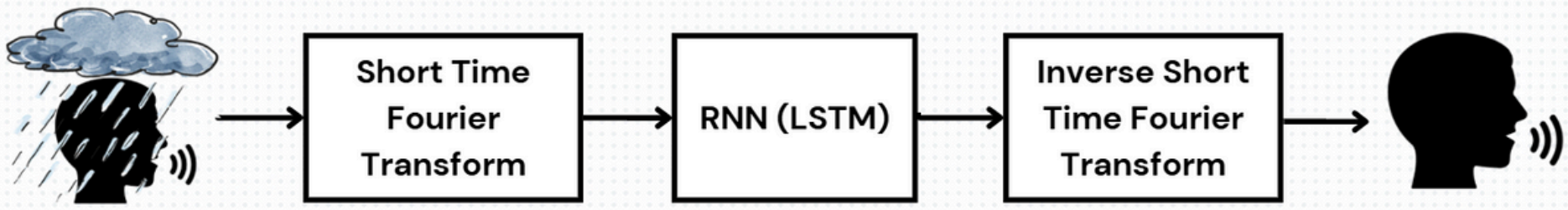
IMPLEMENTATION OF MACHINE LEARNING METHODS BASED ON LSTM AND STFT ARCHITECTURES FOR SPEECH SIGNAL QUALITY ENHANCEMENT

Team

Gabriel Ruben Weslie - 2802404204
Jordan Elishua Wibowo - 2802466865
Rizky Kresnanto Dananjaya - 2802479142
Takeshi Gobstan Lee - 2802540471

Background and Objective

Noise interference is a prevalent issue in voice communication systems, especially in telephone conversations and real-time audio communications. Unwanted noise can compromise speech intelligibility and diminish the quality of communication between users. Consequently, effective strategies are essential to mitigate noise without distorting the primary information within the voice signal.



Data Selection

This study utilizes the Valentini–Botinhao (2017) dataset, which offers pairs of clean and noisy speech signals, thereby facilitating the training of a supervised learning-based speech enhancement model.

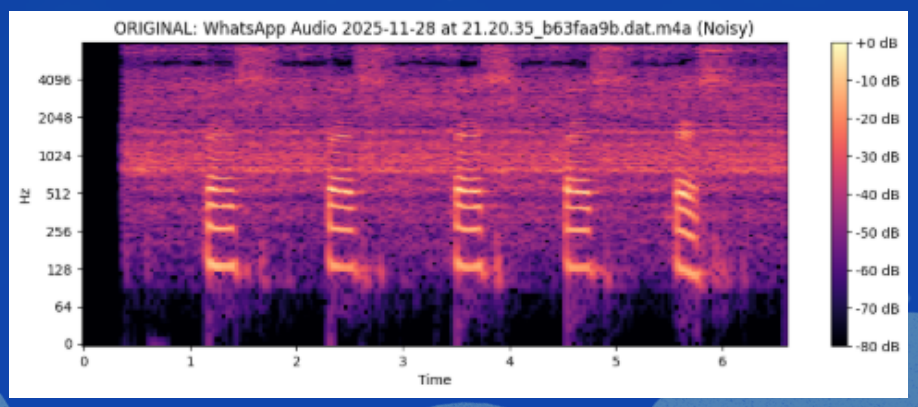
STFT

Preprocessing involved reducing the audio signal's sample rate from 48 kHz to 16 kHz and conducting amplitude normalization. Subsequently, feature extraction was executed using the Short-Time Fourier Transform (STFT) to transition the signal from the time domain to the frequency domain, where the magnitude components of the spectrogram served as input features for the model.

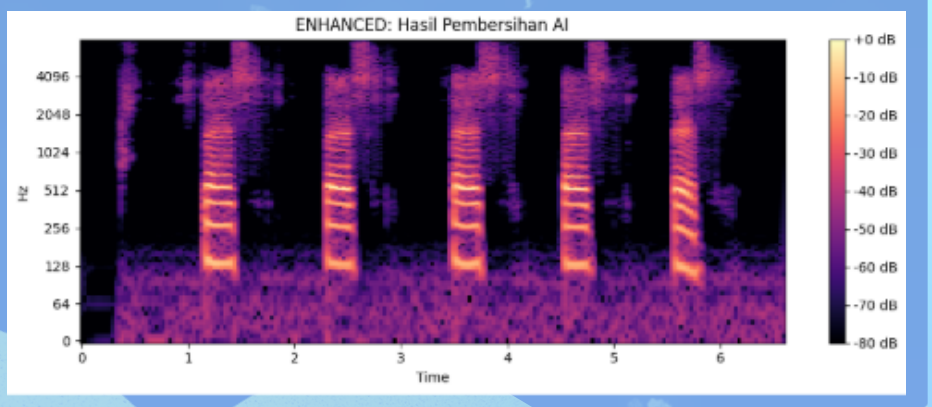
RNN (LSTM)

The developed model employs a Recurrent Neural Network (RNN) architecture utilizing the Long Short-Term Memory (LSTM) technique. Magnitude features extracted from noisy signals serve as the input for the network. The model is trained to correlate the magnitude pattern of the noisy signal with the approximate magnitude pattern of the clean signal. The Loss Function is computed based on the discrepancy between the predicted magnitude produced by the model and the target magnitude derived from the clean data. The predicted magnitude is integrated with phase information from the noisy signal and subsequently transformed back into clean sound using the Inverse Short-Time Fourier Transform (STFT).

Before



After



Result

Spectrogram comparison reveals that the initial speech signal exhibits a bright haze indicative of noise. Following processing by the system, the spectrogram displays enhanced clarity and structure, with the bright horizontal lines representing the human voice preserved. These findings suggest a significant enhancement in the quality of the speech signal.

Conclusion

The findings from the conducted experiments indicate that the STFT-based noise reduction method, when integrated with the LSTM architecture, has demonstrated efficacy in enhancing the quality of speech signals.