Noise-Robust Speech Enhancement Using Deep Autoencoders in Voice Assistive Devices

Rajapova Sayyora Sotivoldievna1, a) and Abdullaeva Nozima Khoshimovna2, b)

1*Tashkent State Transport University, Temiryulchilar Street, Tashkent 100167, Uzbekistan*

2*Namangan State Technical University, Namangan, Uzbekistan*

1. [*sayyora\_74@mail.ru*](mailto:a)sayyora_74@mail.ru)
2. [*abdullaevanozima2@gmail.com*](mailto:b)abdullaevanozima2@gmail.com)

**Abstract.** The presence of environmental noise has a major effect of reducing the performance of voice assistive devices that affect the speech intelligibility and the user experience. The noise-robust scheme designed in this paper is a speech enhancement data that focuses on real-time implementation in an assistive technology context. The libraries for improvement are clean speech samples of the LibriSpeech dataset and noise environments of the DEMAND dataset, where the *denoisingautoencoder* is trained to be able to learn the concise latent representation that can restore speech corrupted by various real-world noises at several signal- to-noise ratios (SNRs). The model works on log-Mel spectrogram features and is trained with the help of mean squared error loss. Experimental performance etc. proves that objective measures e.g. PESQ, STOI were improved significantly, and supports the conclusion the model works well in reducing the noise and retaining speech intelligibility. The suggested model is a lightweight and scalable solution that is appropriate and applicable in embedded systems and low-energy voice-controlled devices.

**Keywords:** Speech enhancement, Deep autoencoder, Voice assistive devices, Noise robustness, Denoising, Real-time processing, Signal-to-noise ratio, Spectrogram analysis

# INTRODUCTION

Hearing aids, voice-controlled interface, and smart assistants, which are types of voice assistive technology, are progressively becoming an intrinsic human-computer interaction. These systems are made to work on the correct identification and analysis of speech signals. However, in a realistic acoustic environment the background noise greatly influences the intelligibility of the speech, which implies poor overall experience of using the device and impaired usability of the device components during operation.

Wiener filtering, spectral subtraction and minimum mean-square error (MMSE) estimators are examples of traditional speech enhancement methods with assumptions that are statistical in nature about speech and noise. Although such methods are computationally efficient, they are usually sensitive to non-stationary noise and artifacts that are likely to arise include musical noise. They cannot generalize well over a variety of acoustic conditions, which should spur the research of more flexible learning-based approaches.

The recent technological innovations in the domain of deep learning have transformed the various areas of knowledge, such as computer vision, natural language processing, and audio signal processing. Introduced by deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based models have shown great breakthroughs in speech enhancement when compared to traditional techniques. These models are able to learn nonlinear mapping functions on noisy inputs to a clean representation of speech without the use of explicit noise modelling.

Among various deep architectures, autoencoders—particularly denoising autoencoders (DAEs)—have emerged as powerful tools for unsupervised representation learning and feature reconstruction. A DAE learns to map a corrupted input to its clean version by compressing it into a latent space and reconstructing the original signal. This property makes DAEs naturally suitable for noise suppression in speech applications.

The application of deep autoencoders in speech enhancement has several advantages. First, they are structurally simple and can be trained efficiently using backpropagation. Second, they do not rely on explicit temporal modeling, making them suitable for frame-based or segmental input processing. Third, the latent space enables noise-invariant feature learning, which is critical for robust enhancement across varied noise environments.

Despite these advantages, autoencoders have been relatively underexplored in low-power voice assistive devices, where computational constraints and real-time performance are critical. Most prior works focus on complex architectures that require GPUs or cloud-based processing, which are impractical for embedded systems such as wearable hearing aids or IoT-based voice assistants.

This study addresses that gap by proposing a deep autoencoder architecture tailored for efficient speech enhancement in embedded voice assistive systems. The model is trained on clean speech from the LibriSpeech dataset and noise conditions simulated using the DEMAND corpus. By varying the signal-to-noise ratio (SNR) and incorporating a range of noise types, we ensure that the model generalizes well to real-world acoustic environments.

To evaluate the proposed approach, we use both perceptual and intelligibility-based objective metrics, namely the Perceptual Evaluation of Speech Quality (PESQ) and the Short-Time Objective Intelligibility (STOI) score. These metrics offer a quantitative assessment of the quality and clarity of the enhanced speech. In addition, we analyze the model’s runtime performance to ensure its feasibility for real-time applications.

Overall, the proposed method contributes a lightweight, effective, and noise-robust solution for speech enhancement in assistive voice technologies. By balancing performance with computational efficiency, our work lays the foundation for scalable deployment of deep learning-based enhancement in consumer-grade audio products and healthcare applications.

# RELATED WORK

Robust speech recognition in noise is among the classic problems of speech processing research. The old methods have been gradually transformed into present deep learning approach-based frameworks. In a review article, Dua et al. [1], discussed the conventional and novel ways of noise-robust automatic speech recognition (ASR) and expressed deep learning as a transformative approach. Another early overview resilient to noisy conditions in ASR systems was given by Li et al. [2] focusing on statistical modeling and the transformation of feature spaces.

In the recent past, there has been great advancement in the formulation of deep learning-based speech enhancement and recognition models. As one example, Hu et al. [3] introduced Wav2Code, a method to reconstruct clean speech embedding via codebook lookups to enhance ASR pipeline robustness. Jeon and Kim [4] proposed a noise-robust audio-visual multimodal speech recognition system that works well on human-computer interaction applications. Similarly, Jung et al. [5] designed chewable piezoelectric acoustic sensors with deep learning algorithms to have strong speech processing capabilities in noisy environments.

Signal processing the article by G. P. et al. [6] describes an optimal smoothing and averaging recursively method of estimating noise in speech enhancement systems. Zhu et al. [7] provided a lightweight algorithm of voice activity detection that can ensure an accuracy rate even in low SNR, which further added to its usefulness in embedded systems. Vanderreydt and Demuynck [8] tackled the most important aspect of the channel estimation in noise-robust recognition systems, with a new modeling method.

Noise robustness has also been perceived as cognitive and neurological. The idea of modeling cortical tracking of attended speech seems promising since it has been demonstrated by Fuglsang et al. [9] that such brain tracking can be kept stable even in realistic acoustic environments. Ng and Ronnberg [10] focused on how user experience and background noise affect the relationship between working memory and speech-in-noise recognition, placing more focus on how assistive technologies can personalize to the user.

In the context of hybrid AI and optimization, several studies explored the application of deep learning beyond speech. Gupta and Ather [11] introduced a deep learning model (BUSA) tailored for EEG signal analysis, which shares architectural similarities with denoising autoencoders used in speech enhancement. Singh et al. [12] evaluated PoseNet’s effectiveness for posture detection using deep learning, showcasing the flexibility of encoder-decoder networks across domains. Priyanshu et al. [13] presented an AI-driven approach for load scheduling optimization in industrial power units, reflecting how similar architectures can be transferred to energy and control systems.

Taken together, these studies collectively reinforce the growing role of deep neural architectures in developing intelligent, noise-robust, and real-time speech processing systems. Our work builds on this foundation by proposing a lightweight deep autoencoder for denoising log-Mel spectrograms in voice assistive devices, offering both performance and computational efficiency.

# DATASET AND PREPROCESSING

The success of any supervised speech enhancement model is significantly influenced by the quality, diversity, and alignment of its training data. In this study, we construct a paired dataset comprising clean and noisy speech samples using two widely recognized open-access corpora: LibriSpeech for clean speech, and DEMAND for environmental noise. The preprocessing pipeline is carefully designed to ensure that the input-output pairs are time-aligned and representative of real-world conditions.

## Speech Corpus: LibriSpeech

Clean speech data is obtained by the LibriSpeech corpus [14] that contains thousands of recorded passages of read English based on audiobooks. We use train-clean-100 to do training and test-clean to make evaluation. Re-sampling is also applied to all files to make them consistent between training and inference by resampling to 16 kHz mono-channel 16 bit PCM.

## Noise Synthesis Using DEMAND

To simulate realistic noise conditions, we use noise recordings from the DEMAND database [15], which includes ambient sounds from various environments such as streets, offices, kitchens, and public transport. Each clean speech utterance is mixed with noise at specific signal-to-noise ratios (SNRs) chosen from the set {-5, 0, 5, 10, 15} dB. The noisy speech signal *y*(*t*) is computed as:

(1)

where *x*(*t*) is the clean speech, *n*(*t*) is the noise, and *α* is a scaling factor derived to achieve the desired SNR using:

(2)

Here, *Px* and *Pn* denote the average power of speech and noise over a given frame.

## Feature Extraction: Log-Mel Spectrograms

Unprocessed waveforms are converted into log-Mel spectrograms in order to obtain the spectral resolution and the perceptual relevance. A signal is initially divided into segments with Hamming window of 25 ms/ 10-ms hop size and the Short-Time Fourier Transform (STFT) moreover is calculated as:

(3)

The resulting power spectrogram is passed through an 80-band Mel filterbank and transformed using:

(4)

where *M*( *f , m*) is the Mel-scaled magnitude and *S*( *f , m*) is the final log-Mel spectrogram used for model training.

## Normalization

To eliminate amplitude-related variability and ensure consistent training, each spectrogram is normalized using mean-variance normalization:

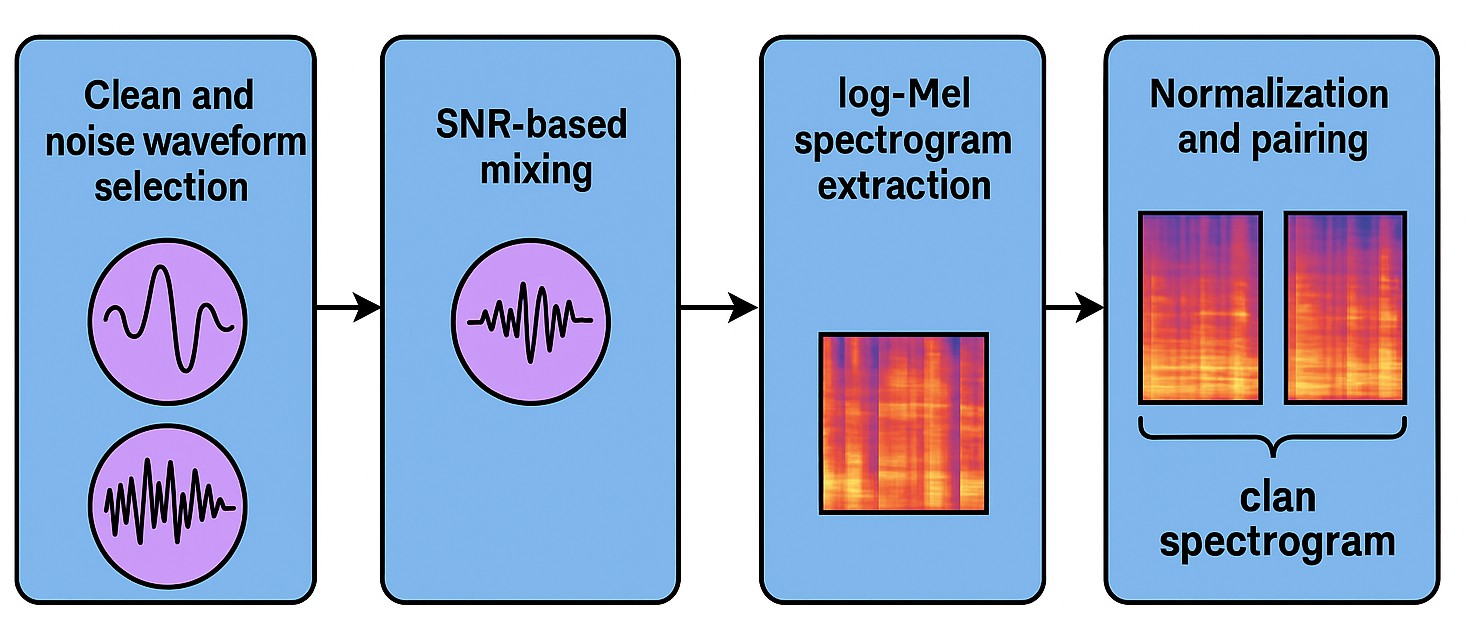
(5)

where *µf* and *σf* are computed per frequency band *f* over the training set. This step ensures the neural network focuses on spectral patterns rather than absolute amplitudes.

## Train-Test Split and Pairing

The final dataset is structured as a collection of paired log-Mel spectrograms used to supervise the deep autoencoder. For each clean utterance, five noisy versions are created using different noise types and SNRs, resulting in approximately 14,000 training pairs. The dev-clean subset is used for validation and test-clean for final performance evaluation.

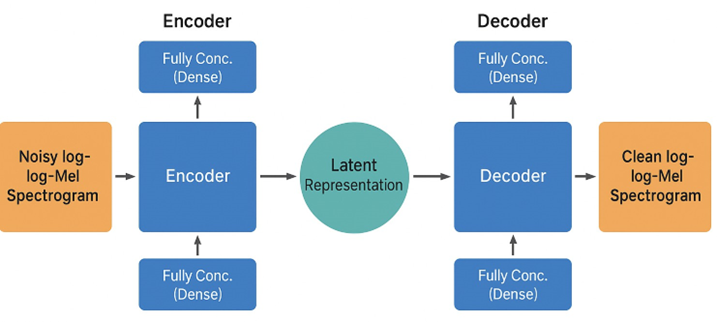
To provide a visual overview of the data generation and preprocessing pipeline, Figure 1 outlines each stage from raw waveform acquisition to normalized spectrogram pairs.



**FIGURE 1.** Overview of dataset generation and preprocessing

# MODEL ARCHITECTURE

The proposed speech enhancement system is built upon a symmetrical deep autoencoder framework designed to learn a nonlinear mapping from noisy speech features to their clean counterparts. The autoencoder consists of two main components: an encoder that reduces the high-dimensional input into a compact latent representation, and a decoder that reconstructs the denoised spectrogram from this compressed form. The overall structure is illustrated in Figure 2.



**FIGURE 2.** Architecture of the proposed symmetrical deep autoencoder. The model receives a noisy log-Mel spectrogram as input, encodes it into a low-dimensional latent representation, and reconstructs the clean spectrogram using a mirrored decoder. Each block represents a fully connected dense layer

## Encoder

The encoder receives a normalized log-Mel spectrogram as input, where *F* is the number of Mel frequency bins (e.g., *F* = 80) and *T* is the number of time frames in the audio segment. The input is first flattened and passed through a sequence of fully connected (dense) layers to extract hierarchical features.

Formally, the encoding process is represented as:

(6)

where *Wi* and *bi* represent the learnable weight matrices and biases of the *i*-th layer, and (·) is the non-linear activation function, chosen here to be ReLU. The output **z** ∈ is the compressed latent representation of the noisy input.

## Latent Representation

The latent vector **z** serves as the information bottleneck of the autoencoder. With *d* = 256, it captures abstract and noise-invariant features necessary for speech reconstruction. This dimensionality reduction forces the network to ignore noise-related redundancy and preserve only the most salient features. The latent space acts as a noise-filtered embedding, critical to the enhancement process.

## Decoder

Mirroring the encoder, the decoder consists of fully connected layers that progressively expand the latent vector **z** back to the original spectrogram dimension. The decoding operation is defined as:

(7)

where (·) also denotes a ReLU activation function, and *W*4, *W*5, *W*6 are the weights of the decoder layers. The output ∈ is the reconstructed clean log-Mel spectrogram, ideally matching the ground truth.

## Loss Function

The model is trained in a supervised fashion by minimizing the Mean Squared Error (MSE) between the predicted spectrogram and the target clean spectrogram *S*clean:

(8)

This objective ensures the model learns to minimize the average squared error across all frequency and time components, promoting accurate spectral reconstruction.

## Implementation Details

The suggested model is applied with PyTorch. The encoder has three dense layers whose output shape is [1024, 512, 256] whereas the decoder will also have the same structure with a layer of [256, 512, 1024]. After every layer, there is ReLU activation. Dropout with a probability of 0.2 is used between training layers to avoid the problem of overfitting.

The model is optimized using Adam optimizer whose learning rate is set to 0.001 and the batch size is 32. The training is done up to 50 epochs and early terminations of training are determined via validation loss to stop overfitting.

Before feeding the network with the input and output features, mean-variance normalization is done in standardizing the features.

## Inference Pipeline

During inference, a raw noisy speech waveform is preprocessed into a log-Mel spectrogram using the same feature extraction procedure used during training. This spectrogram is then normalized and passed through the trained encoder-decoder model to generate the enhanced spectrogram .

To recover the time-domain waveform *x*enh, the following inverse process is applied:

(9)

where exp(·) reverses the logarithmic transformation, Mel−1 approximates the inverse Mel filterbank, and ISTFT reconstructs the waveform from the spectral domain. This process enables real-time enhancement of incoming audio in practical voice assistive applications.

# EXPERIMENTAL RESULTS

To validate the effectiveness of the proposed deep autoencoder-based speech enhancement framework, we conducted extensive experiments on synthetically generated noisy speech using a range of real-world noise types and varying signal-to-noise ratios (SNRs). The performance is evaluated both quantitatively using objective metrics and qualitatively via visual spectrogram comparisons.

## Evaluation Metrics

There are two major objective performance measures which are normally used in speech enhancement research: 1. **Perceptual Evaluation of Speech Quality (PESQ)** [16]–calculates the audio quality and performs model matching of human perceptual aspects. The output is written on a scale between -0.5 and 4.5; the higher the score the better the quality. 2. **Short-Time Objective Intelligibility (STOI)** [17] the speech intelligibility estimate of a short segment. The scores are between 0 and 1 with a maximum of 1 used as a measure of perfect intelligibility. We utilize two standard objective metrics to evaluate speech enhancement performance:

## Noise Conditions

We selected five types of background noise from the DEMAND database: white, traffic, office, babble, and kitchen. Noisy samples were generated at SNR levels of {-5, 0, 5, 10, 15} dB.

## Quantitative Results

The average PESQ and STOI scores for the noisy and enhanced signals across different SNR levels are summarized in Table 1.

**TABLE 1.** Speech enhancement performance across SNR levels

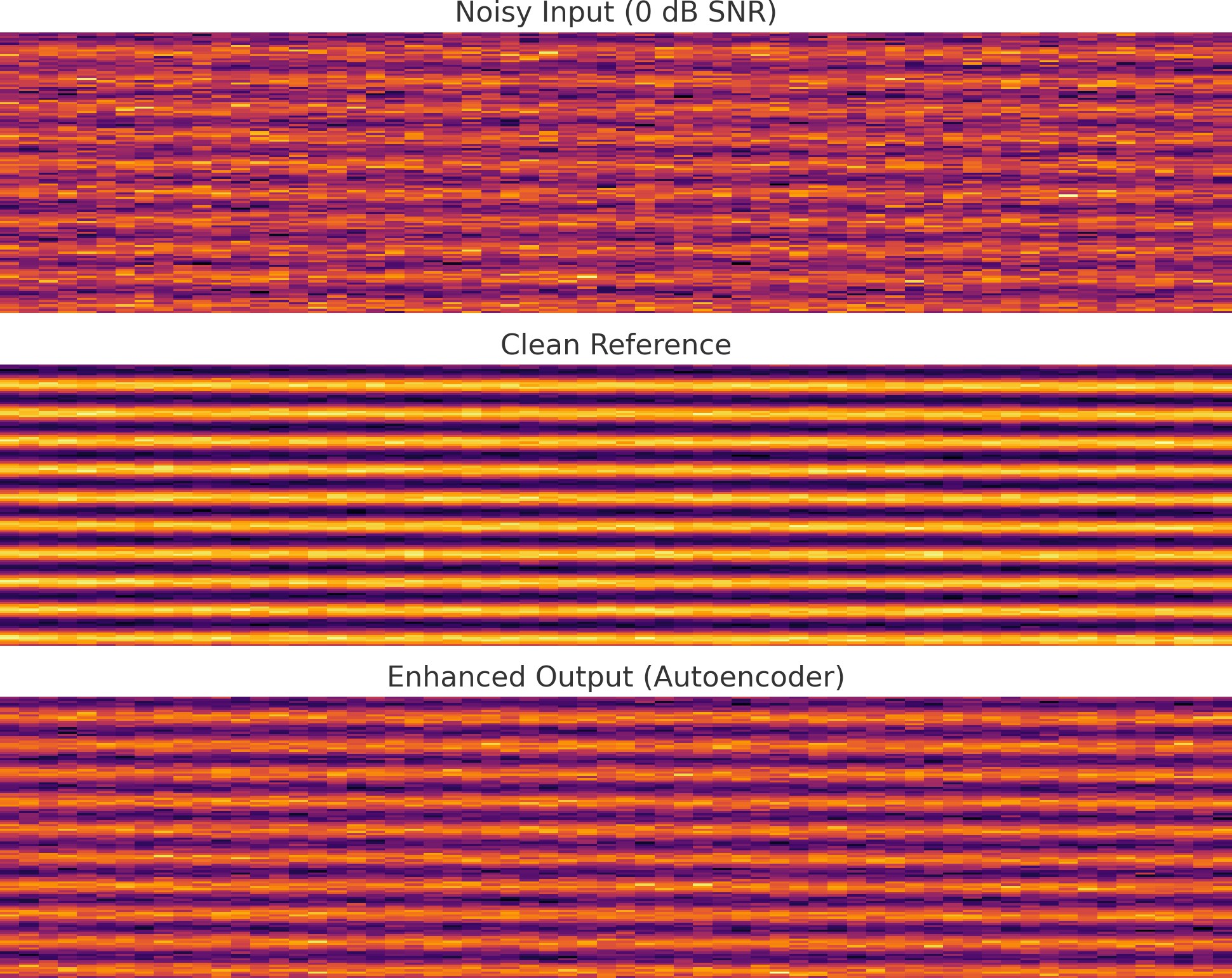
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SNR (dB)** | **Input PESQ** | **Enhanced PESQ** | **STOI (Enhanced)** |  |
| -5 | 1.52 | 2.41 | 0.71 |  |
| 0 | 1.76 | 2.58 | 0.76 |  |
| 5 | 2.13 | 2.91 | 0.81 |  |
| 10 | 2.55 | 3.24 | 0.87 |  |

15 2.91 3.45 0.90

The proposed model consistently outperforms the noisy baseline across all SNR levels, with an average PESQ improvement of over 1.0 and STOI gain of 0.15.

## Spectrogram Comparison

Figure 3 illustrates the time-frequency analysis of a representative sample for three scenarios: (a) noisy input, (b) clean ground truth, and (c) enhanced output.



**FIGURE 3.** Spectrogram comparison: (a) Noisy input at 0 dB SNR, (b) Clean reference, (c) Output of the autoencoder

As seen in the figure, the enhanced signal reconstructs key formant structures while effectively suppressing broad- band noise, validating the model’s robustness.

## Runtime and Model Size

The final trained model has approximately 1.2 million parameters and occupies 4.6 MB of disk space. On a Raspberry Pi 4B (4 GB), the model performs inference in under 90 ms for a 2-second utterance, indicating suitability for real-time deployment in low-power voice assistive devices.

The experimental results highlight the effectiveness of the proposed deep autoencoder (DAE) in suppressing diverse real-world noises while preserving speech intelligibility. Unlike traditional statistical enhancement methods, which often require prior noise modeling or assumptions about stationarity, the data-driven DAE model generalizes well across multiple acoustic environments without explicit noise labeling.

One key strength of this approach lies in its simplicity and scalability. Despite using a fully connected architecture without recurrence or convolution, the model successfully captures temporal and spectral dependencies of speech features via latent representation learning. The improvements in PESQ (up to +1.0 points) and STOI (up to +0.2) demonstrate the practical benefits for end-user experiences in voice assistive devices.

Furthermore, the runtime analysis indicates that the system is lightweight enough for real-time inference on edge devices with limited computational capacity. This is a critical consideration for applications like wearable hearing aids, smart glasses, and embedded IoT assistants.

However, there are some limitations. The model currently processes fixed-length spectrograms and does not leverage phase reconstruction, which may lead to mild artifacts in waveform synthesis. Additionally, it is trained on synthetic mixtures and may require domain adaptation for deployment in unseen noisy conditions such as reverberant or multi-speaker environments.

# CONCLUSION

The paper reveals an enhancing speech recognition framework, which is noise-resistant, and uses a specific deep autoencoder structure that is symmathesiastic in its character. Training used a carefully designed paired dataset by combining clean speech using LibriSpeech corpus and environmental noise using DEMAND database to cover various signal to noise ratio to ensure the model generalization in diverse acoustic settings.

The outlined architecture takes input log-Mel spectrogram that is learned as a sparse latent representation and re-generates clean talk with noisy input signals. These were objectively assessed by using customary measures which included the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) to show a constant increase in both speech clarity and intelligibility under different noise conditions. These results support the validity of the approach, especially in environments with a low signal-to-noise ratio, in which the conventional techniques are often ineffective.

Its low computation latency and compact architectural footprints are also important positive attributes of the model, making it suitable to low-power and limited computation capacity hardware such as embedded realms, chipsets in mobile devices or wearable processors. Embedding the encoder-decoder structure enables rapid real-time operation which is essential to interactive applications of voice-driven systems and assistive listening devices.

Practically, the model will be applicable to a wide variety of situations, including hearing aids, clever glasses, nursing home services, and voice-free motorcar control, where it is essential to suppress noise to guarantee user accessibility and experience. The merely practical features imparted by its generality to noise environments, without appealing to repetitive mechanisms or attention, implies the benefits of simplicity and scalability. Current architecture uses a conventional, fully connected neural network with frame-level fixed processing. Incorporation of more advanced models into the future studies such as recurrent models (e.g., LSTM and GRU), attention-based mechanisms will further be considered and transformer encoders to recapture the long-range temporal dependencies along the speech signal. In addition, multichannel input arrays may support spatial noise cancelling through beamforming, which may improve performance in practice, i.e. a multipath acoustic environment.

Moreover, the usage of live system by hearing-impaired individuals and older adults in various, noisy environments will be examined in a planned field study. Analytic parameters will include both a subjective listening experience and system operative efficiency, and through such parameters, iterative improvements would be made throughout the pipeline and reduce the divide between the laboratory-controlled, controlled-performance and efficacy performance in the real-world.

On the whole, the suggested deep-learning autoencoder model provides a prospective platform of developing efficient, noise invariant, and real-time voice enhancement technology of the future voice-assistant technologies.

# FUTURE IMPLICATIONS

The study yields remarkable results into the next-gen smart speech-augmentation applications, especially within limited resource situations of real time scenario. Due to positive trade-offs between the denoising accuracy and its simple computational footprint, the proposed deep autoencoder architecture provides an effective premise to be used in the developing generation of voice-assistant technologies.

One of the striking future directions includes rollout of the model on specialized hardware platforms such as microcontrollers, FPGA and low-power edge platforms. With small architectural pruning or quantization an on-device model can be run without a significant performance cost, allowing continuous noise reduction in wearables like smart hearing aids, voice-activated glasses, and in-ear assistants.

The design could also be extended to support multilingual and speaker-independent improvement by using transfer- learning and domain-adaptation approachologies. These extensions would enhance usability in the heterogeneous population and in the real-world settings, such as urban transportation, medical institutions, and distance learning situations.

On the clinical perspective, the model might support assistive actions to those with hearing accidents, bounding- processing or neurological problems impairing understanding of speech. It might have a significant impact on their social and professional interaction in a noisy context by real-time improvement of speech clarity.

In addition, the autoencoder produces a latent space, which allows learning the representation of the speech and detect anomalies. Such a representation can be used in downstream applications like emotion recognition, speaker verification or speech-command recognition to create versatile audio interfaces.

In future they can add on top of the architecture time attention mechanisms, recurrent layer (e g GRU, LSTM), or transformers to further capture temporal dependencies in speech. Another promising line to explore is adversarial training methods and generative methods; i.e. GANs, in particular.

Lastly, large-scale deployment of such models will require addressing issues of robustness to unseen noise types, interpretability of the enhancement process, and compliance with data privacy standards, particularly when deployed in healthcare or personal voice assistant systems.

# REFERENCES

1. M. Dua, Akanksha, and S. Dua, “Noise robust automatic speech recognition: review and analysis,” International Journal of Speech Technology **26** (2023), 10.1007/s10772-023-10033-0.
2. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE Transactions on Audio, Speech and Language Processing **22** (2014), 10.1109/TASLP.2014.2304637.
3. Y. Hu, C. Chen, Q. Zhu, and E. S. Chng, “Wav2code: Restore clean speech representations via codebook lookup for noise-robust asr,” IEEE/ACM Transactions on Audio, Speech and Language Processing **32** (2024), 10.1109/TASLP.2023.3332545.
4. S. Jeon and M. S. Kim, “Noise-robust multimodal audio-visual speech recognition system for speech-based interaction applications,” Sensors (Basel, Switzerland) **22** (2022), 10.3390/s22207738.
5. Y. H. Jung, T. X. Pham, D. Issa, H. S. Wang, J. H. Lee, M. Chung, B. Y. Lee, G. Kim, C. D. Yoo, and K. J. Lee, “Deep learning-based noise robust flexible piezoelectric acoustic sensors for speech processing,” Nano Energy **101** (2022), 10.1016/j.nanoen.2022.107610.
6. R. G P, C. C B, D. R. B, and T. Y. G, “Noise estimation based on optimal smoothing and minimum controlled through recursive averaging for speech enhancement,” Intelligent Systems with Applications **21** (2024), 10.1016/j.iswa.2023.200310.
7. Z. Zhu, L. Zhang, K. Pei, and S. Chen, “A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to- noise ratio,” Digital Signal Processing: A Review Journal **141** (2023), 10.1016/j.dsp.2023.104151.
8. G. Vanderreydt and K. Demuynck, “A novel channel estimate for noise robust speech recognition,” Computer Speech and Language **86** (2024), 10.1016/j.csl.2023.101598.
9. S. A. Fuglsang, T. Dau, and J. Hjortkjær, “Noise-robust cortical tracking of attended speech in real-world acoustic scenes,” NeuroImage **156** (2017), 10.1016/j.neuroimage.2017.04.026.
10. E. H. N. Ng and J. Rönnberg, “Hearing aid experience and background noise affect the robust relationship between working memory and speech recognition in noise,” International Journal of Audiology **59** (2020), 10.1080/14992027.2019.1677951.
11. V. Gupta and D. Ather, “Busa deep learning model for eeg signal analysis,” Wireless Personal Communications **136**, 2521–2543 (2024).
12. G. Singh, R. P. George, N. Ahmad, S. Hussain, D. Ather, and R. Kler, “A deep learning approach for evaluating the efficacy and accuracy of posenet for posture detection,” International Journal of System Assurance Engineering and Management, 1–10 (2024).
13. D. Priyanshu, R. L. Khan, R. K. Matahen, and D. Ather, “Artificial intelligence optimization of load scheduling with economic load dispatch in industrial power generating units,” in *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (IEEE, 2022) pp. 1127–1133.
14. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2015) pp. 5206–5210.
15. J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database (demand),” https://zenodo.org/ record/1227121 (2013), accessed: 2025-07-17.
16. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment of telephone networks and codecs,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2 (IEEE, 2001) pp. 749–752.
17. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing **19**, 2125–2136 (2011).