**IMPLEMENTATION OF NEUROMORPHIC HEARING AID WITH NOISE SUPPRESSION**

**PHASE Ⅱ REPORT**

***Submitted by***

**JEROME ALBERT A**

***In partial fulfilment for the award of the degree***

***of***

**MASTER OF ENGINEERING IN**

**VLSI DESIGN AND EMBEDDED SYSTEMS**

**DEPARTMENT OF ELECTRONICS ENGINEERING**

**MADRAS INSTITUTE OF TECHNOLOGY**

**ANNA UNIVERSITY, CHROMPET**

**CHENNAI-600 044**

**APRIL 2025**

**IMPLEMENTATION OF NEUROMORPHIC HEARING AID WITH NOISE SUPPRESSION**

**PHASE Ⅱ REPORT**

***Submitted by***

**JEROME ALBERT A**

***In partial fulfilment for the award of the degree***

***of***

**MASTER OF ENGINEERING IN**

**VLSI DESIGN AND EMBEDDED SYSTEMS**

**DEPARTMENT OF ELECTRONICS ENGINEERING**

**MADRAS INSTITUTE OF TECHNOLOGY**

**ANNA UNIVERSITY, CHROMPET**

**CHENNAI-600 044**

**APRIL 2025**

**ANNA UNIVERSITY, CHENNAI**

**BONAFIDE CERTIFICATE**

Certified that this Report titled “**IMPLEMENTATION OF NEUROMORPHIC HEARING AID WITH NOISE SUPPRESSION**” is the Bonafide work of **JEROME ALBERT (2023617030)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate

SIGNATURE

**Mr. K. VEERAPPAN**

Faculty and Guide

Department of Electronics Engineering

Madras Institute of Technology

Anna University

Chennai – 600 044

SIGNATURE

**Dr. P.T.V. BHUVANESWARI**

Professor and Head

Department of Electronics Engineering

Madras Institute of Technology

Anna University,

Chennai – 600 025

**ABSTRACT**

Speech enhancement is critical for improving speech intelligibility and quality in various audio devices. In recent years, deep learning-based methods have significantly improved speech enhancement performance, but they often come with a high computational cost, which is prohibitive for a large number of edge devices, such as headsets and hearing aids. This work proposes an ultra-low-power speech enhancement system based on the brain-inspired spiking neural network (SNN) called Spiking-FullSubNet. Spiking-FullSubnet follows a full-band and sub-band fusion approach to effectively capture both global and local spectral information. To enhance the efficiency of computationally expensive sub-band modelling, we introduce a frequency partitioning method inspired by the sensitivity profile of the human peripheral auditory system. Furthermore, a novel spiking neuron model that can dynamically control the input information integration and forgetting, enhancing the multi-scale temporal processing capability of SNN, which is critical for speech denoising**.** The solution exploits the efficiency of spiking neural networks (SNNs) to achieve high-performance enhancement with minimal computational overhead—making it suitable for embedded, power-constrained environments**.**

**ACKNOWLEDGEMENT**

I would like to record my sincere thanks to **Dr.K.RAVICHANDRAN**, Dean, Madras Institute of Technology, for having given consent to carry out the project work.

I wish to express my sincere thanks to **Dr.PT.V.BHUVAMESHWARI** , Professor and Head, Department of Electronics Engineering, Madras Institute of Technology, Anna University.

With esteem respect, I express my sincere gratitude to my project guide **Mr.K.VEERAPPAN**, Department of Electronics Engineering, Madras Institute of Technology for her guidance, constant encouragement and support. Her extensive vision, practical inputs and creative thinking has been a source of inspiration throughout this project.

I wish to extend my sincere thanks to the project review panel members Dr.M.KANNAN, Professor, Dr.M.MANIKANDAN, Professor, Dr.S.P.JOY VASANTHA RANI, Professor and Dr.K.MARIAMMAL, Associate Professor for their valuable suggestions and encouragement during the reviews.

I wish to express my gratitude to our project coordinator **Mr.K.VEERAPPAN**, Teaching Fellow, Department of Electronics Engineering and to all the Teaching and Non-teaching faculties in the Department of Electronics Engineering for supporting and extending help whenever needed.

**PLACE:** CHENNAI -44

JEROME ALBERT A

2023617030

**DATE:**

# **TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** | **Ⅴ** |
|  | **LIST OF FIGURES** | **Ⅹ** |
|  | **LIST OF TABLES** | **Ⅺ** |
|  | **LIST OF SYMBOLS, ABBREVATIONS** | **Ⅺ** |
| **1** | **INTRODUCTION** | **1** |
|  | 1.1 Objective | 1 |
|  | 1.2 Hearing Loss and Assistive technology | 1 |
|  | 1.3 Noise Suppression for Hearing Aids | 2 |
|  | 1.4 Biological Inspired System | 3 |
|  | 1.5 Spiking Neural network (SNN) in Audio Processing | 4 |
| **2.** | **LITERATURE REVIEW** | **7** |
| **3.** | **SYSTEM ARCHITECURE** | **16** |
|  | 3.1 High-Level Block diagram | 16 |
|  | 3.1.1 Audio Input | 16 |
|  | 3.1.2 I2S MEMS Microphone | 17 |
|  | 3.1.3 Processing Algorithm | 17 |
|  | 3.1.4 Spiking FullSubnet | 18 |
|  | 3.1.5 Noise Suppressed and Enhanced Speech Output | 19 |
|  | 3.2 Noisy Speech Database for Training Speech Enhancement Algorithms | 19 |
|  | 3.3 PREPROCESSING AND FEATURE EXTRACTION | 20 |
|  | 3.3.1 Normalisation | 21 |
|  | 3.3.2 Spectral Analysis | 21 |
|  | 3.3.3 Ideal Ratio Masking | 23 |
|  | 3.4 FEATURE EXTRACTION | 25 |
|  | 3.4.1 Noisy vs Clean Wave form | 25 |
|  | 3.4.2 Spectrogram Extraction of Noisy and Clean speech | 26 |
|  | 3.4.3 Gammatone -Filter Bank Feature Extraction | 28 |
|  | 3.4.4 Ideal Ratio Mask (IRM) from Gammatone Filtered Outputs | 30 |
| **4.** | **NOISE SUPPRESSION AND SPEECH ENHANCEMENT MODEL TRAINING** | **32** |
|  | 4.1 Speech Enhancement | 32 |
|  | 4.2 Sub-band Modeling in Speech Enhancement | 32 |
|  | 4.3 Neuromorphic Speech Processing | 34 |
|  | 4.4 Spiking Neuron Model | 35 |
|  | 4.5 Proposed Fullsubnet Model | 35 |
|  | 4.5.1 Gated Spiking Neuron | 35 |
|  | 4.5.2 Spiking-FullSubnet Architecture | 37 |
|  | 4.5.2.1 Full-Band Processing | 37 |
|  | 4.5.2.2 Sub-Band Processing | 38 |
|  | 4.5.2.3 Sub-Band Processing Based on Frequency Partitioning | 40 |
|  | 4.6 Implementation Details | 42 |
| **5.** | **RESULTS AND DISCUSSION** |  |
|  | 5.1 Evaluation of Speech Enhancement Using DNSMOS Scores | 43 |
|  | 5.2 SI-SNRi Evaluation Metrics | 45 |
| **7.** | **CONCLUSION AND FUTURE SCOPE** | **47** |
|  | **REFERENCES** | **48** |

# **LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **DESCRIPTION** | **PAGE NO** |
| 1.5 | Illustration of Spiking Neuron Network Model | 5 |
| 3.1 | Block diagram of the proposed neuromorphic speech enhancement system | 16 |
| 3.2 | Spiking Full-subnet with Full band and sub-band Architecture | 18 |
| 3.3 | Frequency vs Time of Noisy and Clean Waveform | 26 |
| 3.4 | Spectrogram Extraction of Noisy and Clean speech | 27 |
| 3.5 | Magnitude Spectrogram of Noisy Spectrogram  (a) clean Spectrogram (b) with Gammatone Filter Bank | 29 |
| 3.6 | IRM Speech Mask (a) IRM Noise Mask (b) | 31 |
| 4.5 | Illustration of the proposed GSN model | 36 |
| 4.6 | Illustration of the sub-band processing in Spiking FullSubNet. | 39 |
| 4.7 | Frequency Partitioning of Input for sub band | 41 |

# **LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **DESCRIPTION** | **PAGE NO** |
| 5.1 | DNSMOS scores of the raw noisy dataset versus the enhanced audio after training using the Spiking FullSubNet model | 43 |
| 5.2 | Signal Invariant-Signal to Noise Ratio Improvement Before and After Enhancement | 46 |

**LIST OF ABBREVATIONS**

|  |  |
| --- | --- |
| SNN | Spiking neural network |
| DSP | Digital signal processing |
| ANN | Artificial Neural Network |
| SRNN | Spiking Recurrent Neural Network |
| DPRNN | Dual-Path Recurrent Neural Network |
| I2S | Inter Integrated Sound |
| MEMS | Micro Electro Mechanical Sensor |
| STFT | Short-Time Fourier Transform |
| RMS | Root Mean Square |
| Istft | Inverse Short-Time Fourier Transform |
| cIRM | Complex Ideal Ratio Masks |
| RSNN | Recurrent networks of spiking neurons |
| GSN | Gated Spiking Neuron |
| DNSMOS | Deep Noise Suppression Mean Opinion Score |
| SI-SNRi | Scale-Invariant Signal-to-Noise Ratio |

# **CHAPTER**

## INTRODUCTION

* 1. **OBJECTIVE**

The project aims to design and implement a real-time hearing aid system that mimics the human cochlea’s biological processing to enhance speech clarity and suppress background noise. Leveraging neuromorphic computing and Spiking Neural Networks (SNNs), the system will replicate the cochlea’s frequency selectivity and dynamic range compression for natural sound perception. By employing biologically inspired signal processing and adaptive noise suppression algorithms, the system will distinguish speech from noise in real time, improving audibility in challenging acoustic environments. The use of SNNs ensures energy-efficient, event-driven processing, closely resembling neural auditory pathways. This approach bridges neuroscience and engineering to deliver a next-generation hearing aid with human-like auditory processing.

* 1. **HEARING LOSS AND ASSISTIVE TECHNOLOGIES**

Hearing loss affects over 466 million people worldwide (WHO, 2021), with projections suggesting this could rise to 900 million by 2050. Conventional hearing aids, while beneficial, struggle in noisy environments due to their reliance on linear amplification and basic noise reduction algorithms. These devices often amplify both speech and background noise equally, reducing clarity in challenging acoustic settings like crowded restaurants or traffic-heavy streets. Studies indicate that 60% of hearing aid users report dissatisfaction in noisy conditions, highlighting the need for advanced signal processing techniques.

The limitations of current technologies stem from their inability to mimic the human auditory system’s selective attention and dynamic filtering. Traditional digital signal processing (DSP) methods, such as spectral subtraction or beamforming, often introduce artifacts or fail to preserve natural speech quality. This creates a demand for biologically inspired solutions that can distinguish speech from noise in real time while maintaining low latency and energy efficiency—critical for wearable devices.

Emerging advancements in neuromorphic computing and machine learning offer promising solutions by replicating the brain’s auditory processing mechanisms. A hearing aid that leverages spiking neural networks (SNNs) and cochlea-like frequency analysis could dynamically suppress noise while enhancing speech intelligibility, bridging the gap between clinical efficacy and user satisfaction. Such innovations could revolutionize assistive hearing technologies, particularly for aging populations and individuals with sensorineural hearing loss.

* 1. **NOISE SUPPRESSION FOR HEARING AIDS**

Hearing aids are assistive devices designed to amplify sound for individuals with hearing loss. While traditional hearing aids have been effective in amplifying speech, they often struggle in noisy environments such as busy streets, restaurants, or crowded public spaces. In these situations, background noise can overwhelm the desired speech signal, making it difficult for users to understand conversations and leading to listening fatigue. One of the core challenges in modern hearing aid technology is the ability to distinguish between speech and noise in real-time and suppress the unwanted background components without distorting the clarity of speech. Conventional noise suppression techniques such as spectral subtraction, Wiener filtering, and beamforming have been used, but they typically require manual tuning, lack adaptability, or introduce audible artifacts. With the rise of deep learning, data-driven models like FullSubNet and DNN-based denoisers have significantly improved speech enhancement quality. However, their computational requirements are often too high for real-time execution on resource-constrained, battery-powered devices like hearing aids.

This has led to increased interest in neuromorphic computing and spiking neural networks (SNNs), which emulate the brain’s energy-efficient way of processing information. By leveraging SNNs for noise suppression, it becomes possible to build low-power, low-latency hearing aids that can intelligently enhance speech even in highly dynamic acoustic environments.

* 1. **BIOLOGICAL INSPIRED SYSTEM**

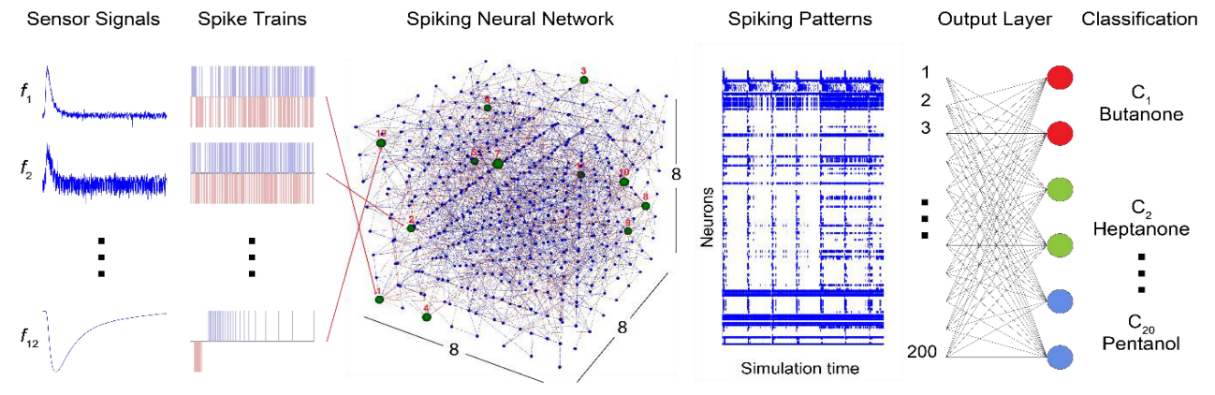
The human auditory system is a sophisticated and highly efficient biological framework that serves as a key inspiration for neuromorphic hearing systems. At the core of this system lies the cochlea—a spiral-shaped organ in the inner ear responsible for breaking down incoming sound waves into their constituent frequency components. This is achieved through a process known as tonotopic frequency analysis, where different regions of the cochlea respond to different frequencies, similar to a filter bank. Within the cochlea, inner hair cells play a critical role in transducing mechanical vibrations into electrical signals. These signals are then relayed via the auditory nerve to the brain, with precise encoding of both temporal (timing) and spectral (frequency-based) characteristics of sound. The auditory nerve uses a spike-based representation, firing action potentials that preserve timing cues and frequency information critical for understanding speech, detecting pitch, and separating background noise.

By mimicking these biological mechanisms, neuromorphic hearing systems can achieve more natural and efficient processing of audio signals, especially in complex and noisy environments. Traditional digital signal processing (DSP) techniques often rely on fixed mathematical operations that do not adapt well to non-stationary real-world noise or variability in speech patterns. In contrast, biologically inspired systems use adaptive, event-driven processing that not only reduces computational overhead but also enhances robustness to noise, latency, and power consumption. These advantages are particularly beneficial for applications such as real-time noise suppression and speech enhancement in wearable or low-power devices like hearing aids. As such, the design of cochlea-inspired architectures and spiking neural networks represents a significant step toward bridging the gap between artificial systems and the remarkable performance of the human auditory system.

* 1. **SPIKING NEURAL NETWORK (SNN) IN AUDIO PROCESSING**

Spiking Neural Networks (SNNs) are a class of brain-inspired computational models that represent the third generation of neural networks. Unlike traditional Artificial Neural Networks (ANNs), which process information in a continuous, frame-based manner, SNNs operate using event-driven computation and temporal coding. In an SNN, neurons communicate through discrete electrical pulses or “spikes,” much like biological neurons in the brain. These spikes are transmitted only when the membrane potential of a neuron reaches a threshold, enabling the network to process data in a highly sparse and energy-efficient way. This approach allows SNNs to naturally encode and preserve temporal information, which is particularly important for time-dependent signals such as speech and environmental sounds.

In the context of audio processing, SNNs offer several advantages over conventional ANNs. While ANNs typically require high sampling rates, dense activations, and constant computation across all layers regardless of input significance, SNNs activate only in response to relevant acoustic events. This not only reduces computational load and power consumption but also enables real-time processing on edge devices like hearing aids or embedded systems. Additionally, because sound is inherently a temporal phenomenon, SNNs are better suited to represent fine-grained timing information—such as the onset and offset of phonemes—thanks to their temporal dynamics and asynchronous updates.



**Fig 1.5 Illustration of Spiking Neuron Network Model**

One of the key motivations for using SNNs in auditory applications is their ability to replicate the spiking behavior of auditory nerve fibers, which transmit sound information from the cochlea to the auditory cortex. These fibers encode both spectral (frequency) and temporal (timing) features of sound through precisely timed spike trains. By emulating this biological behavior, SNNs can more accurately capture the temporal evolution of audio signals and maintain phase and timing cues critical for tasks like speech recognition and noise suppression. This makes them particularly effective in neuromorphic hearing systems, where maintaining the fidelity of the temporal structure of speech can significantly improve intelligibility in noisy or reverberant conditions.

# **CHAPTER 2**

## 2.1 LITERATURE REVIEW

The paper by Claudia Lenk,.et al [1] discusses the design and implementation of a dynamic Micro-Electro-Mechanical Systems (MEMS) cochlea, which functions as an acoustic sensor. designed to improve speech processing in noisy environments. It integrates signal processing at the transduction level and uses real-time feedback to dynamically adjust sensing properties. The system exhibits nonlinear behavior based on a Hopf bifurcation, enhancing signal detection, expanding dynamic range, and adapting to varying acoustic conditions. The MEMS-based cochlea includes piezo-resistive deflection sensing and thermo-mechanical actuation, allowing real-time tuning of response characteristics. Experimental validation confirms its efficiency in filtering and amplifying sounds while managing interference.

The Lightweight Full-band and Sub-band Fusion Network for Real-Time Speech Enhancement [2] by Zhuangqi Chen, Pingjian Zhangpresents a novel lightweight full-band and sub-band fusion network (LFSFNet) aimed at enhancing speech quality in real-time applications. The proposed method achieved superior performance metrics (PESQ=2.99, CSIG=3.8, CBAK=3.5, COVL=3.7) compared to traditional models and other recent approaches. The effectiveness of the Gaussian weights in sub-band and full-band processing was highlighted, improving the overall performance of the model

The work Neuromorphic Lip-Reading with Signed Spiking Gated Recurrent Units by Zhuangqi Chen, Pingjian Zhang [3] addresses the challenge of Automatic Lip-Reading (ALR) tasks, especially for portable edge applications like hearing aids. The authors introduce a new SNN model called the Signed Spiking Gated Recurrent Unit (SpikGRU2+).The proposed model outperforms previous SNNs by 25% and even outperforms the best traditional Artificial Neural Networks (ANNs) by 4% on the DVS-Lip dataset

This study Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware [4] by Nitin Rathi, .et al surveys the rapidly evolving field of neuromorphic computing, focusing on Spiking Neural Networks (SNNs) as a promising paradigm for building energy-efficient, brain-inspired AI systems. The paper explores major algorithmic advancements, including various neuron models such as leaky-integrate-and-fire (LIF), integrate-and-fire (IF), and stochastic models.

The paper introduces DPSNN, a novel spiking neural network architecture designed for low-latency streaming speech enhancement [5] by Tao Sun, and Sander Bohté , targeting real-time applications such as hearing aids and embedded systems. Traditional deep neural network (DNN) solutions, especially those using Short-Time Fourier Transform (STFT), suffer from substantial algorithmic latency (~32ms), which is unsuitable for latency-critical environments. To overcome this limitation, the authors propose a time-domain spiking framework inspired by the Dual-Path Recurrent Neural Network (DPRNN), replacing STFT/iSTFT with a learned convolutional encoder-decoder structure. The DPSNN architecture comprises two key modules: a Spiking Convolutional Neural Network (SCNN) for capturing temporal context and a Spiking Recurrent Neural Network (SRNN) for modeling frequency-related features.

This paper by Abir Riahi and Eric Plourde presents a spiking neural network (SNN) based on a U-Net architecture for single-channel speech enhancement [6]. Traditional artificial neural networks (ANNs) require high computational power and energy, making them unsuitable for real-time applications on low-resource devices. The proposed SNN leverages neuromorphic hardware, allowing energy-efficient speech processing. Unlike previous SNN-based methods that rely on masking approaches, this model directly learns to map noisy speech to clean speech using surrogate gradient-based optimization. This approach demonstrates the potential of SNNs in speech enhancement applications, offering a balance between performance and efficiency.

The work Robust Speaker Recognition Using Speech Enhancement And Attention Model [7] by Yanpei Shi, et al. It aims to improve speaker recognition performance when speech signals are corrupted by noise. Instead of separately processing speech enhancement and speaker recognition, the two modules are integrated into one framework by a joint optimisation using deep neural networks. Furthermore, to increase the robustness against noise, a multi-stage attention mechanism is employed to highlight the speaker related features learned from context information in both time and frequency domains. the use of SE+SID yields better performances for speaker identification. After using multi-stage attention models, SE+SID-MS and SE-MS+SID, about 2∼3% further improvements on Top-1 and Top-5 accuracy are obtained in comparison with the baseline in all noise conditions.

The work Speech Enhancement for Multimodal Speaker Diarization System [8] by Rehan Ahmad, et al, the performance of proposed multimodal speaker diarization system under noisy conditions. Two types of noises comprising additive white Gaussian noise (AWGN) and realistic environmental noise is used to evaluate the system. To mitigate the effect of noise, add an LSTM based speech enhancement block in our diarization pipeline. This block is trained on synthesized data set with more than 100 noise types to enhance the noisy speech. The enhanced speech is further used in multimodal speaker diarization system which utilizes a pre-trained audio-visual synchronization model to find the active speaker.. A subset of AMI corpus consisting of 5.4 h of recordings is used in this analysis.

Modern speech enhancement algorithms achieve remarkable noise suppression by means of large recurrent neural networks (RNNs). In this work TinyLSTMs: Efficient Neural Speech Enhancement for Hearing Aids[9] by Igor Fedorov, .et al ,they use model compression techniques to bridge this gap. We define the constraints imposed on the RNN by the HW. Although model compression techniques are an active area of research, first to demonstrate their efficacy for RNN speech enhancement, using pruning and integer quantization of weights/activations. They also demonstrate state update skipping, which reduces the computational load. Finally, a perceptual evaluation of the compressed models to verify audio quality on human raters. Results show a reduction in model size and operations of 11.9× and 2.9×, respectively, over the baseline for compressed models, without a statistical difference in listening preference and only exhibiting a loss of 0.55dB SDR. The model achieves a computational latency of 2.39ms, well within the 10ms target and 351× better than previous work.

In this Fullsubnet: A Full-Band And Sub-Band Fusion Model For Real-Time Single-Channel Speech Enhancement [10] by Xiang Hao, propose a full-band and sub-band fusion model, named as FullSubNet, for real-time single-channel speech enhancement. This model is designed to integrate the advantages of the full-band and the sub-band models, that is, it can capture the global (full-band) spectral information and the long-distance cross-band dependencies, meanwhile retaining the ability to modeling signal stationarity and attending the local spectral pattern. Regarding the computational complexity, the one STFT frame (32 ms) processing time of the proposed model (PyTorch implementation) is 10.32 ms tested on a virtual quad-core CPU (2.4 GHz) based on Intel Xeon E5-2680 v4, which obviously meets the realtime requirement.

DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement [11] by Yanxin Hu, Conventional time-frequency (TF) domain methods focus on predicting TF-masks or speech spectrum, via a naive convolution neural network (CNN) or recurrent neural network (RNN). In order to train the complex target more effectively, in this paper, we design a new network structure simulating the complex-valued operation, called Deep Complex Convolution Recurrent Network (DCCRN), where both CNN and RNN structures can handle complex-valued operation. The proposed DCCRN models are very competitive over other previous networks, either on objective or subjective metric. With only 3.7M parameters, the DCCRN models submitted to the Interspeech 2020 Deep Noise Suppression (DNS) .

TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation [12] by Zhong-Qiu Wang. The model is a novel deep neural network (DNN) integrating full- and sub-band modeling in the time-frequency (T-F) domain. It stacks several blocks, each consisting of an intra-frame full-band module, a sub-band temporal module, and a cross-frame self-attention module.It obtains a state-of-the-art 23.5 dB improvement in scale-invariant signal-to-distortion ratio (SI-SDR) on WSJ0-2mix, a standard dataset for two-speaker separation. To show its robustness to noise and reverberation, evaluate it on monaural reverberant speaker separation using the SMS-WSJ dataset and on noisy-reverberant speaker separation using WHAMR!, and obtain state-of-the-art performance on both datasets. Extend TF-GridNet to multi-microphone conditions through multi-microphone complex spectral mapping, and integrate it into a two-DNN system with a beamformer in between (named as MISO-BF-MISO in earlier studies), where the beamformer proposed in this article is a novel multi-frame Wiener filter computed based on the outputs of the first DNN. State-of-the-art performance is obtained on the multi-channel tasks of SMS-WSJ and WHAMR!. Besides speaker separation.

Spectro-Temporal SubNet for Real-Time Monaural Speech Denoising and Dereverberation [13] by Feifei Xiong, Preserving the advantages of subband model (SubNet) that processes each frequency band independently and requires small amount of resources for good generalization, the proposed framework named STSubNet exploits sufficient spectro-temporal receptive fields (STRFs) from speech spectrum via a two-dimensional convolution network cooperating with a bi-directional long short-term memory network across frequency bands,increases SRMR by nearly 1.0 on average when compared to the best model-based method STM is summarized via a leave-one-out procedure to compare to the baseline (T = 15, S = 31, biLSTM), i.e., temporal context degraded to T = 2, spectral context degraded to S = 1, or without biLSTM.

The Work Real-Time Denoising And Dereverberation Wtih Tiny Recurrent U-Net [14] By Hyeong-Seok Choi, .Et Al.Modern deep learning-based models have seen outstanding performance improvement with speech enhancement tasks. The number of parameters of state-of-the-art models, however, is often too large to be deployed on devices for real-world applications. To this end, they propose Tiny Recurrent U-Net (TRU-Net), a lightweight online inference model that matches the performance of current state-ofthe-art models. The size of the quantized version of TRU-Net is 362 kilobytes, which is small enough to be deployed on edge devices. In addition, we combine the small-sized model with a new masking method called phase-aware β-sigmoid mask, which enables simultaneous denoising and dereverberation.

Towards spike-based machine intelligence with neuromorphic computing [15] by Kaushik Roy et al. Overview of neuromorphic computing, a brain-inspired approach to artificial intelligence that aims to enhance energy efficiency. It explores the evolution of neuromorphic systems from early silicon circuits mimicking biological neurons to modern spiking neural networks (SNNs). The discussion highlights key differences between conventional computing and brain-like processing, emphasizing event-driven representations and hierarchical neural structures. Additionally, the paper examines challenges in algorithm-hardware co-design and future prospects for neuromorphic intelligence. Overall, it presents insights into the advancements and limitations of neuromorphic computing, positioning it as a promising alternative to conventional AI models.

A Hybrid Neural Coding Approach for Pattern Recognition with Spiking Neural Networks [16] by Xinyi Chen, et al. explores a hybrid neural coding approach for pattern recognition using Spiking Neural Networks (SNNs). It argues that integrating multiple neural coding schemes enhances accuracy, response time, efficiency, and robustness. The proposed framework introduces a diverse set of neural coding strategies and a flexible assignment method tailored to specific tasks. The study demonstrates improvements in image classification and sound localization, achieving high accuracy with reduced inference latency and energy consumption. This research offers valuable insights into optimizing neuromorphic systems for practical applications.

Progressive Tandem Learning for Pattern Recognition with Deep Spiking Neural Networks [17] by Jibin Wu, et al. To compensate for the approximation errors arising from the primitive network conversion, we further introduce a layer-wise learning method with an adaptive training scheduler to fine-tune the network weights. On the MNIST and Cifar-10 datasets, the low-precision SNN models perform exceedingly well regardless of the reduced bit-width and the limited representation space (i.e., Ns = 16). Specifically, when the weights are quantized to 4-bit, the classification accuracy drops by only 0.03% and 0.85% on the MNIST and Cifar-10 datasets, respectively. Therefore, the proposed PTL framework offers immense opportunities for implementing SNNs on the low-precision neuromorphic hardware, for instance with emerging non-volatile memory devices that suffering from limited conductance states.

Fast-SNN: Fast Spiking Neural Network by Converting Quantized ANN [18] by Yangfan Hu, et al. introduces Fast-SNN, a spiking neural network (SNN) conversion method designed for low latency and high accuracy. It addresses challenges in ANN-to-SNN conversion by minimizing quantization and accumulating errors through a novel training approach. The framework leverages a signed IF neuron model and layer-wise fine-tuning to optimize network performance. Experimental results demonstrate state-of-the-art accuracy in computer vision tasks, including image classification, object detection, and semantic segmentation. Fast-SNN offers improved energy efficiency and computational speed, making it suitable for neuromorphic computing applications.

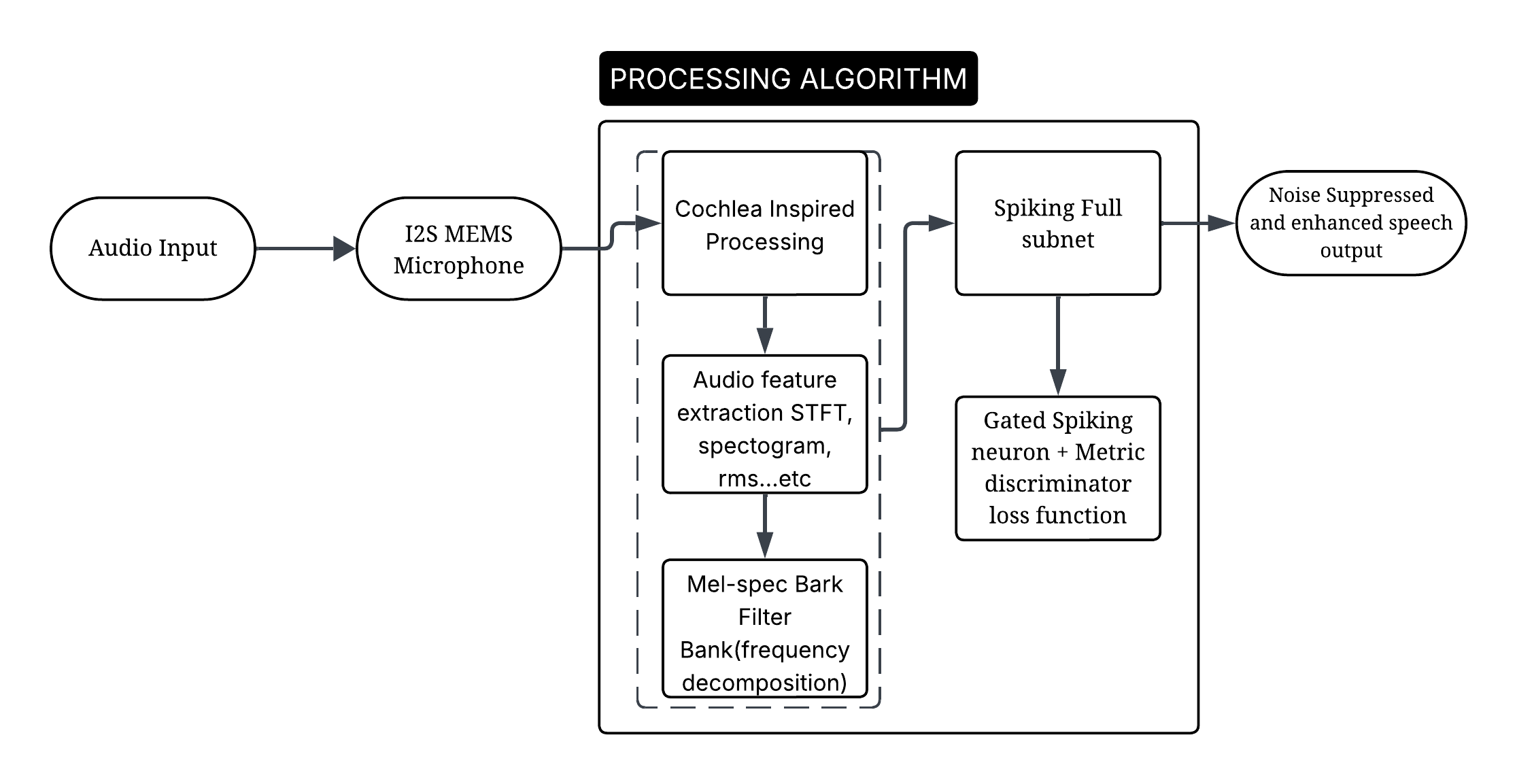
Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition by Jibin Wu,et al[19]. Explores the application of Deep Spiking Neural Networks (SNNs) for Large Vocabulary Automatic Speech Recognition (ASR). Traditional Artificial Neural Networks (ANNs) require significant computational resources, whereas SNNs offer energy-efficient, spike-based computation suitable for neuromorphic hardware. The study evaluates SNN-based acoustic models across various ASR benchmarks, demonstrating competitive accuracy compared to ANN counterparts while significantly reducing computational cost and inference time. The integration of deep SNNs with neuromorphic computing presents a promising solution for on-device speech recognition in mobile and embedded systems. This research highlights the potential of SNNs in advancing ASR technology with improved efficiency and performance.

Weighted Speech Distortion Losses For Neural-Network-Based Real-Time Speech Enhancement [20] by Yangyang Xia , investigates real-time speech enhancement using recurrent neural networks (RNNs). It introduces novel loss functions that balance speech distortion and noise reduction, optimizing intelligibility and quality. The study evaluates feature normalization techniques and sequence lengths, demonstrating their impact on enhancement performance. Experimental results show significant improvements over traditional methods, with better noise suppression and speech clarity. The effect of speech distortion weighting is shown in Fig. 3, where or are changed to search the optimal points for each ob jective measure. Curiously, only STOI and CD agree on the same co efficient in both cases, while both PESQ and SI-SDR suggest smaller weight on speech distortion. The optimal SNR weights for all metrics are concentrated around 20 dB.

## CHAPTER 3

## NEUROMORPHIC NOISE SUPPRESSION METHODOLOGY

## 3.1 HIGH-LEVEL BLOCK DIAGRAM

**** The proposed system implements a real-time noise suppression pipeline for hearing aid applications using a Spiking FullSubNet model. The architecture, as shown in the Figure 3.1, can be broken down into the following key components:

**Figure 3.1 Block diagram of the proposed neuromorphic speech enhancement system.**

## 3.1.1 Audio Input

The system starts with an audio input signal captured from the environment. This input could include both speech and background noise.

## 3.1.2 I2S MEMS Microphone:

The audio signal is acquired using a digital MEMS microphone interfaced over the I2S protocol. MEMS microphones are compact, power-efficient, and widely used in portable devices such as hearing aids and wearables. The digitized audio signal is then passed to the processing unit.

## 3.1.3 Processing Algorithm

This is the core processing block which includes multiple subcomponents designed to simulate biological hearing and perform energy-efficient noise suppression.

**a. Cochlea-Inspired Processing**

This stage mimics the biological cochlea by transforming raw audio into perceptually meaningful features. It captures auditory cues similar to how the human ear filters sound across different frequencies and loudness levels.

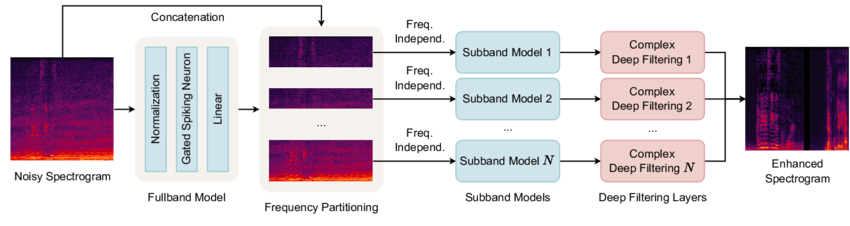
**b. Audio Feature Extraction**

Here, essential features such as Short-Time Fourier Transform (STFT), spectrogram, and RMS energy are computed. These features provide time-frequency representations of the audio signal, crucial for distinguishing between speech and noise.

**c. Mel-Spectrogram Gammatone Filter Bank**

The extracted audio features are further processed using a mel-spectrogram-based Gammatone filter bank. This filter bank decomposes the signal into perceptually relevant frequency bands, reflecting how human hearing is sensitive to different frequencies (critical bands). This step ensures that the input to the neural network is aligned with auditory perception.

## 3.1.4 Spiking FullSubNet

 This module performs noise suppression using a Spiking FullSubNet model, which is a neuromorphic adaptation of FullSubNet, a deep learning-based speech enhancement model. It is tailored to run on low-power hardware by replacing traditional artificial neurons with spiking neurons, which fire only when needed, mimicking biological brains.

**Figure 3.2 Spiking Full-subnet with Full band and sub-band Architecture**

**a. Gated Spiking Neuron + Metric Discriminator Loss Function**

The network leverages Gated Spiking Neurons for better temporal feature modeling. It is trained using a Metric Discriminator Loss Function, which improves the separation between clean speech and noise in the latent space. This results in high speech quality and intelligibility while maintaining computational efficiency.

## 3.1.5 Noise Suppressed and Enhanced Speech Output

Finally, the enhanced, denoised audio signal is reconstructed and outputted. The result is a noise-suppressed speech signal that is clearer and more intelligible, suitable for hearing aid users even in complex acoustic environments.

## 3.2 NOISY SPEECH DATABASE FOR TRAINING SPEECH ENHANCEMENT ALGORITHMS

We utilize the publicly available Noisy-clean VCTK Corpus, Noisy speech database for training speech enhancement algorithms for the speech enhancement performance evaluation. This comprehensive dataset contains a wide array of human speech audio samples in multiple languages, Speech recordings from 109 English speakers with various accents. Each speaker reads approximately 400 sentences, including material from newspapers, the Rainbow Passage, and an accent-eliciting paragraph. Repository provides a synthesizer script that generates clean (ground truth), noise(additive), and noisy (ground truth plus noise) audio segments for both training and validation. The speech samples are high-quality, recorded at 48kHz, and are suitable for both training and evaluating speech enhancement models. The 28-speaker dataset consists of clean and noisy dataset wav files, Noise samples from the 56 speaker DEMAND database (real environmental noise recordings). Speech-shaped noise and babble noise, representative of challenging acoustic environments. synthesize two subsets:

For training and a5-hour subset for validation. All audio samples, synthesized at a sampling rate of 16 kHz, are kept at a consistent duration of 30 seconds. For audios shorter than 30 seconds, we concatenate them with other speech signals from the same speaker, insertinga0.2-second silence interval between clean speech utterances. The noisy audio is simulated using randomly selected speech and noise data, with SNRs ranging from-5 to 20 db. We apply loudness normalization to each noisy audio sample to simulate agnostic input loudness levels from -35to-15 decibels relative to full scale (dBFS). For audio quality evaluation, we employ the metrics, which include SISDR, and Deep Noise Suppression Mean Opinion Score (DNSMOS)

## 3.3 PREPROCESSING AND FEATURE EXTRACTION

### **3.3.1 Normalisation**

Root Mean Square (RMS): RMS Energy estimation is employed as a fundamental metric for analyzing and normalizing speech signals. RMS provides a robust measure of signal power and is calculated using the standard formula:

RMS=

where xi, are individual signal samples.

It is particularly useful for loudness normalization, ensuring that different audio samples maintain consistent perceived volume. Two common strategies are implemented: **Max Normalization** and **RMS Normalization**. Max normalization scales the waveform based on its maximum absolute amplitude, while RMS normalization adjusts it to match a target loudness level, typically set to –25 dBFS (decibels relative to full scale), using the formula

scalar =

where ε is a small constant to avoid division by zero.

**Active RMS Calculation:** ARMS introduced to separately analyze the energy of speech and noise components. This is done by segmenting the audio into 100 ms windows and computing the RMS for each segment. Only segments where the noise RMS exceeds a defined energy threshold of –50 dB are considered "active" and included in the final clean and noise RMS estimates. This process helps isolate meaningful portions of the audio for normalization, avoiding silent or low-energy segments that can distort average energy values. The normalized signals are then scaled based on the calculated active RMS to ensure energy consistency across recordings. These normalization strategies are crucial in enhancing the robustness and perceptual quality of speech enhancement algorithms, especially in noisy environments.

### **3.3.2 Spectral Analysis**

Spectral analysis plays a key role in understanding and enhancing speech signals in the frequency domain using the Short-Time Fourier Transform (STFT). STFT is applied to convert time-domain audio signals into time-frequency representations by segmenting the signal into overlapping frames and analyzing each frame with a Fourier Transform. This provides access to magnitude, phase, real, and imaginary components, enabling deeper analysis and manipulation of spectral characteristics. The system supports flexible output formats—whether complex, real-imaginary pairs, or magnitude-phase pairs—depending on the application. To reconstruct the signal, Inverse STFT (iSTFT) is used, carefully managing the reconstruction based on the input representation. spectral analysis plays a critical role in transforming time-domain audio signals into frequency-domain representations, which are essential for understanding and enhancing speech.

The Short-Time Fourier Transform (STFT) is employed to analyze the signal in overlapping windows, effectively extracting time-localized frequency components. This transformation outputs complex-valued tensors representing each windowed segment of the signal, from which four fundamental components can be derived: magnitude, phase, real, and imaginary parts. The function stft() uses PyTorch's native torch. Stft with a Hanning window, typically defined by three parameters: the number of FFT bins (n\_fft), the hop length (stride between successive windows), and the window length (win\_length). This function is flexible and supports both single- and multi-channel inputs. The complementary inverse function, istft(), reconstructs time-domain signals from the chosen spectrogram representation (e.g., magnitude-phase or real-imaginary), ensuring minimal loss during transformation.

For speech enhancement, integrate Complex Ideal Ratio Masks (cIRM), a widely-used technique for suppressing noise in the spectral domain. The cIRM is computed by comparing the real and imaginary components of the noisy and clean STFTs. The real and imaginary parts of the cIRM are calculated using formulas that account for the alignment of clean and noisy components, normalized by the energy of the noisy input. The resulting mask is further compressed to limit the dynamic range using the function compress\_cIRM, controlled by parameters K = 10 and C = 0.1, following the design proposed in literature (e.g., Williamson et al., 2015, IEEE TASLP).

Additionally, loudness normalization and clipping checks ensure that signals are scaled appropriately before spectral operations. The tune\_dB\_FS() function adjusts the input signal's RMS to a target level of –26 dBFS, standardizing loudness across samples for consistent processing. An optional safety check, is\_clipped(), verifies whether any part of the signal exceeds a clipping threshold (commonly set at 0.999), which would introduce distortion and negatively affect spectral features.

A key innovation in our processing pipeline is the activity detector, which estimates the percentage of speech activity in an audio segment. This is achieved by segmenting the signal into 50 ms frames and computing a smoothed energy probability using a sigmoid-like function p = , where constants a = -1 and b = 0.2 shape the energy thresholding curve. The smoothed probability is updated using an attack-release mechanism (with attack coefficient = 0.8, release coefficient = 0.05), allowing for fast response to increases in signal energy and slower decay otherwise. A frame is considered active if its energy probability exceeds a threshold of 0.13, and the proportion of active frames is returned as an estimate of speech presence. This is particularly useful for evaluating speech versus noise dominance in recordings.

Collectively, this spectral analysis framework enables detailed time-frequency signal decomposition, effective noise suppression via complex masks, loudness standardization, and speech activity estimation. These tools not only enhance speech quality but also ensure the model operates efficiently and robustly across diverse acoustic scenarios.

### **3.3.3 Ideal Ratio Masking**

Ideal Ratio Masking (IRM) is utilized as a foundational technique to separate speech from background noise in the time-frequency domain. The IRM approach operates on complex Short-Time Fourier Transform (STFT) representations of both clean and noisy signals. Given the clean and noise components, the **IRM is computed as a soft mask** using the squared magnitudes of their respective STFTs:

IRMspeech=

IRMnoise​=

where ∣S∣2 and ∣N∣2 are the power spectral densities of the clean and noise signals respectively. These masks have a continuous range between 0 and 1, and are applied multiplicatively to the noisy input to selectively preserve speech-dominant regions and suppress noise-dominant ones.

The implementation allows for multi-channel audio processing, with an optional **reference channel** parameter (typically set to 0) used when computing masks in spatial (multi-microphone) settings. The masks are computed with the same dimensions as the STFT inputs: [B, C, F, T], where B is batch size, C is number of channels, F is number of frequency bins, and T is number of time frames. The final speech and noise masks are shaped as [B, l, F, T], ready to be applied to any single reference channel.

To reconstruct enhanced speech, complex multiplication is employed between the real and imaginary components of the noisy STFT and the estimated masks. This operation is defined by:

Re=noisyr​⋅maskr​−noisyi​⋅maski​, Im=noisyr​⋅maski​+noisyi​⋅maskr​

This effectively applies the mask across both real and imaginary dimensions, preserving phase information while scaling amplitude to reduce noise.

For efficiency and reduced computational cost, especially in models like FullSubNet, the method also supports sub-band frequency selection through a drop\_band() function. This function groups frequencies into num\_groups (e.g., 2 or 4) and drops surplus high-frequency components to ensure consistent dimensionality during parallel training. This is particularly beneficial when dealing with high-resolution STFTs, such as those with 512-point FFTs or more, where computation over all frequency bins becomes redundant or inefficient.

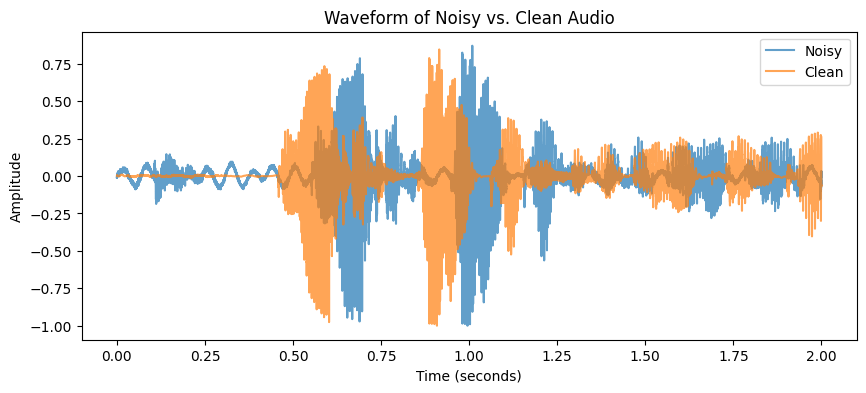
The Mask class also provides a utility for visualizing masks, which is crucial during debugging and evaluation. The generated masks can be plotted using a color map (viridis) to observe how well the model differentiates between speech and noise across time and frequency.

Overall, IRM offers a principled way to suppress noise in the spectral domain and significantly improves the performance of the SNN model by guiding it to focus on clean speech components. This technique, combined with efficient sub-band processing and accurate complex-domain reconstruction, forms a core enhancement mechanism in the proposed speech processing framework.

## 3.4 FEATURE EXTRACTION:

### **3.4.1 Noisy Vs Clean Waveform :**

The waveform plot presents a side-by-side comparison of noisy and clean speech signals in the time domain. The x-axis represents time in seconds, while the y-axis shows the amplitude of the audio signal, which corresponds to the loudness or energy of the sound at any given moment.

This visual comparison highlights the impact of noise on speech signals, which is critical when designing speech enhancement systems. Before applying any enhancement model—such as a spiking neural network (SNN) or traditional DNN-based method—this kind of analysis helps understand how much the noise deviates from the clean signal.

**Figure 3.3 Frequency vs Time of Noisy and Clean Waveform**

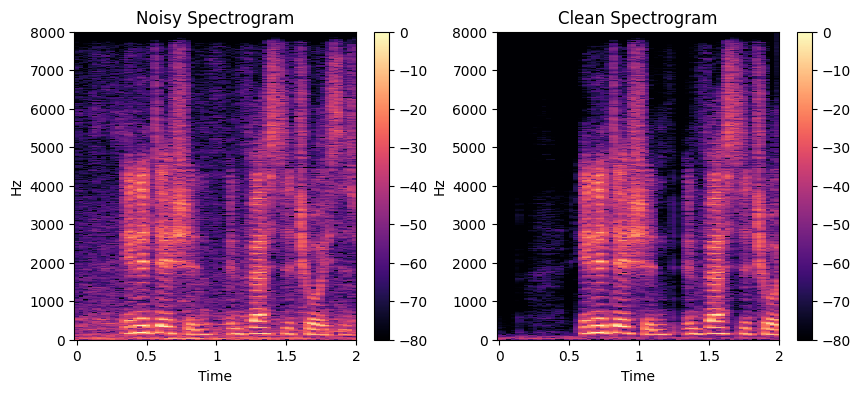
* The **blue waveform** corresponds to the noisy audio, which has been corrupted with environmental or background noise. As seen in the figure, the waveform has larger, inconsistent amplitude variations, especially between the 0.5 to 1.2 second mark. These irregular spikes indicate the presence of external noise sources that distort the original speech.
* The **orange waveform** represents the clean audio, which is the reference or ground truth speech signal without any noise added. Compared to the noisy signal, the clean waveform is smoother and more structured, accurately capturing the natural energy pattern of human speech.

This visual comparison highlights the impact of noise on speech signals, which is critical when designing speech enhancement systems. Before applying any enhancement model—such as a spiking neural network (SNN) or traditional DNN-based method—this kind of analysis helps understand how much the noise deviates from the clean signal.

The difference in waveforms also indicates why time-domain or frequency-domain processing (such as STFT and Ideal Ratio Masking) is necessary: because noise not only adds random amplitude spikes but also alters the temporal dynamics and spectral content of speech. This makes restoring the intelligibility and quality of speech a challenging task in real-time systems like hearing aids. Therefore, this waveform plot provides a baseline visualization to evaluate the effectiveness of the proposed denoising model. Post-enhancement waveforms can later be compared to this to assess how well the model restores the clean speech from noisy inputs.

### **3.4.2 Spectrogram Extraction of Noisy and Clean speech:**

Spectrograms of noisy (left) and clean (right) speech signals, which represent how frequency content evolves over time. The x-axis denotes time (in seconds), the y-axis shows frequency (in Hz), and the color intensity indicates the magnitude of energy (in dB) at each frequency-time point.



**Figure 3.4 Spectrogram Extraction of Noisy and Clean speech**

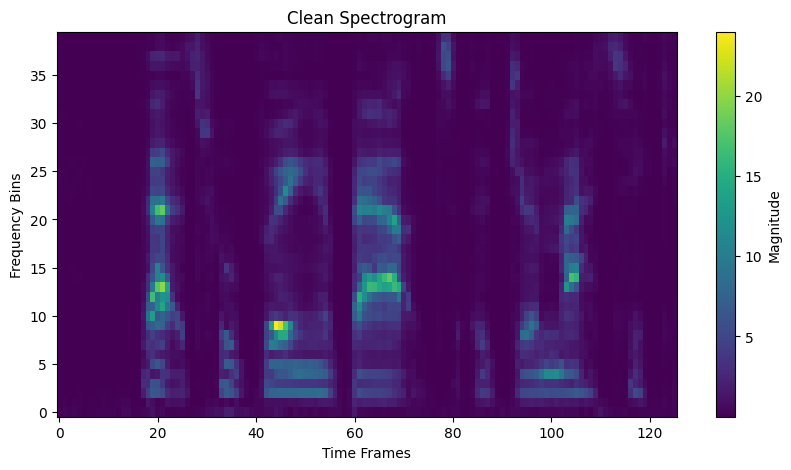
* The **noisy spectrogram** includes extra energy components scattered across various frequency bands, especially in the higher frequencies. These scattered patterns are caused by unwanted background noise, making speech less intelligible and harder to process.
* In contrast, the **clean spectrogram** reveals well-defined harmonic structures and formant patterns typical of human speech. The energy is more concentrated in specific frequency bands (especially below 4000 Hz), indicating the presence of clear vocal information.

This analysis is crucial for noise suppression tasks because it provides a frequency-domain representation of the audio. It helps identify where the noise interferes and guides algorithms like Ideal Ratio Masking (IRM) or Spiking Neural Networks (SNNs) to selectively attenuate noise while preserving important speech features.

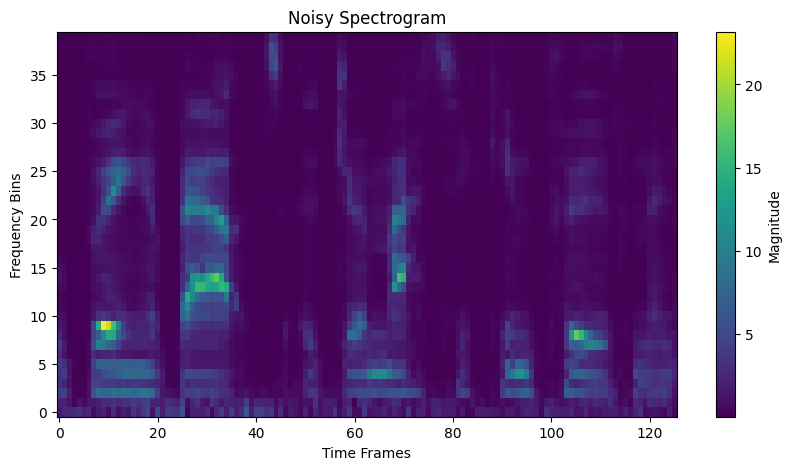
### **3.4.3 Gammatone-Filter Bank Feature Extraction**

The Gammatone filter bank is a perceptual filter bank that maps the linear frequency scale to the Gammatone scale, which better represents how humans perceive sound. Unlike linear or Mel scales, the Gammatone scale is nonlinear, giving more emphasis to low frequencies (where speech energy is concentrated).

* Y-axis: Frequency bins mapped to the Gammatone scale
* X-axis: Time frames
* Color Intensity: Magnitude (energy) of each frequency bin at a specific time



(a)



(b)

**Figure 3.5 Magnitude Spectrogram of Noisy Spectrogram**

**(a) clean Spectrogram (b) with Gammatone Filter Bank**

Parameters used for feature extracting are ,num\_filters: Number of Bark filters to simulate (e.g., 24–64), n\_fft: FFT size, usually a power of 2 like 512 or 1024, sr: Sample rate (commonly 16,000 Hz). low\_freq, high\_freq: Lower and upper frequency limits of analysis (e.g., 0 Hz to 8000 Hz).

It internally uses: hz\_to\_gammatone: Converts physical frequency (Hz) to perceptual Bark scale. gammatone\_to\_hz: Converts back from Bark to Hz.

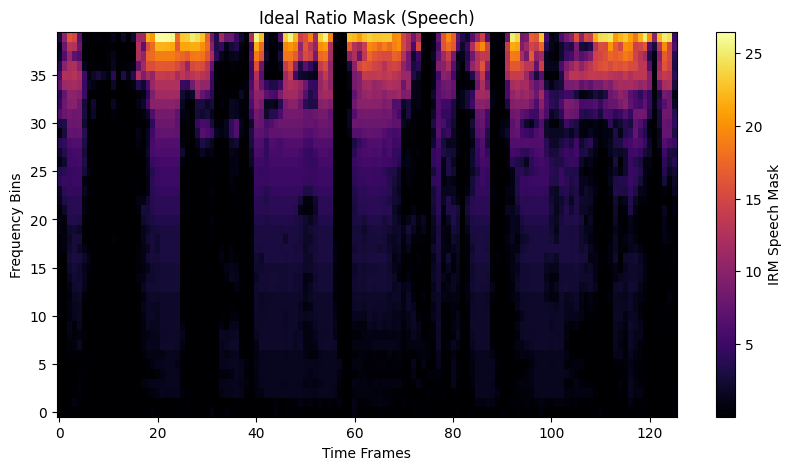
**Noisy Spectrogram:**

The noisy signal contains scattered energy across several frequency bins, especially in non-speech regions, represented by brighter patches in unexpected areas. These high-energy artifacts are due to environmental or background noise, which masks critical speech features.

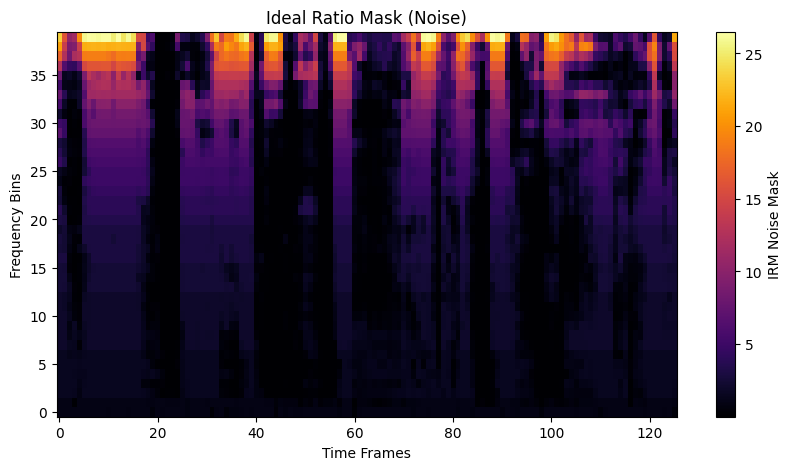
**Clean Spectrogram:**

The clean version shows well-localized and clearly defined bright regions, primarily in low-to-mid frequency bands, corresponding to the core components of human speech. The absence of random high-energy regions indicates that noise has been successfully suppressed, and the important speech patterns are more distinguishable. This comparison demonstrates how the Gammatone filterbank enhances perceptual relevance, enabling better separation of speech from noise for models like SNN-based real-time speech enhancement systems.

### **3.4.4 Ideal Ratio Mask (IRM) from Gammatone Filtered Outputs**

In audio processing, Ideal Ratio Masking (IRM) is a powerful technique used in supervised learning frameworks for speech enhancement and noise suppression. The IRM estimates the ratio of speech to total energy (speech + noise) at each time-frequency bin, helping models distinguish between speech and noise more effectively. The IRM noise mask complements the speech mask by highlighting T-F regions where noise dominates. Bright areas in this mask correspond to **high noise energy** zones. These masks help in training models to learn **what not to retain** during enhancement.

(a)



(b)

**Figure 3.6 IRM Speech Mask (a) IRM Noise Mask (b)**

The IRM speech mask retains time-frequency (T-F) regions where speech is dominant. The brighter regions (yellow/orange) in the image correspond to T-F bins where the speech energy is significantly higher than the noise energy. Darker regions (purple/black) indicate low speech energy or noise-dominant areas that are typically suppressed during enhancement. Frequency bins: 0 to 40, Time frames: 0 to 120+. The maximum mask values (~25+) highlight regions where speech content is strong and clear.

# **CHAPTER 4**

**NOISE SUPPRESSION AND SPEECH ENHACNEMENT MODEL TRAINING**

## 4.1 SPEECH ENHANCEMENT

Speech enhancement aims to improve speech intelligibility and quality which need to remove noise from noisy signals captured by microphones. Existing methods can be divided into time-domain and frequency domain approaches. Time-domain methods directly estimate the clean speech signal, bypassing spectral analysis and waveform synthesis, while frequency-domain methods estimate the spectrogram of the clean speech and then convert it back to the time-domain signal. Between these two approaches, frequency-domain methods have received significant research attention, primarily due to the sparse nature of speech in the frequency domain. Specifically, they can be broadly categorized into spectral magnitude-only enhancement and complex spectrum enhancement. Spectral magnitude-only methods focus on estimating the magnitude of the clean speech spectrum, utilizing the noisy phase for reconstructing the time-domain signal. On the other hand, complex spectrum enhancement methods estimate both the real and imaginary parts of the complex spectrum, which have exhibited greater potential in enhancing speech quality by leveraging the full spectral information.

## 4.2 SUB-BAND MODELING IN SPEECH ENHANCEMENT

In recent years, there has been a notable shift in research focus towards the utilization of sub-band modeling in both single-channel and multi-channel SE In contrast to traditional full-band modeling, sub-band modeling involves the separation of input audio into multiple frequency bands, which are then processed independently. Specifically, each sub-band model takes in a noisy sub-band signal along with its adjacent frequency bands, and then predicts the corresponding clean sub-band signal. This method leverages the distinct stationary characteristics of speech and noise. Speech signals are inherently non-stationary, exhibiting dynamic and variable properties over time. Conversely, many types of noise are relatively stationary, meaning their statistical properties remain more consistent and stable.

In addition, sub-band modeling focuses on the local spectral pattern presented in the current and neighboring frequencies, which has been proven informative for discriminating between speech and other signals. Furthermore, the sub-band models are also effective in modeling the reverberation as the room’s reverberation time (RT60) is frequency-dependent . However, the sub-band modeling approach comes with a trade-off. While it leverages the distinct stationary character istics of speech and noise, it can also result in the loss of the global spectral structure of the speech signal. This global spectral information is also crucial for effective SE. To address this issue, recent works have proposed a full-band and sub band fusion modeling approach In these works, a full-band model and several sub-band models are combined, allowing them to complement each other and capture both the local and global spectral information. Though these fusion-based methods have led to significant improvements, sub-band modeling can still be computationally costly, as it requires processing each frequency band separately.

This poses challenges for real-time edge applications. In this work, we propose a novel approach that applies different granularity levels to various frequency sub-bands, significantly reducing the computational cost while maintaining the speech enhancement performance.

## 4.3 NEUROMORPHIC SPEECH PROCESSING

SNNs have recently emerged as a promising approach for power-efficient speech processing. Compared to traditional ANNs, SNNs can offer significant advantages in terms of reduced computational complexity. Introduce a SOM-SNN framework that in corporates a self-organizing map (SOM) for feature representation, followed by an SNN for pattern classification. Recent studies leverage deeper SNNs and more advanced learning rules to enhance classification performance. For instance, deep convolutional SNNs coupled with the tandem learning rule have achieved significantly improved performance on keyword spotting tasks. Recurrent networks of spiking neurons (RSNNs) are also exploited for speech recognition ,holding enhanced memory capacity and bringing improvements over the feedforward counterparts. These earlier studies focus on small vocabulary speech recognition tasks. More recently, Wu et al. apply deep SNNs to large vocabulary continuous speech recognition tasks and demonstrate competitive accuracy compared to ANN-based systems. Notably, a recent benchmark study compared the performance of neuromorphic keyword spotting systems to ANN-based systems that deployed on conventional computing hardware. The study evaluated metrics such as inference speed, energy cost per inference, and dynamic energy consumption. The findings from this benchmark study suggest that neuromorphic systems can significantly reduce the energy costs per inference while maintaining equivalent inference accuracy compared to their traditional ANN counterparts. However, there is a lack of study of neuromorphic technologies for the speech denoising task.

## 4.4 SPIKING NEURON MODEL

The leaky integrate-and-fire (LIF) neuron is widely used due to its balance of computational efficiency and analytical simplicity. It maintains an internal state, called the membrane potential, which gradually decreases over time at a rate determined by a time constant (τ). Simultaneously, the neuron accumulates input current. When the membrane potential exceeds a predetermined threshold (ϑ), the neuron generates an output spike that is transmitted to connected neurons. After firing, the membrane potential resets to a specific baseline value. This process follows a structured, discrete-time formulation that describes the dynamic interactions governing the neuron’s activity.

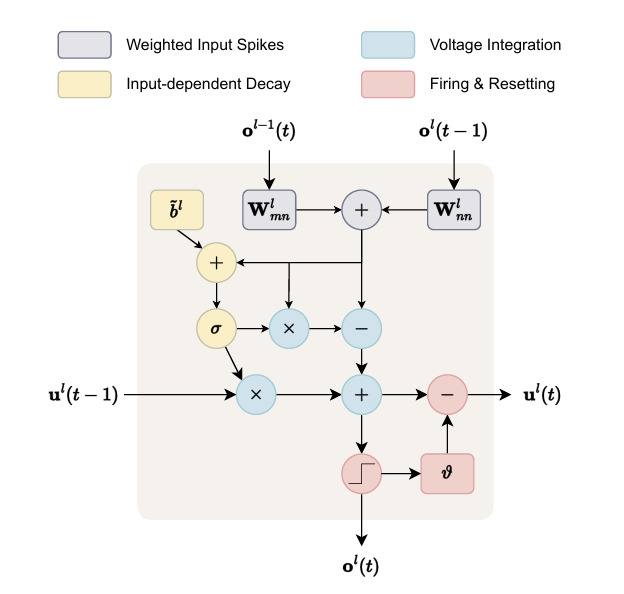
The feed-forward and recurrent weight matrices at the l-th layer, represented as and , respectively, play a crucial role in determining the connections and interactions between neurons. The decay rate of the membrane potential ul(t)is controlled by λ, which is defined as , where τ represents the time constant. The input current is denoted as il(t), contributing to the accumulation of potential. Additionally, bl represents the bias term, which helps refine the neuron’s activation response.

## 4.5 PROPOSED FULLSUBNET MODEL

### **4.5.1 Gated Spiking Neuron.**

Speech enhancement is to remove unwanted noises from audio recordings. Noise can come from a variety of sources, including background sounds, microphone interference, and transmission distortions, and its intensity changes over time. The commonly used LIF model [28] cannot account for dynamic changes due to the constant membrane decay rate. In these conditions, the LIF neurons face two hurdles. The system may struggle to efficiently filter out noise during high-noise times or over-filter informative audio signals during low-noise periods. The fixed decay factor can either offer insufficient attenuation of the noisy signal, resulting in inadequate amplification, or decay the neuron's potential too quickly, causing loss of desired auditory content.

A simple option is to use multiple decay factors at each stage, allowing for flexible adaptability to fluctuating noise levels in input audio signals. However, this would result in a large number of factors, especially for long-term durations. Additionally, it may not perform well for audio signals with varying durations, which are typical in SE tasks. To solve this issue, we present GSN, a novel spiking neuron model that modulates decay rate in an input-dependent manner.



**Figure 4.5 Illustration of the proposed GSN model**

In GSN, input-dependent decay is implemented by modeling it as a function of feed-forward and recurrent input spikes. The sigmoid function σ(·) is used to restrict the decay rate between 0 and 1. The membrane potential dynamics are written as:

In order to save model parameters and reduce overall computation, we employ weight-sharing by using the same weight matrices, denoted as Wl mn and Wl nn, Notably, the GSN model can adaptively modulate the neuron’s membrane potential along the temporal dimension while avoiding the need for a large number of parameters associated with the total time steps. Additionally, GSN bears a resemblance to the forget gate widely used as a critical component of the LSTM architecture. However, this temporal gating mechanism has been underexplored in existing spiking neuron models.

### **4.5.2 Spiking-FullSubNet Architecture**

Spiking-FullSubNet builds upon the GSN model introduced earlier to create an efficient, high-performance speech enhancement system. By leveraging both full-band and sub-band modeling, it effectively captures global and local spectral patterns—important aspects for processing noisy speech signals.To improve efficiency, the model applies varying processing granularity to different frequency partitions, inspired by human auditory perception.

#### **4.5.2.1 Full-Band Processing:**

The full-band model processes magnitude spectral features extracted from noisy speech signals. Each audio frame is represented by an input feature vector x(n), which consists of the magnitudes of complex Fourier coefficients across different frequency bins. The total number of frequency bins is denoted by FF, while x(n,f) represents the Fourier coefficient for frame nn at frequency bin ff, with the magnitude extracted using |.|

To capture the temporal variations across frames, a sequence of feature vectors is defined as:

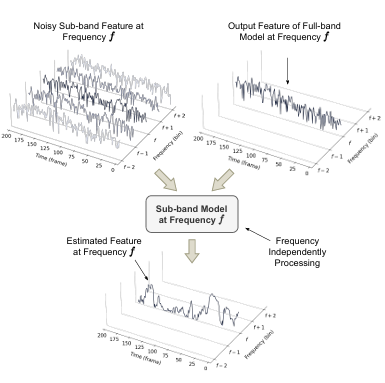
where T represents the total number of discrete time frames. This setup allows the model to analyze both global spectral patterns and variations over time for effective speech enhancement. T is the total number of discrete time frames. We stack GSNs to process X by capturing both the global spectral content and the interactions between frequency bins, yielding a spectral embedding E of the same dimensions as X.

#### **4.5.2.2 Sub-Band Processing:**

The sub-band models operate independently for each frequency band, focusing on differences in stationarity between speech and noise, as well as local spectral patterns and reverberation characteristics. At a given time step n and frequency f, the input to the sub-band model consists of a feature vector Xf(n). This vector includes the noisy magnitude bin at frequency ff, its 2×N times where N neighboring frequency bins, and the corresponding bin from the spectral embedding produced by the full-band model.

This independent frequency processing is inspired by conventional noise reduction techniques like noise density estimation and Wiener filtering, enabling a more detailed examination of local spectral characteristics and stationarity. Research has demonstrated the effectiveness of combining full-band and sub-band models within a unified framework. However, a key computational challenge in existing full-band and sub-band processing techniques lies in the uniform granularity applied to all sub-bands. This uniform approach contrasts with the human auditory system, which exhibits heightened sensitivity to low-frequency sounds while being comparatively less responsive to high-frequency sounds.

To address this issue, a frequency partitioning strategy that applies varying levels of processing granularity across frequency bands. This approach, inspired by human auditory perception, enhances computational efficiency while preserving the quality of speech enhancement.



**Figure 4.6Illustration of the sub-band processing in Spiking FullSubNet.**

#### **4.5.2.3 Sub-Band Processing Based on Frequency Partitioning**

The frequency-independent processing method draws inspiration from traditional signal processing techniques for noise reduction, such as noise density estimation and Wiener filtering. This approach enables an in-depth analysis of the local spectral patterns and stationarity of a signal. Research findings support the integration of full-band and sub-band models within a unified framework. However, a major computational challenge in current full-band and sub-band modeling techniques stems from the fact that all sub-bands are processed at the same granularity. This differs from human auditory perception, where lower frequencies are more sensitive and higher frequencies are comparatively less perceptible.

To address this, we introduce a frequency partitioning strategy, which adjusts processing granularity across different frequency bands. This method considers the varying perceptual significance of frequency bins by dividing the input magnitude spectrum into K non-overlapping partitions:

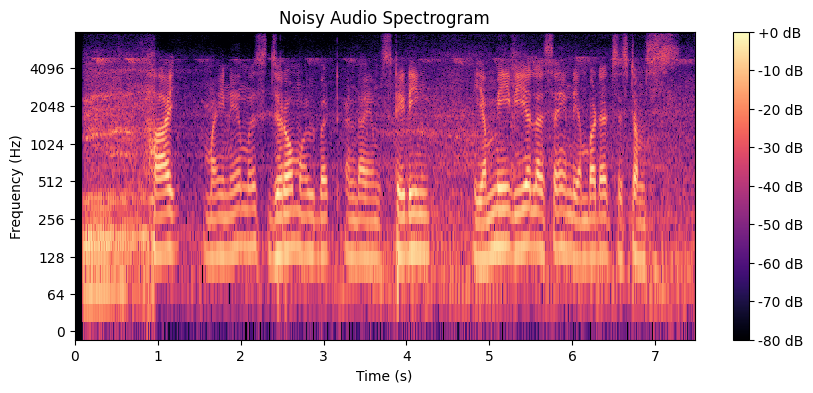
where the total size of all partitions is F, the number of frequency bins. Each partition PkP\_k consists of contiguous frequency bins, covering the entire spectrum. The first partition P₁ contains bins from 1 to fc₁, with fc₁ representing its cutoff frequency. Each subsequent partition Pₖ spans from the previous partition's cutoff frequency fₖ₋₁ + 1 to its respective limit fₖ. The final partition Pₖ extends from fcₖ₋₁ + 1 to the upper bound F, ensuring the entire frequency range is covered.

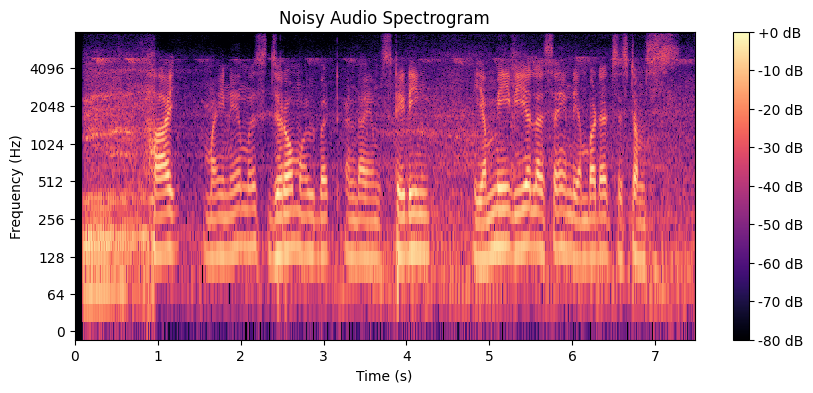
For each partition, the Spiking-FullSubNet dynamically adjusts processing granularity using a grouping parameter gₖ, defining the number of discrete frequency bins processed together in the sub-band model. Each partition includes:

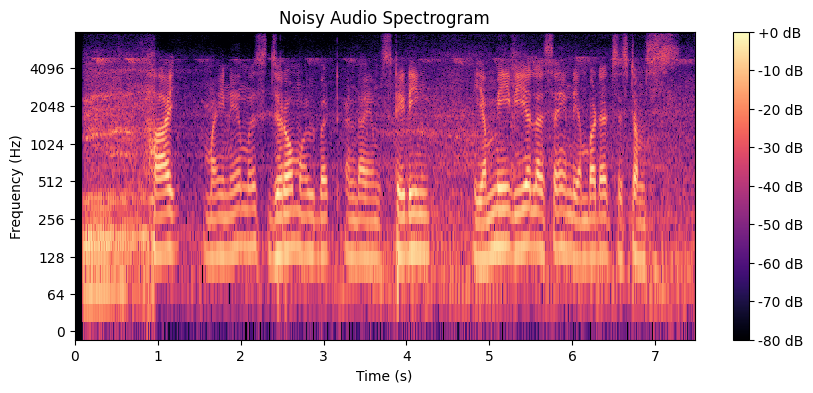
* A group of gₖ + 1 adjacent bins
* A context window of N bins on either side
* Embedding vectors generated by the full-band model

The feature vector for partition Pₖ at time n is:

Here, Xfk(n) represents the grouped feature vector for partition **Pₖ** at time **n**, starting from frequency bin **f** within the partition. The value of **f** falls between the lower and upper bounds of **Pₖ**, ensuring the processing remains within the partition.







**Figure 4.7 Frequency Partitioning of Input for sub band**

This adaptive frequency partitioning technique enables Spiking-FullSubNet to efficiently process spectral information, applying finer granularity to lower-frequency partitions (where speech perception is more sensitive) while reducing complexity for higher frequencies (where detailed spectral features are less crucial).

### **4.6 IMPLEMENTATION DETAILS**

All speech audio data are processed at a sampling rate of 16 kHz. The STFT is set up with a window length of 32 ms (512 samples) and a hop length of 8 ms (128 samples), utilizing a Hanning window and comprising 512 FFT frequency bins. The proposed Spiking-FullSubnet takes the magnitude spectrogram as its input. utilize the AdamW optimizer with a learning rate of 1 × 10−3 and set the gradient norm clipping to 10. In the loss function L, the weights of different terms are set to α = 0.5,γ1 = 0.5,γ2 = 0.001. We set the number of neighboring frequency bins to 15. To further improve energy efficiency, we add the total number of synaptic operations into the loss function to penalize excessive firing. We develop different variants of the Spiking-FullSubNet with varying model sizes. We divide all discrete frequency bins based on the study of human auditory. The fundamental frequency, a crucial characteristic of speech signals, extends up to 1 kHz. Therefore, we group the lowest 32 discrete frequency bins, corresponding to 0 ∼ 1 kHz, applying the smallest grouping size within this range. Then, given that the human voice’s most significant content lies between 1 ∼ 4 kHz, a range particularly sensitive for the human auditory system, we group the following 96 discrete frequency bins, equivalent to 1 ∼ 4 kHz, with uniform processing granularity in this range. Lastly, we group the remaining high-frequency bins, 128 discrete frequency bins corresponding to 4 ∼ 8 kHz, as a separate group. These high-frequency bins have a more minor impact on human auditory perception, allowing for coarser granularity in processing. In summary, we establish three frequency partitions: 0 ∼ 1 kHz, 1 ∼ 4 kHz, and 4 ∼8kHz, each tailored to optimize computational efficiency and performance based on human auditory characteristics.

# **CHAPTER 5**

**RESULTS AND DISCUSSION**

## 5.1 Evaluation of Speech Enhancement Using DNSMOS Scores

We present and interpret the results of speech enhancement using the Spiking FullSubNet model, evaluated with DNSMOS (Deep Noise Suppression Mean Opinion Score)—a non-intrusive, neural network-based metric developed by Microsoft to assess perceived speech quality. DNSMOS provides three distinct scores:

* **OVRL (Overall Quality)**: Measures overall perceptual quality of the speech signal.
* **SIG (Signal Quality)**: Focuses on the naturalness and clarity of the speech component.
* **BAK (Background Noise)**: Quantifies how well the background noise has been removed or suppressed.

**Table 1 DNSMOS scores of the raw noisy dataset versus the enhanced audio after training using the Spiking FullSubNet model**

|  |  |  |
| --- | --- | --- |
| **DNSMOS SCORE** | **NOISY DATASET** | **AFTER TRAINING ENHANCEMENT** |
| **OVRL** | 2.111115 | 3.30644 |
| **SIG** | 3.719177 | 3.88166 |
| **BAK** | 1.631128 | 3.59658 |

1. **Improvement in Overall Speech Quality (OVRL)**

The OVRL score increased from 2.11 to 3.31, indicating a significant enhancement in the overall perceived quality of speech. This metric captures both the clarity of speech and the suppression of background noise. The increase of 1.19 points is a strong indicator that the Spiking FullSubNet effectively enhances the listening experience even in challenging noise conditions. It validates the effectiveness of the cochlea-inspired feature extraction (like Gammatone filtering) and biologically inspired spiking neural architectures.

1. **Stability of Speech Quality (SIG)**

The SIG score started high at 3.72 (even in noisy conditions) and improved further to 3.88 after enhancement. This slight improvement shows that the model preserves the naturalness and intelligibility of speech. It implies that the model avoids introducing artifacts or distortion during the enhancement process. Preserving the semantic clarity of speech is critical for downstream tasks like speech recognition or hearing aids.

1. **Significant Background Noise Reduction (BAK)**

The BAK score improved dramatically from 1.63 to 3.60, an increase of nearly 2 points. This highlights the core strength of the model in suppressing unwanted noise components while maintaining speech integrity. Such improvement demonstrates that the system can function effectively in real-world noisy environments like streets, cafes, or public transport—especially beneficial for hearing-impaired users or for robust speech recognition systems.

The DNSMOS-based evaluation confirms that the Spiking FullSubNet model offers a balanced approach:

* Suppresses noise effectively (high BAK score).
* Preserves speech quality (high and stable SIG score).
* Enhances overall listening experience (improved OVRL score).

These improvements are crucial for real-time applications such as neuromorphic hearing aids, telecommunication, and assistive devices, where both noise suppression and speech clarity must co-exist without sacrificing latency or computational efficiency.

The DNSMOS scores provide quantitative proof that the system offers real-world benefits for enhanced speech communication in noisy environments. The substantial improvement in background noise suppression, along with maintained speech quality, proves that the Spiking FullSubNet model is well-suited for low-power, neuromorphic, real-time speech enhancement systems.

## 5.2 SI-SNRi Evaluation Metrics

SI-SNR (Scale-Invariant Signal-to-Noise Ratio) is a critical objective metric used to evaluate the performance of speech enhancement systems. Unlike traditional SNR, which is sensitive to absolute amplitude, SI-SNR is invariant to signal scaling. This makes it particularly well-suited for neural network-based models that may not preserve signal energy but still reconstruct clean audio effectively.

**Calculate SI-SNR in decibels**:

**Table 2 Signal Invariant-Signal to Noise Ratio Improvement Before and After Enhancement**

|  |  |  |  |
| --- | --- | --- | --- |
| **SI-SNR** | **BEFORE ENHANCEMENT (db)** | **AFTER ENHANCEMENT (db)** | **SI-SNRi (db)** |
| 7.89 | 14.45 | 6.56 |

The noisy input audio had an SI-SNR of 7.89 dB, indicating moderate to heavy degradation in speech quality due to noise.After enhancement using the Spiking FullSubNet model, the SI-SNR increased to 14.45 dB, signifying a clear recovery of the speech signal from noisy conditions.The SI-SNR improvement (SI-SNRi) of 6.56 dB reflects a substantial enhancement, showing that the model significantly suppresses noise while preserving speech content.

The SI-SNR results provide quantitative validation of the effectiveness of the proposed system. By achieving a 6.56 dB gain, the Spiking FullSubNet model demonstrates its ability to retain clean speech features while eliminating background interference, making it well-suited for low-latency, real-time neuromorphic deployment on edge devices. SI-SNR is a preferred metric in speech enhancement competitions like DNS Challenge and real-world applications because it reflects real perceptual gains in intelligibility and quality. A gain of over 5 dB in SI-SNR is considered excellent for real-time systems like hearing aids and communication devices.The observed improvement confirms that the model is capable of high-fidelity speech separation, even under challenging acoustic environments.

# **CHAPTER 7**

# **CONCLUSION AND FUTURE WORK**

The Spiking-FullSubNet exhibits superior performance in audio quality metrics, significantly surpassing the base line models. Spiking-FullSubNet, a ground breaking SNN-based system tailored for real-time audiode noising tasks. This project presents a comprehensive framework for real-time speech enhancement using a biologically inspired Spiking FullSubNet model, integrating multiple signal processing and neuromorphic computing techniques to address the challenges of noisy speech environments. The pipeline incorporates key components including RMS energy estimation, active loudness normalization, Bark-scale spectral analysis, and Ideal Ratio Masking (IRM), all of which contribute to isolating clean speech from background noise. Spectral transformations using STFT and complex ideal ratio masks further improve the separation in the frequency domain. The application of Bark filterbanks, modeled on human auditory perception, enhances the spectral resolution in critical bands while reducing irrelevant noise. Evaluation using DNSMOS perceptual metrics confirms significant gains in overall speech quality (OVRL) and background noise reduction (BAK), while preserving the naturalness of speech (SIG). The Spiking FullSubNet model proves effective in enhancing speech intelligibility under real-world noise conditions without sacrificing latency or energy efficiency. This project not only highlights the power of neuromorphic computing in auditory signal processing but also sets a foundation for deploying spiking models in practical speech applications, paving the way for future research into ultra-efficient, brain-inspired AI systems.

**REFERENCES**

1. Lenk, C., Hövel, P., Ved, K., Durstewitz, S., Meurer, T., Fritsch, T., Männchen, A., Küller, J., Beer, D., Ivanov, T. and Ziegler, M., 2023. Neuromorphic acoustic sensing using an adaptive microelectromechanical cochlea with integrated feedback. *Nature electronics*, *6*(5), pp.370-380.
2. Chen, Z. and Zhang, P., 2022. Lightweight Full-band and Sub-band Fusion Network for Real Time Speech Enhancement. In *INTERSPEECH* (pp. 921-925).
3. Dampfhoffer, M. and Mesquida, T., 2024. Neuromorphic lip-reading with signed spiking gated recurrent units. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2141-2151).
4. Rathi, N., Chakraborty, I., Kosta, A., Sengupta, A., Ankit, A., Panda, P. and Roy, K., 2023. Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware. *ACM Computing Surveys*, *55*(12), pp.1-49.
5. Sun, T. and Bohté, S., 2024. DPSNN: spiking neural network for low-latency streaming speech enhancement. *Neuromorphic Computing and Engineering*, *4*(4), p.044008.
6. Riahi, A. and Plourde, É., 2023, September. Single channel speech enhancement using u-net spiking neural networks. In *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 111-116). IEEE.
7. Taherian, H., Wang, Z.Q., Chang, J. and Wang, D., 2020. Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, pp.1293-1302..
8. Ahmad, R., Zubair, S. and Alquhayz, H., 2020. Speech enhancement for multimodal speaker diarization system. *IEEE Access*, *8*, pp.126671-126680.
9. Fedorov, I., Stamenovic, M., Jensen, C., Yang, L.C., Mandell, A., Gan, Y., Mattina, M. and Whatmough, P.N., 2020. TinyLSTMs: Efficient neural speech enhancement for hearing aids. *arXiv preprint arXiv:2005.11138*..
10. Hao, X., Su, X., Horaud, R. and Li, X., 2021, June. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6633-6637). IEEE.
11. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B. and Xie, L., 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*.
12. Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y. and Watanabe, S., 2023. TF-GridNet: Integrating full-and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, pp.3221-3236.
13. Xiong, F., Chen, W., Wang, P., Li, X. and Feng, J., 2022. Spectro-Temporal SubNet for Real-Time Monaural Speech Denoising and Dereverberation. In *Interspeech* (pp. 931-935).
14. Choi, H.S., Park, S., Lee, J.H., Heo, H., Jeon, D. and Lee, K., 2021, June. Real-time denoising and dereverberation wtih tiny recurrent u-net. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5789-5793). IEEE.
15. Roy, K., Jaiswal, A. and Panda, P., 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, *575*(7784), pp.607-617.
16. Chen, X., Yang, Q., Wu, J., Li, H. and Tan, K.C., 2023. A hybrid neural coding approach for pattern recognition with spiking neural networks. *IEEE transactions on pattern analysis and machine intelligence*, *46*(5), pp.3064-3078.
17. Wu, J., Xu, C., Han, X., Zhou, D., Zhang, M., Li, H. and Tan, K.C., 2021. Progressive tandem learning for pattern recognition with deep spiking neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), pp.7824-7840.
18. Hu, Y., Zheng, Q., Jiang, X. and Pan, G., 2023. Fast-snn: Fast spiking neural network by converting quantized ann. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(12), pp.14546-14562.
19. Wu, J., Xu, C., Han, X., Zhou, D., Zhang, M., Li, H. and Tan, K.C., 2021. Progressive tandem learning for pattern recognition with deep spiking neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), pp.7824-7840.
20. Xia, Y., Braun, S., Reddy, C.K., Dubey, H., Cutler, R. and Tashev, I., 2020, May. Weighted speech distortion losses for neural-network-based real-time speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 871-875). IEEE.