Single-Channel Speech Noise Reduction Using Supervised Learning

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

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Roman Kozulia
dept. name of organization (of Aff.)
name of organization (of Aff.)
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. This work presents a machine learning–based speech enhancement system capable of separating clean speech from noisy inputs in diverse acoustic conditions. The machine learning model we used is trained on mixtures of clean utterances from the LibriSpeech corpus and noise samples from the DEMAND dataset across a range of signal-to-noise ratios.**

***Index Terms*—component, formatting, style, styling, insert**

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, often referred to as the "cocktail party problem" following Colin Cherry's seminal 1953 work [1], where he focused on human's ability to recognize speech while ignoring background noises. His work is considered the starting point of research into selective attention which later involved into "filter" models.

The ability to recognizing speech in noisy environments becomes noticeably fragile for listeners with hearing impairments caused by physical limitations or the effects of aging. Such challenges highlight the increased difficulty these individuals face in understanding speech amid background noise, often leading to reduced communication effectiveness and greater listening effort.

Moreover, 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. This decline is exacerbated depending on the type of noise, with certain modulated noise signals (such as icra5 noise) revealing greater difficulties due to the reduced capacity for 'gap listening'—the ability to utilize brief pauses in noise to understand speech. [4].

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. Modern systems often operate by estimating time–frequency

masks or by reconstructing clean waveforms end-to-end, enabling 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. Their 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(6-12 sources)

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 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[5]. RNNs allow recurrent (feedback) connections, allowing the network to treat input samples as part of a sequence rather than independent observations. A has a temporal structure, where each frame is influenced by preceding frames, and RNNs naturally model these temporal dynamics. In this sense, 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. Prior work[6] evaluating 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, whereas PNCC+GFCC+LOG-MEL is most effective for cochannel separation. 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. 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. Hybrid systems combine the strengths of both: neural networks provide powerful discriminative capabilities for estimating masks or speech characteristics, while model-based components impose physically meaningful constraints that stabilize performance across diverse acoustic conditions. As a result, hybrid methods frequently deliver superior real-world robustness and reliability, making them an increasingly attractive direction for modern speech enhancement research[7].

Recent work has explored architectural innovations to address LSTM's limitations in capturing long-term dependencies. [8] proposed a variable-neurons LSTM with hourglass architecture, attention-gated skip connections, and combined feature sets (MFCC, AMS, GFE, RASTA-PLP), achieving 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.

Erdogan et al. [9] 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:

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

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, confirming that incorporating phase information into the loss function yields more effective speech enhancement [9].

The second major contribution was the adoption of Bidirectional Long Short-Term Memory (BLSTM) networks for modeling temporal dynamics in speech separation. 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. Their experiments showed that BLSTM networks with 384 nodes per layer improved performance by 0.4 dB over comparable unidirectional LSTM baselines on the CHiME-2 dataset containing highly non-stationary noise sources. The authors also explored integrating automatic speech recognition (ASR) outputs as additional input features, demonstrating that phone-level and state-level alignments could further improve separation performance, pointing toward tighter integration between recognition and enhancement as a promising research direction. 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[9].

Deep learning has transformed speech enhancement through sophisticated approaches to time-frequency masking and joint modeling of speech and noise. Bidirectional LSTM networks leverage both past and future temporal context for mask prediction. Pashaian and Seyedin (2024) introduced the constrained phase-sensitive magnitude ratio mask (cPSIRM), which incorporates magnitude and phase information while maintaining bounded values through a ReLU-based phase constraint. Their approach integrated FFT and IFFT transformations directly into a convolutional recurrent network (CRN), allowing the network to estimate mask values with respect to final time-domain signals rather than treating frequency-domain analysis as separate pre-processing. This was combined with a non-linear decoder structure using deep autoencoder layers as alternatives to traditional NMF basis matrices, creating a two-stage architecture where CRN-based masking provides initial speech/noise separation and a joint DNN-decoder performs enhancement by capturing harmonic structures. A hierarchical four-step training strategy progressively mapped from spectral masks to time-domain signals to encoded features to final spectral output, with each step serving as pre-training for subsequent stages. This approach achieved superior performance over single-stage methods and prior two-stage approaches by explicitly incorporating spectral structure knowledge and mask estimation as intermediate training targets within a unified framework[10].

Bidirectional LSTM (BLSTM) networks have proven effective for speech enhancement through their capacity to capture temporal dynamics in both directions. Unlike traditional unidirectional LSTMs that process sequences only forward in time, BLSTMs consist of two separate LSTM layers—one processing the input sequence forward and another processing it backward—with their outputs concatenated at each time step. This bidirectional architecture allows the network to access both past and future context when making predictions at any point in the sequence, which is particularly valuable for speech enhancement where noise characteristics may become apparent only when considering surrounding temporal context. Weninger et al. [11] demonstrated that LSTM-based deep recurrent neural networks (LSTM-DRNN) trained with discriminative objectives substantially outperform traditional methods for single-channel speech enhancement. Their approach used time-frequency masking on magnitude spectrograms, with the network predicting soft masks applied to noisy speech for source reconstruction. They introduced phase-sensitive approximation (PSA) training, where the loss function accounts for phase differences between clean and noisy speech. This phase-aware objective was originally developed by Erdogan et al. [8]. This BLSTM-PSA architecture forms the foundation for the noise removal system developed in this work, extended with SNR-aware training to adaptively adjust denoising aggressiveness based on input noise levels.

III. Methods(2-3)pages

A. Data Preparation

The proposed speech enhancement system was trained and evaluated using the LibriSpeech corpus for clean speech samples and the DEMAND 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:

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

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 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. [**?**]. This architecture operates in the time–frequency domain, leveraging the ability of recurrent networks to model temporal dependencies in spectral representations.

C. 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.

D. 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.

E. 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.

F. 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.

G. 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. The ISTFT parameters (FFT size, hop length, window function) match those used in the forward STFT to ensure perfect reconstruction in the absence of modifications. This yields the final enhanced waveform at 16 kHz.

H. Loss Function and SNR-Aware Training

We employ the Phase-Sensitive Approximation (PSA) loss proposed in [**?**], which minimizes the squared error between the clean magnitude spectrogram and the phase-corrected masked noisy spectrogram:

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

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.

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:

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

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

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

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

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

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.

I. Training Procedure

The model was trained using the Adam optimizer with an initial learning rate $\eta_0 = 1 \times 10^{-4}$, where η denotes the learning rate. A ReduceLROnPlateau scheduler was employed to adaptively decrease the learning rate by a factor of 0.5 whenever the validation loss plateaued for two consecutive epochs. The update rule is defined as

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

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, corresponding to approximately 1,784 batches per epoch from the 28,539-utterance training set. Each 3-second audio segment was randomly cropped from the 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 issue in recurrent architectures with long temporal dependencies—we applied gradient clipping with a maximum ℓ_2-norm of 5.0:

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

where $\mathbf{g}$ denotes the gradient vector and $\|\mathbf{g}\|$ its ℓ_2-norm. This ensures numerical stability during backpropagation through the three-layer bidirectional LSTM while preserving effective gradient flow.

Model Selection: Validation loss was monitored after each epoch, and model checkpoints were saved whenever an improvement was observed. The final model corresponds to the checkpoint achieving the lowest validation loss across all epochs.

J. 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:

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}, \quad (8)$$

where the input and output SNR values are defined as

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

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

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

represents the noise power. 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 I presents the mean ΔSNR achieved by the BLSTM-PSA model with SNR-aware training across different input SNR levels on the LibriSpeech `test-clean` dataset. The model demonstrates consistent noise reduction across all tested conditions, with performance scaling inversely with input SNR—greater improvements are observed at lower (noisier) input SNR levels, while more conservative enhancement is applied at higher (cleaner) input SNR levels.

TABLE I
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	7.911	4.132	518
5	9.950	5.036	486
10	9.251	5.074	516
15	5.559	3.846	532
20	3.227	3.427	568

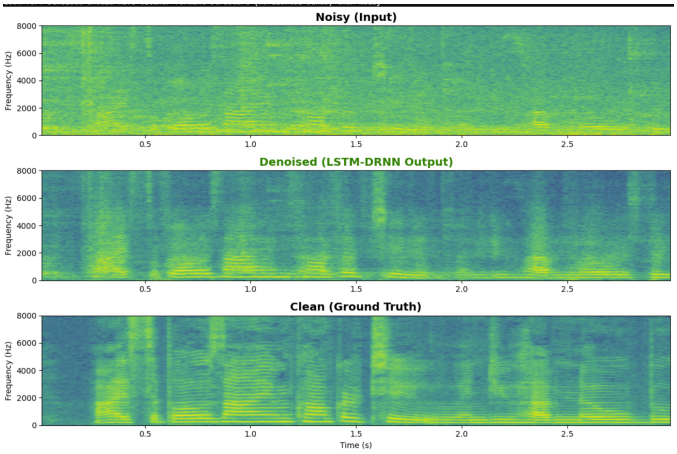

Fig. 1. Comparison of magnitude spectrograms before and after noise removal. The enhanced spectrogram shows significant suppression of noise components while preserving speech structures.

FUTURE PLANS

ETHICAL CONCERNS

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.