Deep Learning and Large Language Models for Speech Denoising and Performance Enhancement

Nipun Agarwal and Sainath Bitragunta, *Senior Member, IEEE*

***Abstract*—Noise suppression serves as a crucial signal processing technique, relying heavily on the precision of estimators and parameters. Traditional deep learning-based speech-denoising methods have historically required noise-free speech signals for training, limiting their applicability to real-time scenarios. Nevertheless, we have spearheaded the development of a pioneering deep learning-based approach that significantly enhances speech denoising in real-world audio environments without the necessity of noise-free speech signals. This innovative approach is centred around self-supervised learning. Following comprehensive qualitative and quantitative evaluations, our research unequivocally demonstrates that our proposed method surpasses Wiener filtering, the standard technique based on processing the magnitude spectrogram. Moreover, we have introduced a groundbreaking large language-based model (LLM) for denoising, showcasing its paramount role in improving the quality of the output sound sample.**

***Index Terms*—Speech signals, spectrogram, noise suppression, denoising, MEL filter bank, Recurrent Neural Network (RNN), mean opinion score, Large language model, performance enhancement.**

I. Introduction and Key Motivation

According to data released by the World Health Organization (WHO), more than a billion people worldwide suffer from varying degrees of hearing loss [1]. Of those who could benefit from hearing aids, only a mere 17% use them. The operation of hearing aids is contingent upon the clarity and quality of speech received, which is significantly compromised due to the presence of background noise [2]. This noise renders communication for users of such devices extremely challenging and sub-optimal [3].

The outbreak of the epidemic has led to an increase in online meetings and remote offices. However, this new way of working has presented new challenges as certain background noises can be generated due to factors, namely, network connectivity, equipment issues, and environmental noise [4], [5]. These noises can lead to a poor communication experience for users. Therefore, it is crucial to have an effective noise suppression system to address all kinds of noise.

The system must be designed to work effectively in low signal-to-noise (SNR) ratio situations while maintaining the integrity of the voice information. To achieve this, noise suppression methods based on deep learning (DL) have shown significant promise. By incorporating DL, one can design, train, and deploy artificial intelligence (AI) models, which could outperform traditional signal processing techniques [6].

Moreover, AI-enabled noise suppression models can learn from vast data and adapt to new environments, making them more effective than traditional methods. These models can also be customized to suit specific use cases and continually improve with the latest data. Therefore, with the help of AI-based noise suppression models, we can ensure that online meetings and remote offices are more productive and efficient for users, even in noisy environments [7].

Neural Network (NN)-Powered, Intelligent Speech Signal Processing System (ISSPS): In the realm of artificial intelligence, the integration of NN has revolutionized various domains, including speech signal processing. With the continuous advancement of technology, neural network-powered intelligent systems are reshaping how we interact with and understand speech signals. Traditional speech signal processing systems have long relied on handcrafted features and algorithms to analyze and interpret audio data. However, these methods often need help handling complex variations in speech patterns and environmental conditions. Neural networks offer a promising solution by enabling machines to learn directly from data, allowing for more robust and adaptive speech signal processing systems. One of the critical advantages of neural network-powered speech signal processing systems is their ability to extract meaningful features from raw audio data automatically.

In speech recognition, NNs have achieved remarkable accuracy levels, rivaling and sometimes surpassing human performance. By training on vast amounts of annotated speech data, NNs can learn to recognize phonetic patterns, words, and even entire sentences, making them invaluable for applications such as virtual assistants, transcription services, and language translation tools. Moreover, neural networks excel in speaker identification tasks by capturing unique characteristics of individuals' voice.s, enabling systems to accurately identify speakers across various scenarios and languages. This capability finds applications in security systems, authentication processes, and personalized user experiences.

Emotion recognition from speech is another area where neural networks demonstrate their prowess. By analyzing subtle acoustic cues such as pitch, intensity, and tempo, neural network-powered systems can accurately infer emotional states from speech signals. This capability has diverse applications, including mental health monitoring, customer sentiment analysis, and human-computer interaction.

Furthermore, neural network-based speech synthesis systems, often called text-to-speech (TTS) systems, have signifi-

Nipun Agarwal was a former student of Birla Institute of Technology and Science (BITS) Pilani. Sainath Bitragunta is with the Department of Electrical and Electronics Engineering, BITS Pilani, Pilani, Rajasthan, 333031 India.

Emails: agarwalnipun2@gmail.com, sainath.bitragunta@pilani.bits-pilani.ac.in

cantly generated natural-sounding speech from text inputs. By learning the complex mapping between text and corresponding speech waveforms, these systems can produce human-like voices with varying accents, intonations, and emotions. Such advancements have enhanced accessibility for individuals with speech disabilities and enabled the development of lifelike virtual assistants and interactive entertainment experiences.

Integrating neural networks into speech signal processing systems also enables real-time and adaptive processing capabilities, making them suitable for applications in noisy environments, mobile devices, and Internet of Things (IoT) devices. These systems can dynamically adjust their processing parameters based on the input data and environmental conditions, ensuring reliable performance in diverse scenarios. However, despite their remarkable capabilities, neural network-powered speech signal processing systems still face particular challenges. Data privacy concerns, bias in training data, and computational resource requirements must be addressed to ensure these technologies' ethical and equitable deployment.

Large language models (LLMs) have completely revolutionized speech denoising by providing robust solutions for significantly improving speech clarity in noisy environments. LLMs effectively separate speech from background noise using deep learning architectures, producing remarkably cleaner audio signals. Additionally, LLMs can be expertly trained to enhance speech in real-time directly from the raw waveform, eliminating various types of background noises and room reverb. This groundbreaking technology in speech denoising is particularly beneficial for applications such as voice assistants, telecommunication systems, and hearing aids, where clear speech is crucial for effective communication. The seamless integration of LLMs in speech denoising elevates user experience and sets the stage for more natural and accessible human-computer interactions.

II. Related Works, Comparisons and Novelty

In speech signal processing (SSP) systems, noise suppression is important for better performance. Existing methods can be classified based on whether they rely on clean speech signals. Traditional methods such as Wavelet threshold denoising and Wiener filtering [8] don't require clean data, but they have a high time complexity, which makes them unsuitable for real-time applications. On the other hand, Noisy2Noisy signal mapping [9] is a relatively new technique that does not require clean speech signals. It uses two noisy realizations of the same speech signal as input and target data for a fully convolutional neural network (CNN). This method is designed to reduce noise in speech signals in real-world audio environments effectively.

Various techniques are available to denoise speech signals in noisy and reverberant environments. One of the most popular approaches involves using deep learning networks. For instance, Nils L. Westhausen's DTLN technique [10] utilizes STFT for signal pretreatment and trains the network with LSTM. Similarly, Umut Isik's PoCoNet method [11] employs a large U-Net with DenseNet and self-attention blocks, which rely on frequency-positional embeddings, with Short-time Fourier Transform (STFT) used to obtain input data.

Other methods, such as Dario Rethage's Wavenet [12] and Hyeong-Seok Choi's phase-aware single-stage speech denoising method [13], consider the speech signals' phase. Wavenet predicts target fields with non-causal, dilated convolutions. At the same time, Choi's method employs a unique masking technique called phase-aware b-sigmoid mask (PHM) [13], which reuses the estimated magnitude values to estimate the clean phase. Jean-Mare Valin's proposal is a hybrid digital signal processing (DSP)/deep learning approach [14] that combines DSP techniques with deep learning to achieve superior results.

This paper's novelty and key contributions are as follows.

- It introduces a framework for developing an automatic connected speech recognition system that utilizes wavelet transformation (WT) denoising into conventional Mel-frequency cepstral coefficients (MFCC).
- We focus on analyzing the effectiveness of different wavelet families and thresholding rules to remove environmental noise from speech signals and reduce the effect of non-stationary noise. This process is particularly important in the presence of the disturbance of environmental noise, which can significantly impact speech recognition.
- We introduce a new large language-integrated speech denoising model that works on the inputs extracted from the input speech spectrogram.

To evaluate our denoising procedure, we use the Microsoft DNS challenge [15] speech corpus, which includes $10,000$ pieces of voice data in English. The length of each voice data ranges from 5 to 20 seconds, and the sampling rate is uniformly set to 8 KHz. Using this dataset, we build a speaker-independent acoustic model and focus on thresholding the approximation components and detail components of the noise signal in our denoising procedure. Additionally, using a speaker adaptive training model, the system is designed to tolerate the variability of different speaker characteristics.

We show that using a wavelet denoising scheme in the MFCC significantly improved recognition accuracy, especially at low SNRs. However, at high SNR values, the recognition accuracy remained almost unchanged. Significantly, the proposed algorithm does not impact the processing speed or require more computational resources than the traditional version.

The paper is structured as follows. Section III describes the design of the acoustic feature extraction intelligent SSP system. Section IV discusses the use of speech materials, presents experimental results, and provides a discussion. Further, section V presents a large language model (LLM) for speech denoising and evaluates its performance. Finally, Section VI offers concluding remarks.

III. SSP System: Modeling and Implementation

In this section, we first present the overview of the Speech Signal Processing (SSP) system model and the core module within the model. The core module is a neural network for training. Later, we present the neural network for testing. The rest of the section includes key details: NN module design, feature extraction, spectrum gain analysis, and insights.

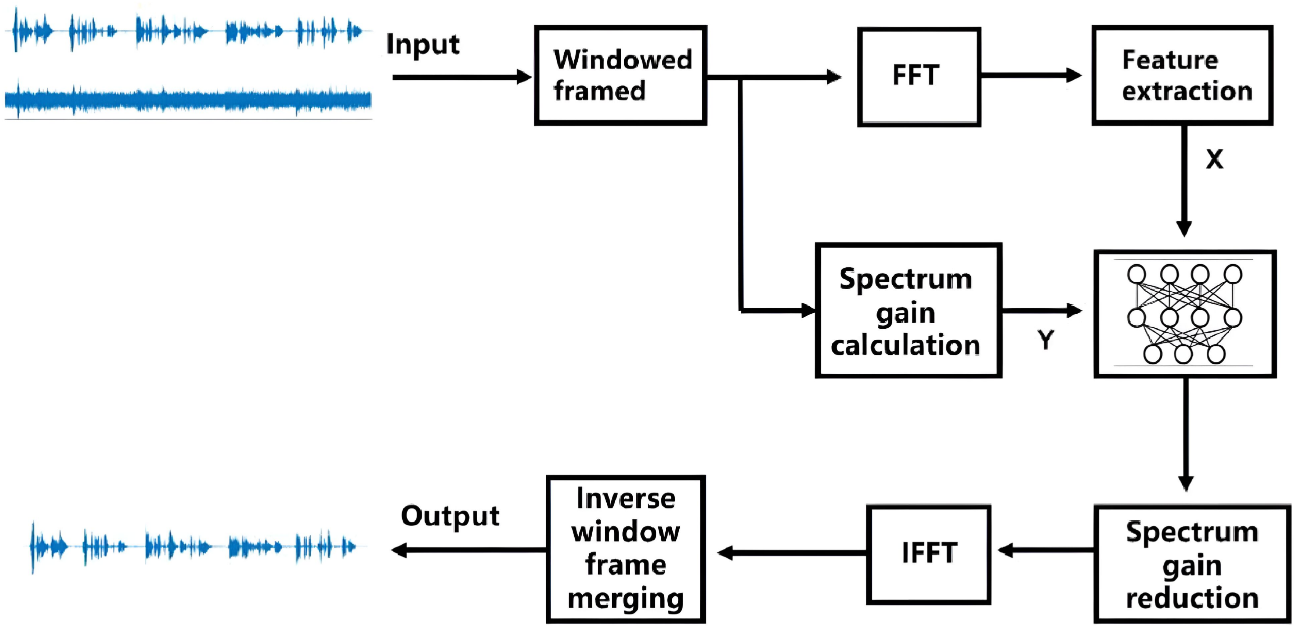


Fig. 1: Functional building blocks of SSP system.

A. *SSP System Model: Training Neural Network*

The system we adopt and update follows the framework of RNNoise [16], a popular noise suppression tool for audio signals. However, we incorporate several changes to the original structure to enhance its performance and adapt it to our needs. To better understand the SSP system, we include a detailed representation of the overall structure in Figure 1. The diagram illustrates the different components of the system and how they interact with each other to achieve the desired outcome. Additionally, we include a comprehensive depiction of the network training process in Figure 2. The training process involves feeding the network with a dataset of input and output pairs and adjusting the weights and biases of the network iteratively until it learns to generalize well on new input data.

Improving speech that is affected by noise involves various steps. Initially, both the clear speech and the same noisy speech with Gaussian noise are subjected to windowing and framing processing. This process divides the speech signals into small frames of the same duration, which usually range from 20 to 40 milliseconds (ms). Each frame is then individually analyzed using the Fast Fourier Transform (FFT) algorithm to obtain its frequency spectrum. The frequency spectrum provides information about the different frequencies present in each frame.

Next, the spectral gain of speech is calculated by comparing the frequency spectrum of the noisy speech with that of the clear speech. This spectral gain represents the discrepancy between the frequency domain's noise-free speech versus the noisy speech. At the same time, MFCCs are computed for each frame of the noisy speech. MFCCs represent the short-term power spectrum of an acoustic sound signal based on the cosine transform of a logarithmic power spectrum on a nonlinear Mel scale of frequency [17]. These coefficients serve as input for the DNN.

The DNN takes the MFCC coefficients of the noisy speech as input and produces the corresponding spectral gain for the clear speech. Doing so predicts the spectral gain for each frame of the noisy speech. The predicted spectral gain is then used to modify the noisy speech to obtain an improved version of the speech signal.

B. *Testing Neural Network Module*

After the NN has undergone training, as shown in figure 3, it is necessary to test and apply it in real-world scenarios. The testing process involves taking a new piece of noisy speech, similar to the training data, and subjecting it to windowed and framing operations. The Mel-frequency cepstral coefficients are calculated for each processed frame and fed as input to the trained NN. Based on the learned patterns from the training data, the NN outputs the spectral gain value of the frame.

Simultaneously, an FFT is performed on the frame to obtain its frequency spectrum. The spectrum gain from the network output is then applied to the frequency spectrum of the noisy speech frame, allowing for the reconstruction of the clean speech frame frequency spectrum. Finally, an Inverse Fast Fourier Transform (IFFT) is performed to restore the frame's time domain data. Once each frame has been processed, we restore inverse window and frame overlap to achieve the reconstructed denoised speech. This reconstructed speech is free from noise and is the desired output of the denoising process.

1) Windowing and Framing Procedures: When we speak, the sound of our voice changes as we form different words

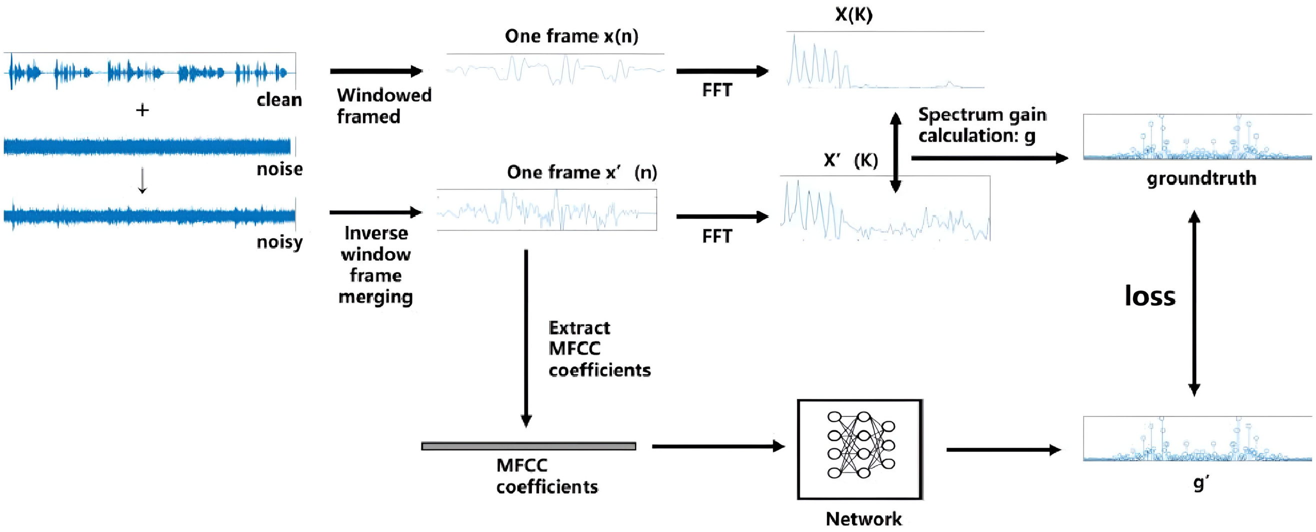


Fig. 2: An illustration on training Process.

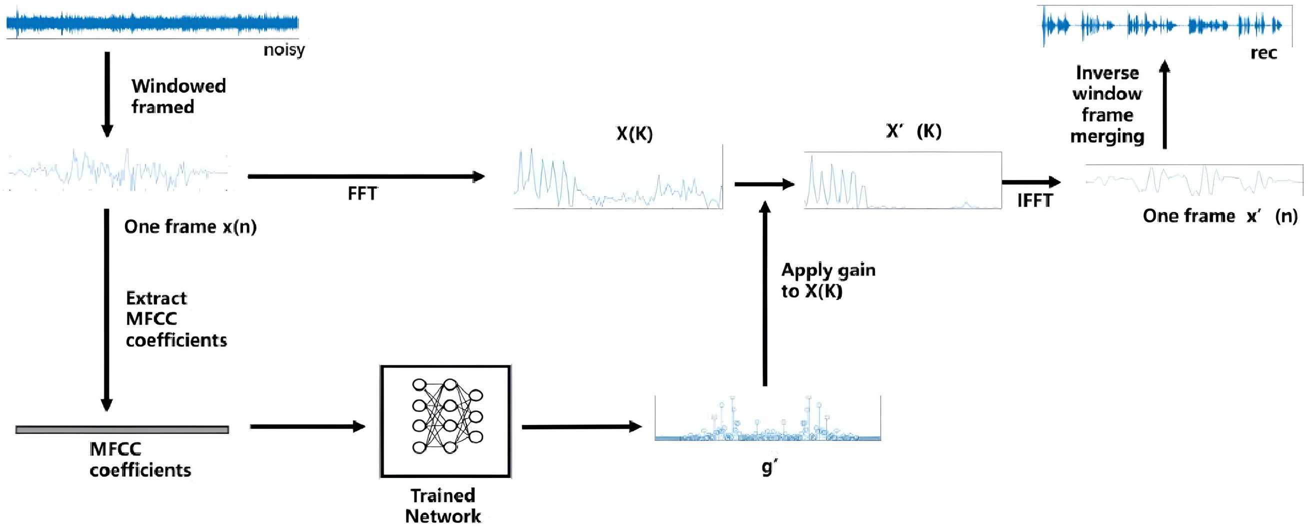


Fig. 3: An illustration of the testing process.

and sounds. However, if we examine the voice signal over a short time period, its properties remain relatively consistent. This nature of the signal indicates that the voice signal can be considered a quasi-steady state process, which means it possesses short-term stability.

We focus on obtaining feature parameters from the voice signal. For it, we divide the signal into frames. However, if non-overlapping frames are used, important information regarding the smooth changes occurring in the voice signal over time may be lost. Some overlapping frames can be incorporated between non-overlapping frames to circumvent the loss. This technique enables the extraction of feature parameters from the overlapping sections between adjacent frames, ensuring smoother transitions between feature parameters. This approach helps to capture more information about the voice signal and can enhance the accuracy of voice-related applications such as speech recognition.

In a specific implementation, we add a first-order sine window to each frame of speech, which is given by:

$$a(m) = \sin\left(\frac{\pi(m+1)}{N+1}\right),\ 0 \leq m \leq M, M \leq N. \qquad (1)$$

The specific parameters of windowed and framed are as follows: i) Audio sampling rate $=$ 8 KHz, ii) Frame length $=$ 20 ms (160 points) iii) Overlap between frames $=$ 10 ms (80 points).

2) Feature Extraction: We aim to simplify network input while preserving speech characteristics. For it, we extract speech features from each frame, which involves computing MFCC for each speech frame. The Mel scale is a frequency scale based on the human ear's response to sound. It approximates the relationship between frequency and perceived pitch using a non-linear scale. The rationale behind this is as follows: the human ear is more sensitive to changes in pitch at lower frequencies than at higher frequencies. The Mel scale reflects this non-linear relationship, with equal distances on the scale corresponding to equal distances in pitch perception.

We adopt the following steps to obtain MFCC coefficients. The speech signal is initially divided into short frames, usually $20-30$ ms in length. Each frame of speech is then transformed into the frequency domain using the Fourier transform. The Mel filterbank is then applied to each frame's frequency spectrum, grouping together perceptually similar frequency components. Finally, the logarithm of the filterbank energies is taken, and the resulting cepstral coefficients are transformed using the discrete cosine transform (DCT). The MFCC coefficients allow for efficient and compact representation of the speech signal, which is crucial for various speech processing applications such as speech recognition, speaker identification, and speech synthesis. Due to their effectiveness in capturing the unique characteristics of speech, the MFCC coefficients are widely used in speech processing. These are defined by

$$\mathrm{Mel}(f) = 2595 \times \log\left(1 + \frac{f}{700}\right). \tag{2}$$

Figure 4 illustrates the process of calculating the MFCC, with the Mel filter bank, logarithmic energy calculation, and DCT being the focus of this section. It should be noted that the aforementioned previous process has already been implemented during windowing and framing.

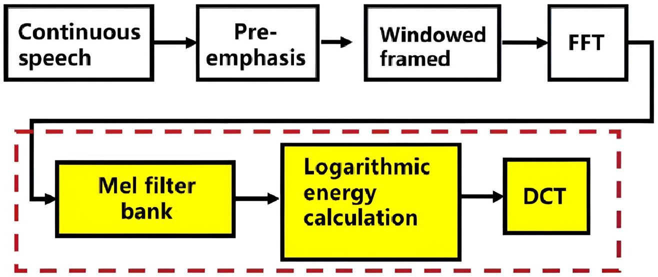


Fig. 4: MFCC coefficient calculation process

Define $(f(m) - f(m-1)) = \Delta f_1$ and $(f(m+1) - f(m-1)) = \Delta f_2$. The Mel filter bank is defined as follows:

$$H_m(k) = \begin{cases} 0, & k < f(m-1), \\ \frac{2(k-f(m-1))}{\Delta f_2 \Delta f_1}, & f(m-1) < k < f(m), \\ \frac{2(f(m+1)-k)}{\Delta f_2 \Delta f_1}, & f(m) < k < f(m+1), \\ 0, & f(m+1) < k. \end{cases} \tag{3}$$

Once the audio signal is passed through the Mel filter bank, the resulting output is a series of Mel-frequency coefficients. These coefficients are then subjected to logarithmic energy calculation, which involves computing the logarithm of the sum of squares of the coefficients. Then, the DCT is applied to the logarithmic energy coefficients to obtain a compact representation of the audio signal in the frequency domain given by

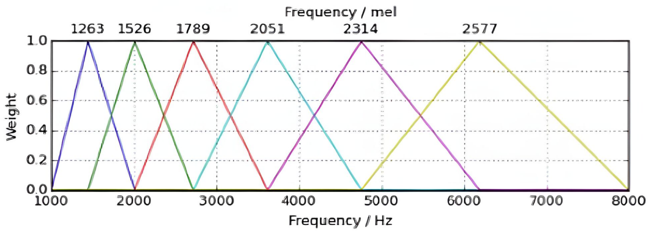


Fig. 5: Mel filter bank: Weight as a function of frequency.

$$s(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), \ 0 \le m \le M, \tag{4}$$

$$c(n) = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1, 2, \ldots, L. \tag{5}$$

The addition of differential coefficients of MFCC during training captures the dynamic characteristics of speech. This methodology enhances the precision of speech recognition systems by incorporating contextual changes within the speech signal. By adopting this technique, the system can effectively differentiate between speech sounds, resulting in improved accuracy.

$$d_t = \begin{cases} C_{t+1} - C_t, & t < K \\ \frac{\sum_{k=1}^{K} k(C_{t+k} - C_{t-k})}{\sqrt{2\sum_{k=1}^{K} k^2}}, & K \le t \le L - K \\ C_t - C_{t-1}, & t > L - K \end{cases} \tag{6}$$

In the actual model calculation, for each frame: i) MFCC coefficient order $L = 16$ ii) MFCC differential coefficient order $K = 8$ Therefore, the input feature of each frame is $16 + 8 = 24$ points.

3) Spectrum Gain Computation: This system offers two viable methods for calculating spectrum gain: gain calculation by frequency and gain calculation by band energy. Both methods can be employed as per the specific requirements of the user. Gain calculation by frequency is a method that calculates the gain based on the frequency range, while gain calculation by band energy is a method that calculates the gain based on the band's energy. These methods can be helpful in accurately measuring the gain of the spectrum and making informed decisions based on the results obtained.

If we calculate the gain for the frequency point, we shall set clean speech frame spectrum $X(k)$ and noisy speech frame spectrum $X^{'}(k)$,/ Therefore, the gain g_k is given by

$$g_k = \frac{X(k)}{X'(k)}. \tag{7}$$

Currently, the network's output dimensions directly correlate to the frame length, resulting in a precise gain calculation. However, this comes at the cost of a relatively high computational load. An alternative approach involves computing the gain by considering the energy of the frequency band

and utilizing the frequency band definition found in the Opus diagram, illustrated in Figure 6. This method holds promise in reducing the computational load while preserving the desired level of accuracy. The gain at this time is given by

$$g_b = \left(\frac{E_s(b)}{E_x(b)} \right)^{0.5}, \tag{8}$$

where $E_x(b)$ is the energy of the b^{th} frequency band of the clean speech frame spectrum. Further, $E_s(b)$ is the noisy spectrum when

$$E(b) = \sum_k w_b(k) |X(k)|^2. \tag{9}$$

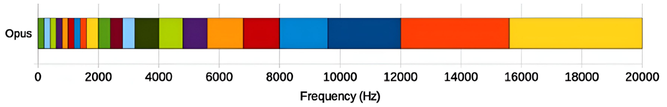


Fig. 6: An illustration of opus frequency.

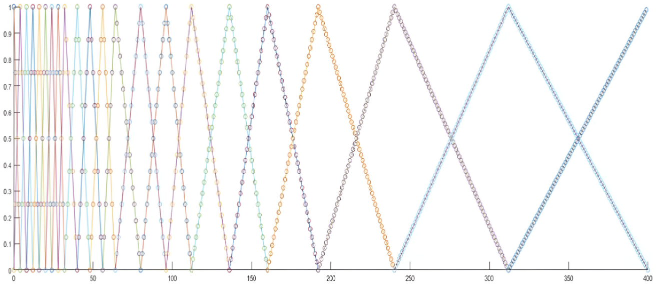


Fig. 7: An illustration on the definition of w_b.

Here, w_b denote a set of triangular filters, satisfying $\sum_b w_b(k) = 1$, defined as shown in Figure 7.

Two methods were employed to compute the gain during the model calculation process to ensure precise calculations: the frequency band method and the frequency point method. The frequency band method determines the gain for a range of frequencies, while the frequency point method computes the gain for a specific frequency point. Although the frequency point method had a slightly better noise reduction effect, the frequency band method significantly reduced computation by computing gain over a range of frequencies. This reduction in computational burden made it a more effective and practical method for this particular computation.

Both methods were utilized to guarantee accuracy, with the frequency point method serving as a comparison to the frequency band method. The number of frequency bands was defined as 17, taking into consideration the dataset's sampling rate. This consideration enables a detailed computation of the gain across a range of frequencies while ensuring computational efficiency.

C. Network Module Design and Analysis

Our system utilizes the impressive capabilities of a time series processing network specifically designed for voice signals. The RNNoise architecture, which heavily uses the gated recurrent unit (GRU) [18], has been integrated into our system. Refer to Figure 8 for a visual representation of the network configuration. The input comprises the MFCC and their differential coefficient, while the output is the band gain.

D. Spectrum Gain Reduction

After obtaining the ideal frequency band gain g_b of the network output, we can interpolate it to the dimension of the frame and get the denoise spectrum.

$$r(k) = \sum_b w_b(k) g_b, \tag{10}$$

$$X(k) = X^{'}(k) \times r(k). \tag{11}$$

That is, transform the reconstructed $X(k)$ to the time domain by IFFT to obtain the reconstructed frame data in the time domain.

E. Reverse Window Frame Restoration

When analyzing voice data, it's common practice to segment it into smaller frames for easier processing. Typically, each frame is twice the length of the frame shift to allow for overlap between adjacent frames. A sine window is applied to each frame to ensure precision during analysis.

During the reconstruction process, the sine window serves a vital function. By utilizing its characteristics, the frames can be overlapped and combined directly, resulting in the restored voice data. This step is critical for achieving accurate results and restoring the data to its original form, emphasizing the importance of correctly utilizing the sine window. The operational details on the window are summarized below:

- A window is added before the positive transformation to suppress the side lobes on the spectrum.
- After the inverse transformation and before the overlap addition, a window is added to avoid abrupt changes in the signal after the synthesis.
- The sine window multiplied by the sine window is exactly equal to the Hanning window, and the overlap of the half of the Hanning window is exactly equal to 1, so if we do not do any processing after the forward transformation and directly inverse transformation synthesis, we can get the original signal.

IV. Result and Numerical Evaluation

This section aims to demonstrate the effectiveness of our system in reducing noise from multiple perspectives. The foremost measurement is to analyze the performance of time domain data and spectrogram graph.

A. Denoising Effect

Our system is trained on SNR $= 0$ dB noise level. Hence, we conducted the noise reduction performance test under the same condition of 0 dB noise. The results of the time domain data and spectrogram measurements are presented in Figure 9.

The noise reduction effect has been significantly enhanced at this stage. The spectrogram analysis reveals that, except for

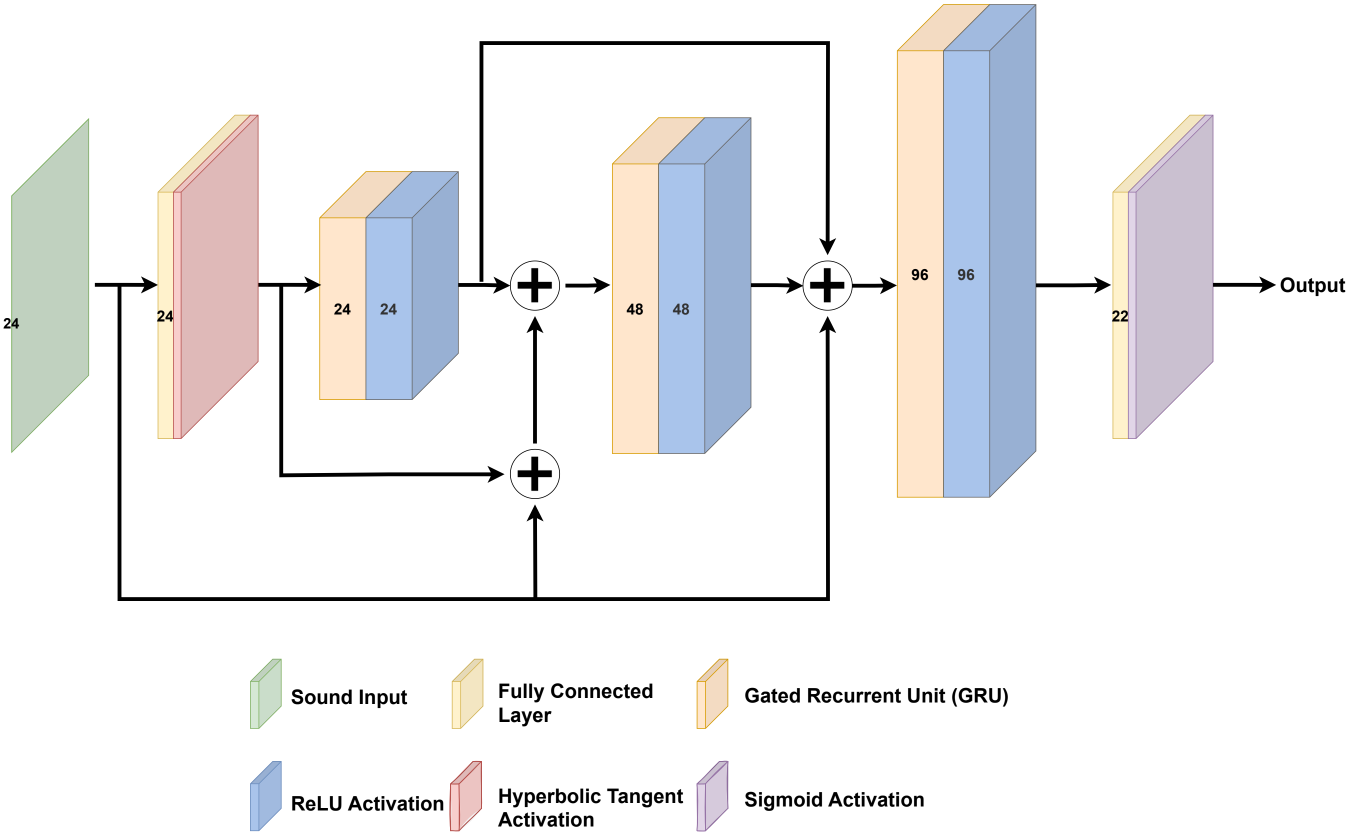


Fig. 8: An illustration of Network Module Structure, with numbers indicating the size of each layer.

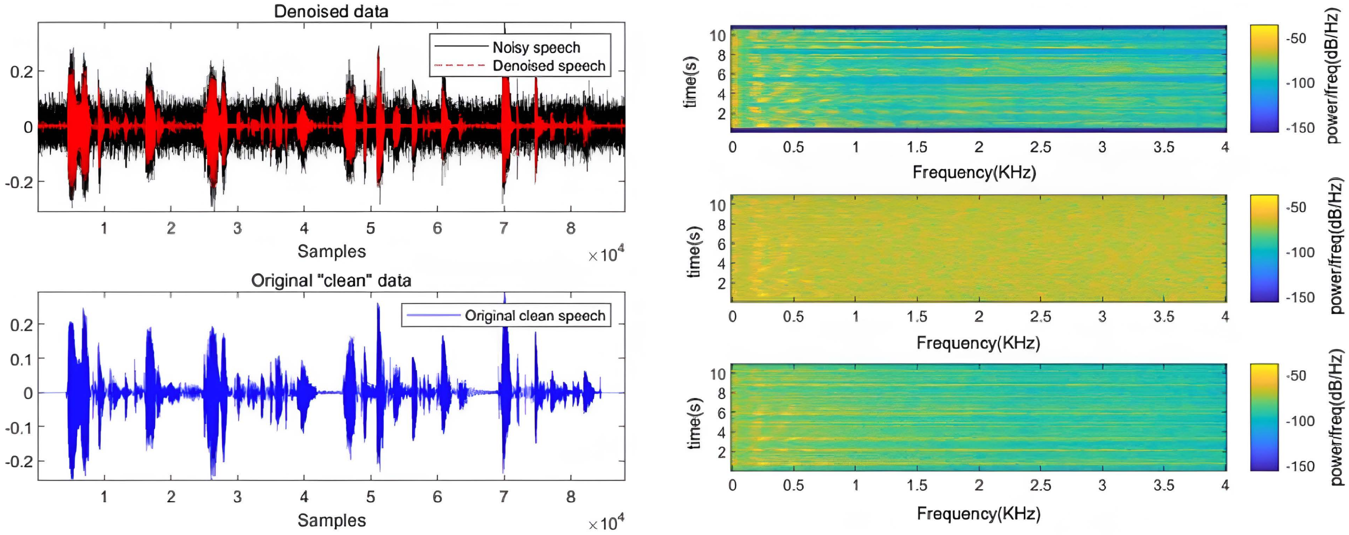


Fig. 9: Denoise effect in SNR $= 0$ dB

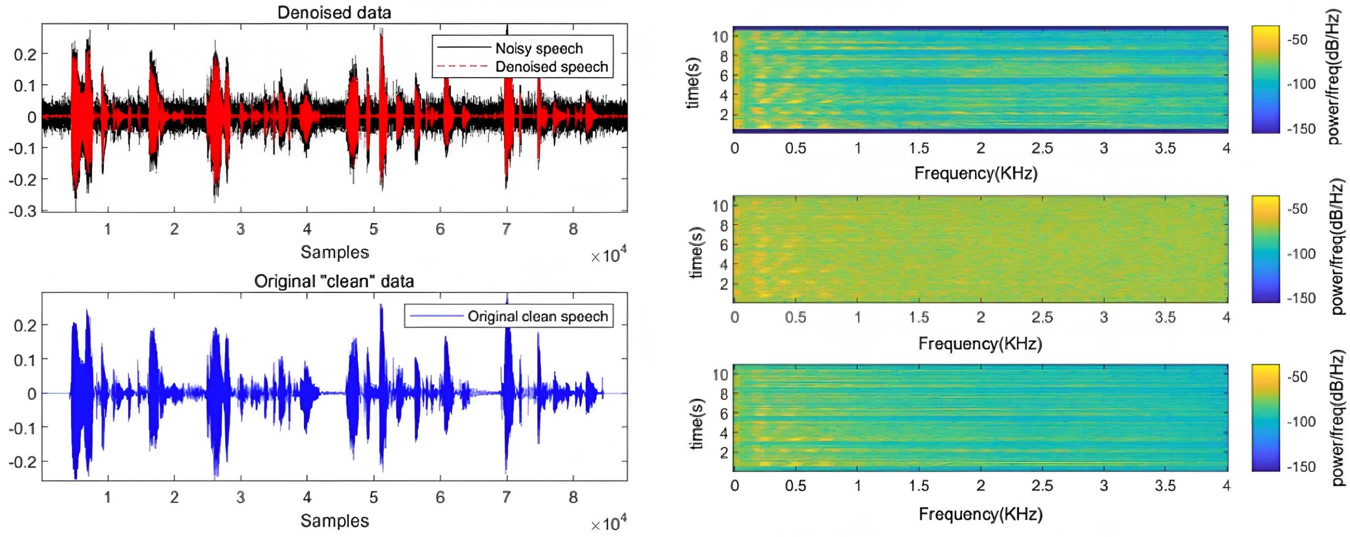


Fig. 10: Denoise effect in SNR $= 5$ dB

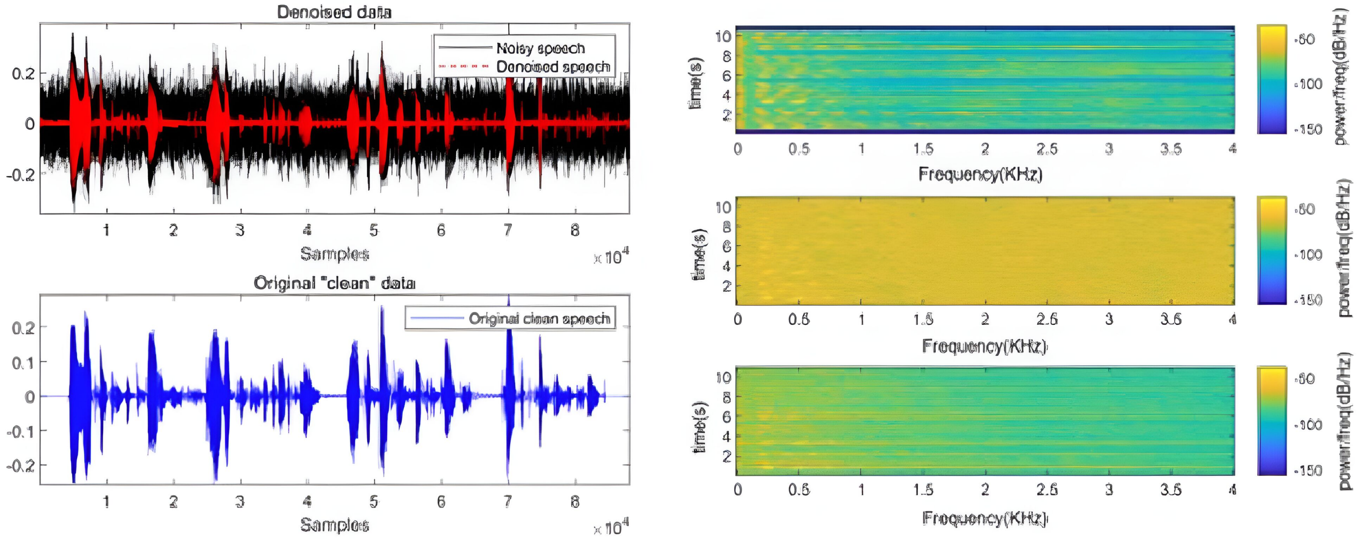


Fig. 11: Denoising effect in SNR $= -5$ dB.

some distortion in the high-frequency range, the remaining parts of the speech have successfully recovered their original characteristics. Moreover, when subjected to weak noise (SNR $= 5$ dB), the noise reduction effect is further augmented, as portrayed in Figure 10.

When operating under severe noise conditions with an SNR of -5 dB, the SSP system can reduce the impact of noise to a certain extent. However, it should be noted that the original voice may experience greater distortion in such situations, as evidenced by Figure 11.

B. Comparison of Objective Indicators

We first present details on various indicators, namely, SNR, Segmental SNR (SSNR), log-spectral distortion (LSD), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI).

1) Review of Performance Indicators:

- *SNR:* The ratio of the power of signals and noises. The definition is as follows:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=0}^{L} x^2(n)}{\sum_{n=0}^{L}(x^2(n)-\hat{x}^2(n))}. \tag{12}$$

  where L is the length of the signal, $x^2(n)$ is the clean speech signal, and $\hat{x}^2(n)$ is the noisy or processed signal.
- *SSNR:* Segmental SNR is the average SNR per frame. Both SNR and SSNR range from $-\infty$ to $+\infty$, indicating a closer relationship to the original signal when bigger. They reflect the overall effect and have a high correlation coeffcient with the listener's subjective auditory perception.
- *LSD:* A measure of the distance between two spectra, ranging from 0 to $+\infty$, indicates a closer relationship to

the original signal when smaller.

- *PESQ:* An objective indicator close to the subjective evaluation score, ranging from -0.5 to 4.5, and indicates a closer relationship to the original signal when they are bigger.
- *STOI:* Since a word in a speech signal can only be understood or not understood, intelligibility can be regarded as binary from this perspective. Therefore, STOI ranges from 0 to 1, representing the percentage of words correctly understood, and 1 indicating that speech can be fully understood.

2) Evaluation of Different Methods: We compared our system with the origin method, fully connected method, convolution method and frequency bin method by indicators mentioned upfront. Each value needs two speech signals to calculate. The noisy signal and clean signal are used to get the indicators before the treatment, and the processed signal and clean signal are used to get the indicators after the treatment. The results are shown in figure 12. The results show that the frequency bin method produces the best performance, and the second one is our system. The origin method has a pretty PESQ, but its SNR is even worse than the SNR before the process. This degradation is because PESQ is an indicator closer to the listening feel, but SNR is just the ratio of power.

V. Large Language Model for Speech Denoising

Multimodal large language models (LLMs) have shown remarkable abilities in visual perception by connecting with image encoders. However, their performance on auditory tasks has yet to be widely explored. Automatic speech recognition (ASR) and automatic audio captioning (AAC) are achieved separately, resulting in incomplete auditory perception abilities. In this regard, we propose a model that simultaneously separates background noise from audio events in clips containing mixed speech and background audio events. This updated model incorporates a step toward more complete machine auditory perception. Figure 13 presents our model. Our model unequivocally employs Whisper-Large [19] as the speech encoder, finetuned BEAT [20] as the noise sample encoder, and SALM [21] as the backbone LLM.

The output dimensions of Whisper and BEATs encoders are 1280 and 768, respectively. By adding these two values, we get the input dimension of the cosine analysis, which is equal to 2048. The output of the cosine analysis is a 768-dimensional vector, which is projected to a 5120-dimensional vector using a fully connected layer. In all experiments, the cosine analysis consists of two Transformer blocks that share the same multi-head self-attention layers, randomly initialized before training. The rank and scale of LoRA are set to 10 and 5. The number of trainable parameters is $\sim$ 30M, $\sim 0.22\%$ of the total parameters of the entire model.

Figures 14 and 15 show our model's denoising performance. It is worth noting that when we evaluate Speech SALM, which is trained with context but without providing the context during inference, we obtain a Noise Error Rate (NER) of 8.31%, which corresponds to a slight NER gap compared to the model trained without context.

VI. Conclusions and possible research extensions

Through this work, we designed and developed an intelligent and efficient SSP system model as a novel solution for reducing audio noise by combining signal processing with DL techniques. The SSP system was rigorously designed and tested using MATLAB and the RNNoise concept. Our approach involves merging traditional signal processing methods with deep neural networks, which has resulted in superior real-time processing performance compared to other existing methods. We used various objective evaluation indicators to measure the system's performance, including SNR, MOS, and PESQ. Our evaluation showed that the proposed SSP system model achieved remarkable improvements: noise suppression, reduced signal distortion, and speech quality enhancement. Furthermore, we evaluated and analyzed the proposed denoising procedure with large language models (LLMs) and showed its significance. Adding LLM to the denoising procedure and its performance evaluation brings more novelty and is a useful benchmark for future denoising SSP systems.

Future research directions: In the future, we shall explore the potential of incorporating additional input features, such as the cepstrum of the far-end signal or the filtered far-end signal and leakage estimates, to expand the application of our technique. This extension could enable us to suppress residual echo and microphone array post-filtering. We are confident that our innovative approach offers a promising solution for audio noise reduction that could have far-reaching implications for various industries, including telecommunications, entertainment, broadcasting, and healthcare.

References

[1] D. McDaid, A.-L. Park, and S. Chadha, "Estimating the global costs of hearing loss," *Int. J. Audiol.*, vol. 60, no. 3, pp. 162–170, Mar. 2021.

[2] P. Tremblay, V. Brisson, and I. Deschamps, "Brain aging and speech perception: Effects of background noise and talker variability," *Neuroimage*, vol. 227, p. 117675, Feb. 2021.

[3] S. Li, M. O. Yerebakan, Y. Luo, B. Amaba, W. Swope, and B. Hu, "The effect of different occupational background noises on voice recognition accuracy," *J. Comput. Inf. Sci. Eng.*, vol. 22, no. 5, p. 050905, Mar. 2022.

[4] P. Aumond, A. Can, M. Lagrange, F. Gontier, and C. Lavandier, "Multidimensional analyses of the noise impacts of COVID-19 lockdown," *J. Acoust. Soc. Am.*, vol. 151, no. 2, p. 911, Feb. 2022.

[5] L. M. Thibodeau, R. B. Thibodeau-Nielsen, C. M. Q. Tran, and R. T. d. S. Jacob, "Communicating during COVID-19: The effect of transparent masks for speech recognition in noise," *Ear Hear.*, vol. 42, no. 4, pp. 772–781, 2021.

[6] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions," *IEEE Access*, vol. 9, pp. 72 894–72 936, 2021.

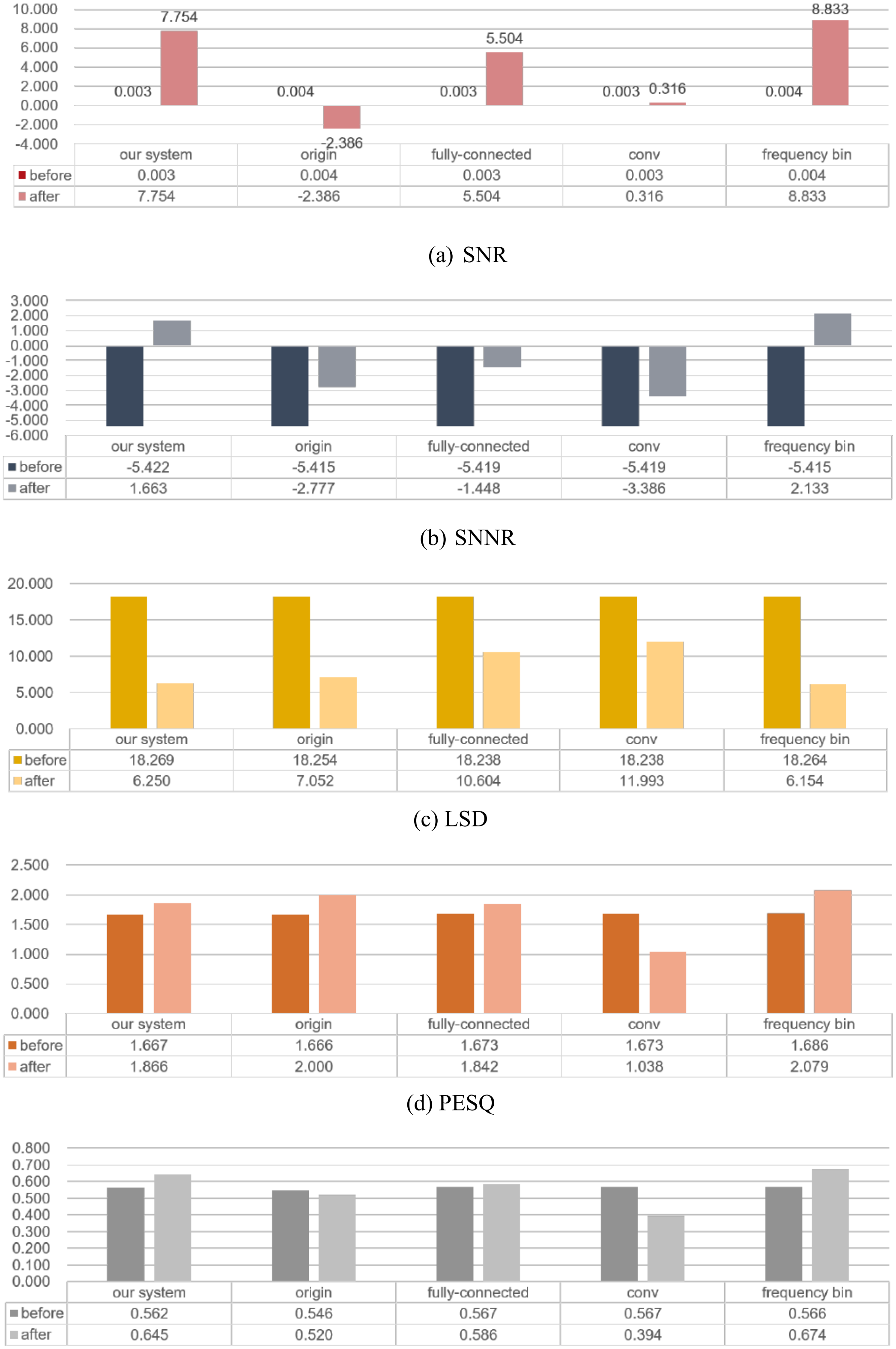
[7] C. A. Duarte-Salazar, A. E. Castro-Ospina, M. A. Becerra, and E. Delgado-Trejos, "Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: An overview," *IEEE Access*, vol. 8, pp. 15 983–15 999, 2020.

[8] A. Z. Alsheibi, K. P. Valavanis, A. Iqbal, and M. N. Aman, "Speech enhancement framework with noise suppression using block principal component analysis," *Acoustics*, vol. 4, no. 2, pp. 441–459, May 2022.

[9] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by Noisy2Noisy signal mapping," *Appl. Acoust.*, vol. 172, p. 107631, Jan. 2021.

[10] N. L. Westhausen and B. T. Meyer, "Dual-Signal transformation LSTM network for Real-Time noise suppression," May 2020.

[11] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better speech enhancement with Frequency-Positional embeddings, Semi-Supervised conversational data, and biased loss," Aug. 2020.



	our system	origin	fully-connected	conv	frequency bin
■ before	0.003	0.004	0.003	0.003	0.004
■ after	7.754	-2.386	5.504	0.316	8.833

(a) SNR

	our system	origin	fully-connected	conv	frequency bin
■ before	-5.422	-5.415	-5.419	-5.419	-5.415
■ after	1.663	-2.777	-1.448	-3.386	2.133

(b) SNNR

	our system	origin	fully-connected	conv	frequency bin
■ before	18.269	18.254	18.238	18.238	18.264
■ after	6.250	7.052	10.604	11.993	6.154

(c) LSD

	our system	origin	fully-connected	conv	frequency bin
■ before	1.667	1.666	1.673	1.673	1.686
■ after	1.866	2.000	1.842	1.038	2.079

(d) PESQ

	our system	origin	fully-connected	conv	frequency bin
■ before	0.562	0.546	0.567	0.567	0.566
■ after	0.645	0.520	0.586	0.394	0.674

(e) STOI

Fig. 12: Output on different metric

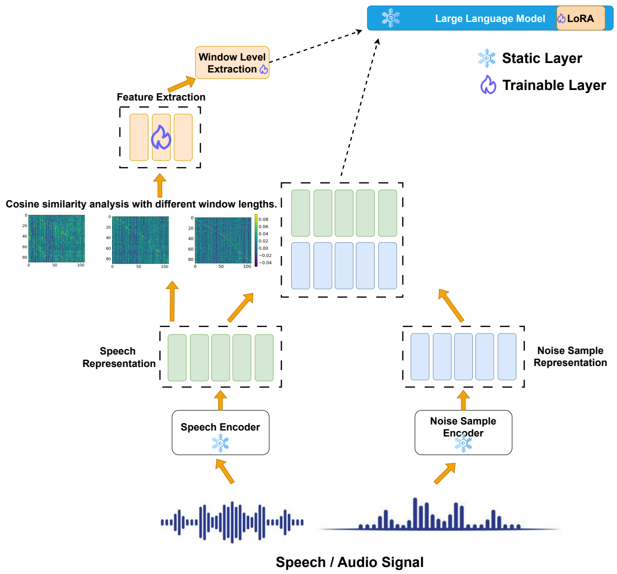


Fig. 13: Multimodal LLM structure with dual auditory encoders with spectrogram feature extraction.

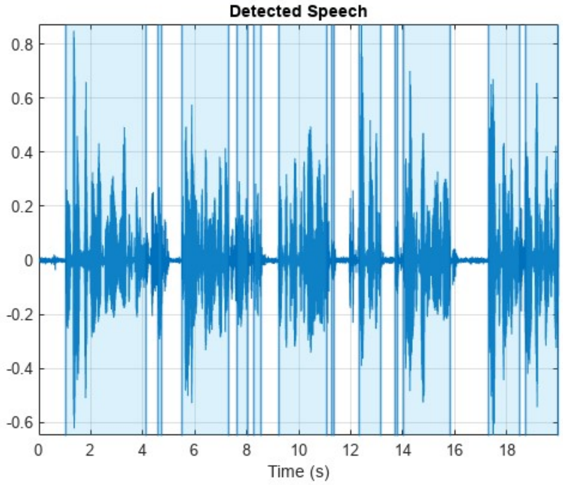


Fig. 14: Denoising effect LLM Output

[12] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. IEEE, Apr. 2018, pp. 5069–5073.

[13] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with U-Net," Jun. 2020.

[14] J.-M. Valin, "A hybrid DSP/Deep learning approach to Real-Time Full-Band speech enhancement," in *2018 IEEE 20th Int. Workshop on Multimed. Signal Process. (MMSP)*. IEEE, Aug. 2018, pp. 1–5.

[15] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, and others, "Icassp 2023 deep speech enhancement challenge," *arXiv preprint arXiv*, 2023.

[16] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated approach for Low-Complexity, Real-Time enhancement of fullband speech," Aug. 2020.

[17] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[18] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN Encoder-Decoder for statistical machine translation," Jun. 2014.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th Int. Conf. on Mach. Learn.*, ser. Proceedings of Mach. Learn. Research, vol. 202. PMLR, Jul 2023, pp. 28 492–28 518.

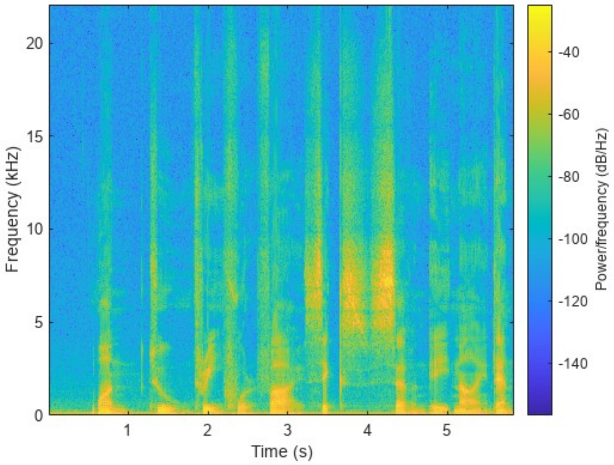


Fig. 15: Denoising effect LLM Spectrogram Output

[20] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th Int. Conf. on Mach. Learn.*, ser. Proceedings of Mach. Learn. Research, vol. 202. PMLR, Jul 2023, pp. 5178–5193.

[21] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada, J. Li, S. Ghosh, J. Balam, and B. Ginsburg, "Salm: Speech-augmented language model with in-context learning for speech recognition and translation," in *ICASSP 2024 - 2024 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr 2024, pp. 13 521–13 525.