Implementation of Machine Learning Methods Based on LSTM and STFT Architectures for Speech Signal Quality Enhancement

Gabriel Ruben Weslie
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
gabriel.weslie@binus.ac.id

Jordan Elishua Wibowo
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
jordan.wibowo001@binus.ac.id

Rizky Kresnanto Dananjaya
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
rizky.dananjaya@binus.ac.id

Takeshi Gobstan Lee
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
takeshi.lee@binus.ac.id

Winda Astuti, S.T., M.Sc., Ph.d.
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
winda.astuti@binus.ac.id

Dr. Nur Afny Catur Andryani, S.Si., M.Sc.
BINUS ASO School of Engineering
Tangerang Selatan, Indonesia
nur.afny@binus.ac.id

Abstract*— Noise interference is a common problem in speech communication systems that degrades audio clarity and quality. This study aims to enhance speech signal quality by applying a noise reduction approach based on the Short-Time Fourier Transform (STFT) and the Inverse Short-Time Fourier Transform (ISTFT), combined with a Long Short-Term Memory (LSTM) architecture. The experiments utilize the VoiceBank-DEMAND dataset, which consists of paired clean and noisy speech signals. The process begins with data preprocessing, followed by feature extraction using STFT, modeling with an LSTM network, and audio signal reconstruction through ISTFT. Performance evaluation is conducted objectively using the Signal-to-Noise Ratio (SNR) metric and visually through spectrogram analysis. Experimental results demonstrate that the proposed system achieves an average SNR improvement of 6.32 dB, indicating effective noise reduction without degrading the harmonic structure of human speech. Furthermore, generalization tests on out-of-dataset data show that the model shows good generalization to various noise conditions. Therefore, the proposed method is effective and potentially suitable for real-time speech communication applications to improve audio quality in noisy environments.

***Keywords* —** *STFT, LSTM, Noise Reduction, Spectral Domain, ISTFT*

I. Introduction

Noise interference is a common problem in speech communication systems, particularly in telephone conversations and real-time audio communication. Undesired noise can degrade speech intelligibility and reduce the overall quality of communication between users [1]. Therefore, an effective method is required to reduce noise without degrading the essential information contained in the speech signal.

Various noise reduction techniques have been developed in the field of speech signal processing, one of which utilizes signal transformation into the frequency domain through the Short-Time Fourier Transform (STFT) [2]. The STFT method enables the analysis of non-stationary signals by segmenting the signal into short time frames, allowing the characteristics of noise and the desired speech signal to be more effectively separated.

In this study, a noise reduction method based on STFT and the Inverse Short-Time Fourier Transform (ISTFT) is implemented to suppress noise interference in speech signals. This approach is expected to enhance speech clarity and be applicable to speech communication systems such as telephone calls and real-time audio applications [3].

II. Theoretical Background

A. Digital Signal Processing

Digital signal processing is a data processing stage in which analog signals are sampled and converted into digital signals [2]. Analog signals are continuous in nature, meaning that their variations are smooth and uninterrupted. In contrast, digital signals are discrete, characterized by abrupt changes and represented by numerical values, typically in binary form (0 and 1).

Digital signal processing can be applied to sample various types of analog signals, such as speech, light, and electrical voltage [2]. The sampled analog signals are converted into digital form at a specific sampling frequency. These digital signal values are then processed so that they can be efficiently interpreted and manipulated by computational systems.

1. Framing

Framing refers to the partitioning of an analog signal into discrete, uniform segments known as frames. Each frame represents a short-duration interval derived from the signal through equidistant temporal division. This process is a fundamental prerequisite for executing feature extraction on analog signals [1]

2. Fourier Transform

The Fourier Transform (FT) is a mathematical technique employed to transform a signal from the time domain into the

frequency domain [2]. It is utilized to analyze the frequency components and their respective amplitudes within a signal. As an integral component of signal processing, the FT facilitates various signal conversions and feature extraction processes.

In the context of Digital Signal Processing (DSP), the FT is applied to discrete signals, enabling computational analysis and the processing of signal values. The Discrete Fourier Transform (DFT) is a specific variant of the FT designed for digital signals. The DFT formula is expressed as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] \, . \, e^{-\frac{j2\pi kn}{N}} \qquad (1)$$

In (1), the value n represents the sample index within a single frame. The value of n ranges from 0 to N-1 where N denotes the total number of samples per frame. The variable k represents the frequency range of the signal within each respective frame.

3. *Short-Time Fourier Transform*

The Short-Time Fourier Transform (STFT) represents an evolution of the DFT specifically designed to address non-stationary signals, which are defined as signals with time-varying frequency content [1]. Consequently, the STFT is of paramount importance in the analysis of real-world signals, such as human speech. The STFT formula is defined as follows:

$$X[m, k] = \sum_{n=0}^{N-1} x[n + mH] \, . \, e^{-\frac{j2\pi kn}{N}} \qquad (2)$$

Equation (2) represents a specific implementation of the DFT that incorporates the variable m within its mathematical formulation. The variable m serves as the frame index, indicating the temporal sequence of the segment currently undergoing analysis, whereas H denotes the hop size, which specifies the number of samples shifted between consecutive segments.

B. *Machine Learning*

Machine Learning (ML) is a computational paradigm that enables systems to autonomously derive patterns and relationships from data [3]. By constructing mathematical models that characterize the underlying data, an ML system is capable of performing autonomous prediction and classification. The development of an ML system typically involves a training phase, where the model is optimized using input datasets, and a testing phase to evaluate the model's generalization performance on previously unseen data [4].

1. *Neural Network*

Neural Networks (NNs) constitute a class of ML models comprising an architecture of interconnected neurons designed to facilitate the representation of complex mathematical functions. Within this structure, individual neurons are interconnected via specific weights to process input data. Each neuron executes a weighted summation of its inputs and subsequently applies an activation function to generate an output. This mechanism enables the model to effectively characterize non-linear relationships between input and output datasets.

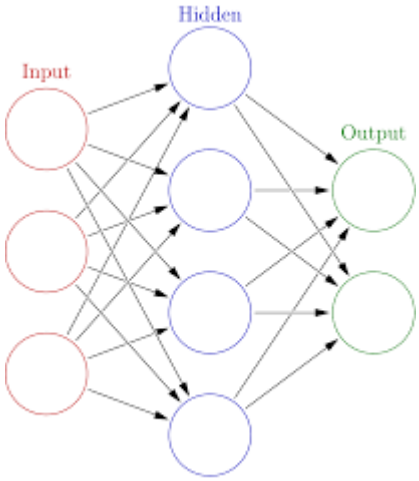

Fig 1. Visualization of a Neural Network

2. *Recurrent Neural Network*

Recurrent Neural Networks (RNNs) represent a specialized class of neural networks designed to process sequential data by accounting for temporal dependencies within preceding inputs [5]. RNNs incorporate feedback loops that enable information from previous time steps to be retained and utilized in subsequent iterations, thereby facilitating the modeling of temporal dynamics within the dataset.

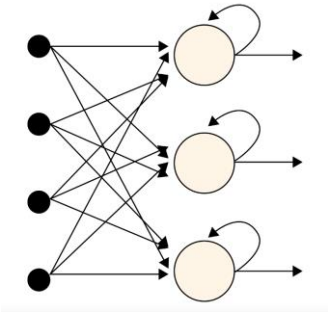

Fig 2. Feedback Loop of RNN

3. *Long-Short Term Memory*

Long Short-Term Memory (LSTM) is an extension of the RNN architecture designed to mitigate the limitations of standard RNNs in capturing long-term dependencies within data. In conventional RNNs, information from distant time steps is frequently difficult to retain due to the exploding gradient or vanishing gradient phenomena resulting from continuous feedback loops. LSTM addresses these issues by introducing a specialized memory cell structure that enables the preservation of salient information over extended temporal intervals [5].

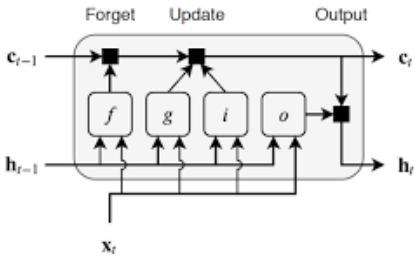

Fig 3. Structure of an LSTM Model

The LSTM architecture incorporates a forget gate, an input gate, and an output gate. These gating mechanisms regulate the flow of information that is ingested, retained, and emitted from the LSTM cell state based on the current input data and the previous hidden state [5]. Through this architecture, the LSTM is capable of selectively retaining or processing specific information, thereby achieving greater stability in modeling long-term dependencies.

III. METHODOLOGY

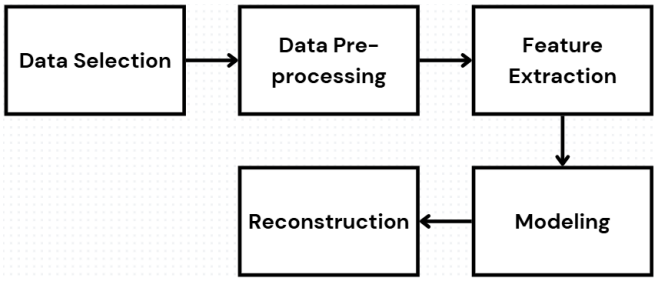

Fig 4. Research Block Diagram

The methodology of this research comprises five primary, interconnected stages, ranging from data selection to reconstruction. Each stage is meticulously designed to ensure a systematic noise filtering process, ultimately achieving a significant enhancement in audio signal quality.

A. Data Selection

The research data is sourced from the Valentini–Botinhao (2017) dataset, which contains voice recordings across various noise conditions [4]. This dataset provides parallel data pairs consisting of clean speech signals as the target and noisy speech signals as the input. This paired data structure facilitates the application of supervised learning methods to train the speech enhancement (noise reduction) model.

B. Data Pre-Processing

The preprocessing stage is executed to standardize the data format and optimize computational efficiency. All raw audio data with an initial sampling rate of 48 kHz is downsampled to 16 kHz. This reduction in sampling rate aims to alleviate memory overhead during the training process while maintaining sufficient frequency resolution to accurately represent human speech. Furthermore, the audio signals are loaded and automatically normalized to an amplitude range of -1 to 1 using the Librosa library.

C. Feature Extraction

The audio signals are transformed from the time domain to the frequency domain utilizing the Short-Time Fourier Transform (STFT). This process segments the signal into short-duration frames of 32 ms (equivalent to 512 samples) with a hop length of 16 ms (equivalent to 256 samples), resulting in a 50% overlap. The complex-valued STFT output is subsequently decomposed into magnitude and phase components. In this research, only the magnitude components are employed as the input features for the model.

D. Modeling

The developed model utilizes a Recurrent Neural Network (RNN) architecture incorporating Long Short-Term Memory (LSTM) units. The magnitude features derived from the noisy signals serve as the network input. The model is trained to map the magnitude patterns of the noisy signals to approximate those of the clean signals. The loss function is computed based on the residual between the predicted magnitudes generated by the model and the target magnitudes from the clean dataset.

E. Reconstruction

The reconstruction stage aims to transform the model's predictions back into the time-domain waveform. Since the model exclusively predicts the magnitude components, the reconstruction process utilizes the phase information derived from the original noisy signal (noisy phase). The predicted magnitude is recombined with this phase component and subsequently converted back to the time domain via the Inverse Short-Time Fourier Transform (ISTFT). To ensure signal integrity, the windowing and overlap parameters employed are kept consistent with those used in the feature extraction stage.

IV. EXPERIMENT RESULTS

A. Evaluation Methodology

The system evaluation aims to assess the performance of the trained Long Short-Term Memory (LSTM) model in separating human speech signals from noise interference. All experiments were conducted within the Google Colab computational environment. The evaluation procedure consists of two primary stages:

1. Objective Evaluation: Conducted using the official test set from the VoiceBank-DEMAND corpus, which constitutes unseen data not utilized during the training phase. This evaluation employs the Signal-to-Noise Ratio (SNR) metric to quantitatively measure signal quality improvement.
2. Subjective and Visual Evaluation: Involves spectral analysis (spectrograms) and testing on real-world recordings outside the dataset (out-of-distribution data) to assess the model's generalization capabilities.

B. Objective Evaluation Results

Objective testing was performed on randomly selected samples from the Noisy Test Set. Each file was processed by the model and subsequently compared against the Clean Test Set (ground truth) to calculate the SNR values. The statistical summary of the results is presented in Table I.

TABLE I. SUMMARY OF SNR IMPROVEMENT RESULTS

Evaluation Metric	Average Value (dB)
Initial SNR (Input)	8.51
Final SNR (Output)	14.83
Improvement (Gain)	6.32

As summarized in Table I, the model achieved an average signal quality improvement of 6.32 dB. This gain indicates that the model can reduce the power of the noise signal.

While the average output SNR of 14.83 dB may appear lower than theoretical studio-grade audio standards (typically >30 dB), it represents a substantial technical success in the context of noise suppression from severely degraded signals. This result is constrained by several deliberate design choices: the use of 16 kHz down sampling to optimize the LSTM's temporal learning window, and the reliance on original noisy

phase during the ISTFT reconstruction process. In speech enhancement research, these factors are known to cap the absolute SNR value; however, the consistent Gain of 6.32 dB confirms that the model is highly effective at increasing speech intelligibility and suppressing background interference without requiring complex end-to-end architectures

To analyze the distribution of model performance in greater detail, a scatter plot is presented in Fig. 5.

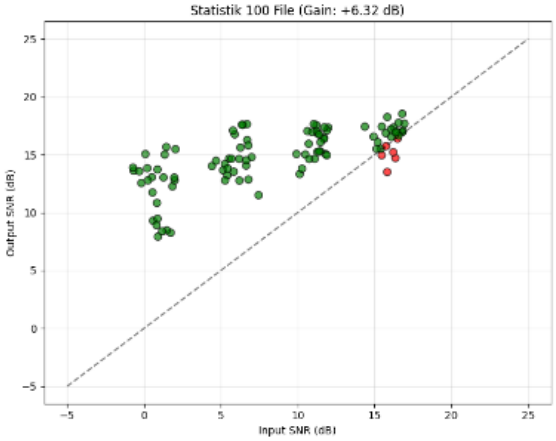

Fig. 5. Distribution of Input vs. Output SNR on the test set.

Based on Fig. 5, the following analysis can be drawn:

- Effective Region (Lower Left): For data with low input quality (Input SNR < 5 dB), most data points are green and positioned well above the diagonal baseline. This demonstrates that the model operates highly effectively under extreme noise conditions.
- Consistency: The vast majority of test data exhibits quality improvement (green points), demonstrating the model's stability across various noise profiles.

C. *Visual Analysis (Spectrogram)*

In addition to quantitative metrics, analysis was conducted in the frequency domain using spectrograms to observe how the model suppresses noise without degrading the speech structure. The sample utilized for this analysis is file p232_001.wav.

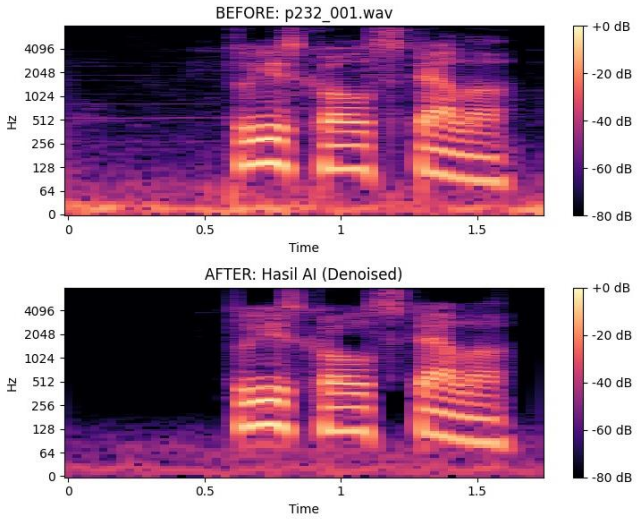

Fig. 6. Spectrogram Comparison: Noisy Input (Top) vs. Denoised Output (Bottom).

In Fig. 6 (Top), visual "fog" or brightness in the background is evident, representing broadband noise. After processing (Bottom), the background area becomes dark (black), indicating the suppression of the noise floor. Crucially, the bright horizontal lines representing human speech formants (harmonics) remain intact. This proves that the model does not merely perform frequency cutting but successfully preserves the distinct patterns of human speech.

D. *Generalization Testing*

To verify the model's robustness and ensure the absence of overfitting (where the model merely memorizes the dataset), testing was conducted using direct recordings of the author's voice mixed with synthetic background noise.

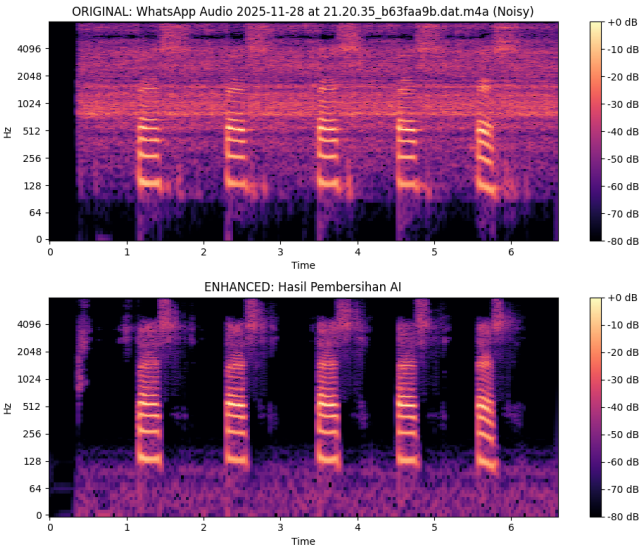

Fig. 7. Test results on independent recordings (non-dataset).

The results in Fig. 7 demonstrate that the model remains capable of separating noise (e.g., fan or environmental noise) from the speaker, despite differences in microphone characteristics and vocal attributes compared to the training data. This indicates that the model possesses strong generalization capabilities.

E. Discussion

Based on the experimental results, the LSTM architecture utilizing STFT features proves effective for speech enhancement tasks. The model successfully learns temporal correlations between audio frames to predict accurate noise suppression.

However, a trade-off phenomenon was observed in Fig. 5 (upper right quadrant), where several samples with high initial SNR (already clean) experienced a slight decrease in SNR values. This is attributed to the model's aggressive noise reduction strategy, which occasionally introduces minor distortions to signals that are inherently clean. Nevertheless, in terms of auditory perception, the audio quality remains preserved, and noise interference is reduced.

V. Conclusion

Based on the experimental results, the noise reduction method integrating STFT and ISTFT with an LSTM architecture demonstrated efficacy in enhancing speech signal quality. Objective evaluations indicate an average Signal-to-Noise Ratio (SNR) improvement of 6.32 dB, signifying the system's capability to reduce noise interference.

Visual analysis via spectrograms demonstrates that background noise is effectively suppressed without compromising the primary harmonic structure of the human voice. Furthermore, generalization testing on out-of-distribution data indicates that the model performs robustly across varying acoustic conditions and voice characteristics. Although a marginal degradation in quality was observed in certain signals with high initial SNRs, the system consistently improves overall speech intelligibility. Consequently, the LSTM-based STFT ISTFT method is viable for integration into real-time voice communication applications, such as telephony, to enhance audio quality in high-noise environments.

References

[1] A. Chottera and G. A. Jullien, "A linear programming approach to the design of FIR digital filters," IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113–120, Apr. 1979

[2] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, 3rd ed. Boston, MA, USA: Pearson, 2010.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. 32,no.6,pp. 1109–1121, Dec. 1984.

[4] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR), 2017. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2791

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997