Comparative Analysis of Pretrained Models for Speech Enhancement in Noisy Environments

Cheegiti Mahesh, Runkana Durga Prasad, Epanagandla Asha Bibi, Abhinav Dayal(✉), and Sridevi Bonthu

Department of CSE, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
{21pa1a0528,21pa1a05e8,21pa1a0544,abhinav.dayal,
sridevi.b}@vishnu.edu.in

Abstract. Speech Enhancement is the set of techniques and algorithms aimed at enhancing the overall quality of speech signals across diverse conditions both qualitatively and quantitatively. Speech enhancement aims to enhance voice signals whose quality has been diminished by various kinds of noise or distortion. Different techniques were adopted in previous years. Researchers have started working with Machine Learning techniques recently, prior to which they have followed traditional methods like Wiener Filtering, Spectral Subtraction, etc. The advancement of machine learning techniques day by day has laid the path for our work. Our work is to investigate the performance of three models viz., ESPNet-SE, SpeechBrain MetricGAN+ and SpeechBrain SepFormer models on a mixed dataset namely *VoiceBank* and *Demand*, which has added noise on clean signals. Among all the models, SpeechBrain MetricGAN+ performed well by approximately **30.05%** on ESPNet-SE and **10.29%** on SpeechBrain SepFormer models. Trained models are publicly available.

Keywords: Speech Enhancement · ESPNet-SE · Generative Adversarial Network · SepFormer · SpeechBrain MetricGAN+ · SpeechBrain SepFormer

1 Introduction

Speech Enhancement is a specialized domain within signal processing that strives to enhance the perceptual quality and intelligibility of speech signals that are affected by external factors such as noise, reverberation, or other forms of interference. It involves the development and application of algorithms and techniques to estimate and separate underlying clean speech from corrupting components, while minimizing distortion and preserving essential characteristics of original speech signal. The primary goal of speech enhancement techniques is to improve speech intelligibility and improve the overall user experience across a broad spectrum of applications [1]. The need for speech enhancement comes when working on activities such as speech recognition [2], speaker diarization [3, 4], hearing aids, speaker identification [5], telecommunications and voice assistants to

C. Mahesh, R. D. Prasad, E. A. Bibi—UG Student.

P. Pareek et al. (Eds.): IC4S 2023, LNICST 536, pp. 274–287, 2024.
https://doi.org/10.1007/978-3-031-48888-7_23

improve quality and intelligibility of speech that has been degraded by the presence of background noise. The development of speech enhancement models is crucial because they reduce listener fatigue, especially when the listener is subjected to loud noise levels. Speech enhancement algorithms work to reduce background noise to a certain degree, or suppress noise that occurs when a speaker is engaged in communication with others.

There exist numerous scenarios where the enhancement of speech signals is desired [6]. Consider a scenario where students learning through online meeting platforms typically suffer from background noise, microphone sensitivity, audio distortions, audio feedback and compression artifacts. This is where the speech enhancement algorithms come into picture, such algorithms can be applied at the receiving end to improve the speech. People who are impaired to listening use hearing aids, experience difficulty in hearing when the background noise is too high. Techniques like spectral subtraction were used in such hearing aids to preprocess or clean the audio signals before amplification [7].

The paper follows a structured flow as outlined. Section 2 provides an overview of the related work conducted by other researchers in the field of speech enhancement. It discusses the traditional and existing approaches that have been employed for this task. Section 3 outlines the methodology employed in the study, including the process of dataset creation, preprocessing techniques applied, description of the models used, and evaluation metrics. Section 4 details the experimental setup, including information about the hyperparameters of the pre-trained models utilized. Section 5 presents the results of the experiments, where the performance of the models on the dataset is analyzed.

1.1 Contributions

- A dataset which consists of mixed audio signals from *voice bank* corpus dataset and *demand* dataset, containing various types of noise along with the clean signals. The curated dataset includes audio recordings with 16 kHz and 8 kHz sampling rates to accommodate the requirements of different models.
- Evaluating the prepared dataset on ESPNet-SE, SpeechBrain MetricGAN+, SpeechBrain SepFormer models.
- The prepared data and the source code of model evaluation are publicly available to benefit the researchers working in this area.

2 Related Work

In this section, we present a comprehensive overview of the recent advancements and research conducted in the field of speech enhancement. The focus is on summarizing the significant work carried out in this area. Researchers adopted various techniques to address this task. Most of the techniques fall into machine learning and deep learning approaches.

2.1 Traditional Methods

The Spectral Subtraction method proposed by Boll et al. is widely studied and adopted for speech enhancement, aiming to reduce the influence of background noise on the

speech signal. This technique was introduced in 1979 and it works based on the subtraction of an estimated noise spectrum from the magnitude spectrum of the noisy speech. This approach involved dividing the noisy speech signal into frames of short duration and calculating the magnitude spectrum of each frame using the Fast Fourier Transform (FFT). This method assumes that every frame consists of stationary noise only, which is not the case in real-life scenarios. This can lead to imprecise noise estimation, consequently diminishing the effectiveness of spectral subtraction. To evaluate the performance of the method, metrics such as signal-to-noise ratio (SNR), perceptual evaluation of speech quality (PESQ), and mean opinion score (MOS) were utilized [8–10].

Adaptive wiener filtering was introduced in 2008 by Abd El-Fattah, M., et al. This approach utilizes a Short-Time Fourier transform (STFT) for the conversion of the noisy speech signal into frequency domain representation. This method estimates the statistical properties of the noisy input signal and then filters noisy components in the frequency domain. This method uses SNR to calculate filter coefficients to create a balance between noise suppression and speech preservation. The performance of the method was evaluated using signal-to-noise ratio improvement (SNRI) and PESQ [11].

Due to their interpretability and simplicity, traditional approaches like spectral subtraction and Wiener filtering have been frequently employed, however they can have difficulties when dealing with non-stationary and dynamic noise circumstances. Deep learning models, on the other hand, offer the promise for enhanced generalization and flexibility to various noise environments, but they demand a significant amount of labeled data and may be difficult to comprehend. Additionally, research is needed to enhance the adaptability of deep learning models to handle limited data scenarios, improve their interpretability, optimize their computational efficiency for real-time applications, and develop techniques for reliable uncertainty estimation.

2.2 Deep Learning Models

Wave-U-Net: Macartney et al. proposed a deep learning model that is specially designed for the audio source separation and enhancement tasks. It is built upon the U-Net architecture, a widely used model in image segmentation. The Wave-U-Net works directly in the time domain and takes raw audio waveforms as input. The performance of the model is evaluated on the benchmark datasets using SNR, PESQ, and STOI. The computational complexity of the Wave-U-Net model may require significant resources for training and interference [12].

Dual-Path RNN (DPRNN): Luo et al. A novel approach has been proposed for effectively modelling long sequential inputs using a combination of different types of recurrent neural network (RNN) layers. The idea revolves around dividing the input sequence into smaller segments and employing two RNNs: one within each segment (intra-chunk RNN) and one that considers the relationships between segments (inter-chunk RNN). This enables both local and global modelling of the input data. During training, the model is optimized by reducing the disparity between the improved speech signal, and the original clean speech signal using loss functions like mean square error (MSE) or SNR. While DPRNN has proven to be successful in modelling long sequential inputs, it may not be the best option for all types of speech processing tasks. A drawback of

DPRNN is its reliance on a substantial amount of training data to achieve optimal performance. This can pose a challenge in certain applications where obtaining a sufficient quantity of labeled data is difficult [13].

ESPnet-SE: ESPnet-SE is a toolkit for speech enhancement and separation that provides a unified framework for developing systems to improve speech quality and separate sources, by Li, Chenda, et al. It supports various models, including time-frequency masking networks, neural beamformers, and time-domain models, and provides multiple loss functions for training these models. The toolkit also includes evaluation metrics for assessing the performance of both the speech enhancement and the speech separation systems. Additionally, ESPnet-SE offers optional downstream speech recognition integration, allowing users to evaluate the impact of the speech enhancement and the separation on speech recognition performance. The toolkit has been evaluated on several benchmark datasets, and its implementations have achieved promising results on these datasets [14].

The SpeechBrain MetricGAN+: This model was proposed by Fu, Szu-Wei, et al. It enhances speech signals by removing noise and other distortions while preserving their quality. The model uses a combination of various training techniques that include multi-task learning, adversarial training, and a learnable sigmoid function to optimise objective metrics such as PESQ and STOI. Experimental results have shown that MetricGAN+ was effective in improving speech quality and achieved state-of-the-art results [15].

The SpeechBrain SepFormer: This model mainly focuses on speech separation along with speech enhancement, which reduces background noise and distortions. The authors Subakan, Cem, et al. have introduced a novel structure of a neural network model for speech separation called SepFormer. Unlike the traditional RNNs and LSTM models, the SepFormer employs a masking technique that is composed of transformers only. The masking technique is commonly utilized to separate the speech signal from background noise or distortion. This approach has demonstrated exceptional performance on standard datasets, establishing itself as state-of-the-art in the field [16].

Table 1. Comparison table for various models

Models	PESQ	STOI	SSNR
Spectral Subtraction	2.13	0.897	4.52
Adaptive Weiner Filtering	2.22	0.902	5.07
Wave-U-Net	2.40	0.915	15.41
DPRNN	3.07	0.928	15.53
ESPnet-SE	**3.25**	**0.953**	**15.94**
MetricGAN+	**3.15**	**0.972**	**18.23**
SepFormer	**3.08**	**0.923**	**17.61**

A sample of 10 audio signals is taken and are fed as inputs to the various models above and Table 1. was formulated with the results from various models specified above.

It is evident from Table 1. that ESPNet-SE, SpeechBrain MetricGAN+, and SepFormer have exhibited remarkable performance and achieved state-of-the-art results in the realm of speech enhancement techniques. Therefore, the objective of our study is to evaluate the performance of these three models on our custom *Voicebank+Demand* dataset that we have created.

3 Methodology

This section of the paper is divided into three subparts, where the first subpart provides a clear view of the processes followed while creating the corpus and sampling the audio signals to meet the requirements of the various models. While the second subpart provides a detailed description of the pre-trained models that have been chosen to evaluate on our custom dataset, the third subpart of this section gives insights into the evaluation metrics that were taken into account to compare the enhanced signal and the corresponding clean signal.

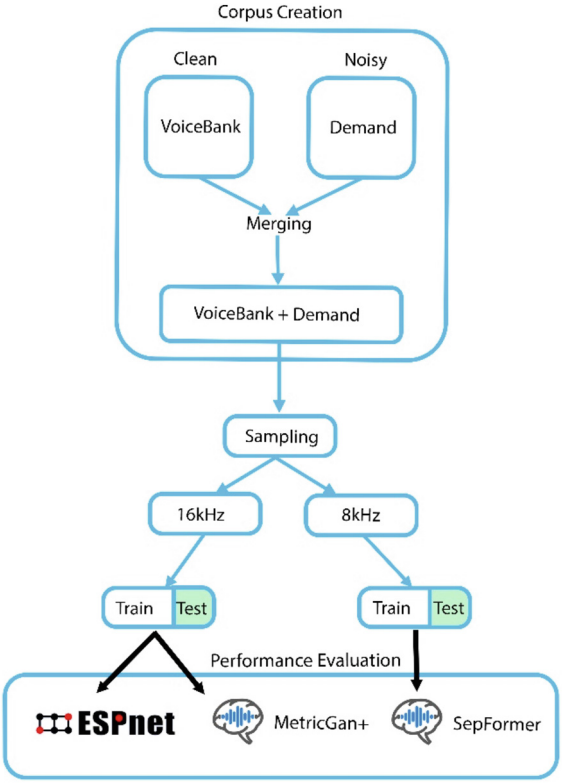

Fig. 1. Architectural Diagram

Figure 1 depicts the creation of the corpus from voice bank and demand datasets, which are then merged to create a *VoiceBank + Demand* dataset. To meet the model's

requirements, the sampling rates of the audio signals are converted to 16 kHz and 8 kHz. The pre-trained models are loaded into the Google Colaboratory Notebooks for evaluating the performance of the model against the dataset, which was split into training and testing sets.

Our work leverages the *Voice Bank + Demand* dataset to generate the denoised speech signals. Three models, ESPNet-SE, SpeechBrain MetricGAN+, and SpeechBrain SepFormer, are tested for speech enhancement on this dataset, and their performance is evaluated.

3.1 Dataset

The dataset consists of 12,396 mixed audio records sourced from the Voice Bank corpus, and the Demand dataset. It combines speech signals with various types of noise and the corresponding clean signals to evaluate the performance of different models. Both datasets are available publicly and are used for research purposes. The 300 h of speech data in the Voice Bank dataset of 500 healthy speakers with a range of accents and speaking styles are represented in the recordings [17]. The audio signals are crystal clear and of good quality. The speech signals are offered in WAV format, 16-bit, 48 kHz. Each speaker's audio is kept in a separate folder to make it easy to access their own audio. The Demand dataset comprises a diverse collection of 16 environmental noise recordings, including sounds such as car noise, babble noise, factory noise, and more. The noise signals are provided in 16-bit, 16 kHz WAV format.

Various speakers' audios from the voice bank dataset are merged with various noisy signals from the demand dataset and added into a new dataset, *VoiceBank+Demand*. The merging process is done with the help of the librosa[1] library in Python as shown in Fig. 2; the audio recordings sum up to 12396 noisy recordings. The generated dataset consists of 12,396 noisy recordings along with corresponding clean signals.

Among our three models, two of them, namely ESPNet-SE and SpeechBrain MetricGAN+, use 16 kHz sampling rate audio signals, and the SpeechBrain SepFormer model uses 8 kHz sampling rate audio signals. To meet this requirement, our dataset is made in such a way that it contains recordings of both 16 kHz and 8 kHz sampling rates in separate folders. The sampling rates are converted to the desired values using the Scipy[2] library in Python. Figure 2 shows a clear process for the creation of the dataset. The dataset is now split into training and testing sets, with 11572 recordings for the training set and 824 recordings for the test set for the models that require 16 kHz and 8 kHz sampling rates [18].

3.2 Trained Models

ESPnet- SE

The ESPnet speech enhancement model utilizes a modified version of the Conv-TasNet architecture [19]. Conv-TasNet was initially designed for the task of speech separation,

[1] Librosa—librosa 0.10.0 documentation.

[2] SciPy documentation—SciPy v1.10.1 Manual.

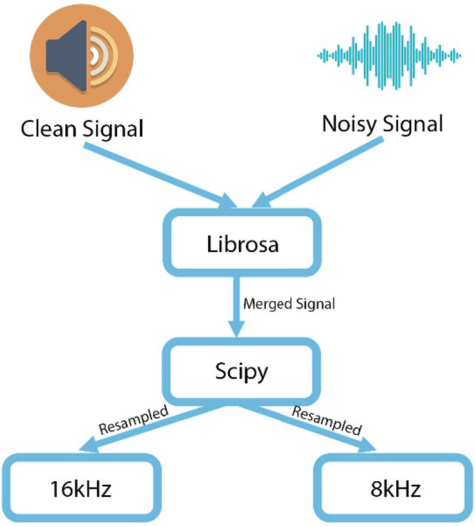

Fig. 2. Merging process followed to obtain mixed dataset VoiceBank + Demand

but it has been adapted for the speech enhancement task in this case. The encoder, separator, and decoder are the three primary parts of the model. By lowering noise and interference, these parts combine to enhance the quality of the incoming speech signal.

Encoder: This component in ESPNet takes an input noisy audio waveform and processes it through 1D convolutional layers. These layers capture different aspects of the audio signal, such as spectral and temporal characteristics.

Decoder: This component in ESPNet-SE takes enhanced audio signals that have been separated from noise and reconstructs clean speech. It uses 1D convolutional layers to un-sample the masked signal and generate the final output. The decoder's purpose is to remove noise and interference.

Separator Component: The separator component uses the encoded features generated by the encoder to separate the speech and the noise components of the input signal. It typically consists of additional 1D convolutional layers that further process masked features to remove any remaining noise and interference.

Training: This model is trained to improve speech quality by reducing noise and interference. This process makes use of a loss function called SI-SNR, which measures the similarity between the predicted and clean speech signals. The training data is divided into batches for efficient processing. This toolkit was trained using the CHiME-4 dataset, which consists of a large number of noisy speech recordings made in a variety of everyday environments. The model learns to map a noisy input to a clean speech output during training. It adjusts its parameters using gradient descent and backpropagation to minimise the SI-SNR loss. The model goes through multiple epochs, where it processes the training data, computes the loss, and updates its parameters. Throughout training, the model learns to extract features, separate speech from noise, and reconstruct clean

speech. The goal is to enhance speech quality in real-world situations with noise and reverberation [14].

SpeechBrain MetricGAN+
The SpeechBrain MetricGAN+ uses a generative adversarial network (GAN) [20] to generate high-quality signals from degraded or noisy signals. The generator and the discriminator are the two main parts of the model.

Generator: The noise-filled input signal is converted into a clean, better signal with fewer noise and distortion using the generator.

Discriminator: The discriminator plays a major role in differentiating the optimised output signal from the corresponding clean, high-quality signal. It also contains a deep neural network that classifies if the output is actually a clean signal or not.

Training: It uses a combination of adversarial and metric-based training. The generator and discriminator models are trained in such a way that one outperforms the other. Adversarial training focuses on enhancing the performance of the model by training it to generate more realistic samples that can fool a discriminator. The feedback mechanism comes from the discriminator. Discriminator, which provides feedback to the generator by classifying the input signal to it. The generator then adjusts to give a more realistic output. Adversarial training involves optimising the generator and discriminator simultaneously. It also involves a loss function that measures the difference between generated samples and real samples. During training, the parameters of the generator and discriminator are updated using the learnable sigmoid loss function. During the training phase, stochastic gradient descent (SGD) with a learning rate of 2e-5 was used to optimise the model's parameters [15].

SpeechBrain Sepformer
Model Architecture: The SpeechBrain SepFormer model architecture consists of an encoder, a transformer-based decoder [21], and a mask estimator.

Encoder: This component receives the noisy speech signal as input and transforms it into a high-dimensional feature representation. Typically, the encoder consists of convolutional layers followed by a self-attention mechanism that captures long-range dependencies in the input signal.

Transformer-Based Decoder: The transformer-based decoder takes the encoded features as input and generates the clean speech signal as output. The decoder is composed of multiple layers of transformer blocks, where each block contains a multi-head self-attention mechanism and a feed-forward neural network. The transformer blocks enable the model to capture complex dependencies between different parts of the input signal and generate a high-quality, clean speech signal.

Mask Estimator: The mask estimator is responsible for estimating the binary masks that indicate which parts of the input signal are noise and which parts are speech. The mask estimator is typically composed of a feed-forward neural network that takes the encoded features as input and generates a binary mask for each time-frequency bin in the input signal.

Training: The SpeechBrain SepFormer speech enhancement model was trained on a dataset of noisy and clean speech signals called the DNS Challenge dataset. The dataset consists of 16,000 audio clips with a duration of 3 s each. The audio clips were recorded in four different noisy environments: babble, car, street, and train. STFT features were used for pre-processing the audio signals. The MSE function was employed to quantify the dissimilarity between the predicted clean speech signal and the ground truth clean speech signal. Additionally, binary cross-entropy loss was utilized to minimize the disparity between the estimated binary mask and the ground truth binary mask. During training, the model was optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The training procedure spanned 100 epochs, and the model with the lowest validation loss was chosen as the final model. [16].

3.3 Evaluation Metrics

All three models were evaluated against various metrics for speech enhancement based on the training and testing sets. STOI, SSNR, PESQ, and composite objective metrics (COVL, CBAK, and CSIG) were used to assess the models' performance [22, 23].

Segmental Signal-to-Noise Ratio(SSNR): A voice or audio signal's energy is compared to the energy of the background noise in the same signal segment to calculate the SSNR. A higher SNR value indicates that the model has succeeded in reducing the noise, thereby producing quality output [22, 23].

Perceptual Evaluation of Speech Quality(PESQ): It is used to evaluate the improvement of the signal quality achieved by the model. The range of the PESQ score is from -0.5 to 4.5, where higher values indicate better quality of speech [22, 23].

Short-Term Objective Intelligibility(STOI): It is a measure of similarity between the enhanced and noisy speech signals. It measures the intelligibility of degraded speech signals relative to the corresponding clean signal. It ranges from 0 to 1 [22, 23].

COVL: It is a measure of how well the enhanced speech signal matches the clean speech signal [23, 24].

CBAK: It quantifies how well the algorithm is able to suppress or remove the background noise in the enhanced speech signal [23, 24].

CSIG: It measures the amount of residual echo or reverberation that is present in the enhanced speech signal [23, 24].

The higher the values of COVL, CBAK, and CSIG, the better the enhanced speech signal.

4 Experiment

We conducted our experimentation using Google Colaboratory (Colab)[3] notebooks, a cloud-based Python environment. The experiments were implemented and executed within Colab, leveraging its built-in libraries and computing resources. The hyperparameters employed in our work are tabulated in Table 2.

Table 2. Parameters of the adapted models

Models →	ESPnet-SE	MetricGAN+	SepFormer
No of Parameters	8,645,184	1,895,514	25,613,569
Loss Functions	MSE, SI-SNR	MSE, PESQ	SI-SNR
Model Size	9695.76 MB	151.85 MB	40353.41 MB
Memory Usage	252.932 MB	209.630 MB	942.521 MB
Activation Function	ReLU, Sigmoid & tanh	Learnable Sigmoid Function	GeLU

From Table 2, EPSNet-SE, MetricGan+, and SepFormer differ in terms of their number of parameters, loss functions, model size, memory usage, and activation functions.

EPSNet-SE, with 8,645,184 parameters, is a model that utilises various loss functions for different objectives. The mask approximation loss employs MSE and cross-entropy (CE) cost functions. The signal approximation loss is based on MSE, and the metric-based loss uses scale-invariant SNR as the metric. The model has a size of 9695.76 MB and requires 252.932 MB of memory. The activation functions used in EPSNet-SE include ReLU, sigmoid, and tanh.

MetricGan+, on the other hand, has 1,895,514 parameters. The model primarily utilises SI-SNR as the loss function. It has a smaller model size of 151.85 MB and a memory usage of 209.630 MB. The activation function used in MetricGan+ is a learnable sigmoid function.

Lastly, SepFormer has 25,613,569 parameters. The model utilises MSE and PESQ as the loss functions. SepFormer has the largest model size among the three, with 40353.41 MB, and requires 942.521 MB of memory. The activation function used in SepFormer is GELU (Gaussian Error Linear Units).

These models are designed for speech enhancement tasks, aiming to enhance the quality and intelligibility of speech signals. The differences in their architectures, parameters, loss functions, model sizes, memory usage, and activation functions provide options for researchers and practitioners to choose the model that suits their specific requirements and constraints in terms of computational resources and performance

[3] Welcome To Colaboratory - Colaboratory (google.com).

objectives. Trained models are publicly available in Github and Hugging Face Platforms ESPnet-SE[4], SpeechBrain MetricGan+[5]and SpeechBrain SepFomer.[6]

The following figures depict the waveforms of clean, noisy and enhanced audio signals.

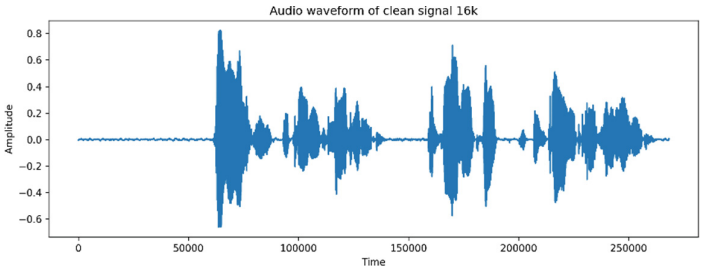

Fig. 3. Audio waveform of clean signal

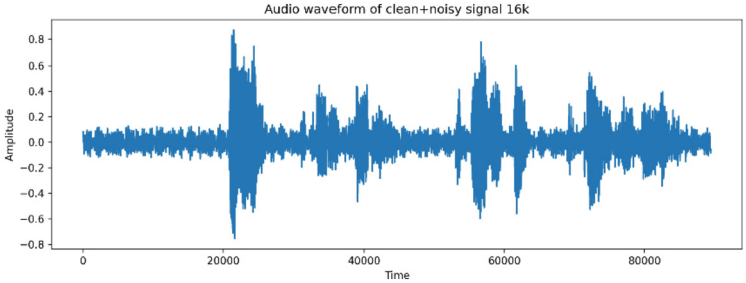

Fig. 4. Audio waveform of clean+noisy signal

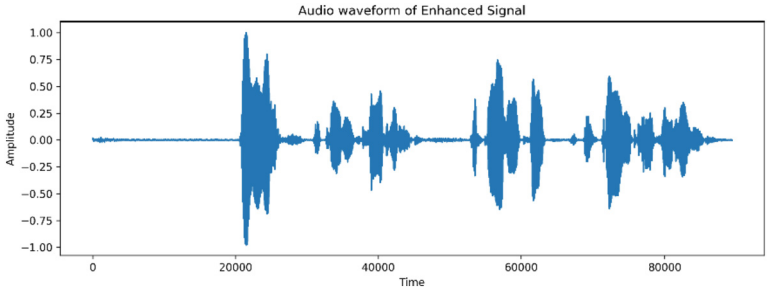

Fig. 5. Audio waveform of an Enhanced signal

Figure 3 presents audio waveform of clean audio signal i.e., when there is no noise in the speech signal, Fig. 4 presents the waveform for an audio signal that consists of

[4] ESPNet-SE: https://github.com/ESPNet/ESPNet.

[5] SpeechBrain MetricGAN+: https://huggingface.co/speechbrain/metricgan-plus-voicebank.

[6] SpeechBrain SepFormer: https://huggingface.co/speechbrain/sepformer-whamr-enhancement.

speech along with the noise while Fig. 5 presents the general waveform for an enhanced audio signal when passed to a specific model (*MetricGAN+* here).

5 Results

From the Table 3, the performance of three existing speech enhancement models, namely ESPNet-SE, SpeechBrain-MetricGan+, and SpeechBrain- SepFormer, was evaluated using various metrics, including SSNR, PESQ, STOI, CSIG, CBAK, and COVL. In terms of SSNR, all models achieved negative values, indicating a reduction in noise. Among the models, SpeechBrain-MetricGan+ demonstrated the highest PESQ scores, outperforming both ESPNet-SE and SpeechBrain-SepFormer by approximately 39.05% and 10.29%, respectively. Regarding STOI, SpeechBrain-MetricGan+ achieved the highest scores on the test set. Additionally, SpeechBrain-MetricGan+ exhibited the best performance in terms of CBAK, CSIG, and COVL metrics, surpassing the other models. These results highlight the superior performance of SpeechBrain-MetricGan+ in enhancing speech quality and reducing noise compared to the other evaluated models.

Table 3. Performance Evaluation of models using metric

Existing Models	Train/Test	SSNR	PESQ	STOI	CSIG	CBAK	COVL
ESPnet-SE	Train	−7.304	1.892	0.869	2.585	1.811	2.197
	Test	−7.258	2.233	0.932	3.140	2.031	2.668
SpeechBrain-MetricGAN+	Train	**−0.797**	**2.731**	**0.866**	**3.303**	**2.609**	**2.983**
	Test	**−0.191**	**3.108**	**0.925**	**3.725**	**2.894**	**3.403**
SpeechBrain-SpeFormer	Train	−4.133	2.617	0.873	3.109	2.321	3.110
	Test	−4.167	2.818	0.893	3.307	2.810	3.890

6 Conclusion

This paper investigated and compared the performance of three speech enhancement models, namely ESPNet-SE, SpeechBrain-MetricGan+, and SpeechBrain-SepFormer. The evaluation was based on various metrics, including SSNR, PESQ, STOI, CSIG, CBAK, and COVL. The results indicate that SpeechBrain's MetricGan+ outperforms the other models across multiple metrics. It achieves higher PESQ scores, indicating better perceived speech quality, compared to both ESPNet-SE and SpeechBrain-SepFormer. Furthermore, SpeechBrain-MetricGan+ demonstrates superior performance in terms of STOI, CBAK, CSIG, and COVL metrics, highlighting its effectiveness in reducing noise and enhancing speech signals. These findings suggest that SpeechBrain-MetricGan+ is a promising model for speech enhancement tasks, offering improved speech quality and noise reduction capabilities. The superior performance of SpeechBrain and MetricGan+ underscores their potential for real-world applications that require high-quality speech processing.

References

1. Benesty, J., Makino, S., Chen, J. (eds.): Speech Enhancement. Springer Science & Business Media, Heidelberg (2006). https://doi.org/10.1007/3-540-27489-8
2. Bai, Z., Zhang, X.-L.: Speaker recognition based on deep learning: an overview. Neural Netw. **140**, 65–99 (2021)
3. Park, T.J., et al.: A review of speaker diarization: recent advances with deep learning. Comput. Speech Lang. **72**, 101317 (2022)
4. Arla, L.R., Bonthu, S., Dayal, A.: Multiclass spoken language identification for indian languages using deep learning. In: 2020 IEEE Bombay Section Signature Conference (IBSSC), pp. 1–2. IEEE (2020)
5. Tirumala, S.S., et al.: Speaker identification features extraction methods: a systematic review. Expert Syst. Appl. **90**, 250–271 (2017)
6. Gogate, M., et al.: CochleaNet: a robust language-independent audio-visual model for real-time speech enhancement. Inf. Fusion **63**, 273–285 (2020)
7. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press (2013)
8. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
9. Kumar, B.: Spectral subtraction using modified cascaded median based noise estimation for speech enhancement. In: Proceedings of the Sixth International Conference on Computer and Communication Technology 2015, pp. 1–2 (2015)
10. Kumar, B.: Mean-median based noise estimation method using spectral subtraction for speech enhancement technique. Indian J. Sci. Technol. **9**, 1–6 (2016)
11. Abd El-Fattah, M., et al.: Speech enhancement using an adaptive wiener filtering approach. Prog. Electromagnet. Res. M **4**, 167–184 (2008)
12. Macartney, C., Weyde, T.: Improved speech enhancement with the Wave-u-Net. arXiv preprint arXiv:1811.11307 (2018)
13. Luo, Y., Chen, Z., Yoshioka, T.: Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–2. IEEE (2020)
14. Li, C., et al.: ESPnet-SE: end-to-end speech enhancement and separation toolkit designed for asr integration. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 1–2. IEEE (2021)
15. Fu, S.-W., et al.: MetricGAN+: an improved version of MetricGAN for speech enhancement. arXiv preprint arXiv:2104.03538 (2021)
16. Subakan, C., et al.: Attention is all you need in speech separation. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–2. IEEE (2021)
17. Veaux, C., Yamagishi, J., King, S.: The voice bank corpus: design, collection and data analysis of a large regional accent speech database. In: 2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), pp. 1–2. IEEE (2013)
18. Koyama, Y., et al.: Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks. arXiv preprint arXiv:2005.11611 (2020)
19. Luo, Y., Mesgarani, N.: Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech and Lang. Process. **27**(8), 1256–1266 (2019)
20. Creswell, A., et al.: Generative adversarial networks: an overview. IEEE Signal Process. Mag. **35**(1), 53–65 (2018)

21. Li, M., Zorilă, C., Doddipatla, R.: Transformer-based online speech recognition with de-coder-end adaptive computation steps. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 1–2. IEEE (2021)
22. Dong, X., Williamson, D.S.: Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks. J. Acoust. Soc. Am. **148**(5), 3348–3359 (2020)
23. Hu, Y., Loizou, P.C.: Evaluation of objective measures for speech enhancement. In: Ninth International Conference on Spoken Language Processing, pp. 1–2 (2006)
24. Kumar, B.: Comparative performance evaluation of greedy algorithms for speech enhancement system. Fluctuation Noise Lett. **20**(02), 2150017 (2021)