrad, SGD) for different numbers of epochs (50, 100, and 150) with mean absolute error (MAE) for both training and testing datasets. As it is clear from Fig. 3b the Training MAE values decrease on increasing number of epochs from 0.028876 (for Adam with 50 epochs) to 0.021579 (for Adam with 150 epochs). It indicates that the model is improving its predictions with more training. On the other hand, the testing MAE scores initially increase slightly, when epochs are increased from 50 to 100 which may indicate overfitting of the model to the training data. Though after 100 epochs, testing MAE values are decreasing on increasing epochs to 150. Finally, for MAE objective function lowest value is achieved at 150 epochs for testing dataset.



(a) (b) (c)

Fig. 3. Performance analysis for three optimizers: (a) using MSE, (b) using MAE, and (c) Huber Loss (AM, AD, SG stand for Adam, Adagrad and Stochastic Gradient Descent respectively).

In addition, Fig. 3c compares the performance of three optimization algorithms for training a speech enhancement model for different numbers of epochs (50, 100, and 150) for Huber loss. Huber Loss is less sensitive to outliers compared to mean squared error, making it a robust loss function. A lower Training Huber Loss value indicates a better fit of the model to the training data. The Training Huber Loss values in the graph range from 0.001429 (for Adam with 50 epochs) to 0.001207 (for Adam with 150 epochs). Testing Huber Loss, on the other hand measures the average difference between the predicted values and the actual values of the target variable on a separate test dataset, which is not used during the training process. The Testing Huber Loss values in the graph range from 0.001777 (for Adam with 50 epochs) to 0.0016829 (for Adam with 150 epochs). A lower Testing Huber Loss value indicates a better generalization performance of the model to unseen data.

Further, simulation results indicate that for Adagrad optimizer, increasing the number of epochs leads to lower data particularly from 50 epochs to 100 epochs for all three objective functions. After this when epochs are increased to 150 then a slightly increment is observed in Training and Testing values. Thus, it can be concluded that Adagrad optimizers is not showing good performance for training data which leads to towards worst performance among three optimizers. Further, SGD optimizer shows performance in between Adam and Adagrad optimizers. The training values are continuously decreasing on increasing number of epochs from 50 to 150, whereas, testing values show the same pattern as in case of Adam optimizers with high values of objective functions. It can be inferred simulation results that Adam optimizer outperforms the other two optimizers for objective function Huber Loss at 150 epochs.

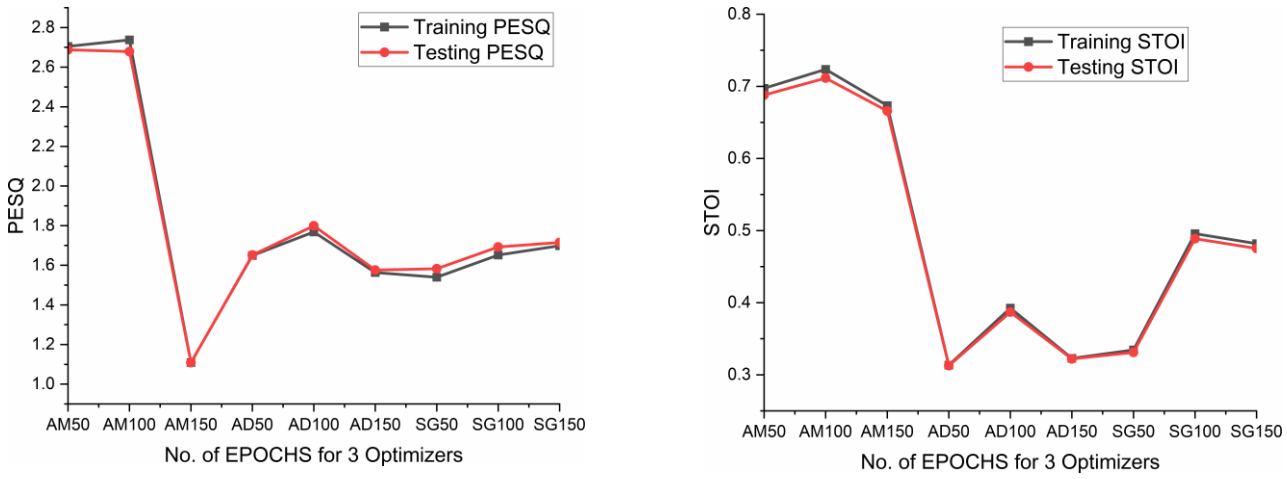
Furthermore, the performance of the optimizer can be evaluated by examining the PESQ and STOI scores of the training and testing data at different epochs. PESQ is a measure of the quality of speech signals, with higher scores indicating better quality. Moreover, new research demonstrates a strong correlation between speech intelligibility and STOI predictions of noisy speech improved using DNN-based speech enhancement systems. As a result, STOI is now the speech intelligibility estimate that is arguably most frequently used for objectively assessing how well speech enhancement systems work [27]. In addition, comparative analysis of evaluation metrics PESQ and STOI for different number of epochs of speech enhancement model using U-Net is carried out. Training simulation results shown in Fig. 4a indicate that, initially, PESQ score of the training data is improving as the number of epochs are increased up to 100 for each optimizer. This demonstrates a gain in performance is obtained on increasing number of epochs. After that there is a significant degrade in performance of Adam and Adagrad optimizers, whereas in case of SGD optimizer, it keeps on increasing up to 150 epochs. Further, in SGD optimizer PESQ score is lower than other two optimizers.

Moreover, the testing PESQ score decreases with increasing number of epochs but slightly lower than the training PESQ score. This indicates that the model performed better on the training data than on the unseen testing data, which could be a result of overfitting. The highest PESQ score was achieved by the Adam optimizer at epoch 100with a score of 2.7374. Here, it is observed that overall, Adam optimizer outperforms the other two optimizers. SGD shows better performance in comparison to Adagrad optimizer as performance is not degraded on increasing epochs beyond 100 for training. Furthermore, the lowest PESQ score was achieved by the Adam optimizer at epoch 150, with a score of 1.1087. The results suggest that the choice of optimizer and the number of epochs can have a significant impact on the performance of the model.

Additionally, Fig. 4b shows the performance of three different optimizers (Adam, Adagrad, and SGD) in terms of STOI scores during different epochs. It can be observed from simulation results that the training and testing STOI scores vary among different optimizers and at different epochs. For Adam, the training STOI score reaches a maximum value of 0.72353 at 100 epochs. However, the testing STOI score is the highest at 0.711516 at 100 epochs. There is no noticeable gap between the training and testing STOI scores which avoid overfitting and improving generalization ability. For Adagrad, the training STOI score increases gradually from 0.313359 to 0.392552 on increasing epochs up to 100 and then decreases to 0.322799 at 150 epochs. The testing STOI score is almost comparable with the training STOI score at epoch 100 which is 0.386919.

However, it decreases to 0.322008 at epoch 150. For the optimizer SGD, the training and testing STOI scores are close to each other, with the training STOI score being slightly higher than the testing STOI score. The training STOI score increases from 0.33448 at epoch 50 to 0.495665at epoch 100 then shows a slight decrease at 150 epoch with 0.481966 score which is almost comparable with the score obtained at epoch 100. Further, the testing STOI scores depict the same trend as shown by training data. It increases from 0.33104at epoch 50 to 0.488662 at epoch 100. Then at epoch 150 it obtains a slightly lower score of 0.475163, which is almost equal to the score attained at epoch 100. This shows that model is good and working well for SGD optimizer but overall, Adam optimizers supersedes the other two optimizers. Besides, it is evident from Fig. 4 that Adagrad and SGD optimizers have better performance in terms of generalization ability than Adam optimizer.

Further, Tab. 1 presents a comparison of the proposed model with other state-of-the-art models.



(a) (b)

Fig. 4. Simulation results with three optimizers for evaluation metric:
(a) PESQ, (b) STOI (AM, AD, SG stand for Adam, Adagrad and Stochastic Gradient Descent respectively).

Table 1. Comparison of proposed model with state-of-the-art models

Model	PESQ	STOI
Noisy [16]	1.97	0.916
Wiener [13]	2.22	0.914
Wave-U-Net [20]	2.40	-
Attention-wave-U-Net [28]	2.62	-
Proposed model	**2.68**	**0.71**

This shows that the proposed model renders an increased value of PESQ which shows better performance of the model, whereas STOI shows a decrement in the value. At last, it can be concluded that the best training and testing performance was achieved by Adam optimizer at 150 epochs with Huber Loss considered as objective function. Overall, using spectral gating along with U-Net model speech enhancement performance has been improved.

5 Conclusion and future scope

In this work, spectral gating with U-Net for speech enhancement is proposed for noisy signals. Based on simulation results Huber Loss shows optimal performance as an objective function with Adam optimizer. Furthermore, comparing evaluation metrics, best training and testing performance was achieved by Adam optimizer at 100 epochs The best testing performance was achieved by the Adam optimizer at 100 epochs with a PESQ of 2.678575 and a STOI of 0.711516 which shows a room for improving STOI. Simulation results on STOI scores emphasize that the SGD optimizer has shown better generalization ability compared to the other two optimizers. However, it should be noted that the best training performance does not always translate to the best testing performance, which highlights the importance of considering both training and testing scores when evaluating the performance of a model. Additionally, optimizing DNN by changing the number of layers and utilizing various optimizers might open up a wide range of new research opportunities for enhancing precision, speed, and computational costs in speech enhancement systems based on real-time application.

References

[1] Y. Masuyama, M. Togami and T. Komatsu, "Consistency-aware multi-channel speech enhancement using deep neural networks", Proceedings 2020 IEEE International Acoustics, Speech and Signal Processing Conference (ICASSP), pp. 821-825, 2020. DOI: 10.1109/ICASSP40776.2020.9053501

[2] P. C. Loizou, Speech enhancement: theory and practice, 1st ed. Boca Raton: CRC press, pp. 1-10, 2007.

[3] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A consolidated perspective on multi microphone speech enhancement and source separation", IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 25, no. 4, pp. 692-730, 2017. DOI: 10.1109/TASLP.2016.2647702

[4] C. Rascon, "Characterization of Deep Learning-Based Speech-Enhancement Techniques in Online Audio Processing Applications", Sensors, vol. 23, no. 9, p. 4394, 2023. DOI: https://doi.org/10.3390/s23094394

[5] H. Garg, B. Sharma, S. Shekhar and R. Agarwal, "Spoofing detection system for e-health digital twin using Efficient Net Convolution Neural Network", Multimedia Tools and Applications, vol. 81, no. 16, pp. 26873-26888, 2022. DOI: https://doi.org/10.1007/s11042-021-11578-5

[6] D. Agarwal and A. Bansal, "Fingerprint liveness detection through fusion of pores perspiration and texture features", J. King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4089-4098, 2020. DOI: https://doi.org/10.1016/j.jksuci.2020.10.003

[7] G. Gosztolya and T. Grósz, "Domain adaptation of deep neural networks for automatic speech recognition via wireless sensors", Journal of Electrical Engineering, vol. 67, no. 2, pp. 124-130, 2016. DOI: https://doi.org/10.1007/s11042-022-13056-y

[8] S. Shekhar, D. K. Sharma, M. M. Sufyan Beg, "Hindi Roman linguistic framework for retrieving transliteration variants using boot-strapping", *Procedia Computer Science*, vol. 125, pp. 59-67, 2018. DOI: 10.1016/j.procs.2017.12.010

[9] R. Martinek, M. Kelnar, J. Vanus, P. Bilik and J. Zidek, "A robust approach for acoustic noise suppression in speech using ANFIS", Journal of electrical engineering, vol. 66, no. 6, pp. 301-310, 2015. DOI: https://doi.org/10.2478/jee-2015-0050

[10] Y. Tsao and Y. H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement", Speech Communication, vol. 76, pp. 112-126, 2016. DOI: https://doi.org/10.1016/j.specom.2015.10.003

[11] J. Cheng, R. Liang and L. Zhao, "DNN-based speech enhancement with self-attention on feature dimension", Multimedia Tools and Applications, vol. 79, pp. 32449-32470, 2020. DOI: https://doi.org/10.1007/s11042-020-09345-z

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on acoustics, speech, and signal processing, vol. 27, no. 2, pp. 113-120, 1979. DOI: 10.1109/TASSP.1979.1163209

[13] P. Scalart, "Speech enhancement based on a priori signal to noise estimation", Proceedings 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 629-632, 1996. DOI: 10.1109/ICASSP.1996.543199

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", IEEE Trans. on acoustics, speech, and signal processing, Vol. 32, no. 6, pp. 1109-1121, 1984. DOI: 10.1109/TASSP.1984.1164453

[15] C. Lan, Y. Wang, L. Zhang, C. Liu and X. Lin, "Research on Speech Enhancement Algorithm of Multiresolution Cochleagram Based on Skip Connection Deep Neural Network", Sensors, vol. 2022, 2022. DOI: https://doi.org/10.1155/2022/5208372

[16] Z. Kang, Z. Huang and C. Lu, "Speech Enhancement Using U-Net with Compressed Sensing", App. Sciences, vol. 12, no. 9, p. 4161, 2022. DOI: https://doi.org/10.3390/app12094161

[17] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", Proceedings 2015 International Conference on Medical image computing and computer-assisted intervention, (Springer Cham.), pp. 234-241, 2015. DOI: https://doi.org/10.1007/978-3-319-24574-4_28

[18] C. Geng and L. Wang, "End-to-end speech enhancement based on discrete cosine transform", Proceedings 2020 IEEE International Artificial Intelligence and Computer Applications Conf. (ICAICA), pp. 379-383, 2020. DOI: 10.1109/ICAICA50127.2020.9182513

[19] D. Stoller, S. Ewert and S. Dixon S, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation", arXiv preprint arXiv:1806.03185, 2018. DOI: https://doi.org/10.48550/arXiv.1806.03185

[20] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net", arXiv preprint arXiv:1811.11307, 2018. DOI: https://doi.org/10.48550/arXiv.1811.11307

[21] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications", Proceedings of the IEEE, vol. 63, no. 12, pp. 1692-1716, 1975. DOI: 10.1109/PROC.1975.10036

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit", arXiv preprint arXiv:2106.04624, 2021. DOI: https://doi.org/10.48550/arXiv.2106.04624

[23] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books", Proceedings IEEE International Acoustics, Speech and Signal Processing Conference (ICASSP), pp. 5206-5210, 2015. DOI: 10.1109/ICASSP.2015.7178964

[24] P. Loizou and Y. Hu, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms", Speech Communication vol. 49, pp. 588-601, 2007. DOI: 10.1016/j.specom.2006.12.006

[25] I. T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", Rec. ITU-T. P. 862, 2001.

[26] M. Al-Akhras, K. Daqrouq and A. R. Al-Qawasmi, "Perceptual evaluation of speech enhancement," In 2010 7th International Multi-Conference on Systems, Signals and Devices, pp. 1-6, IEEE, 2010. DOI: 10.1109/SSD.2010.5585514

[27] M. Kolbaek, Z. H. Tan and J. Jensen, "On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 283-295, 2018. DOI: 10.1109/TASLP.2018.2877909

[28] R. Giri, U. Isik and A. Krishnaswamy, "Attention wave-u-net for speech enhancement", IEEE Workshop 2019 Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 249-253, 2019. DOI: 10.1109/WASPAA.2019.8937186

Jharna Agrawal received her M.Tech. degree in 2008 from MNIT, Jaipur, India. She worked as Research Investigator in the Departament of Microelectrònica i Sistemes Electrònics Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. She is currently pursuing a Ph.D. in Electronics and Communication Engineering Department at GLA University, Mathura, India. Her research interest includes VLSI signal processing and artificial intelligence.

Manish Gupta received his Ph. D. in 2015 from University Engineering College, Rajasthan Technical University, Kota, Rajasthan. He is presently working as Associate Professor in Department of Electronics and Communication Engineering, GLA University, Mathura, India. He has more than 30 research papers in the international journals/conference of repute. His research areas are image processing, and signal processing.

Hitendra Garg did his Ph.D. (CSE) from Motilal Nehru National Institute of Technology, Allahabad and Masters (Software Systems) from BITS-Pilani. He is presently working as Professor in Department of Computer Engineering and Applications, GLA University, Mathura, India. He has total experience of more than 20 years in the field of academics/ research. He has more than 50 research papers in the international journals/conference of repute. His research areas are image processing, cryptography, 3D data processing, data security.

Received 26 July 2023