Single-Channel Speech Noise Reduction Using Supervised Learning

*Note: Sub-titles are not captured in Xplore and should not be used

Vivian Philips
Rochester Institute of Technology
Rochester, NY
vp6254@rit.edu

Mohammad Raziuddin Chowdhury
Rochester Institute of Technology
Rochester, NY
mc6044@r.t.edu

Roman Kozulia
Rochester Institute of Technology
Rochester, NY
rk8318@rit.edu3

***Abstract*—Intelligible speech is essential for effective communication and for the performance of modern speech-driven systems such as virtual assistants, teleconferencing platforms, and hearing devices. However, real-world recordings are often contaminated by environmental noise, which reduces intelligibility and degrades downstream speech processing. To address this challenge, this work presents a machine-learning–based speech enhancement system designed to separate clean speech from noisy inputs across diverse acoustic conditions. The system uses a BLSTM-PSA (Bidirectional Long Short-Term Memory with Phase-Sensitive Approximation) model trained on mixtures of clean utterances from the LibriSpeech corpus and noise samples from the DEMAND dataset spanning a wide range of signal-to-noise ratios. On the LibriSpeech test-clean set, the model demonstrates strong improvements in signal quality, with mean SNR gains of approximately 8–10 dB at low input SNRs and progressively smaller gains at higher SNR levels.**

I. Introduction

Humans possess an extraordinary ability to recognize speech even amid multiple overlapping voices and background noise. The task of separating target speech from interfering sounds is often referred to as the "cocktail party problem," a term popularized following Colin Cherry's seminal 1953 work [1], which focused on humans' ability to recognize speech while ignoring background noise. His research is considered the starting point of selective attention studies, which later evolved into various "filter" models.

Exposure to noisy speech has been shown to increase listener fatigue and cognitive load, leading to higher levels of stress and reduced concentration [2].

Noisy speech has a particularly adverse impact on individuals with hearing impairments. Hearing-impaired listeners often exhibit reduced speech intelligibility in noisy environments, not only because they have limited access to acoustic speech cues but also due to a decreased ability to adapt to background noise. This reduced adaptability can contribute to as much as 10% of the speech reception threshold deficit observed in hearing-impaired listeners[3].

The effects of noise on speech perception are even more pronounced in older adults. Research indicates a significant decline in speech recognition in noise as age increases, with measurable impairments beginning as early as the 51–60 year age group and worsening thereafter. Older listeners require higher signal-to-noise ratios (SNRs) to achieve 50% speech recognition accuracy, reflecting diminished ability to extract speech from background noise [4].

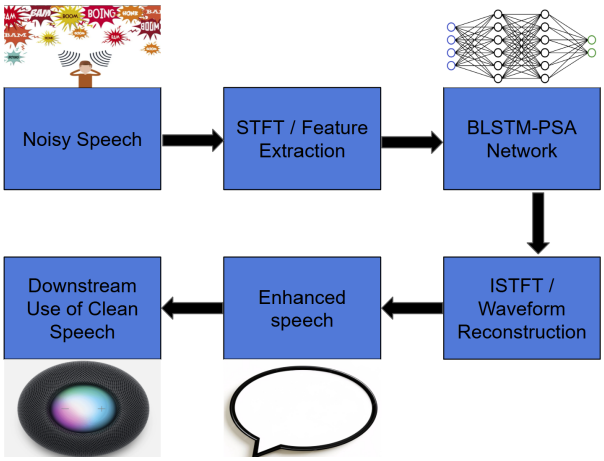

Fig. 1. System Overview

Effective speech noise reduction has become critical across a wide range of modern technologies and industries. Hearing aids and cochlear implants rely heavily on noise suppression algorithms to restore clarity for users in everyday acoustic environments directly influencing communication and quality of life. In mobile communication and teleconferencing platforms such as Zoom or Microsoft Teams, background noise can degrade call quality, reduce productivity, and increase listener fatigue—issues that have become even more prominent with the rise of remote work. Noise reduction also plays a central role in automatic speech recognition systems and smart assistants like Siri and Alexa, where background noise can drastically reduce command accuracy and user satisfaction. Furthermore, robots and human–robot interaction systems increasingly operate in cluttered, noisy environments and require robust speech enhancement to interpret commands reliably. As speech interfaces continue to expand across consumer, medical, and industrial settings, improving noise suppression

is essential to ensuring accessibility, usability, and dependable performance.

The rapid advancement of machine learning techniques has significantly accelerated progress in speech noise reduction. Deep neural networks—including convolutional, recurrent, and more recently transformer-based architectures—have demonstrated substantial improvements over traditional signal processing by learning complex noise patterns directly from data. [5] Modern systems often operate by estimating time–frequency masks or by reconstructing clean waveforms end-to-end that results in more flexible and accurate separation of speech from diverse and highly non-stationary noise sources. Despite these advances, current learning-based approaches still fall short of users' expectations in many real-world settings. Noise reduction systems' performance can degrade in unseen acoustic conditions, in extremely low signal-to-noise ratios, or when trained on limited or biased datasets. Many state-of-the-art models demand considerable computational resources making them impractical for real-time applications such as hearing aids, mobile devices, or embedded systems. As a result, there remains a strong need for continued research to develop noise reduction methods that are more robust, generalizable, lightweight, and capable of delivering consistent improvements in the unpredictable acoustic environments encountered by everyday users.

II. Literature Review

In recent years, the focus of speech enhancement research has shifted from traditional signal processing techniques toward data-driven machine learning approaches. Classical methods such as spectral subtraction, Wiener filtering, and statistical noise estimators rely on explicit mathematical assumptions about noise behavior, which often break down in real-world, highly non-stationary environments. Modern supervised learning approaches instead treat noise reduction as a pattern recognition problem, learning discriminative representations of speech and noise directly from large training corpora. This paradigm shift, accelerated by the introduction of deep neural networks, has led to substantial improvements in intelligibility, perceptual quality, and robustness of noise reduction systems across diverse acoustic conditions. As a result, deep learning–based speech enhancement has become the dominant direction in current research.

Neural networks have gained significant popularity over the past decades for solving a variety of complex problems, including pattern recognition, classification, and regression. However, due to the inherently sequential and time-dependent nature of speech signals, Recurrent Neural Networks (RNNs) appear to be the most suitable choice for speech noise reduction [6]. RNNs include recurrent (feedback) connections that makes the network to treat input samples as elements of a sequence rather than independent observations. Speech has a temporal structure, where each frame is influenced by preceding frames, and RNNs naturally model these temporal dynamics. RNNs introduce a flexible and extensible time dimension that feedforward networks cannot capture, regardless of their depth.

Feature selection is a critical step for training an RNN model for noise removal because the choice of acoustic features directly affects separation performance across different noise and reverberation conditions. The study by Delfarah et al.[7] evaluates a broad range of feature types for masking-based speech separation shows that contextual information substantially improves performance, which aligns with the temporal modeling strengths of RNNs. When feature combinations are considered, the study shows that PNCC+GF+LOG-MEL provides the best overall performance for both matched and unmatched noise. These findings demonstrate that carefully selecting features—particularly complementary sets that capture both spectral and temporal properties—is essential for maximizing the noise-removal capability of RNN-based systems.

Hybrid speech enhancement approaches have emerged as a compelling alternative to purely model-based or purely data-driven methods. [8] While deep neural networks often achieve state-of-the-art results under controlled and well-matched training conditions, their performance can degrade sharply when faced with unexpected or out-of-domain noise. Classical signal-processing techniques, in contrast, offer predictable behavior and strong robustness derived from physical and statistical models, but they lack the flexibility needed to handle highly non-stationary or complex noise environments. [9]

Recurrent models such as Long Short-Term Memory (LSTM) networks are widely used to model temporal dependencies in speech, but standard LSTMs can struggle with very long-range context and often require large models [10]. Recent work has therefore explored architectural innovations to address LSTM's limitations in capturing long-term dependencies. For instance, [11] proposed a variable-neurons LSTM with an hourglass architecture, attention-gated skip connections, and combined feature sets (MFCC, AMS, GFE, RASTA-PLP), achieving a 16.41% STOI improvement on LibriSpeech through a 7-layer architecture. While demonstrating the value of architectural complexity, such approaches significantly increase model size and training requirements.

While these architectural refinements to LSTMs can be effective, another line of work focuses on improving the underlying recurrent formulation and training objectives rather than merely increasing architectural complexity. In this direction, Erdogan et al. [12] introduced significant improvements to deep learning-based speech separation through two key innovations: phase-sensitive objective functions and bidirectional recurrent neural networks. The authors demonstrated that traditional magnitude-only approaches were suboptimal because they ignored the interaction between phase errors and amplitude reconstruction when using noisy phase information. They proposed the Phase-Sensitive Approximation (PSA) loss function, which directly minimizes the error in the complex

spectrum domain:

$$L_{\mathrm{PSA}} = \|S_{\mathrm{clean}} - M \cdot S_{\mathrm{noisy}} \cdot \cos(\theta_{\mathrm{clean}} - \theta_{\mathrm{noisy}})\|^2 . \quad (1)$$

This formulation allows the predicted time–frequency mask M to compensate for phase discrepancies by targeting amplitudes that maximize SNR rather than simply approximating clean speech magnitudes. Oracle experiments on the CHiME-2 benchmark showed that phase-sensitive filtering outperformed the ideal ratio mask (IRM) by approximately 2 dB in SDR. The PSA objective consistently outperformed magnitude spectrum approximation (MSA) across all SNR conditions. This confirms that phase information into the loss function yields more effective speech enhancement.

The second major contribution, introduced by Erdogan et al., was the adoption of Bidirectional Long Short-Term Memory (BLSTM) networks for modeling temporal dynamics in speech separation [12]. While previous works demonstrated that unidirectional LSTMs outperformed deep feedforward networks by capturing temporal dependencies, Erdogan et al. showed that BLSTMs provided further gains by leveraging both past and future context. The bidirectional architecture processes input sequences in both forward and backward directions, with hidden states from both directions concatenated at each timestep before mask prediction. This enables each frame's enhancement to benefit from contextual information spanning the entire utterance. The combination of PSA loss with BLSTM architecture established a strong foundation for subsequent speech enhancement research, achieving state-of-the-art results that surpassed non-negative matrix factorization (NMF) methods by at least 2.8 dB across all SNR conditions [12].

Building on the BLSTM dirrection, Pashaian and Seyedin (2024) proposed the Constrained Phase-Sensitive Ideal Ratio Mask (cPSIRM), a masking strategy that incorporates both magnitude and phase information while ensuring mask values remain physically meaningful using a Rectified Linear Unit (ReLU)-based phase constraint [13]. A key innovation in their design was to integrate the Fast Fourier Transform (FFT) and Inverse FFT (IFFT) directly into a Convolutional Recurrent Network (CRN). This allowed the model to estimate mask values with direct consideration of the final time-domain waveform, rather than handling frequency analysis as a separate preprocessing step. They further replaced traditional Non-Negative Matrix Factorization (NMF)-based decoder structures with deep autoencoder layers, forming a two-stage system: the CRN first produces an initial speech/noise separation, followed by a joint deep neural network (DNN) decoder that captures harmonic structure to refine speech quality.

Weninger et al. [14] expanded upon the contributions of Erdogan et al. [8] by showing that Long Short-Term Memory-based Deep Recurrent Neural Networks (LSTM-DRNNs) trained with the Phase-Sensitive Approximation (PSA) loss achieve significantly better performance than earlier single-channel speech enhancement methods. While Erdogan et al. introduced PSA as a way to incorporate phase information into mask estimation, Weninger et al. demonstrated that applying PSA within a Bidirectional LSTM (BLSTM) architecture further improves reconstruction quality by leveraging both past and future temporal context during mask prediction. This BLSTM-PSA framework has since become a reference baseline for deep learning–based speech enhancement. The system developed in this work adopts this foundational architecture as its starting point and extends it with additional improvements tailored to our experimental setup.

III. Method

A. Data Preparation

The proposed speech enhancement system was trained and evaluated using the LibriSpeech corpus [15] for clean speech samples and the DEMAND [16] database for background noise. LibriSpeech is a large-scale corpus of read English speech sampled at 16 kHz, derived from audiobooks in the LibriVox project. For this study, three subsets were utilized: train-clean-100 for training (28,539 utterances), dev-clean for validation (2,703 utterances), and test-clean for testing (2,620 utterances). These subsets provide speaker-independent evaluation, as speakers in the test set do not appear in the training set. The DEMAND database contains diverse environmental recordings captured in 18 different scenarios, including domestic environments (kitchen, living room, washing machine), nature settings (field, park, river), office spaces (hallway, meeting room, office), public areas (cafeteria, restaurant, station, square), and transportation contexts (bus, car, metro). All noise recordings were resampled to 16 kHz to match the LibriSpeech sampling rate.

To generate noisy training mixtures, clean speech utterances were artificially degraded by adding noise at five signal-to-noise ratio (SNR) levels of 0, 5, 10, 15, and 20 dB. The SNR-controlled mixing process followed a standard energy-based approach. For a given clean speech signal $s(t)$ and noise signal $d(t)$, the noise was first scaled to achieve the target SNR before being added to the clean speech. The scaling factor α was computed as:

$$\alpha = \frac{\mathrm{RMS}(s)}{\mathrm{RMS}(d)} \cdot 10^{-\frac{\mathrm{SNR}}{20}}$$

where $\mathrm{RMS}(\cdot)$ denotes the root-mean-square energy of the signal. The noisy mixture $y(t)$ was then constructed as

$$y(t) = s(t) + \alpha \cdot d(t).$$

This formulation ensures that the desired SNR relationship is maintained:

$$\mathrm{SNR} = 20 \log_{10} \left(\frac{\mathrm{RMS}(s)}{\mathrm{RMS}(\alpha \cdot d)} \right).$$

For each training utterance, a noise type was randomly selected from the DEMAND database, and a random SNR level was chosen from the five available options. All audio segments were processed using fixed-length segmentation with a duration of 3.0 seconds (48,000 samples at 16 kHz). If a clean speech utterance was shorter than 3.0 seconds, it was zero-padded to the required length. For longer utterances, a random

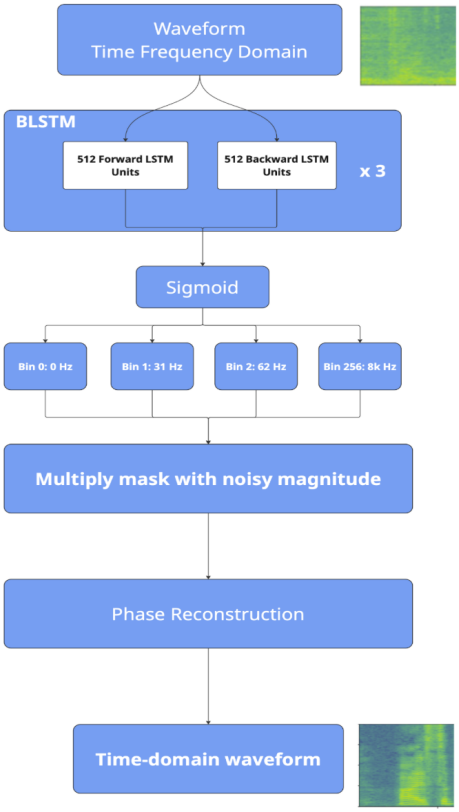

Fig. 2. System Architecture

3-second segment was extracted during each training epoch to provide variability and prevent overfitting. Noise signals were similarly processed, with repetition applied when the noise recording was shorter than the required duration. This data preparation strategy produced a diverse training corpus spanning multiple speakers, acoustic environments, and noise conditions, allowing the model to generalize effectively to unseen speakers and noise types during evaluation.

B. Model Architecture

Our speech enhancement system implements a five-stage processing pipeline based on the BLSTM-PSA (Bidirectional Long Short-Term Memory with Phase-Sensitive Approximation) architecture proposed by Weninger et al[11]. This architecture operates in the time–frequency domain, leveraging the ability of recurrent networks to model temporal dependencies in spectral representations.

1) Stage 1: Short-Time Fourier Transform (STFT): The input waveform is first transformed into the time–frequency domain using the STFT, which decomposes audio into its constituent frequencies over time by applying the Fourier transform to overlapping, windowed segments of the signal. We employ an FFT size of 512 samples (32 ms at 16 kHz), a hop length of 128 samples (8 ms), yielding 75% overlap between consecutive frames, and a Hann window to minimize spectral leakage caused by abrupt discontinuities at frame boundaries. This configuration produces 257 frequency bins spanning 0–8 kHz with 31.25 Hz resolution.

The complex-valued STFT output $Y(f,t)$ is decomposed into magnitude and phase components:

$$|Y(f,t)| \quad \text{and} \quad \angle Y(f,t),$$

where $Y(f,t)$ represents the noisy STFT at frequency bin f and time frame t, $|Y(f,t)|$ denotes the magnitude spectrogram, and $\angle Y(f,t)$ denotes the phase spectrogram.

2) Stage 2: Bidirectional LSTM Processing: The magnitude spectrogram $|Y(f,t)|$ serves as input to a three-layer bidirectional LSTM network with 512 hidden units per direction (1,024 total). Bidirectionality enables each frame to incorporate context from both preceding and following frames, aiding in distinguishing speech from noise and resolving transient speech events. Dropout regularization (rate $= 0.2$) is applied between layers.

The LSTM processes the sequence $\{|Y(f,t)|\}_{t=1}^{T}$, where T is the total number of time frames, and outputs hidden representations $\{h_t\}_{t=1}^{T}$, where h_t encodes learned spectro-temporal patterns at time frame t.

3) Stage 3: Mask Prediction: The LSTM outputs h_t are passed through a fully connected layer followed by a sigmoid activation to produce a time–frequency mask

$$M(f,t) \in [0,1],$$

where $M(f,t)$ represents the estimated ratio of speech to total signal energy at frequency bin f and time frame t. Values near 1 indicate speech-dominated regions, whereas values near 0 indicate noise-dominated regions. This soft mask preserves weak speech components that binary masks would suppress.

4) Stage 4: Spectral Masking: The predicted mask is applied element-wise to the noisy magnitude spectrogram to produce the enhanced magnitude:

$$\hat{S}(f,t) = M(f,t) \cdot |Y(f,t)|.$$

The original phase $\angle Y(f,t)$ is preserved and combined with the enhanced magnitude to form the enhanced complex spectrogram:

$$\hat{S}(f,t) \cdot e^{j\,\angle Y(f,t)},$$

where $j = \sqrt{-1}$ is the imaginary unit.

The Short-Time Fourier Transform (STFT) produces a complex-valued representation of audio, where each point contains both magnitude and phase information. The phase represents the timing or position of frequency components and is essential for accurately reconstructing the audio signal. When applying the mask, the enhanced magnitude is combined with the original phase to preserve this timing information.

5) Stage 5: Inverse Short-Time Fourier Transform (ISTFT): The enhanced complex spectrogram is transformed back to the time domain via the ISTFT, producing the waveform $\hat{s}(t)$. The ISTFT applies the inverse Fourier transform to each frame and reconstructs the signal using overlap-add synthesis.

C. Loss Function and SNR-Aware Training

We employ the Phase-Sensitive Approximation (PSA) loss proposed by Erdogan et al. [9], which minimizes the squared error between the clean magnitude spectrogram and the phase-corrected masked noisy spectrogram:

$$L_{\text{PSA}} = \|S_{\text{clean}}(f,t) - M(f,t) \cdot |Y(f,t)| \cdot \cos(\theta_{\text{clean}}(f,t) - \theta_{\text{noisy}}(f,t))\|^2 \tag{2}$$

Here, $S_{\text{clean}}(f,t)$ denotes the clean magnitude spectrogram, $M(f,t)$ is the predicted mask, $|Y(f,t)|$ is the noisy magnitude, and $\theta_{\text{clean}}(f,t)$ and $\theta_{\text{noisy}}(f,t)$ represent the clean and noisy phase spectra, respectively. The cosine term compensates for phase differences and yields a more accurate training target than magnitude-only objectives such as the Signal Approximation (SA) loss.

The PSA loss measures how close the estimated clean speech is to the true clean speech in the frequency domain by comparing both magnitude and phase. It applies the predicted mask to the noisy magnitude but adjusts it using the phase difference between clean and noisy signals. This phase correction helps the model better capture important timing information, resulting in more accurate and natural speech reconstruction compared to magnitude-only losses.

SNR-Aware Regularization: To mitigate over-suppression at high input SNR levels—where aggressive denoising may distort the speech signal—we introduce an SNR-aware regularization term that adapts the objective based on input signal quality. For utterances with SNR $>$ 12 dB, we add the following penalty:

$$L_{\text{penalty}} = w_{\text{SNR}} \cdot \frac{1}{FT} \sum_{f,t} \max(0,\, M_{\text{target}} - M(f,t))^2\,, \tag{3}$$

where F and T denote the number of frequency bins and time frames. The target mask threshold is fixed at

$$M_{\text{target}} = 0.85.$$

The SNR-dependent penalty weight is defined as

$$w_{\text{SNR}} = \text{clamp}\left(\frac{\text{SNR}_{\text{in}} - 12}{10},\, 0,\, 1\right), \tag{4}$$

such that $w_{\text{SNR}} = 0$ at SNR $= 12$ dB, $w_{\text{SNR}} = 0.5$ at SNR $= 17$ dB, and $w_{\text{SNR}} = 1$ for SNR ≥ 22 dB.

The overall loss function is given by

$$L_{\text{total}} = L_{\text{PSA}} + \alpha \cdot w_{\text{SNR}} \cdot L_{\text{penalty}}, \tag{5}$$

with $\alpha = 0.5$ controlling the relative weight of the regularization term. This framework provides strong denoising at low SNRs (0–12 dB), while avoiding unnecessary suppression at higher SNRs (15–20 dB), thereby preserving speech quality in cleaner conditions.

TABLE I
TRAINING HYPERPARAMETERS USED

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Scheduler	ReduceLROnPlateau
ReduceLROnPlateau	factor 0.5 patience 2
Batch Size	16 (3-second segments)
Epochs	10
Loss Function	Phase-sensitive loss + SNR-aware penalty
Gradient Clipping	Max-norm = 5.0
STFT Window Length	512 samples
Hop Length	128 samples
FFT Size	512
Model Architecture	3-layer BLSTM, 512 units per direction
Dropout	0.2 (between LSTM layers)
Validation Set	LibriSpeech dev-clean
Model Selection	Best model based on validation loss

D. Training Procedure

The model was trained using the Adam optimizer with an initial learning rate of $\eta_0 = 1 \times 10^{-4}$. To adaptively reduce the learning rate when the training loss plateaued, a `ReduceLROnPlateau` scheduler was employed, which halves the learning rate if the validation loss does not improve for two consecutive epochs. Formally, the update rule is

$$\eta_{t+1} = \begin{cases} 0.5\,\eta_t, & \text{if } \mathcal{L}_{\text{val}}^{(t)} \geq \mathcal{L}_{\text{val}}^{(t-1)} \text{ and } \mathcal{L}_{\text{val}}^{(t-1)} \geq \mathcal{L}_{\text{val}}^{(t-2)}, \\ \eta_t, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $\mathcal{L}_{\text{val}}^{(t)}$ represents the validation loss at epoch t. Training was performed for 10 epochs with a batch size of 16 utterances, resulting in approximately 1,784 batches per epoch from the 28,539-utterance training set. Each 3-second audio segment was randomly cropped from full LibriSpeech utterances and mixed with noise at randomly sampled SNR levels from $\{0,5,10,15,20\}$ dB.

Gradient Stabilization: To prevent exploding gradients—a common challenge when training recurrent neural networks with long temporal dependencies—we applied gradient clipping with a maximum ℓ_2-norm of 5.0. Specifically, gradients g are scaled down when their norm exceeds this threshold, as defined by

$$\mathbf{g}_{\text{clipped}} = \begin{cases} \mathbf{g}, & \text{if } \|\mathbf{g}\| \leq 5.0, \\ 5.0 \cdot \dfrac{\mathbf{g}}{\|\mathbf{g}\|}, & \text{otherwise,} \end{cases} \tag{7}$$

This technique maintains numerical stability during backpropagation through the three-layer bidirectional LSTM while preserving meaningful gradient updates.

Model Selection: During training, we tracked the model's performance on a separate validation set after each epoch by measuring the validation loss. Whenever the validation loss improved—that is, decreased compared to previous epochs—we saved a checkpoint of the model. After training

completed, we selected the final model as the one corresponding to the lowest validation loss observed, ensuring the best generalization to unseen data.

E. Evaluation Metrics

Model performance was evaluated using SNR improvement (ΔSNR) computed on the LibriSpeech `test-clean` subset. The metric quantifies the enhancement in signal-to-noise ratio achieved by the denoising system:

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}, \tag{8}$$

where the input and output SNR values are defined as

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right). \tag{9}$$

Here, $P_{\text{signal}} = \mathbb{E}[\, s^2_{\text{clean}}(t)\,]$ denotes the clean speech power.

$$P_{\text{noise}} = \mathbb{E}\big[(\, s_{\text{noisy}}(t) - s_{\text{clean}}(t)\,)^2\big]$$

A small stability constant $\epsilon = 10^{-8}$ was added to all power computations to avoid numerical issues. ΔSNR is expressed in decibels (dB), where positive values indicate noise reduction and negative values indicate degradation.

Evaluation was conducted separately for each input SNR level in the test set $\{0, 5, 10, 15, 20\}$ dB to assess robustness across noise conditions. For each SNR level, we computed the mean ΔSNR, standard deviation, and sample count across all test utterances. The test set consisted of 2,620 utterances from LibriSpeech `test-clean`, each mixed with randomly selected DEMAND noise segments at the specified SNR levels.

IV. Results

Table II summarizes the enhancement performance of the model across five input SNR conditions (0–20 dB). The largest improvement was observed at 0 dB input SNR, where the system achieved a mean ΔSNR of 13.277 dB (SD = 6.018, $n = 533$). Performance remained strong at 5 dB and 10 dB, with mean improvements of 11.126 dB and 8.005 dB, respectively. As input SNR increased, the magnitude of enhancement gradually declined, with mean ΔSNR values of 5.308 dB at 15 dB and 3.123 dB at 20 dB. This pattern is consistent with the expectation that cleaner inputs offer less opportunity for improvement. Standard deviations were largest at lower SNRs, reflecting greater variability in denoising difficulty under more severe noise conditions.

TABLE II
Evaluation Results: Mean ΔSNR, Standard Deviation, and Sample Count across Input SNR Levels on LibriSpeech test-clean

Input SNR (dB)	Mean ΔSNR (dB)	Std Dev (dB)	Sample Count
0	13.277	6.018	533
5	11.126	5.353	526
10	8.005	4.232	495
15	5.308	3.548	541
20	3.123	3.113	525

From Table II we observe that across all tested conditions, the proposed system:

- Improves SNR reliably for every input SNR level
- Maintains speech structure while removing broadband and transient noise components
- Generalizes across >2,600 test samples, demonstrating stable performance

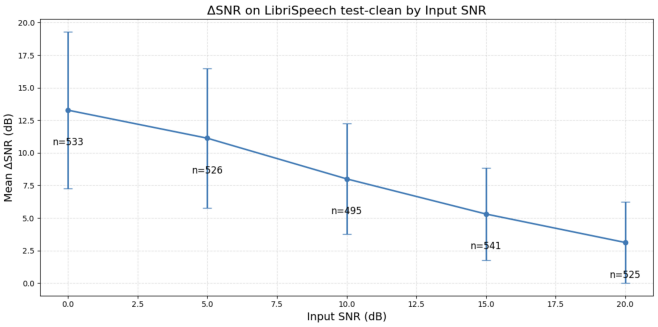

Fig. 3. Mean SNR improvement achieved by the model on test-clean set across input SNR levels from 0 to 20 dB. Error bars represent standard deviation, and the number of test samples (n) for each SNR condition is shown below each point.

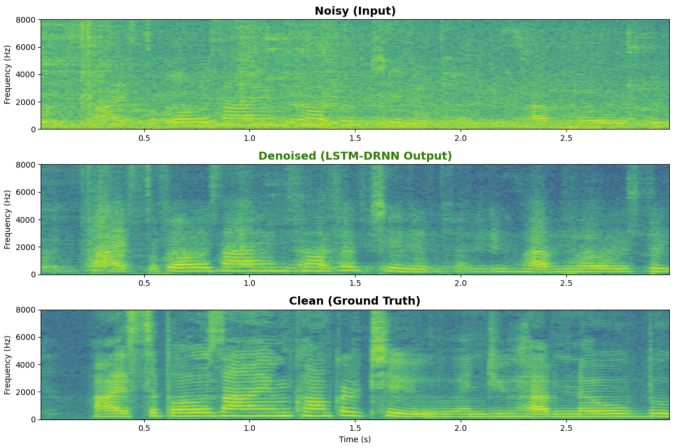

Fig. 4. Comparison of magnitude spectrograms before and after noise removal. The enhanced spectrogram shows significant suppression of noise components while preserving speech structures.

Figure 4 presents a qualitative comparison of the noisy input spectrogram, the model's denoised output, and the clean ground-truth waveform for a representative 3-second utterance.

The noisy input spectrogram shows energy smearing across both low and high frequencies, typical of real-world environmental noise. Compared with the clean ground truth, the reconstructed spectrogram preserves most major speech characteristics, although some residual noise and minor spectral smoothing are still visible. These qualitative patterns are consistent with the quantitative SNR improvements reported above.

Ethical Concerns

Enhanced speech could unintentionally improve the intelligibility of recordings captured without consent, raising con-

cerns about surveillance and unauthorized monitoring. Care must also be taken to ensure that models trained on publicly available datasets, such as LibriSpeech, do not inadvertently embed or leak sensitive information. Finally, deploying speech enhancement systems in real-world applications—such as teleconferencing, healthcare settings, or assistive technologies—should prioritize user transparency, informed consent, and secure handling of audio data to protect individual rights and maintain trust.

Discussion, Limitations and Future Direction

This study has several limitations. First, the system operates exclusively on single-channel audio and therefore does not leverage spatial cues available in multi-microphone arrays. Second, model performance was evaluated using only SNR change; more perceptually relevant metrics such as PESQ, STOI, or human listening tests were not included. The BLSTM architecture, while effective, is computationally heavy and not yet optimized for real-time or on-device processing. Finally, the model was trained on read speech from LibriSpeech, and its generalizability to spontaneous or conversational speech remains untested.

Future work will incorporate perceptual evaluation metrics, including PESQ, STOI, and subjective listening studies, to better assess real-world speech quality. Lighter and more efficient architectures—such as CRN models, Conv-TasNet, or compact transformer variants—will be explored to enable real-time deployment. Additionally, domain adaptation experiments will extend the model to conversational speech, other languages, and a wider range of acoustic environments.

V. Conclusion

This study demonstrates that an BLSTM-PSA speech enhancement system can substantially improve the quality of noisy single-channel speech across a wide range of input SNR levels. The model achieved its highest gains at moderate noise conditions (5–10 dB), while still providing meaningful improvements even when the input was already relatively clean. Qualitative spectrogram analysis further confirmed that the network effectively suppresses broadband noise while preserving key speech structures. Although current evaluation relied primarily on SNR change, the results collectively indicate that supervised, mask-based enhancement remains a strong approach for practical denoising tasks. Continued development—incorporating perceptual metrics, lighter architectures, and broader speech domains—will help advance this system toward real-world, real-time applications.

References

[1] Cherry EC. Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America. 1953;25:975–979.

[2] J. Kristiansen, S. P. Lund, R. Persson, H. Shibuya, P. M. Nielsen, and M. Scholz, “A study of classroom acoustics and school teachers’ noise exposure, voice load and speaking time during teaching, and the effects on vocal and mental fatigue development,” *Int. Arch. Occup. Environ. Health*, vol. 87, no. 8, pp. 851–860, Nov. 2014, doi: 10.1007/s00420-014-0927-8.

[3] M. I. Marrufo-Pérez, M. J. Fumero, A. Eustaquio-Martín, *et al.*, “Impaired noise adaptation contributes to speech intelligibility problems in people with hearing loss,” *Sci. Rep.*, vol. 14, p. 8, 2024.

[4] T. Rahne, T. M. Wagner, A. C. Kopsch, S. K. Plontke, and L. Wagner, “Influence of age on speech recognition in noise and hearing effort in listeners with age-related hearing loss,” J. Clin. Med., vol. 12, no. 19, p. 6133, 2023, doi: 10.3390/jcm12196133.

[5] Y. H. Lai et al., “Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients,” Ear and Hearing, vol. 39, no. 4, pp. 795–809, 2018.

[6] Wang and J. Chen, “Supervised Speech Separation Based on Deep Learning: An Overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702–1726, Oct. 2018, doi: 10.1109/TASLP.2018.2842159.

[7] M. Delfarah and D. Wang, “Features for Masking-Based Monaural Speech Separation in Reverberant Conditions,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 1085–1094, May 2017, doi: 10.1109/TASLP.2017.2687829.

[8] R. Haeb-Umbach, T. Nakatani, M. Delcroix, C. Boeddeker and T. Ochiai, “Microphone Array Signal Processing and Deep Learning for Speech Enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering,” IEEE Signal Processing Magazine, vol. 41, no. 6, p. 23, Nov. 2024, doi: 10.1109/MSP.2024.3451653.

[9] Gómez-Vilda, P., Gómez-Rodellar, A. (2023). Data-Driven Vs Model-Driven Approaches in Cognitive Speech Processing. In: Lopez-Soto, T., Garcia-Lopez, A., Salguero-Lamillar, F.J. (eds) The Theory of Mind Under Scrutiny. Logic, Argumentation & Reasoning, vol 34. Springer, Cham. https://doi.org/10.1007/978-3-031-46742-4$_2$1

[10] A. Shewalkar, “Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU,” J. Artif. Intell. Soft Comput. Res., vol. 9, no. 4, pp. 235–245, 2019, doi: 10.2478/jaiscr-2019-0006.

[11] J. Wang, N. Saleem, and T. S. Gunawan, “Towards efficient recurrent architectures: A deep LSTM neural network applied to speech enhancement and recognition,” Cognitive Computation, vol. 16, pp. 1221–1236, 2024, doi: 10.1007/s12559-024-10288-y.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 708–712, doi: 10.1109/ICASSP.2015.7178061.

[13] M. Pashaian and S. Seyedin, “Speech Enhancement Based on a Joint Two-Stage CRN+DNN-DEC Model and a New Constrained Phase-Sensitive Magnitude Ratio Mask,” IEEE Access, vol. 12, pp. 98567–98583, 2024, doi: 10.1109/ACCESS.2024.3427854.

[14] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in Proc. Lecture Notes in Computer Science, vol. 9237, 2015, pp. —. doi: 10.1007/978-3-319-22482-4$_1$1.

[15] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, ”Librispeech: An ASR corpus based on public domain audio books,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[16] J. Thiemann, N. Itoand E. Vincent, “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments”. Zenodo, Jun. 09, 2013. doi: 10.5281/zenodo.1227121.