Master's Thesis

Study on speech detection method under heavy noise conditions using spectro-temporal modulation analysis

TRAN, Quynh Nhu

Supervisor UNOKI, Masashi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

August, 2025

Abstract

Detecting and locating survivors in disaster areas remains a critical challenge, especially in Japan, where natural disasters such as earthquakes, landslides, and floods occur frequently. Deploying drones equipped with microphone arrays offers a promising approach for rapidly scanning dangerous areas. By capturing human voices in disaster sites, such systems can significantly reduce search times and increase survival rates. However, the extremely adverse acoustic conditions present in disaster environments-often characterized by high-intensity noise from wind, rain, machinery, or collapsing structures-pose significant difficulties for conventional speech detection methods. In these scenarios, the signal-to-noise ratio (SNR) can drop well below -10 dB, sometimes reaching -20 dB, severely limiting the performance of traditional speech processing techniques.

Existing approaches, such as direction of arrival (DOA) estimation and voice activity detection (VAD), are practical under moderate noise conditions but often fail to function reliably in extremely low SNR environments. In particular, these methods experience a sharp decline in detection accuracy once the SNR falls below -10 dB, making them unsuitable for the most challenging disaster scenarios. This limitation underscores the urgent need for a speech detection system capable of robust performance under extreme noise conditions.

The objective of this study is to develop a speech detection method specifically designed for the heavy noise environments encountered during disaster response operations. The proposed framework leverages spectro-temporal modulation (STM) analysis combined with deep neural networks to detect speech at SNRs spanning from 20 dB down to -20 dB, with a special emphasis on the challenging range below -10 dB. STM features capture both spectral and temporal modulation patterns, which exhibit distinctive structures for speech compared to noise. These patterns remain relatively stable even under severe noise interference, making STM a strong candidate for robust feature representation. By exploiting these unique characteristics, the proposed method aims to enable reliable detection.

The novelty of this research is the focus on extremely low SNR conditions down to -20 dB and its systematic integration of STM analysis with attention-augmented deep learning models. Unlike conventional spectrogram-based approaches, STM represents audio signals in a joint modulation frequency domain, offering a unified multi-resolution view of the signal's temporal and spectral fluctuations. Speech signals typically manifest coherent

spectro-temporal patterns, whereas noise—particularly stationary noise—tends to produce more random or diffuse modulation structures. By capitalizing on these differences, the system can discriminate between speech and non-speech even when noise energy dominates the signal. Additionally, the incorporation of the Convolutional Block Attention Module (CBAM) into the ResNet18 architecture enhances the model's ability to focus on modulation regions that are most indicative of speech presence. This attention mechanism operates both spatially and channel-wise, directing the network's resources toward salient spectro-temporal regions and suppressing irrelevant noise-dominated areas.

The methodology is divided into two main stages: feature extraction and classification. In the first stage, clean speech signals and noise are mixed at predefined SNR to generate noisy speech samples. Spectrograms are computed using multiple filterbank configurations—Short-Time Fourier Transform (STFT), Constant-band, Mel, and Gammatone filterbanks—to examine their impact on STM quality and detection performance. The spectrograms are then transformed into STM representations via a two-dimensional Fast Fourier Transform, capturing modulation content across both time and frequency. Various normalization and compression strategies are applied to enhance robustness.

In the classification stage, a set of deep learning models are employed. A baseline convolutional neural network (CNN) serves as an initial benchmark, followed by the ResNet18 architecture, which offers greater depth and residual connections for improved feature learning. To further refine detection in low SNR conditions, an attention mechanism named CBAM is integrated into ResNet18. The resulting model learns to emphasize speech-relevant modulation patterns and suppress noise-dominated features. The system is trained and evaluated on the JVS corpus for clean speech, with white noise used as a controlled stationary noise source to simulate disaster-site interference. Performance is assessed across a wide range of SNR values using accuracy, equal error rate (EER), and additional detection metrics.

Several experimental results in this study demonstrate the effectiveness of the proposed method. The ResNet18-CBAM model trained with STM features derived from the Gammatone filterbank achieved the best overall performance. The system attained an accuracy of 92.79% across all SNR conditions, with near-perfect detection at high SNRs. A comparative analysis against typical VAD methods, rVAD-fast and WebRTC VAD, highlighted the superior performance of the proposed method. While these conventional methods showed a sharp decline in accuracy and a rapid increase in EER as SNR decreased, the proposed method maintained a high degree of robustness. Remarkably, at -10 dB SNR—where conventional methods typically

fail—the system achieved an EER of less than 10%. Even at −15 dB, a highly challenging condition, the EER remained around 20%, far outperforming the approximately 50% EER of the other methods. While performance declined at −20 dB due to the overwhelming dominance of noise, the method still maintained a measurable detection capability, highlighting its robustness relative to existing approaches.

The comparative analysis of filterbanks revealed that the Gammatone filterbank consistently outperformed STFT, Constant-band and Mel alternatives. This superiority is attributed to its auditory-inspired frequency resolution, which emphasizes lower-frequency components crucial for speech perception. Such emphasis is particularly beneficial in noisy environments, where low-frequency speech cues are often more resilient to masking. The findings indicate that the selection of filterbank has a significant impact on STM representation quality and, consequently, on detection accuracy.

This study makes several contributions to the field of robust speech detection. First, it provides a comprehensive evaluation of STM features under multiple filterbank configurations, offering practical guidance on optimal feature extraction for noisy environments. Second, it demonstrates the value of integrating attention mechanisms into deep residual networks for speech detection, particularly in scenarios with extreme noise interference. Third, it presents a scalable framework capable of operating across a wide SNR range, making it adaptable to varying disaster-site conditions.

In conclusion, this dissertation addresses the need for reliable speech detection in disaster environments where noise power are high. By combining spectro-temporal modulation analysis with attention-enhanced deep learning, the proposed system achieves high detection accuracy even at SNRs well below the threshold where conventional methods fail. While the framework demonstrates strong performance, especially with Gammatone-based STM features, challenges remain at extremely low SNRs such as −20 dB. Future work may explore the incorporation of non-stationary noise profiles, real-world disaster soundscapes, and multimodal fusion with visual or thermal data to enhance detection reliability further. The findings of this research have direct implications for the development of autonomous drone-based survivor detection systems, potentially accelerating rescue operations and saving lives in disaster-stricken regions.

Contents

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Ho Chi Minh University of Science, Vietnam National University.

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research background, the potential issues, research objectives, and the organization of this thesis.

1.1 Research background

Japan is an attractive tourist destination with majestic natural beauty and a rich cultural history. Moreover, this is also a country that is resilient in the face of natural challenges. Due to its geographical location on the Pacific Ring of Fire, Japan experiences a high frequency of natural disasters, including earthquakes, tsunamis, typhoons, and floods. In the immediate aftermath of such events, one of the most critical and time-sensitive operations is the search and rescue (SAR) of survivors who may be trapped, injured, or disoriented. The success of these missions is often measured in hours, as the probability of survival decreases significantly with each passing moment [1].

The effectiveness of traditional SAR operations, which depend on human teams and rescue dogs, is severely limited by the scale of the disaster and the physical dangers at a site, as collapsed structures, unstable ground, and hazardous materials can render areas completely inaccessible. To overcome these limitations, recent years have seen the emergence of Unmanned Aerial Vehicles (UAVs), which are commonly known as drones, as a technology

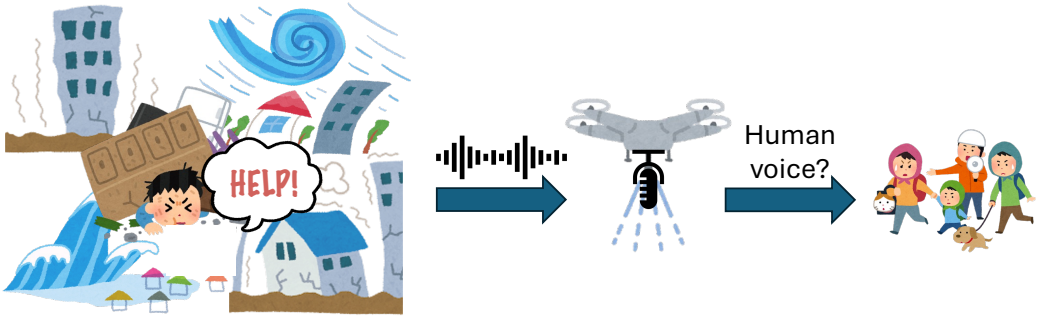

Figure 1.1: Illustration of speech detection by drone microphones in disaster scenarios.

in disaster management. Equipped with various sensors, including high-resolution cameras, drones can rapidly survey large and dangerous areas.

Using UAVs equipped with microphone arrays is an efficient approach for detecting and locating people in need of rescue in disaster areas. This application of drones can help SAR teams locate survivors by detecting crucial signs of life, such as calls for help from individuals who may be hidden from traditional visual sensors. This method, specifically designed to detect human voice, significantly enhances the speed and effectiveness of locating people who must evacuate from dangerous sites, making it a critical tool for modern disaster response.

1.2 Research issues

The primary issue in disaster zones is the extreme noise, making it challenging for acoustic surveillance to function effectively in SAR. Things like wind, rushing water, collapsing buildings, fires, and heavy machinery create a chaotic mix of sounds that can drown out cries for help. This acoustic environment results in a very low signal-to-noise ratio (SNR), which is a measure of the power of the desired signal (speech) relative to the power of the background noise. In many disaster scenarios, the SNR can fall to critically low values, often below -10 dB, and especially at -20 dB. At such low SNRs,

the noise power completely dominates the power of the human voice.

In such high noise environments where SNR is below -10 dB, conventional speech detection techniques such as direction of arrival (DOA) and voice activity detection (VAD) face significant limitations. DOA via drone microphone array is one of the techniques for locating the sound source of people in need of rescue [2]. However, it is uncertain whether the sound detected by DOA via drone microphone array is truly a human voice or not, since there are various sources of sounds due to disasters at SNR below -10 dB. Robust VADs are also the techniques for detecting speech or non-speech sections in the recorded sounds [3, 4]. Nevertheless, these techniques can only work well under conditions with SNRs above 0 dB.

1.3 Research objectives

To address the challenges outlined above, the primary objective of this dissertation is to develop a robust speech detection method capable of operating reliably under high noise conditions, specifically in heavy noise conditions with an SNR below -10 dB and especially at -20 dB. Our approach is based on the hypothesis that the spectral and temporal modulation patterns of human speech contain resilient information that can be leveraged for detection even when conventional features fail.

To achieve this goal, this research aims to achieve the following specific objectives:

- To propose a robust feature extraction method for human speech based on spectro-temporal modulation analysis. This study will investigate the spectro-temporal modulation feature extraction, which represents the joint spectral and temporal modulation of a signal. These features are hypothesized to capture characteristics unique to the human voice.

- To design and implement a deep learning model for a speech detection system. A deep neural network will be developed to learn the complex,

high-dimensional relationships within the extracted spectro-temporal modulation features. The model will be trained to act as a speech and non-speech classifier, effectively distinguishing feature patterns corresponding to human speech from those corresponding to heavy background noise.

- To systematically evaluate the performance of the proposed method. The effectiveness of the complete system will be tested on a dataset of speech signals mixed with noise signals at various SNR conditions from 20 dB down to -20 dB. Performance will be measured using standard metrics such as accuracy and equal error rate.

1.4 Organization of the thesis

The outline of this dissertation is described as follows.

- Chapter 2 provides a literature review of the foundational concepts relevant to this research. It begins by surveying existing approaches to speech detection, including direction of arrival (DOA) and voice activity detection (VAD) techniques. The chapter then delves into the theoretical principles of spectro-temporal modulation (STM) analysis, our core feature representation. Finally, it reviews the deep learning techniques leveraged in this study, namely Convolutional Neural Networks (CNNs), Residual Networks (ResNet), and an attention mechanism, Convolutional Block Attention Module (CBAM).

- Chapter 3 details the proposed method. This chapter outlines the specific pipeline for extracting robust STM features from audio signals and presents the deep learning architecture designed for the final speech/non-speech classification task.

- Chapter 4 describes the experimental framework used to validate the proposed method. It details the clean speech and noise datasets used

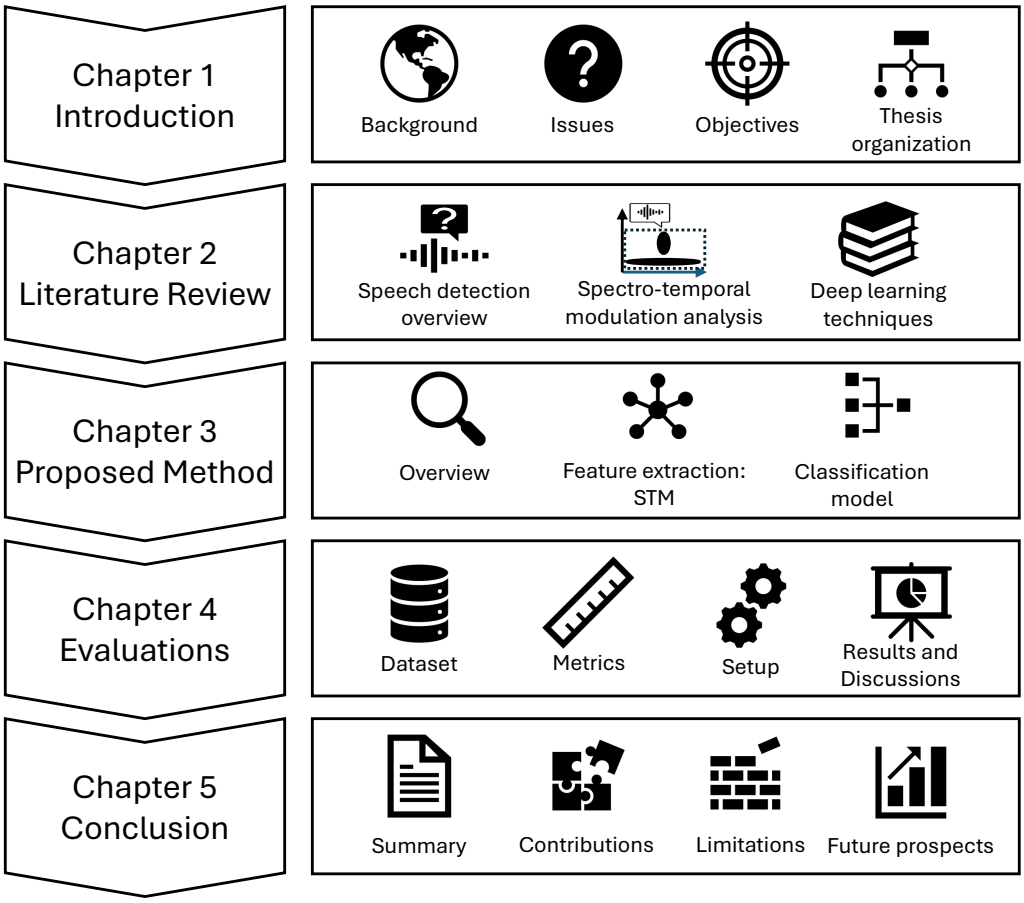

Figure 1.2: Organization of the thesis.

to construct the speech/non-speech dataset under various SNRs, the evaluation metrics for measuring performance, and the complete experimental setup. This chapter presents the quantitative results and provides a detailed discussion of the findings.

- Chapter 5 concludes the dissertation. It offers a summary of the work and its key contributions, acknowledges the study's limitations, and proposes potential directions for future research.

Chapter 2

Literature Review

This chapter provides an overview of the existing literature on speech detection in noisy environments, spectro-temporal modulation analysis, and deep learning techniques for audio classification. This review aims to cover fundamental and current state-of-the-art methods, then identify critical research gaps that the proposed methodology seeks to address.

2.1 Speech detection overview

Speech detection in noisy conditions, such as disasters [5], serves as a foundational and critical step for numerous signal processing applications, including automatic speech recognition [6], telecommunications [7], and security systems [8]. The primary objective of speech detection is to accurately identify the presence or absence of human speech in an input signal. Misclassifying noise as speech (false positives) or missing genuine speech (false negatives) will reduce the performance of the speech detection system under noisy conditions.

The key challenge in speech detection in disaster scenarios is the diversity of real-world environmental noise. Noise can vary from stationary rain sounds to sudden, transient sounds like sirens or collapsed buildings. This problem becomes more serious because, during disaster sites, background noise often

dominates human speech, thereby complicating the task of accurately detecting its presence. The complexity of this issue has led to several speech detection methods, from traditional, statistical algorithms to advanced deep learning techniques that can adapt to these challenging, noisy conditions.

To tackle these challenges, various studies and approaches have been developed, including direction of arrival (DOA) and voice activity detection (VAD).

2.1.1 Direction of arrival

Direction of arrival (DOA) is the process of estimating the angle at which a signal arrives at an array of sensors. It helps in determining the location or direction of sound sources relative to the receiver. In disaster search and rescue (SAR) operations, DOA estimation is employed to locate sources of sounds, such as human voices or distress signals, enabling SAR teams to identify and rescue survivors [9].

Subspace methods, including Multiple Signal Classification (MUSIC) [10] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [11], represent high-resolution approaches that infer signal characteristics from noisy data. Beamforming-based methods, including Delay-and-Sum [12] and Linearly Constrained Minimum Variance (LCMV) [13], utilize microphone arrays to achieve spatial filtering, thereby enhancing sensitivity toward a specific direction while attenuating noise. Deep learning techniques such as convolutional neural networks [14] and recurrent neural networks [15] demonstrate state-of-the-art results in DOA estimation, particularly when there are multiple sound sources and noisy conditions.

In scenarios such as disasters, DOA techniques face significant limitations. First, these include difficulties in handling noise and reverberation. Second, it is uncertain whether the sound detected by DOA via the drone microphone array is truly a human voice, as various sound sources present in disaster environments are present at SNRs below -10 dB. Furthermore, DOA methods experience performance degradation when encountering unseen or high noise

characteristics. These factors limit the robust and accurate application of DOA techniques in real-world noisy environments.

2.1.2 Voice activity detection

Voice activity detection (VAD) is the process of determining whether human speech is present or absent within a given segment of an audio signal, which is often considered as a speech/non-speech classification problem [16]. This approach is crucial for distinguishing speech from background noise, thereby enhancing the performance of speech recognition systems and reducing computational costs by enabling downstream tasks to operate on speech segments [17].

Historically, VAD algorithms have evolved through various techniques, each with distinct advantages and limitations, particularly under noisy environments. Early approaches, such as energy-based methods [18] are relatively simple but suffer from instability in noise, as fluctuations in background sound severely disrupt their ability to identify speech boundaries accurately. Statistical models, including Hidden Markov Models (HMMs) [19] and Gaussian Mixture Models (GMMs) [20], offer better discrimination by modeling the temporal and harmonic structures of speech and noise.

The beginning of Deep Neural Networks (DNNs) has improved VAD performance, particularly in highly challenging noisy conditions. Frameworks utilizing DNN-based speech enhancement preprocessing have demonstrated superior accuracy over traditional methods [21]. Architectures like Convolutional Recurrent Neural Networks (CRNNs) [22] and Long Short-Term Memory (LSTM) networks [23] are widely employed to classify audio events with high precision, often outperforming classic models by leveraging auditory speech features such as energy, zero crossing rate (ZCR), and Mel Frequency Cepstral Coefficients (MFCCs). These deep learning models excel at extracting robust features and adapting to diverse acoustic environments.

Among existing VAD algorithms, rVAD [4] is one of the state-of-the-art methods that combines denoising, spectral flux analysis, and decision

smoothing to detect speech boundaries in low-SNR environments accurately. It has been widely adopted in speech processing tasks due to its balance between accuracy and computational efficiency, with a faster variant (rVAD-fast) available for real-time applications. Another widely used VAD tool is WebRTC VAD [24], developed by Google as part of the WebRTC project. This open-source implementation employs Gaussian Mixture Model–based classification, providing reliable speech detection for streaming applications while maintaining low computational overhead.

In general, VAD methods perform well in high-SNR environments, where speech signals are relatively clean and easily distinguishable from background noise. However, their performance degrades significantly under heavy noise conditions, particularly at SNRs below -10 dB, where speech is heavily masked by interfering sounds. In addition, many of these methods involve high computational complexity, making them less suitable for real-time SAR operations in challenging acoustic environments like disasters.

2.1.3 Other approaches

Beyond DOA and VAD, advanced techniques such as speech enhancement [25] and noise suppression [26] are employed to improve speech detection performance in noisy environments. Nevertheless, their effectiveness drops significantly at SNRs below -10 dB and becomes especially unreliable at -20 dB, where speech is almost entirely masked by noise. Under such extreme conditions, these methods often fail to preserve essential speech cues, highlighting the need for a robust feature capable of distinguishing speech from non-speech.

2.2 Spectro-temporal modulation analysis

2.2.1 Amplitude modulation

In communication theory, modulation is a key process where a carrier signal's characteristics (its amplitude, frequency, or phase) are adjusted based on the information we want to send, called the modulating wave [27]. This complex process involves mathematical operations that embed the modulating signal onto the carrier, thereby producing the modulated signal. This modulated signal is then ready for transmission to a receiver via various communication channels, including cables, fiber-optic lines, or wireless networks. Amplitude modulation is one of the primary modulation techniques employed in practical communication systems. Figure 2.1 provides a clear demonstration, showing how a message signal modifies the amplitude (as in AM) of a carrier signal. It specifically alters the amplitude of the carrier signal to encode the message, while its phase and frequency remain constant. Figure 2.1 (c) illustrates an AM signal with a carrier frequency f_c of 50 Hz, which carries a 5-Hz sinusoidal message. The mathematical equation of the AM signal $s_{AM}(t)$ is described as

$$s_{AM}(t) = A\left(\frac{m(t)+1}{2}\right)\cos(2\pi f_c t), \tag{2.1}$$

where

- A is the peak amplitude of the carrier signal,
- $m(t)$ is the message signal, normalized to range from -1 to 1,
- f_c is the carrier frequency (in Hertz).

After transmission, the receiver needs to extract the original message signal from the incoming signal; this process is called demodulation. In a broadcasting system with multiple simultaneous transmitters, each employs a distinct carrier frequency to enable signal differentiation. The receiver "tunes in" to a specific transmitter by filtering around its carrier frequency,

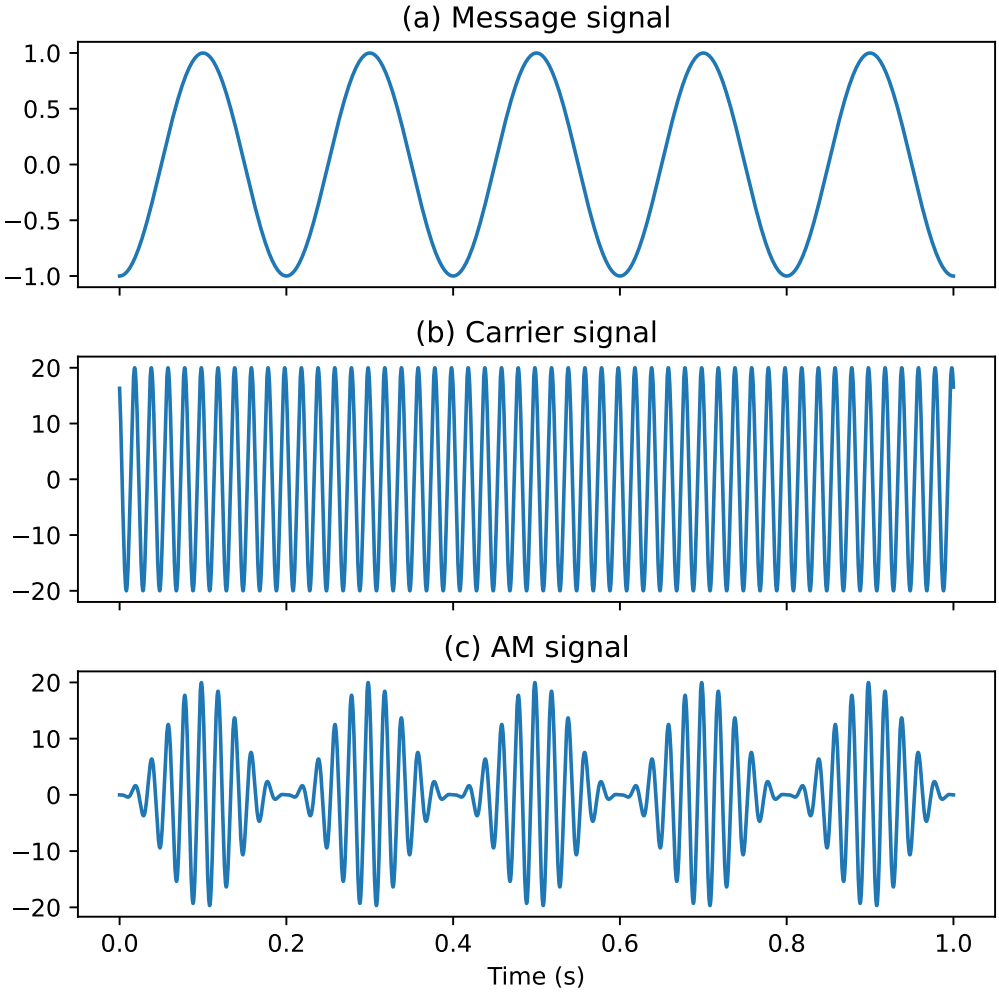

Figure 2.1: Illustration of an AM signal with a carrier frequency of $f_c = 50$ Hz carrying a 5 Hz message m(t): (a) message signal, (b) carrier signal, (c) AM signal.

permitting only signals within a defined frequency band to pass. Then, using AM demodulation techniques (for Amplitude Modulation), the receiver extracts the message signal.

2.2.2 Spectro-temporal modulation of speech

The speech signal undergoes amplitude modulation along both the spectral and temporal axes [28, 29, 30, 31]. Overall, spectral temporal modulations are fluctuations in the amplitude of a speech signal that occur across time and frequency (temporal and spectral domains). [32]

The spectro-temporal modulation (STM) analysis is a computational model of auditory analysis that is based on psychoacoustical and neurophysiological findings in the early and central stages of the auditory system [33]. It offers a unified multi-resolution spectral and temporal representation of sound. The model consists of two main stages:

1. Early Stage: This stage captures monaural processing from the cochlea to the midbrain, which converts the acoustic stimulus into an auditory time-frequency spectrogram-like representation. This auditory spectrogram is an enhanced and noise-robust estimation of the Fourier-based spectrogram.

2. Cortical Stage: This stage models the more complex spectro-temporal analysis believed to occur in the mammalian primary auditory cortex. It derives the spectral and temporal modulation content of the auditory spectrogram using a bank of filters, each tuned to specific spectro-temporal modulation parameters (rates and scales).

Figure 2.2 is an example of the spectro-temporal modulation of speech. The energy separates into two distinct regions along the spectral modulation axis, which describes the source-filter information of human speech [34]. The first region, a triangular area at low spectral modulation frequencies, represents the broad spectral amplitude variations imposed by the upper vocal tract, specifically the formants and their transitions. The second region, located at

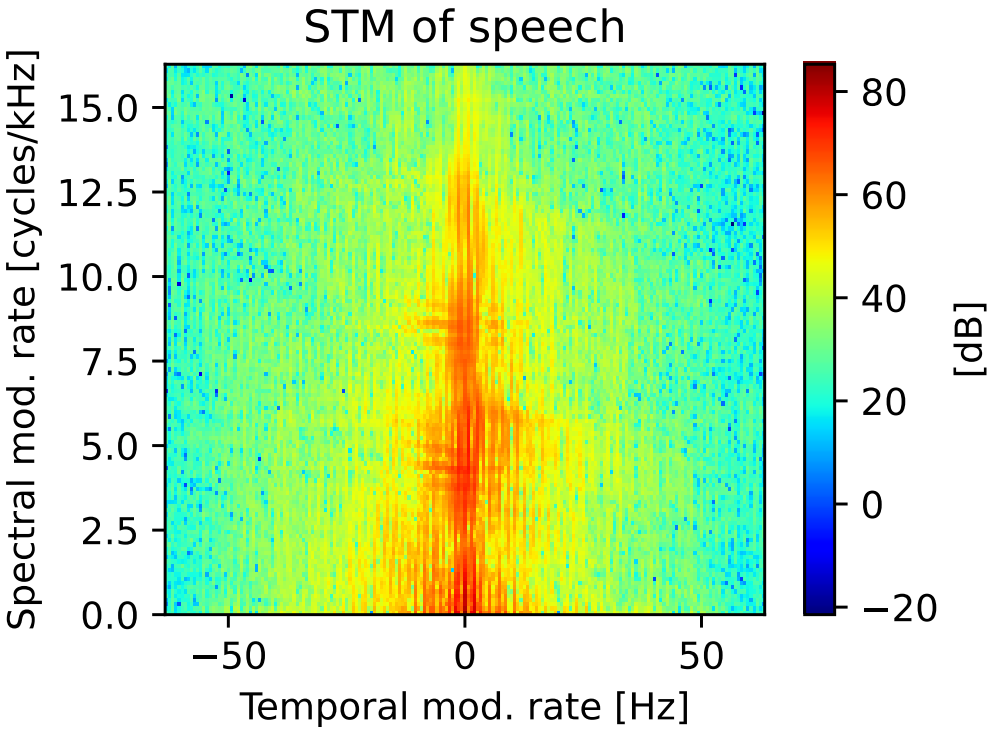

Figure 2.2: Example of STM of speech.

higher spectral modulation frequencies, corresponds to the harmonic structure of vowel sounds generated by glottal pulses [35]. These specific ranges of spectral and temporal modulation rates carry important information about the speech.

Temporal modulation

Temporal modulation represents variations in the amplitude of a sound signal over time, which captures how fast the intensity of a signal fluctuates. Previous studies show that temporal modulation frequencies below 64 Hz are considered the most important for speech perception [36]. The significance of this low-frequency modulation range is directly correlated with the movements of the vocal articulators – the lips, jaw, and tongue. These articulatory gestures, which occur at quasi-syllabic rates, induce systematic changes in the acoustic signal [37]. Specifically, these low-frequency movements are evident in speech acoustics in two primary ways. Firstly, there are oscillations of energy at quasi-syllabic rates, reflecting the opening and closing movements of the lips and jaw. These energy fluctuations correspond to the rhythmic pulsing of speech, providing critical cues for syllable segmentation and prosody. Secondly, there is significant variation in formant patterns, which directly reflect the changes in vocal tract resonances associated with tongue movement. These emphasize the critical role of low-frequency temporal modulations in speech understanding.

Spectral modulation

Spectral modulation involves variations in the frequency content of a sound signal. The energy separates into two distinct regions along the spectral modulation axis, which describes the source-filter information of human speech. Previous works [38, 39] suggest that the achievable fundamental frequency F0 of human speech ranges from approximately 60 to 800 Hz. Consequently, the corresponding spectral modulation rates of speech can span from approximately $\frac{1000}{800} = 1.25$ to $\frac{1000}{90} \approx 11.11$ cycles per kHz, reflecting the inverse

relationship between F0 and modulation rate.

2.2.3 Filterbanks

Various signal processing techniques have been developed to extract the time-frequency spectrogram for STM analysis, each with distinct underlying principles and implications for resolution in both time and frequency domains. The choice of analysis method significantly impacts how well the fine-grained details of speech features can be captured.

Short-Time Fourier Transform

Short-Time Fourier Transform (STFT) is one of the most widely used techniques for analyzing the time-varying spectral content of a signal. Its principle involves dividing the continuous speech signal into short, overlapping frames, applying a windowing function to each frame, and then computing the Discrete Fourier Transform (DFT). The STFT formula is described as:

$$\mathbf{STFT}\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}dt, \qquad (2.2)$$

where $w(\tau)$ is a window function, such as a Hann or Gaussian window centered at zero, and $x(t)$ is the signal to be transformed. A primary characteristic of the STFT is its linear frequency resolution, meaning that frequency bins are equally spaced in Hertz across the entire spectrum.

While straightforward to implement and interpret, the STFT faces inherent limitations regarding its temporal resolution, primarily due to framing effects. The fixed window length imposes a fundamental trade-off: a shorter window provides better temporal resolution but poorer frequency resolution, while a longer window offers better frequency resolution but blurs rapid temporal changes according to Heisenberg's uncertainty principle. This trade-off leads to two standard configurations:

- Narrowband STFT: Uses a longer window ($>$ 20 ms), resulting in high

frequency resolution and low temporal resolution.

- Wideband STFT: Uses a shorter window (< 20 ms), resulting in high temporal resolution and low frequency resolution.

Constant-Band filterbank

Constant-Band filterbank (CBFB) analyzes the signal through a collection of bandpass filters, each with a consistent bandwidth across the entire frequency range. This design provides consistent frequency resolution regardless of the center frequency, which can be advantageous for specific analyses where a uniform resolution across all bands is desired.

Mel filterbank

The Mel filterbank (MelFB) based on the Mel scale, a scale for the measurement of psychological magnitude pitch [40], is widely used in several speech processing tasks. It consists of a set of triangular bandpass filters whose center frequencies and bandwidths are arranged according to the Mel scale. The formula of the Mel scale is described as

$$\text{Mel}(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right), \tag{2.3}$$

where f in Hertz is the frequency. The Mel filterbank can be implemented in either the frequency or time domain:

- Frequency domain implementation: This is the most common approach. The process involves:

    1. Computing the power spectrum of a signal frame (for example, using an STFT).
    2. Applying a set of triangular, overlapping filters to the power spectrum. The filter center frequencies and bandwidths are arranged according to the Mel scale.

3. Summing the energy within each filter band to produce a set of Mel-filtered coefficients.

- Time domain implementation: This approach uses a bank of digital filters (IIR or FIR) that directly operate on the time-domain signal. Each filter is designed to have a frequency response that approximates a Mel-scaled triangular filter.

Gammatone filterbank

Similar to the Mel filterbank, the Gammatone filterbank (GTFB) [41] is also a type of auditory filterbank, particularly the filtering characteristics of the human cochlea. It employs filters with a specific impulse response whose amplitude follows the Gammatone distribution. The spacing of center frequencies in the Gammatone filterbank follows the Equivalent Rectangular Bandwidth (ERB) scale, which is defined as

$$\mathrm{ERB}(f_k) = 24.7 \times \left(\frac{4.37 f_k}{1000} + 1 \right), \tag{2.4}$$

where f_k is the k-th center frequency of the filterbank. Gammatone filterbanks are commonly implemented in several ways:

- Finite Impulse Response (FIR) filter: The Gammatone filters can be implemented as FIR filters by approximating their impulse response with a discrete-time finite-length sequence. This typically involves sampling the continuous-time Gammatone function, which consists of a gamma distribution envelope modulated by a sinusoid, and truncating it to a fixed number of taps.

- Slaney's implementation [42]: A popular and efficient method uses a bank of IIR filters, often a cascade of four second-order sections, to approximate the Gammatone impulse response. This approach is computationally efficient and widely used in psychoacoustic modeling.

2.3 Deep learning techniques for audio classification

Deep learning plays a crucial role in modern life, powering applications such as natural language processing (NLP) and computer vision (CV) tasks. In the field of signal processing, it has opened new directions for analyzing and understanding signals. One key application of signal processing is speech processing, particularly speech detection, which can be considered as a speech/non-speech classification problem, serving as a fundamental front-end step for several speech processing tasks such as voice activity detection [43], speech recognition [44], and other robust audio classification tasks.

Several deep learning techniques have been developed for classifying audio signals [45]. Convolutional Neural Networks (CNNs) excel at extracting local spectro-temporal features from two-dimensional representations like spectrograms, making them highly effective for classifying audio signals [46, 47]. For tasks requiring the understanding of temporal dependencies and sequential information, Recurrent Neural Networks (RNNs) [48] and Long Short-Term Memory (LSTM) networks [49] are employed, capable of processing variable-length sequences. More recently, Transformers [50] have revolutionized audio classification by employing self-attention mechanisms, allowing them to capture global dependencies across an entire audio sequence simultaneously, without relying on sequential processing.

2.3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is one of the fundamental deep learning models, particularly well-suited for analyzing visual and audio data. CNNs are highly effective in capturing local spectro-temporal patterns, which are crucial for capturing two-dimensional feature representations and comprehending audio signals. They have been widely adopted in many audio classification tasks, including the classification of musical instruments [52] and environmental sounds [53]. The core architecture of a CNN often in-

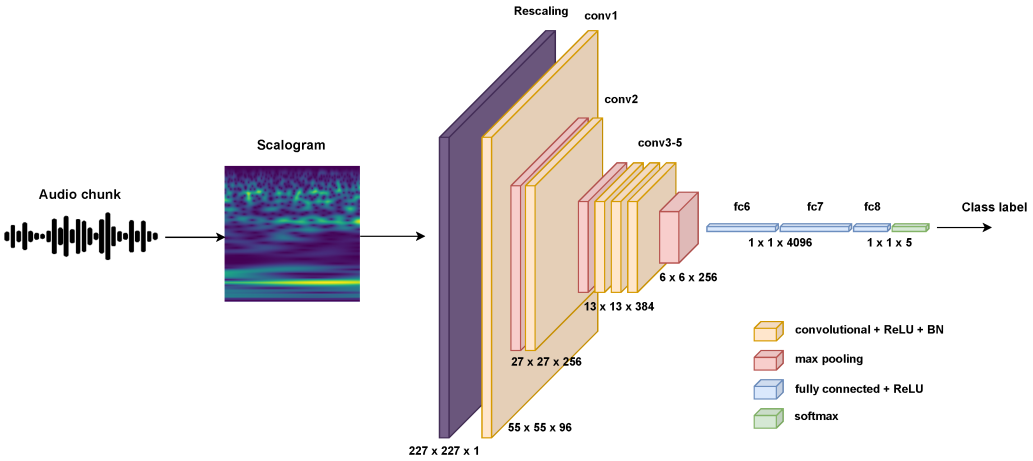

Figure 2.3: Example of CNN architecture for audio classification (from: M. Scarpiniti, R. Parisi, and Y.-C. Lee. "A Scalogram-based CNN approach for audio classification in construction sites". In: Applied Sciences 14.1 (2023), p. 90 [51].)

cludes several convolutional layers, followed by mean/max pooling layers, Rectified Linear Unit (ReLU) activation functions, and some fully-connected layers for final classification. While demonstrating effectiveness in various applications, CNNs can encounter challenges as network depth increases. This includes the "degradation problem", which can lead to slower convergence and potentially poorer performance compared to shallower networks.

2.3.2 ResNet

ResNet [54] is one of the CNN variants that has revolutionized deep learning, particularly in visual tasks [55], and its principles are highly transferable to audio classification [56]. While standard CNNs build hierarchical feature representations through stacked convolutional and pooling layers, they encounter a significant limitation as they deepen: the degradation problem. This makes it challenging to train very deep, effective CNNs, as gradients can vanish during backpropagation, hindering learning in earlier layers.

Figure 2.4 shows the architecture of the ResNet model, which consists of several deep convolution blocks and layers. ResNet addresses this by intro-

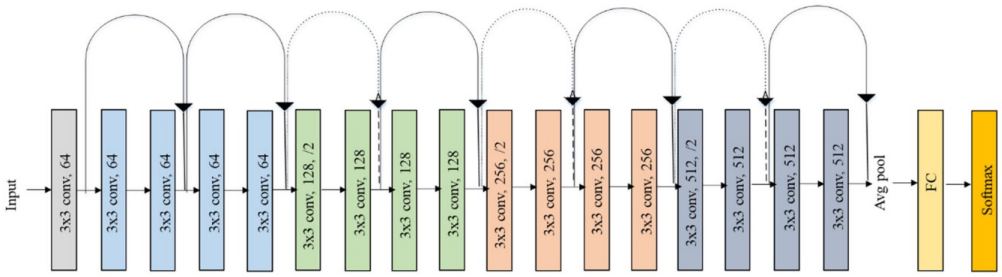

Figure 2.4: Block diagram of ResNet18 architecture (from: F. Ramzan et al. (2020) [55].)

ducing residual connections, also known as "skip connections". Instead of directly learning a mapping from input x to output $H(x)$, a residual block learns a residual function $F(x) = H(x) - x$. The output of the block then becomes $F(x) + x$. This seemingly simple modification allows the network to effectively learn incremental changes from the input, making it easier to optimize deeper architectures. Suppose a layer (or block of layers) is unnecessary. In that case, the network can learn to set $F(x)$ to zero, effectively bypassing that block via the identity mapping, thus preserving information flow and mitigating the vanishing gradient problem.

For complex audio classification tasks like speech detection under heavy noise conditions, deeper networks are often crucial. ResNet's ability to train much deeper networks allows for the learning of progressively more abstract features through its many layers, which is vital for distinguishing subtle speech cues from overwhelming background interference. This makes ResNet a more suitable and powerful choice compared to simpler CNNs for achieving high performance in such demanding real-world audio applications.

2.3.3 Attention mechanism

While ResNet provides a strong backbone for feature extraction, not all features or spatial locations are equally important for distinguishing speech from noise, particularly under varying SNR conditions. Furthermore, another drawback of CNNs and their variants is the spatial invariance problem,

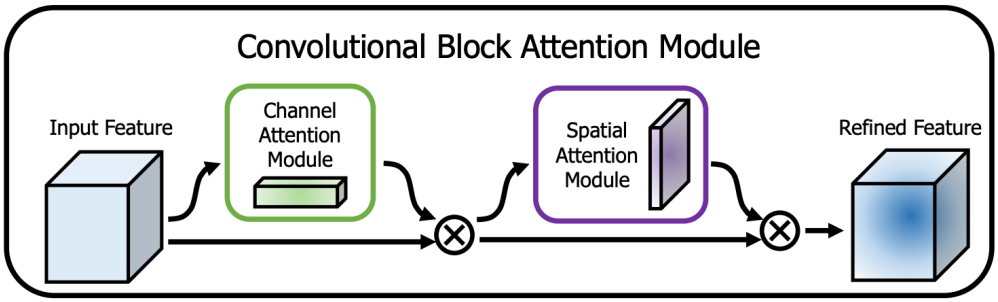

Figure 2.5: The overview of CBAM architecture (from: S. Woo et al. (2018) [57].)

which is the ability to recognize an object regardless of its position in an image. Since the STM of speech is located in a specific region, an attention mechanism is needed to learn the fixed nature of STM features.

The attention mechanism was significantly advanced in 2017 by Google Brain and Google Research teams in the paper "Attention is all you need" [58], which introduced the Transformer architecture — a model relying entirely on self-attention mechanisms to capture global dependencies in sequences without the use of recurrent or convolutional structures. Following this development, attention-based methods have been widely adopted in several tasks, often integrated with convolutional neural networks (CNNs) to enhance feature representation. These approaches enable networks to effectively focus on the most informative spatial regions or feature channels, improving performance in tasks such as image classification, detection, and segmentation. Among these methods, the Convolutional Block Attention Module (CBAM) [57] stands out as an effective and lightweight attention mechanism, refining intermediate feature maps for better representation learning. CBAM architecture, which is illustrated in Figure 2.5, sequentially infers attention maps along two independent dimensions: channel and spatial.

- Channel Attention: This module focuses on "what" is meaningful in the input by performing global average pooling and max pooling oper-

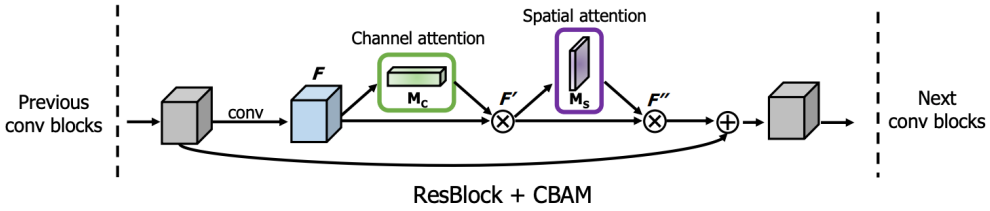

Figure 2.6: Example of CBAM integrated with ResNet (from: S. Woo et al. (2018) [57].)

ations on the input feature map. The pooled features are then fed into a shared multi-layer perceptron (MLP) to generate channel attention weights. These weights are then multiplied by the input feature map, scaling the importance of each channel.

- Spatial Attention: This module focuses on "where" the informative part is by aggregating channel-wise information. It applies average-pooling and max-pooling operations along the channel axis to generate two 2D feature maps. These are then concatenated and convolved to produce a spatial attention map, which highlights relevant spatial locations. This spatial attention map is then multiplied by the channel-refined feature map. Spatial attention can help the model to focus on specific regions in STM features, as speech-related modulations are known to consistently manifest in fixed, characteristic spatial patterns within the spectro-temporal domain.

Chapter 3

Proposed Method

In this chapter, the proposed speech detection method under heavy noise conditions is described in detail, beginning with the extraction of spectro-temporal modulation (STM) features, followed by the complete speech detection system architecture leveraging advanced deep neural network (DNN) models integrated with an attention mechanism to focus on important modulation ranges.

3.1 Overview

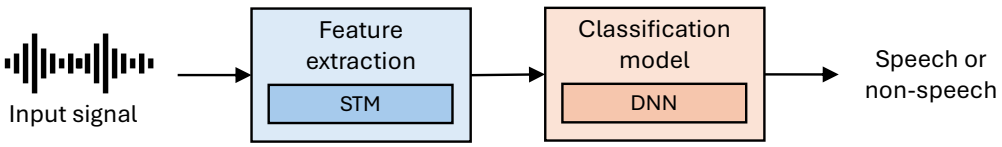

Figure 3.1: Overview of the proposed speech detection method.

In the field of signal processing, the speech detection task can be considered as a speech/non-speech classification problem. The speech detection system's objective is to determine whether an input signal contains human speech or not. When addressing speech detection under heavy noise conditions, "speech" refers to noisy speech signals, while "non-speech" refers to noise-only signals. Therefore, the speech detection method proposed in

this study functions as a speech/non-speech classification system. Figure 3.1 illustrates the overview of the proposed speech detection method, which contains two main stages: feature extraction and classification model.

- **Feature extraction using STM analysis:** The proposed method leverages STM analysis, which involves applying a 2D Fourier transform to a spectrogram, and offers several benefits over traditional time-frequency representations, particularly in the context of speech analysis:
  - Captures joint temporal and spectral modulations: While a spectrogram shows how energy varies across frequency and time, STM quantifies how the energy patterns themselves modulate along both axes, which captures temporal modulation (syllabic and prosodic rhythms), and spectral modulation (formant structure and harmonic spacing)
  - Improved feature discrimination: STM analysis highlights key modulation patterns characteristic of different phonemes and syllables of human speech, which can distinguish between voiced and unvoiced segments, and between speech and noise.
  - More robust to noise: STM representations are more robust to background noise since they focus on modulation patterns rather than raw energy levels. Speech-relevant modulations tend to occupy specific regions in STM space (e.g., -64 to 64 Hz temporal and 0 to 16.67 cycles/kHz spectral), which allows filtering out irrelevant components.
  - Biological plausibility: STM closely aligns with how the human auditory cortex processes sound, suggesting that STM is a perceptually meaningful representation of speech signals.
- **Deep-learning-based classification model:** The use of deep learning models is particularly advantageous for this task because of their

ability to learn complex, hierarchical feature representations directly from the input data. Unlike traditional machine learning methods that often require hand-crafted features, deep learning models can automatically discover the discriminative patterns in the STM features, which is crucial for speech detection under noisy conditions. A CNN architecture is first used as the baseline, then extended to a deeper ResNet model, and further enhanced with an attention mechanism named Convolutional Block Attention Module (CBAM) to improve performance under heavy noise conditions. CNNs are well-suited for learning hierarchical features from image-like data, making them suitable for processing the STM features produced in the feature extraction stage. The ResNet architecture allows for the training of deeper networks, which enables the model to learn more complex and robust representations of speech and noise modulation patterns. Furthermore, the integration of CBAM helps the network to emphasize and focus on dominant regions of speech features while suppressing irrelevant ones of noise in the STM domain.

Together, these stages enable robust speech detection in challenging acoustic environments.

3.2 Feature extraction: STM

3.2.1 STM calculation process

The calculation of STM involves a sequence of calculation steps, which convert a signal from a one-dimensional acoustic waveform into a two-dimensional representation of its spectro-temporal characteristics. Figure 3.2 describes the process of the STM analysis. Initially, the raw input signal is decomposed from the time domain into a time-frequency representation, typically a spectrogram. The spectrogram of the input signal is extracted using one of the four filterbanks: STFT, Constant-band filterbank (CBFB), Mel filter-

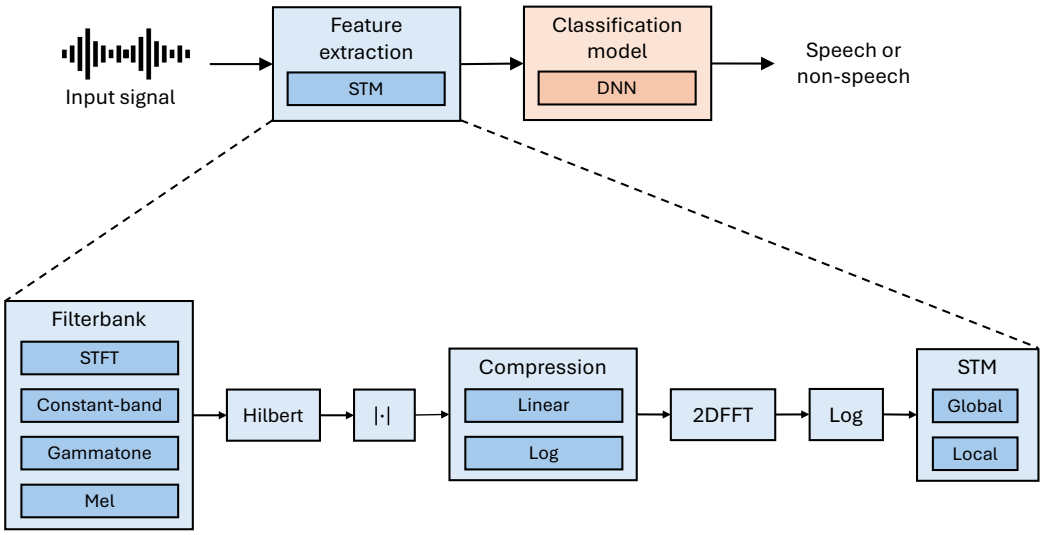

Figure 3.2: Block diagram of the STM calculation process.

bank (MelFB), and Gammatone filterbank (GTFB). The outputs of these filterbanks are sub-band signals, which are obtained as

$$y_k(t) = g_k(t) * s(t), \tag{3.1}$$

where

- $s(t)$ is the input signal,
- $g_k(t)$ is the impulse response of the k-th channel of the filterbank,
- $*$ is the convolution operation,
- $y_k(t)$ is the output of the k-th channel of the filterbank.

The collection of outputs $y_k(t)$ forms the spectrogram with shape (K, T), where K is the number of frequency channels and T is the number of time frames. After that, to extract the instantaneous amplitude envelope of each frequency band, the Hilbert transform is applied to each sub-band signal from the filterbank output as

$$e_k(t) = |y_k(t) + j\text{Hilbert}(y_k(t))|, \tag{3.2}$$

where

- $y_k(t)$ is the output of the k-th channel of the filterbank,
- Hilbert$(\cdot)$ is the Hilbert transformation,
- $|\cdot|$ is the absolute value,
- $e_k(t)$ is the temporal amplitude envelope of the k-th channel.

Taking the absolute value of the analytic signal's real or imaginary part yields the amplitude envelope for each frequency channel, which captures the energy fluctuations in each frequency band. Finally, a two-dimensional fast Fourier transform (2DFFT) is applied to the resulting time-frequency representation, yielding the spectral-temporal modulation, which captures the joint modulation patterns across time and frequency domains and is computed as

$$\text{STM} = \text{2DFFT}(e_k(t)). \tag{3.3}$$

The result of applying 2DFFT is a matrix comprising complex numbers; therefore, the absolute value is then applied to obtain the STM representation. The resulting STM can then be converted to the log-scale (in dB) by applying a logarithmic operation, which helps to reduce the dynamic range of its energy.

3.2.2 STM configurations

As described in Figure 3.2, STM analysis has various configurations, which are dependent on different parameters. These configurations involve the selection of the initial filterbank, the type of amplitude envelope compression, and the range of spectro-temporal modulations considered for analysis.

Filterbanks

The first configuration investigates the effectiveness of different filterbanks used for time–frequency spectrogram decomposition. Four distinct filter-

banks are considered in this study:

- Short-Time Fourier Transform (STFT): This is the most fundamental and widely used technique to convert the input raw waveform to the time-frequency spectrogram representation. This method applies discrete Fourier transforms over short, overlapping frames of the signal. It provides a linear frequency scale with equally spaced bins; however, it faces a limitation in temporal resolution due to the framing effect.

- Constant-band filterbank: To address the poor temporal resolution of STFT due to the framing effect, the constant-band filterbank is employed, as this time-domain filterbank provides detailed temporal resolution. These filters have a fixed bandwidth across the entire frequency range, offering a consistent frequency resolution on a linear scale.

- Mel filterbank: This time-domain filterbank is also one of the most commonly used filterbank in several signal processing tasks. Exhibiting narrower bandwidths at low frequencies and wider bandwidths at higher frequencies, this filterbank is based on the Mel scale, whose triangular filter shape approximates the perceptual frequency resolution of the human auditory system.

- Gammatone filterbank: To mimic human auditory perception, the Gammatone filterbank is leveraged. The Gammatone filters are based on the ERB scale with narrower bandwidths at low frequencies and wider bandwidths at higher frequencies. The narrow low-frequency bands can better capture important speech features, as much of the speech energy is concentrated in this range.

The choice of filterbanks results in a different time-frequency spectrogram representation, producing distinct STM patterns that can affect the feature extraction stage and the performance of the classification model.

Compression

The second configuration is the compression applied to the amplitude envelopes after the Hilbert transform. Two types of compression are examined:

- Linear: The 2DFFT is applied directly after taking the absolute value of the amplitude envelopes, producing the linear STM that preserves the original dynamic range of the envelopes.
- Logarithm: By applying a logarithmic function, this method compresses the high-amplitude variations more strongly than low-amplitude ones, thereby reducing the dynamic range in the resulting STM.

Global STM and global STM

The final STM analysis can be configured in two ways: global STM and local STM, which differ in the resolution and specific modulation ranges:

- Global STM: This configuration preserves the full range of temporal modulation frequencies. Meanwhile, the spectral modulation range is based on the fundamental frequency (F0) range of human speech, which typically spans from around 60 Hz to 800 Hz; therefore, the important spectral modulation rates are considered to be from $\frac{1}{800} = 1.25$ cycles per Hz to $\frac{1}{60} \approx 16.67$ cycles per Hz. Due to the computational complexity of the filterbanks, the spectral modulation of global STM is limited to cover this specific range, which results in the identical spectral modulation resolution between global STM and local STM.
- Local STM: The local STM focuses on a specific range of temporal modulations. For speech, important temporal modulations usually fall within the range of -64 Hz to 64 Hz. This configuration enhances the discriminability of speech by emphasizing the most informative modulation cues while potentially ignoring irrelevant components that are outside this range.

By investigating all these configurations, the proposed method aims to determine the optimal STM feature for robust speech detection under heavy noise conditions.

3.2.3 STM validation

To validate the efficiency of the STM features under various noisy conditions, a comprehensive visualization of the spectrograms derived from four different filterbanks and the corresponding STM representations is performed, demonstrating how different filterbank choices affect the initial time-frequency representation and how the final STM features capture distinct patterns for clean speech, noise, and speech mixed with noise across various SNRs.

Spectrograms of speech derived from four filterbanks

Figure 3.3 illustrates various spectrograms of a clean speech utterance derived from the four different filterbanks: STFT, Constant-band filterbank, Gammatone filterbank, and Mel filterbank. The speech sections display clear harmonic structures. While the STFT and Constant-band spectrograms show uniform frequency resolution, the Gammatone and Mel representations emphasize lower-frequency harmonics with finer resolution.

STM of speech

The core of the STM validation involves examining the STM spectrograms themselves. For each of the four filterbank configurations, a set of STM visualizations of clean speech is shown in Figure 3.4. The difference in the spectrograms of each filterbank leads to distinct speech patterns in the STM representation.

STM of white noise

Figure 3.5 demonstrates the spectrogram and STM of white noise. It is notable that white noise exhibits a peak at the origin in the STM. Origin is

the point at 0 Hz in temporal modulation and 0 cycles per Hz in spectral modulation, which corresponds to the averaged power in the spectrogram and reflects the noise power.

STM of noisy speech at various SNRs

The STMs of noisy speech at five distinct SNRs will be presented: 20 dB, 10 dB, 0 dB, −10 dB, and −20 dB. At higher SNRs, the speech patterns are visible, while at gradually lower SNRs, the effect of the noise modulations will become more dominant, challenging the extraction of speech-specific features. The comparison across SNRs visualized in Figure 3.6 demonstrates how the speech information is masked or preserved within the STM representation as noise levels increase.

Global STM and local STM

Figure 3.7 compares the global STM and the local STM of a noisy speech signal at SNR 0 dB. The global STM preserves the full temporal and spectral modulation resolution, demonstrating high energy at low temporal modulation and across spectral modulation. Regarding local STM, temporal modulation is limited to the range of −64 to 64 Hz, which highlights the high-energy region to reveal the speech distinct patterns.

3.3 Classification model

The final stage of the proposed speech detection system involves classification of the STM representations obtained from the feature extraction stage as either "speech" (noisy speech signals) or "non-speech" (noise-only signals), which is illustrated in Figure 3.8. The classification model starts with a foundational Convolutional Neural Network (CNN), progressing to a Residual Network (ResNet) for deeper feature learning, and finally incorporating the Convolutional Block Attention Module (CBAM) to enhance feature discrimination.

As an initial approach, a standard CNN is employed as a baseline classification model. CNNs are capable of learning image-like data, such as spectro-temporal representations, leveraging their ability to learn hierarchical features. The CNN architecture in the proposed method consists of two convolutional layers, each followed by a two-dimensional batch normalization (BatchNorm2d) layer, a Rectified Linear Unit (ReLU) activation function, and a max pooling layer. The final layer of the system is a single neuron with a sigmoid activation function, which outputs a probability in the value range $[0, 1]$, indicating the likelihood of the input signal being speech.

The CNN baseline provides initial insights into the separability of speech and noise features in the STM domain. However, deeper networks often capture better complex patterns, which led to the exploration of ResNet architectures. The "skip connections" of ResNet allow gradients to flow directly through the network, enabling the training of deeper models without compromising performance and allowing for the learning of better, more robust representations from the input data. The specific ResNet architecture implemented for this study is the standard ResNet18, where the last layer is also a single neuron with a sigmoid activation function for binary classification.

Speech features are fixed at a specific location in the STM representation, which is at lower spectral and temporal modulations. To guide the speech detection model to focus on the speech-dominant region in the STM, the CBAM is integrated into the ResNet architecture. CBAM is an attention mechanism that sequentially applies the channel and spatial attention modules to intermediate feature maps, allowing the network to focus on "what" and "where" to emphasize. In the proposed method, the CBAM module is inserted within the ResNet architecture, which guides the network to pay more attention to speech-discriminative patterns while suppressing noise-dominant features.

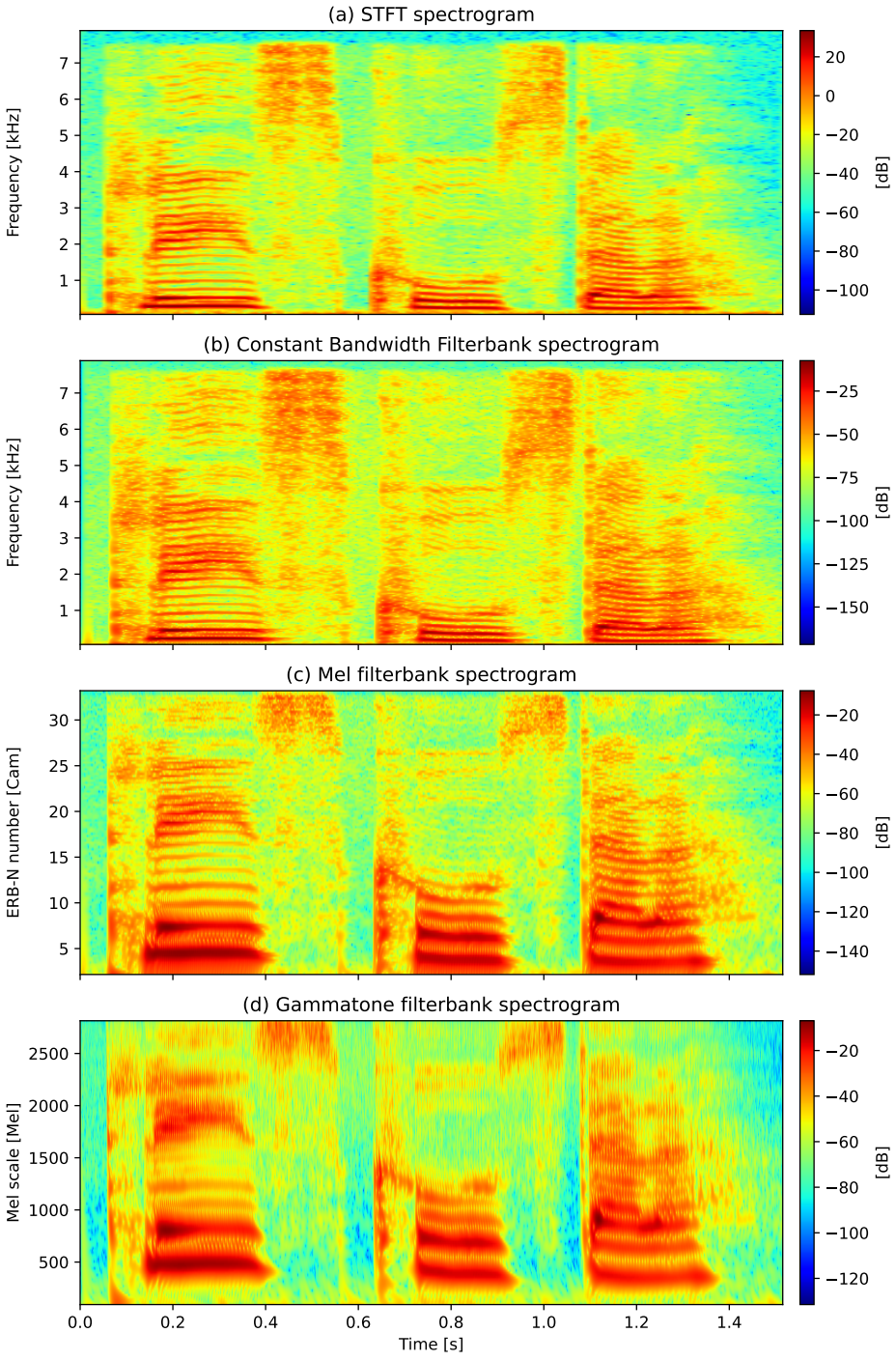

Figure 3.3: Spectrograms of clean speech derived from four filterbanks: (a) STFT spectrogram, (b) Constant-bandwidth filerbank spectrogram, (c) Mel filterbank spectrogram, and (d) Gammatone filterbank spectrogram.

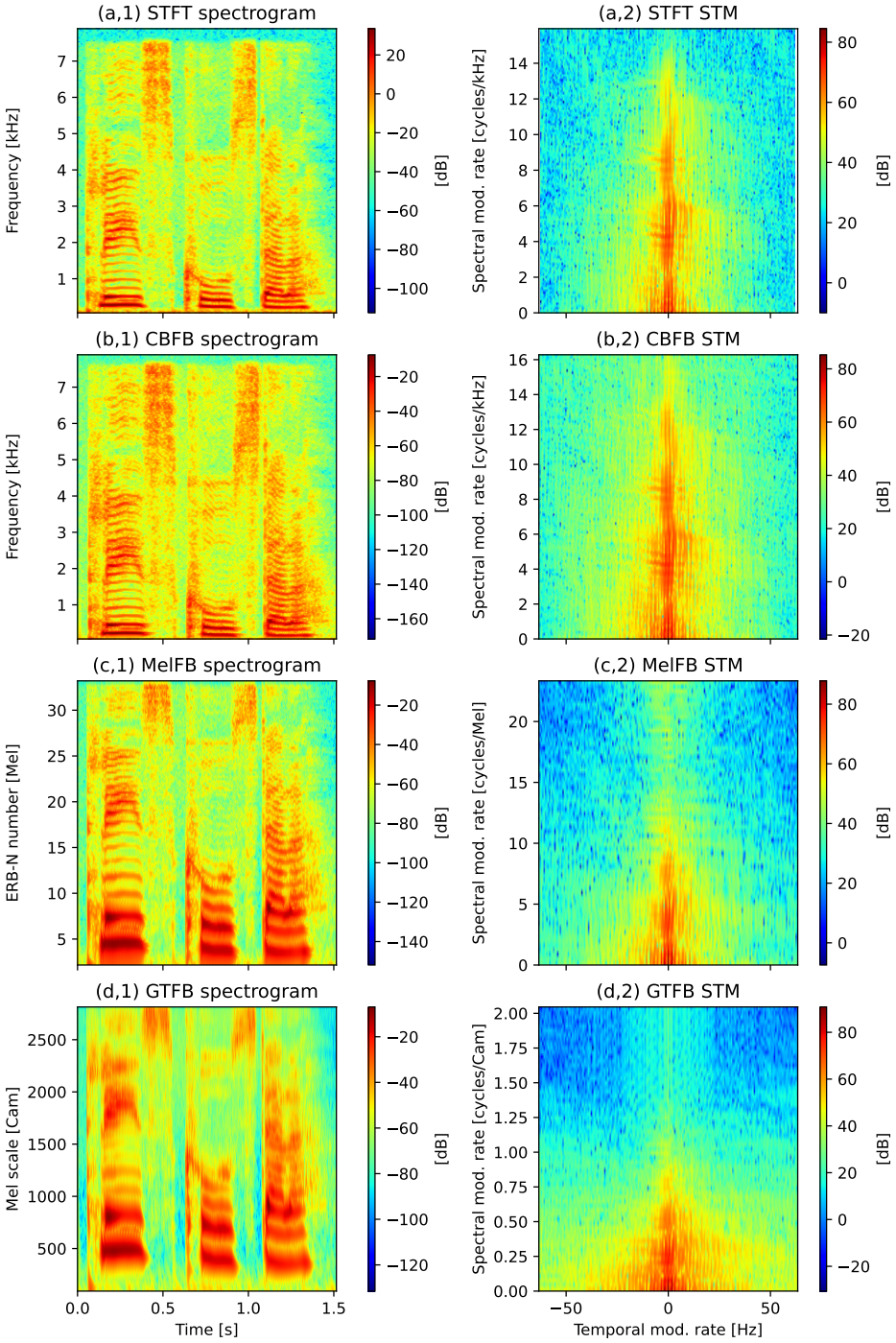

Figure 3.4: Spectrograms and STM of clean speech derived from four filterbanks, where (a,*), (b,*), (c,*), and (d,*) corresponds to STFT, Constant-band, Mel, and Gammatone filterbanks, while (*,1) and (*,2) represents spectrograms and STMs of clean speech, respectively.

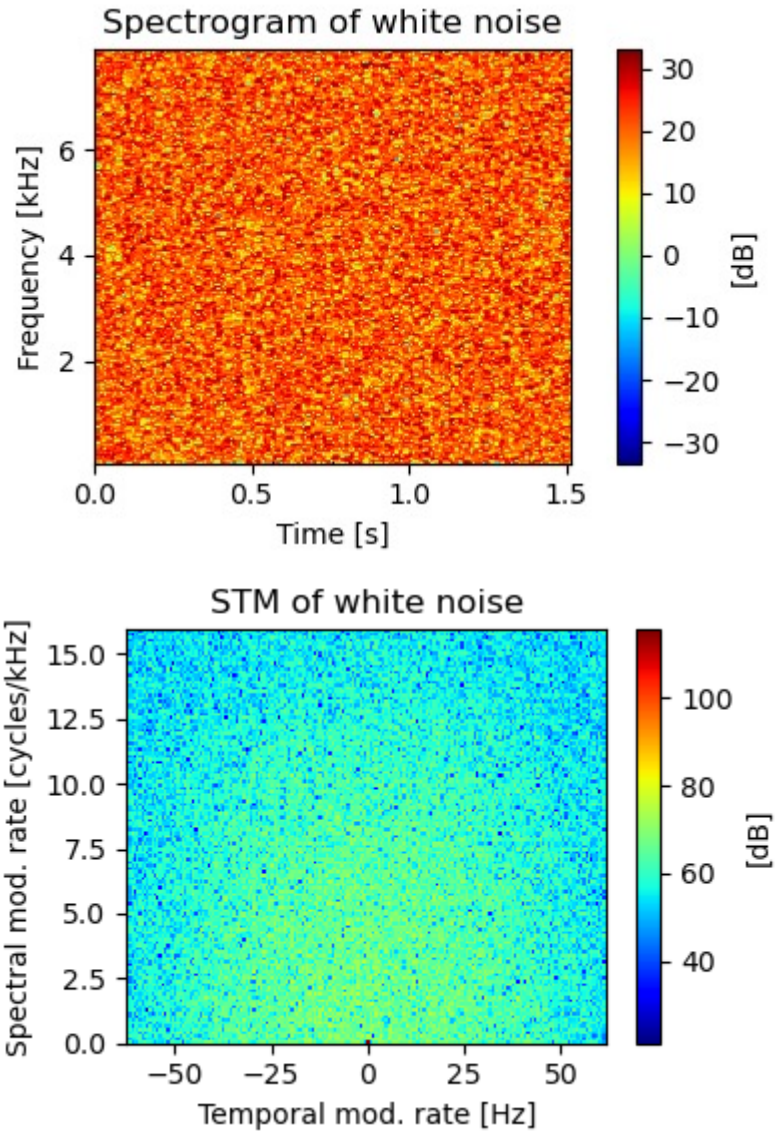

Figure 3.5: Example of STFT spectrogram and STM of white noise.

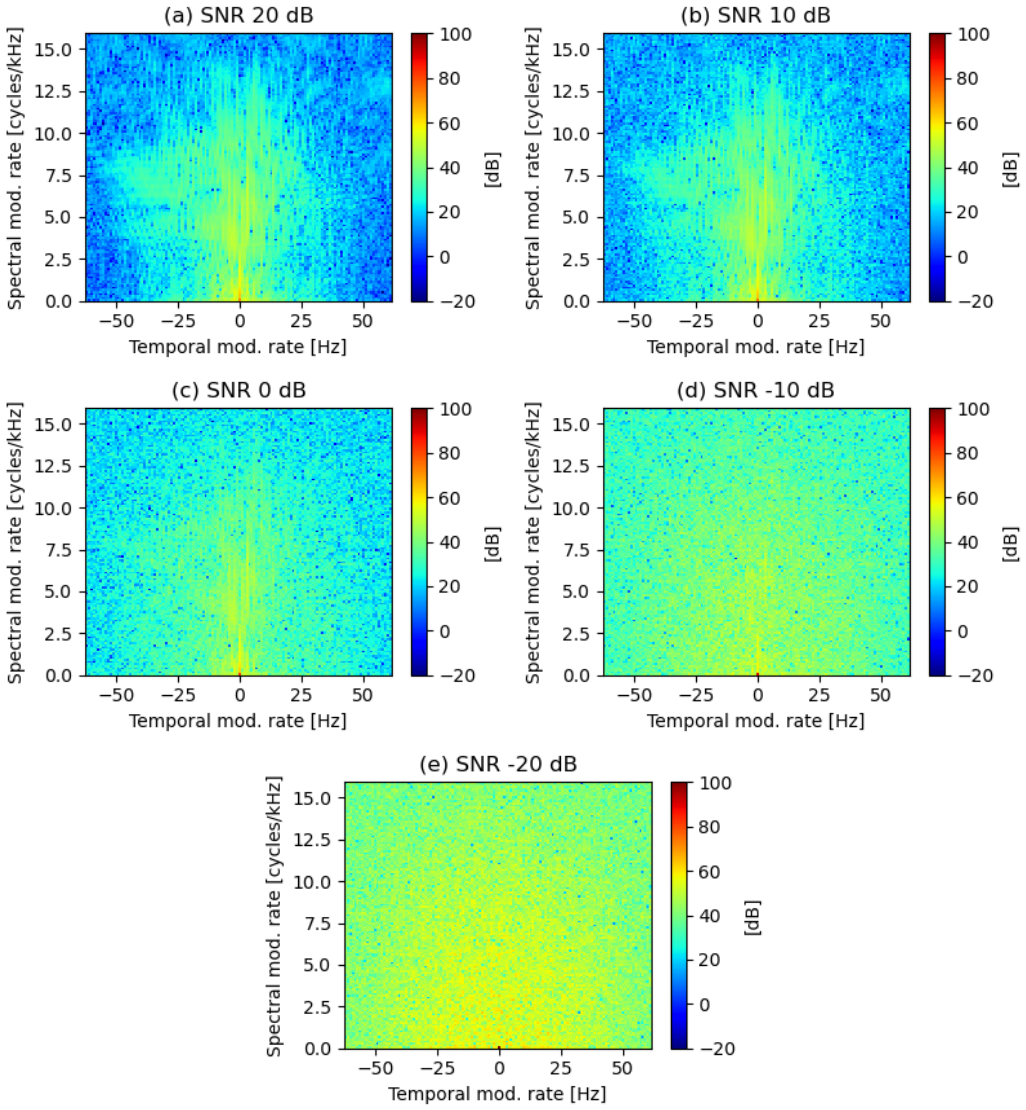

Figure 3.6: STMs of noisy speech at various SNRs: (a) SNR 20 dB, (b) SNR 10 dB, (c) SNR 0 dB, (d) SNR −10 dB, and (e) SNR −20 dB.

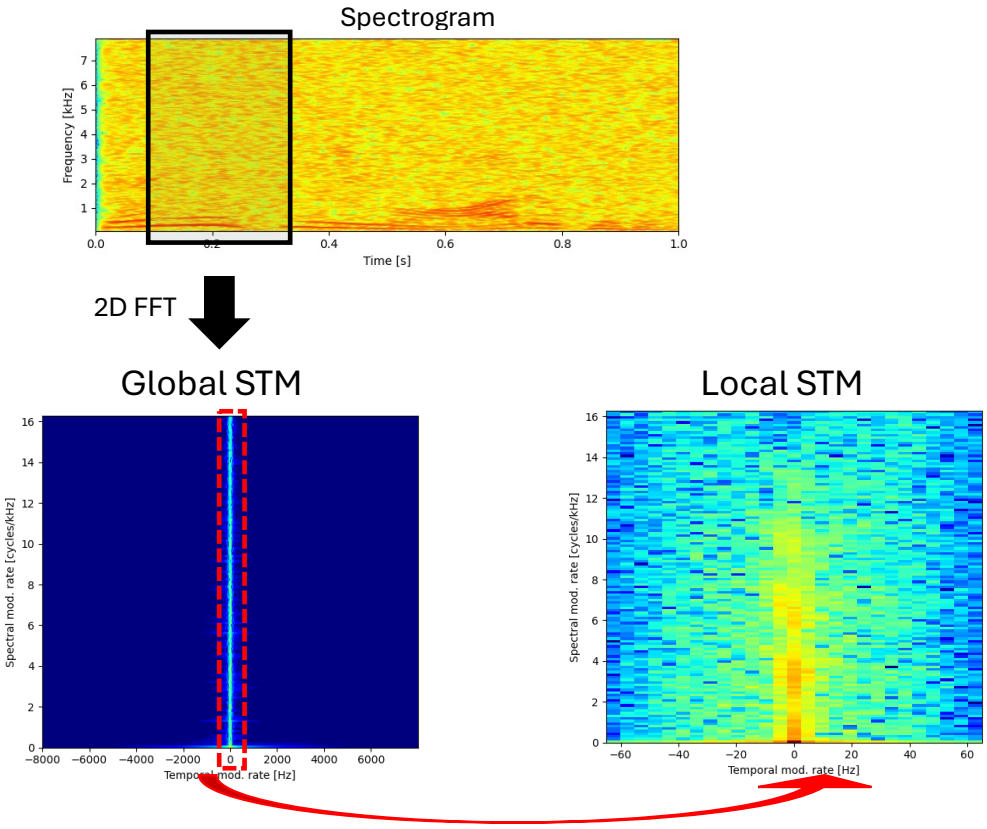

Figure 3.7: Comparison between global STM and local STM of noisy speech at SNR 0 dB.

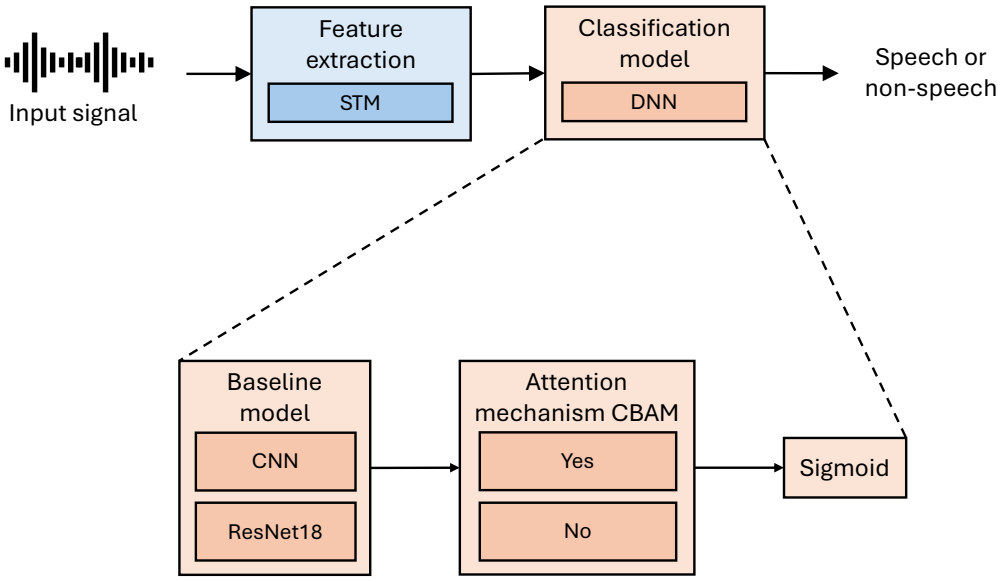

Figure 3.8: Block diagram of the classification model.

Chapter 4

Evaluations

This chapter presents the experimental evaluations of the proposed speech detection method under heavy noise conditions. It describes the datasets, evaluation metrics, and experimental setup, followed by analyses of the results. Comparisons with baseline methods and different system configurations are also provided to assess performance and identify the most effective approach.

4.1 Dataset

To develop a speech detection model, it is essential to utilize both speech and non-speech datasets, which include noisy speech and noise-only signals, respectively. Specifically, a clean speech dataset and a noise dataset were utilized in this research.

4.1.1 Clean speech dataset

The Japanese Versatile Speech (JVS) corpus [59] is utilized as the clean speech dataset in this study. The corpus is a collection of high-quality recordings from 100 native Japanese professional speakers, including 49 males and 51 females. Each speaker contributed 30 different utterances, and all recordings were sampled at 24 kHz.

During the training pipeline, the JVS corpus is partitioned into train, validation, and test sets, and the audios in the corpus are downsampled to a 16,000 Hz sampling rate. This partitioning ensures that the balanced characteristics of speakers and utterances are maintained across each subset. The specific details of the clean speech dataset used for the training pipeline are summarized in Table 4.1.

Table 4.1: Description of speech dataset.

Type	**Number of utterances**	**Description**
Train	1740	29 males and 29 females
Validation	600	10 males and 10 females
Test	600	10 males and 10 females

4.1.2 Noise dataset

In real disaster scenarios, background noise is usually complex and non-stationary, such as machinery, alarms, and building collapse. However, in this study, the evaluation begins with a controlled noise environment to establish a baseline for performance analysis. Specifically, the noise dataset consists exclusively of white noise, a stationary broadband signal with equal energy distributed across all frequencies. The white noise samples are synthesized using a random process with a uniform distribution. The number of noise samples equals the number of clean speech utterances in each train, validation, and test dataset to create a balanced evaluation. This setup enables an initial assessment of the proposed method's ability to extract and detect speech features before extending to more realistic disaster noise conditions in future work.

4.1.3 Speech/non-speech dataset construction

A data augmentation strategy is employed to construct the speech/non-speech dataset under various noisy conditions. The speech/non-speech dataset

for the speech detection system development is constructed using an on-the-fly augmentation approach.

For each training or validation sample, a clean speech utterance is randomly selected from the clean speech dataset, and a white noise sample is randomly synthesized. Both the selected speech and noise signals were normalized to a fixed duration of 200ms. This duration corresponds to the syllable and phoneme rates of human speech [60].

To simulate noisy conditions, the clean speech signal is mixed with the noise sample at various signal-to-noise ratios (SNRs). For each speech-noise pair, an SNR value was randomly selected from a continuous range of -20 dB to 20 dB. SNR measures the ratio between the speech signal power and the noise signal power. The lower the value of the SNR, the greater the power of noise in the noisy speech signal. SNR is calculated as

$$\mathrm{SNR} = 10 \log_{10} \frac{\mathrm{P}_{\mathrm{speech}}}{\mathrm{P}_{\mathrm{noise}}}, \tag{4.1}$$

where

- $\mathrm{P}_{\mathrm{speech}}$ is the power of speech signal,
- $\mathrm{P}_{\mathrm{noise}}$ is the power of noise signal.

The resulting noisy signals are labeled as speech (1), and the noise-only signals are labeled as non-speech (0). Eventually, the speech/non-speech dataset is partitioned into the train, validation, and test datasets for model development.

4.2 Evaluation metrics

4.2.1 Equal error rate

In a binary classification task such as speech detection, false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER) are the commonly used metrics to evaluate the performance of the system.

The FAR is the proportion of non-speech mistakenly classified as speech. A lower FAR indicates a more precise system because it minimizes the instances where non-speech is wrongly accepted as speech. FAR is calculated as:

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{4.2}$$

where FP and TN denote the number of false positives (non-speech incorrectly classified as speech) and true negatives (non-speech correctly classified as non-speech), respectively.

The FRR quantifies the percentage of genuine speech instances misclassified as non-speech. A lower FRR reflects a more sensitive system where genuine speech is not frequently dismissed. FRR is calculated as:

$$\text{FRR} = \frac{\text{FN}}{\text{FN} + \text{TP}}, \tag{4.3}$$

where FN and TP denote the number of false negatives (speech incorrectly classified as non-speech), true positives (speech correctly classified as speech), respectively.

The EER is defined as the point where the FAR equals the FRR with respect to different decision thresholds. A lower EER signifies a more accurate and robust system, effectively balancing the acceptance of speech samples and the rejection of non-speech samples.

4.2.2 Accuracy

Accuracy serves as a fundamental and overall measure of the system's correctness, reflecting the ratio of accurate predictions to total predictions. This includes both correctly recognized speech and non-speech elements. It is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{4.4}$$

A higher accuracy value indicates a greater number of correct classifica-

tions by the system. Accuracy is typically considered alongside FAR, FRR, and EER for a comprehensive evaluation, especially in noise-heavy conditions where the balance between correctly identifying speech and rejecting noise is critical.

4.3 Experimental setup

4.3.1 Filterbanks configurations

For feature extraction, raw audio signals are transformed into STM representations. This involves generating spectrograms using four filterbanks described in this research. The frequency range for analysis spans from a minimum frequency of 60 Hz up to half of the sampling rate of 8000 Hz. The specific configurations for each filterbank are as follows:

- STFT: The FFT size is set to 512, with the window length of 32 milliseconds and hop length of 8 milliseconds. This configuration balances frequency and temporal resolutions, ensuring sufficient overlap for smooth transitions between frames while maintaining computational efficiency.

- Constant-band filterbank: A total of 256 filters are employed, each designed as a 4th-order Butterworth bandpass filter. The bandwidth of each filter is determined by dividing the total frequency range by the number of filters, with a minimum bandwidth of 55 Hz enforced to prevent excessively narrow filters. The center frequencies are linearly spaced, and for each center frequency, a Butterworth bandpass filter of the specified order is designed.

- Gammatone filterbank: This filterbank utilizes 128 filters. The design principle involves spacing filter center frequencies linearly on the ERB-rate scale. For each center frequency, an IIR Gammatone filter is designed, typically corresponding to a 4th order to model human hearing.

- Mel filterbank: This filterbank is configured with 128 filters, each implemented as a 2nd-order Butterworth bandpass filter. The design principle involves spacing filter center frequencies linearly on the Mel scale. The bandwidth of each filter is set such that it is proportional to its center frequency on the Mel scale.

4.3.2 Local STM determination

As described in the previous chapter, the local STM is limited to the temporal modulation range of (-64, 64) Hz and spectral modulation range of (1.25, 16.67) cycles per kHz. The configuration of local STM modulation is achieved by selectively cropping the global STM to ensure that the temporal modulation values align within this specified range. It is important to note that the selection of the filterbank configurations (the number of channels, the center frequency values, and the bandwidth of each channel) is based on the spectral modulation characteristics of human speech. Specifically, the number of channels in each filterbank is determined to effectively encompass the lowest and highest achievable fundamental frequencies of human speech, which are approximately 60 Hz and 800 Hz, respectively. This consideration leads to a spectral modulation range that consistently maintains the values of (1.25, 16.67) cycles per kHz for both local and global STMs.

Furthermore, due to the careful selection of parameters and the framing effects of the STFT, it is observed that the global STFT-based STMs match the determined spectral and temporal modulation range. Therefore, the configurations of global and local STMs obtained from STFT are the same.

4.3.3 Speech detection model parameters

The speech/non-speech classification models in the proposed method begin with a standard CNN architecture, which is then upgraded to the ResNet18 model integrated with the CBAM. The input to the system is the two-dimensional STM features, which have different shapes depending on the

specific filterbank configuration used in the feature extraction stage. The final layer of each network is a fully-connected layer, followed by a sigmoid activation function, designed to output a probability indicative of speech presence within the input.

The model undergoes training for 100 epochs, using a batch size of 64 samples for each iteration. The Binary Cross-Entropy (BCE) loss function is employed, which is defined as

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \tag{4.5}$$

where N is the total number of training samples, $y_i \in \{0, 1\}$ is the ground truth label indicating non-speech or speech, and $\hat{y}_i \in [0, 1]$ is the predicted probability output by the model for the i-th sample. The Adam optimizer was chosen for the optimization process, initialized with a learning rate of $1e-6$. A ReduceLROnPlateau scheduler is implemented, which reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 10 consecutive epochs. The model that yields the lowest validation loss during training is saved as the best-performing checkpoint for later testing.

4.4 Results and Discussions

4.4.1 Overall performace

Table 4.2 details the overview performance of the proposed method across various local-STM configurations. This assessment considers four types of filterbanks (STFT, Constant-band, Gammatone, and Mel), two compression types (linear and logarithmic), and three classification models (CNN, ResNet18, and ResNet18 integrated with the CBAM attention mechanism). As described in the previous chapter, the global and local STFT-based STM configurations are the same due to the selection of STFT parameters and framing effect.

Table 4.2: Accuracy of the proposed method with different STM configurations.

Filterbank	**Compression**	**Model**	**Accuracy (%)**	
			Local STM	Global STM
STFT	Linear	CNN	-	87.77
STFT	Linear	ResNet	-	85.88
STFT	Linear	ResNet+CBAM	-	88.34
STFT	Log	CNN	-	82.61
STFT	Log	ResNet	-	82.87
STFT	Log	ResNet+CBAM	-	84.38
CBFB	Linear	CNN	85.65	87.74
CBFB	Linear	ResNet	85.90	90.36
CBFB	Linear	ResNet+CBAM	87.83	90.65
CBFB	Log	CNN	78.86	81.59
CBFB	Log	ResNet	81.35	88.30
CBFB	Log	ResNet+CBAM	84.47	88.17
MelFB	Linear	CNN	75.81	84.30
MelFB	Linear	ResNet	75.24	84.36
MelFB	Linear	ResNet+CBAM	78.00	83.43
MelFB	Log	CNN	64.06	71.77
MelFB	Log	ResNet	75.04	86.64
MelFB	Log	ResNet+CBAM	83.16	85.13
GTFB	Linear	CNN	81.62	88.34
GTFB	Linear	ResNet	86.64	91.14
GTFB	Linear	ResNet+CBAM	**88.88**	**92.79**
GTFB	Log	CNN	74.95	76.76
GTFB	Log	ResNet	75.76	88.70
GTFB	Log	ResNet+CBAM	82.15	90.24

The experimental results in Table 4.2 reveal several important trends. First, the Gammatone filterbank (GTFB) consistently achieves the highest accuracies for both Local and Global STM, with the best performance obtained using linear compression and the ResNet augmented with CBAM model (88.88% and 92.79%, respectively). This indicates that the auditory-inspired frequency resolution of the GTFB, which emphasizes low-frequency components, is effective for speech/non-speech discrimination under noisy conditions. Second, linear compression generally outperforms logarithmic

compression across most configurations, suggesting that preserving the original amplitude dynamics benefits STM-based feature extraction. Third, incorporating the CBAM into ResNet leads to accuracy improvements in nearly all cases, demonstrating the advantage of the attention mechanism in focusing on the important modulation regions of speech in the STM representation. Furthermore, STFT-based features showed stable performance with all its configurations, demonstrating an accuracy above 82%. The final observation is that the global STM typically yields higher accuracies than local STM, implying that preserving entire STM results in richer feature representations and stronger discriminative power. In conclusion, the optimal STM configuration of the proposed method involves the Gammatone filterbank, linear compression, global STM, and ResNet enhanced with the CBAM model.

4.4.2 Comparison with other methods

In order to assess the effectiveness of the speech detection method proposed in this research, a comparative analysis was conducted alongside two established methods, rVAD-fast [4] and WebRTC VAD [24], using the same test dataset. The configuration used in the comparison is the proposed method's optimal configuration, with the STM features derived from the Gammatone filterbank, linear compression, global modulation range, and ResNet integrated with the CBAM classification model. This comparison was based on accuracy and EER across a range of signal-to-noise ratio (SNR) conditions.

The outputs generated by rVAD-fast consist of speech segments rather than explicit labels indicating speech or non-speech. Consequently, to assess the performance of speech and non-speech classification, the feature extraction phase and the segment classification process of rVAD-fast are utilized for comparison purposes. In contrast, the outputs produced by WebRTC VAD are direct classifications of either speech or non-speech, which facilitates a more straightforward comparison with the proposed method.

Accuracy comparison

Figure 4.1 (a) presents the accuracy comparison between the proposed method and two typical methods, rVAD-fast and WebRTC VAD, across various SNRs from 20 dB to –20 dB. The proposed method consistently outperforms the other two methods across all tested SNRs, proving that the proposed speech detection method performs better than the other VAD and fund. While other methods show reasonable accuracy at high SNRs but quickly drop to around 50% at extremely noisy conditions, the proposed method achieves perfect accuracy down to 10 dB and maintains high performance even in extremely low-SNR environments, with 74.75% at –15 dB and 54.58% at –20 dB.

In addition to evaluating accuracy, the comparative analysis also examines the Equal Error Rate (EER) performance of the proposed method in relation to the other two techniques. It is important to note that the classifier component in the rVAD-fast method outputs a spectral flatness score, determining a segment to be classified as speech if the score is less than or equal to a threshold of 0.5. Conversely, the output from the WebRTC VAD system categorizes the input signals directly into speech or non-speech, which complicates the calculation of EER. Consequently, the FAR and FRR are computed, with EER being derived as the average of FAR and FRR.

EER comparison

Figure 4.1 (b) presents the EER comparison between the proposed method and two typical VAD approaches, rVAD-fast and WebRTC VAD, across SNR conditions from 20 dB to -20 dB. The proposed method consistently achieves the lowest EER at all tested SNRs, indicating superior reliability in distinguishing speech from non-speech under noisy conditions. While rVAD-fast and WebRTC VAD show relatively high EER values even at high SNRs (e.g., 9.67% and 24.83% at 20 dB, respectively) and rapidly degrade to around 50% in low-SNR environments, the proposed method maintains an EER close to 0% down to 5 dB and remains below 4% at -10 dB. Even under extremely adverse noise conditions of -15 dB and -20 dB, it achieves substantially

lower EERs (21.00% and 41.50%, respectively) compared to the two typical methods, further demonstrating its robustness and effectiveness in challenging acoustic environments.

4.4.3 Filterbanks comparison

A comparative evaluation of different types of filterbanks is also conducted in this study. Figure 4.2 shows the performance comparison of four filterbanks across various SNRs. The configurations used in this assessment are the optimal ones for each type of filterbank:

- STFT: Linear compression, ResNet with CBAM model, and global STM.
- Constant-band filterbank: Linear compression, ResNet with CBAM model, and global STM.
- Mel filterbank: Log compression, ResNet model without CBAM, and global STM.
- Gammatone filterbank: Linear compression, ResNet with CBAM model, and global STM.

At high SNRs (10 to 20dB), all configurations operate nearly ideally, achieving accuracies close to 100%. For SNRs above 0dB, the accuracies of all filterbanks remain above 90%, but a sharp degradation is observed as the SNR decreases further, dropping below 60% at SNR -20 dB. Among the tested filterbanks, the Gammatone filterbank consistently yields the highest accuracy across the entire SNR range, maintaining more than 90% accuracy at SNR -10 dB and over 70% even at SNR -15 dB. In contrast, the Mel filterbank exhibits the lowest performance across SNRs, with noticeably larger accuracy reductions under severe noise conditions.

Similar to the accuracy trends, all configurations achieve near-optimal results (EER $\approx 0\%$) at high SNRs (10 to 20 dB). When SNRs remain above

−5 dB, the EER for all filterbanks stays below 10%, but it increases sharply as the noise level becomes more severe, reaching approximately 40% at SNR −20 dB. The Gammatone filterbank again demonstrates the most robust performance, maintaining an EER of around 20% at SNR −15 dB and keeping the EER below 40% even at SNR −20 dB. The Mel filterbank remains the weakest performer, showing consistently higher EER values across SNRs.

4.4.4 Local vs Global STM comparison

Figure 4.3 compares the performance of local and global STM across various SNRs, using the same configuration of Gammatone filterbank, linear compression, and a ResNet model with CBAM integration. Across the entire SNR range, global STM consistently outperforms local STM, suggesting that reserving entire STM results in richer feature representations and stronger discriminative power. At SNRs above −5 dB, both methods perform well, with global STM achieving accuracy above 90% and EER below 10%. The performance drops when the SNR falls below −10 dB. Nevertheless, global STM retains an accuracy above 80% and an EER below 20% at SNR −10 dB, indicating greater robustness to noise compared to local STM. These results highlight the advantage of global STM configuration under low-SNR environments.

Figure 4.4 demonstrates the differences between the Gammatone filterbank global STM and local STM of noisy speech and white noise at different SNRs from 20 down to −20 dB. As described in the previous chapter, global STM is obtained by applying 2DFFT to the amplitude envelope of the input signal. Local STM, which is cropped from the resulting global STM to value range (−64, 64) Hz in temporal modulation and $(\frac{1}{800}, \frac{1}{60})$ cycles per Hz in spectral modulation, contains valuable information of speech and is expected to be powerful at distinguish speech and noise. However, the proposed method of global STM achieves better results than that of local STM. From SNR 20 dB down to 10 dB, the results are the same, but from SNR 10 dB down to −20 dB, the result of global STM is better than that of local

STM.

To better understand why global STM achieves better results than local STM, observing the outside area of local STM is necessary. This can be achieved by two approaches:

- In the global STM, the local area is preserved while the outside area of the local STM is removed. Then, the inverse 2DFFT is applied to obtain the time-frequency spectrogram. This approach provides a better understanding of the local area in the global STM if the outside area is removed.

- The local area is removed from the global STM, and the outside area is preserved. After that, the inverse 2DFFT is also applied to obtain the time-frequency spectrogram. This approach provides a better understanding of the outside-the-local-STM area in the global STM without the local area.

Figure 4.5 illustrates the global STM of speech with only the local area, and its corresponding spectrogram. In this figure, the local area in the global STM is still preserved, and the outside area is set to zero values. Its spectrogram, which is obtained by taking the inverse 2DFFT, shows formants and harmonics of speech. Meanwhile, figure 4.6 describes the global STM of speech without the local area, and its corresponding spectrogram. In this figure, the local area is set to zero values, while the remaining area remains unchanged. Its corresponding spectrogram reveals periodicity of speech.

In conclusion, while the local STM preserves important speech features, the outside area in the global STM still contains valuable information about human speech. This explains why the proposed method of global STM yields better results than local STM.

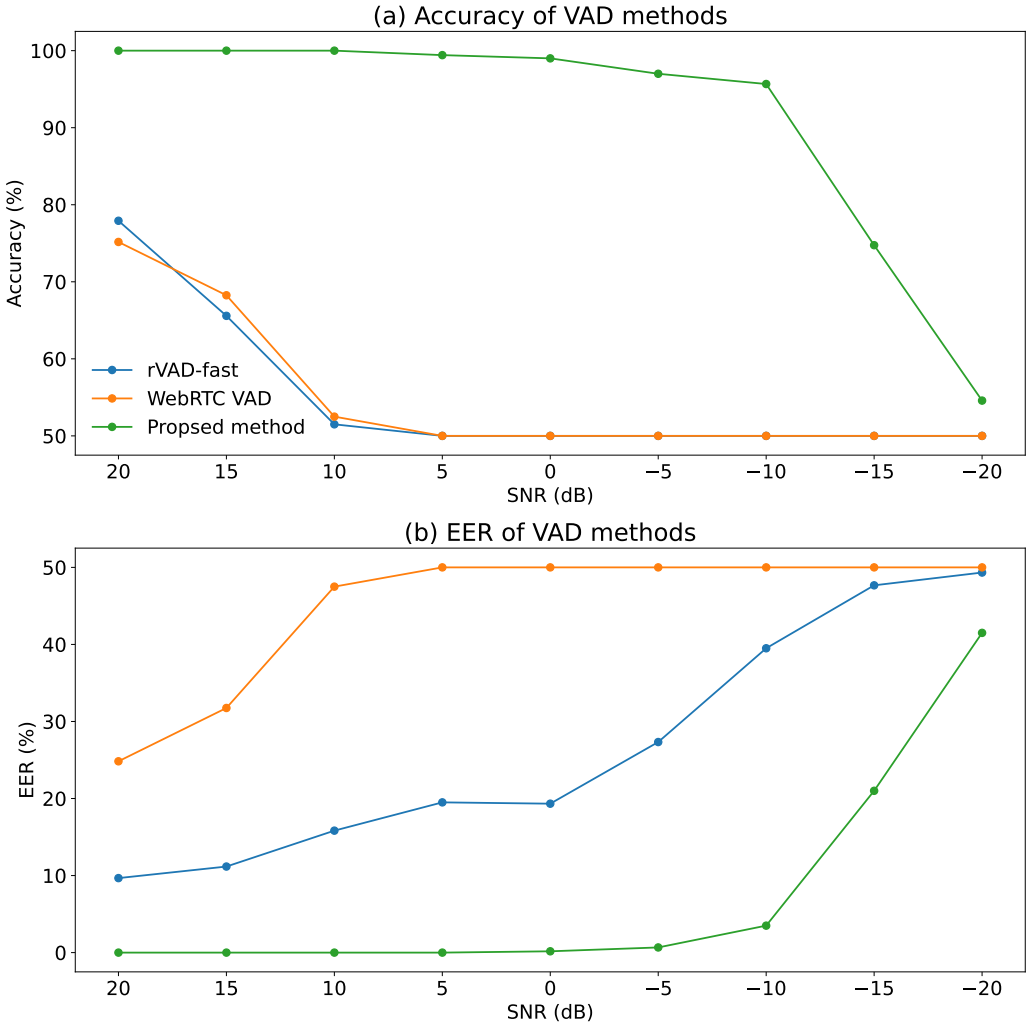

Figure 4.1: Accuracy and EER comparison between the proposed method and other methods at various SNRs: (a). Accuracy of VAD methods, and (b) EER of VAD methods.

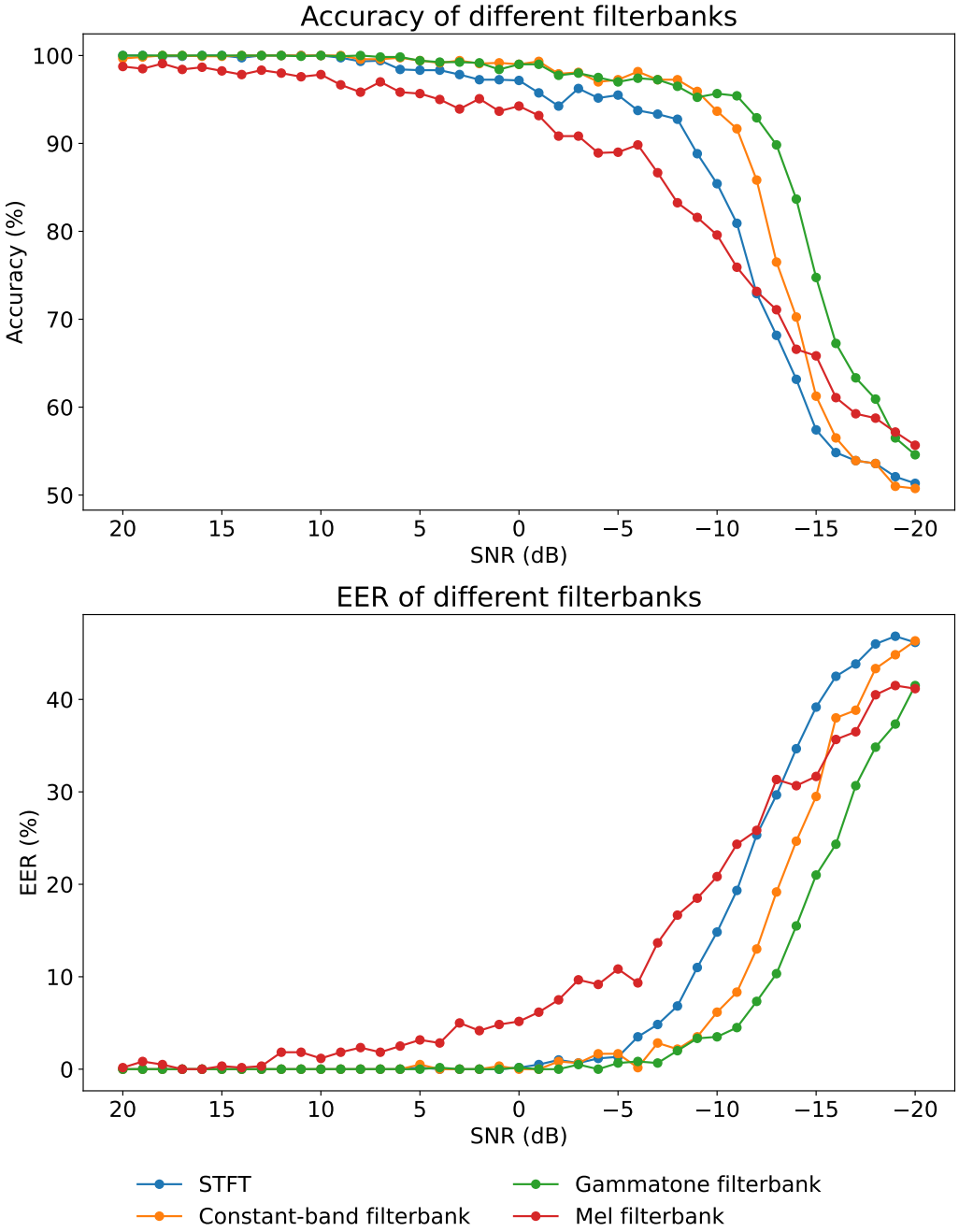

Figure 4.2: Performance comparison between local STM and global STM of the optimal configuration across different SNRs.

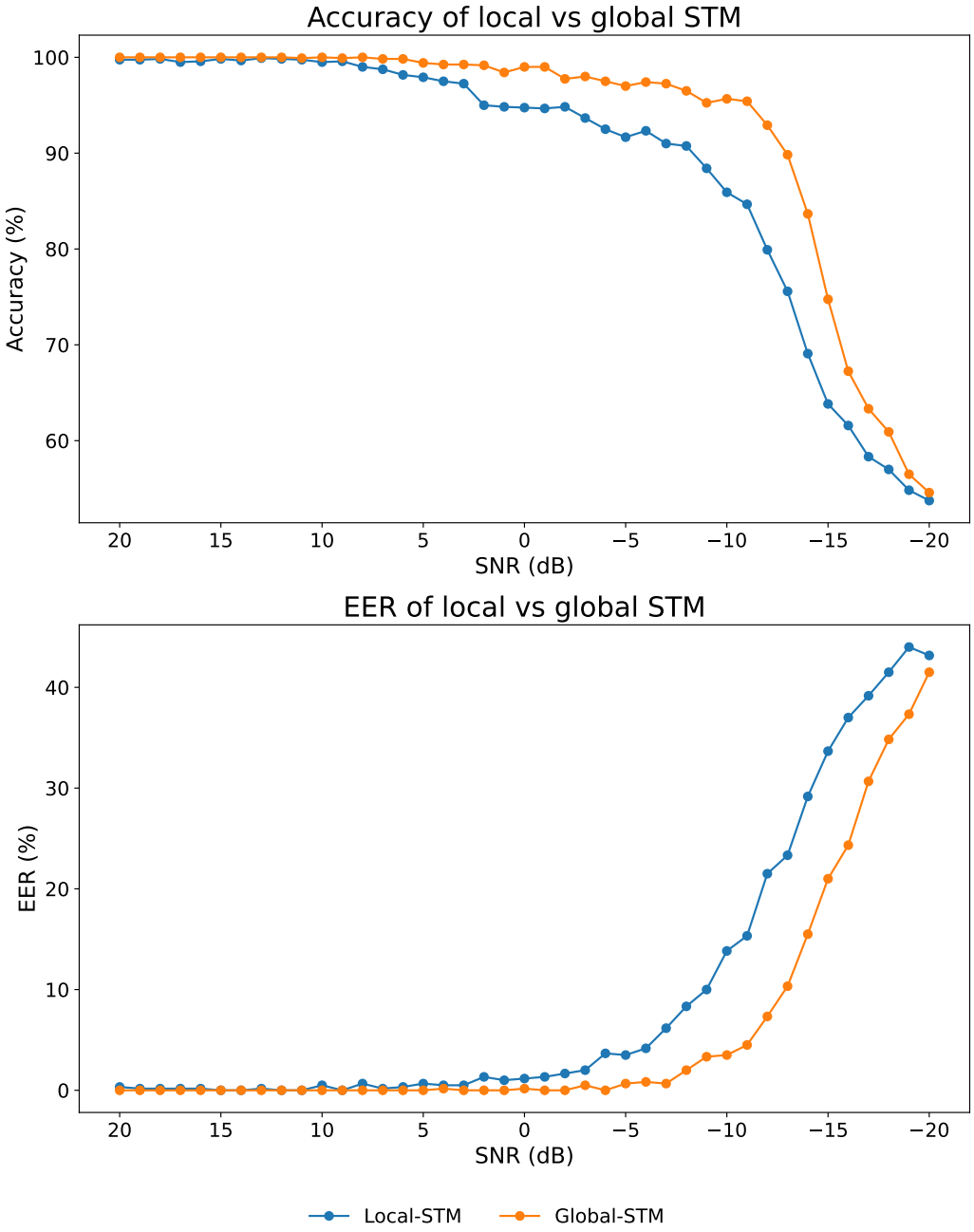

Figure 4.3: Performance comparison between local STM and global STM of the optimal configuration across different SNRs.

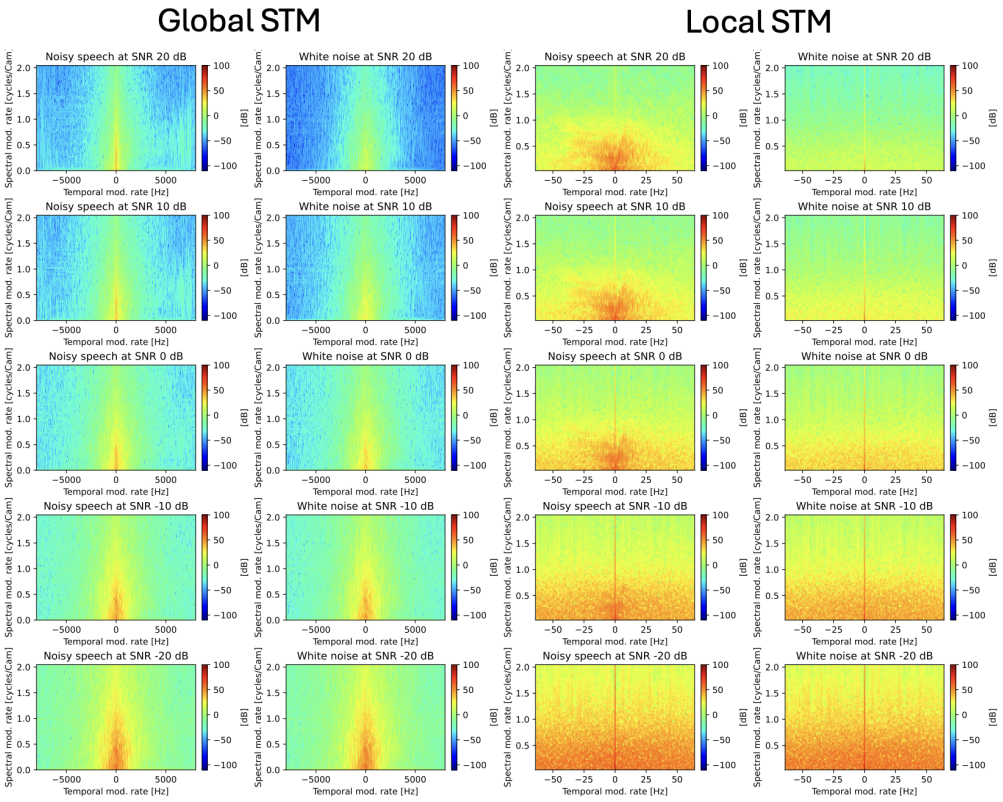

Figure 4.4: Gammatone Filterbank global STM and local STM of noisy speech and white noise at different SNRs.

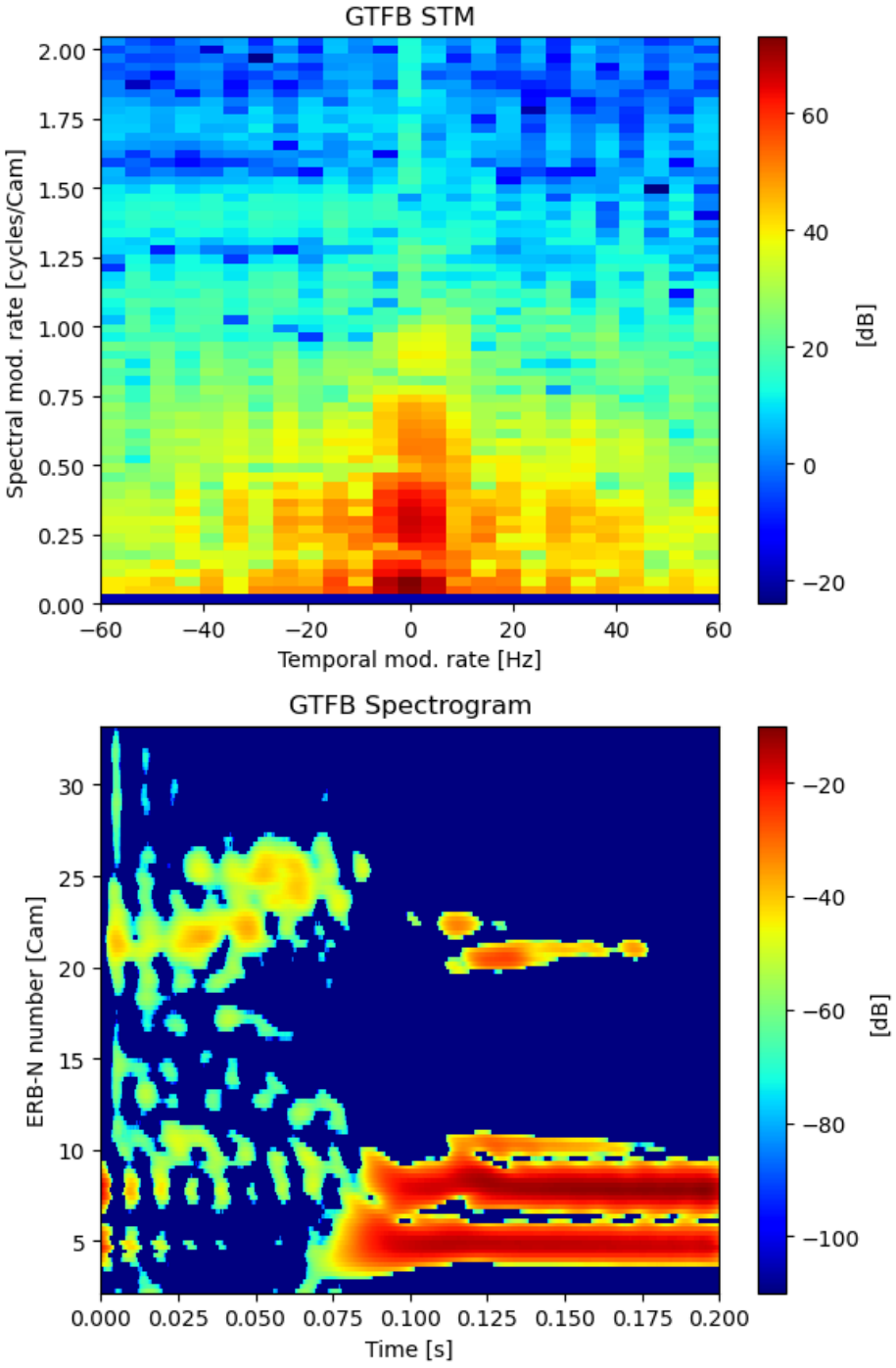

Figure 4.5: Global STM of speech with only the local area and the corresponding spectrogram.

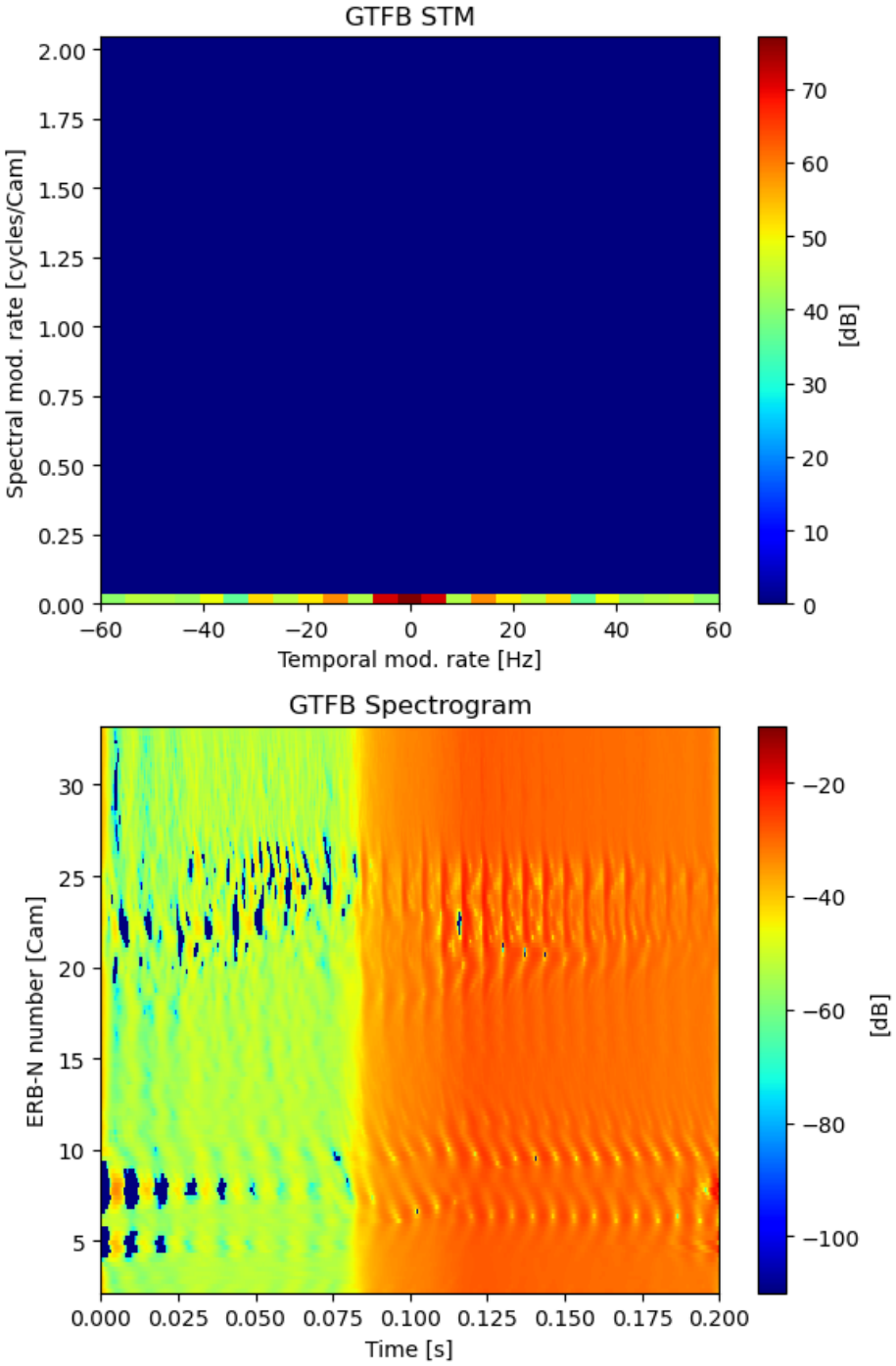

Figure 4.6: Global STM of speech without the local area and the corresponding spectrogram.

Chapter 5

Conclusion

This chapter summarizes the research's findings, highlights its key contributions, discusses the limitations encountered, and outlines potential directions for future research.

5.1 Summary

This dissertation addressed the critical challenge of robust speech detection under heavy noise conditions, particularly those relevant to disaster scenarios. The core of the proposed framework involved leveraging Spectro-Temporal Modulation (STM) features as a rich and discriminative representation of audio signals. STM features are derived by applying a 2D Fast Fourier Transform to spectrograms, which are decomposed using multiple filterbanks: STFT, Constant-band, Gammatone, and Mel filterbanks.

For speech/non-speech classification, a deep learning approach was adopted, employing the CNN as a baseline, then enhanced with ResNet18 architecture and augmented with a Convolutional Block Attention Module (CBAM). CNN and ResNet18 provided a robust backbone for learning hierarchical features from the STM representations. At the same time, the integrated CBAM enhanced the model's ability to focus on salient spectro-temporal regions and channels, which is crucial for distinguishing speech cues from overwhelming

noise.

Experiments were systematically conducted across a wide range of SNRs from −20 dB to 20 dB, utilizing the JVS corpus for clean speech and a controlled stationary noise environment (white noise). The performance was evaluated using accuracy and equal error rate (EER). The results demonstrated that the proposed method achieved 92.79% overall accuracy and showed nearly ideal performance at high SNR. The optimal configuration of the proposed method (Gammatone filterbank, linear compression, global STM with the ResNet integrated with CBAM model) surprisingly achieved an EER of less than 10% at an SNR of −10 dB and approximately 20% at −15 dB. Furthermore, a comparative analysis against established VAD methods, rVAD-fast and WebRTC VAD, highlighted the superior performance of the proposed method. While rVAD-fast and WebRTC VAD showed a sharp decline in accuracy and a rapid increase in EER as SNR decreased, the proposed method maintained a high degree of robustness, achieving perfect accuracy down to 10 dB and outperforming the other methods even in extremely challenging conditions (e.g., 74.75% accuracy and 21.00% EER at −15 dB). This comparison demonstrates the framework's strong discriminative power under difficult conditions. While performance degradation was observed at extremely low SNRs, particularly at −20 dB, this highlights the inherent difficulty of the task in such adverse environments. Furthermore, the experimental results also indicated that the Gammatone filterbank outperformed other types of filterbanks at various SNRs, showcasing that emphasizing low-frequency components makes it effective for distinguishing between speech and non-speech in noisy environments.

5.2 Contributions

The research presented in this dissertation makes several contributions to the field of noise-robust speech detection and deep learning-based audio analysis:

- Systematic evaluation of STM configurations: This study provides a

comprehensive comparative analysis of Spectro-Temporal Modulation (STM) features derived from various filterbank types (STFT, Constant-band, Gammatone, Mel), transformations (linear vs. logarithmic), and modulation range (local vs. global). The results offer valuable insights into the optimal STM configurations for speech/non-speech classification under diverse noise conditions, demonstrating the benefits of specific feature choices.

- Effective integration of attention mechanisms in ResNet: This innovative combination allows the model to robustly extract and focus on the most discriminative spectro-temporal patterns, showcasing the power of attention mechanisms in enhancing noise robustness for audio classification tasks.

- Performance characterization across various SNRs: The research systematically evaluates the proposed framework's performance across a wide range of SNRs (from 20 dB down to -20 dB) with white noise. This provides critical benchmarks for the system's capabilities and identifies the performance boundaries at extremely low SNRs, contributing to an understanding of current deep learning limitations in these severe environments.

- Development of a robust prototype for Japanese speech detection in noise: By utilizing the JVS corpus, this work contributes to the development of a robust speech detection prototype tailored for Japanese speech, which can enhance search and rescue operations by detecting individuals calling for help amidst, thereby improving response times and potentially saving lives.

5.3 Limitations

This study proposes a speech detection framework under heavy noise conditions, leveraging spectro-temporal modulation analysis and a ResNet18

architecture augmented with an attention mechanism, CBAM. While the proposed method demonstrates promising results, particularly at higher and moderate SNRs, it is important to mention certain limitations of the current study. Identifying these boundaries not only provides a realistic assessment of the system's current capabilities but also highlights crucial avenues for future research.

The first limitation of the current research lies in the restricted diversity of the noise environments used for the speech detection system development. Specifically, only stationary noise (white noise) is utilized. Although artificial, white noise can approximate certain real-world conditions with broadband and steady characteristics, such as rainfall and wind. This controlled setup allows the method to focus primarily on detecting speech features under stationary noise before extending to more complex and non-stationary environments.

Second, consistent with the difficulty of speech processing in extremely noisy conditions, the model exhibited a significant degradation in performance, as evidenced by high EER at very low SNRs, particularly at -20 dB. At such severe noise levels, the distinctive spectro-temporal cues of speech become obscured. While the attention mechanism of CBAM and the deep feature extraction of ResNet18 helped to mitigate this, the current architecture still faces substantial challenges in reliably distinguishing speech from dominant noise at these extreme SNRs.

Beyond these primary limitations, other technical aspects warrant consideration:

- Fixed sample duration: The consistent input duration of 0.2 seconds, while simplifying model architecture and training, might not capture long-range temporal dependencies or contextual information within an utterance.

- Computational efficiency for real-time deployment: While ResNet18 is relatively efficient among deep networks, the current model's computational requirements, especially combined with on-the-fly STM feature

extraction, may still pose challenges for real-time system deployment.

Acknowledging these limitations is important for providing current findings and for guiding future research directions aimed at developing even more robust and applicable speech detection systems.

5.4 Future prospects

Several promising directions for future work emerge to enhance further the robustness, generalizability, and practical applicability of the proposed framework.

First, it is essential to diversify the noise corpus beyond stationary sounds like white noise. Real-world environments, especially during disasters, feature complex non-stationary noises such as sirens, machines, and natural sounds. Future work will incorporate these varied environmental noises to help the model learn more robust and noise-invariant speech features.

Second, improving performance at extremely low SNR, particularly at -20 dB, remains a significant challenge and a key area for future investigation. Techniques such as speech enhancement, noise suppression, or other advanced deep learning models could be explored to further disentangle speech components from dominant noise at these critical SNRs.

Publications

[1] Tran Quynh Nhu, Nguyen Huy-Quoc, and Unoki Masashi. “A Comparative Study of Methods for Detecting Speech Signals under Heavily Noisy Conditions”. In: *2025 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing* (2025).

Bibliography

[1] Alhasan Hakami et al. “Application of soft systems methodology in solving disaster emergency logistics problems”. In: *International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering* 7.12 (2013), pp. 2470–2477.

[2] Kotaro Hoshiba et al. “Design of UAV-embedded microphone array system for sound source localization in outdoor environments”. In: *Sensors* 17.11 (2017), p. 2535.

[3] Shota Morita et al. “Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments”. In: *Journal of Signal Processing Systems* 82.2 (2016), pp. 163–173.

[4] Zheng-Hua Tan, Najim Dehak, et al. “rVAD: An unsupervised segment-based robust voice activity detection method”. In: *Computer speech & language* 59 (2020), pp. 1–21.

[5] Angelo J Soto-Vergel et al. “Transforming ground disaster response: Recent technological advances, challenges, and future trends for rapid and accurate real-world applications of survivor detection”. In: *International Journal of Disaster Risk Reduction* 98 (2023), p. 104094.

[6] Mishaim Malik et al. “Automatic speech recognition: a survey”. In: *Multimedia Tools and Applications* 80.6 (2021), pp. 9411–9457.

[7] Lawrence R Rabiner. “Applications of speech recognition in the area of telecommunications”. In: *1997 IEEE workshop on automatic speech recognition and understanding proceedings*. IEEE. 1997, pp. 501–510.

[8] Phillip L De Leon et al. “Evaluation of speaker verification security and detection of HMM-based synthetic speech”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.8 (2012), pp. 2280–2290.

[9] Andreas Depold et al. “A direction-of-arrival estimation system for UAV-assisted search and rescue: locating mobile phones to improve the survival chance of disaster victims”. In: *IEEE microwave magazine* 24.3 (2023), pp. 59–64.

[10] Ralph Schmidt. “Multiple emitter location and signal parameter estimation”. In: *IEEE transactions on antennas and propagation* 34.3 (1986), pp. 276–280.

[11] Richard Roy, A Paulraj, and Thomas Kailath. “Estimation of signal parameters via rotational invariance techniques-ESPRIT”. In: *MILCOM 1986-IEEE Military Communications Conference: Communications-Computers: Teamed for the 90's*. Vol. 3. IEEE. 1986, pp. 41–6.

[12] Lal C Godara. “Application of antenna arrays to mobile communications. II. Beam-forming and direction-of-arrival considerations”. In: *Proceedings of the IEEE* 85.8 (2002), pp. 1195–1245.

[13] Samuel D Somasundaram. “Linearly constrained robust Capon beamforming”. In: *IEEE Transactions on Signal Processing* 60.11 (2012), pp. 5845–5856.

[14] Yuhang He and Andrew Markham. “SoundDoA: Learn Sound Source Direction of Arrival and Semantics from Sound Raw Waveforms.” In: *Interspeech*. 2022, pp. 2408–2412.

[15] Zixuan Li, Shulin He, and Xueliang Zhang. “Robust Target Speaker Direction of Arrival Estimation”. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.

[16] Mayank Sharma et al. “A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows”. In: *Neurocomputing* 494 (2022), pp. 116–131.

[17] Rajesh Maharudra Patil and CM Patil. “Unveiling the state-of-the-art: A comprehensive survey on voice activity detection techniques”. In: *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. IEEE. 2024, pp. 1–5.

[18] Philippe Renevey and Andrzej Drygajlo. “Entropy based voice activity detection in very noisy conditions.” In: *Interspeech*. 2001, pp. 1887–1890.

[19] Hadi Veisi and Hossein Sameti. “Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement”. In: *IET signal processing* 6.1 (2012), pp. 54–63.

[20] Dongwen Ying et al. “Voice activity detection based on an unsupervised learning framework”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.8 (2011), pp. 2624–2633.

[21] BG Nagaraja and G Thimmaraja Yadava. “Enhancing Voice Activity Detection in Noisy Environments Using Deep Neural Networks”. In: *Circuits, Systems, and Signal Processing* (2025), pp. 1–15.

[22] Guan-Bo Wang and Wei-Qiang Zhang. “An rnn and crnn based approach to robust voice activity detection”. In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2019, pp. 1347–1350.

[23] Florian Eyben et al. “Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies”. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 483–487.

[24] WebRTC Project Authors. *Voice Activity Detector (VAD)*. WebRTC audio processing module, BSD-licensed, WebRTC Project (Google). 2015.

[25] Xu Tan and Xiao-Lei Zhang. “Speech enhancement aided end-to-end multi-task learning for voice activity detection”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 6823–6827.

[26] MR Prasad et al. “Integrated noise suppression techniques for enhancing voice activity detection in degraded environments”. In: *International Journal of Speech Technology* 27.4 (2024), pp. 987–995.

[27] W. M. Hartmann. *Signals, Sound, and Sensation*. Springer Science & Business Media, 2004.

[28] Homer Dudley. “The carrier nature of speech”. In: *Bell System Technical Journal* 19.4 (1940), pp. 495–515.

[29] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers”. In: *The Journal of the Acoustical Society of America* 102.5 (1997), pp. 2892–2905.

[30] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. “Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration”. In: *The Journal of the Acoustical Society of America* 102.5 (1997), pp. 2906–2919.

[31] Stephan D Ewert, Jesko L Verhey, and Torsten Dau. “Spectro-temporal processing in the envelope-frequency domain”. In: *The Journal of the Acoustical Society of America* 112.6 (2002), pp. 2921–2931.

[32] Haowei Cheng et al. “Analysis of spectro-temporal modulation representation for deep-fake speech detection”. In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2023, pp. 1822–1829.

[33] Taishih Chi, Powen Ru, and Shihab A Shamma. “Multiresolution spectrotemporal analysis of complex sounds”. In: *The Journal of the Acoustical Society of America* 118.2 (2005), pp. 887–906.

[34] Taffeta M Elliott and Frédéric E Theunissen. “The modulation transfer function for speech intelligibility”. In: *PLoS computational biology* 5.3 (2009), e1000302.

[35] Taishih Chi et al. “Spectro-temporal modulation transfer functions and speech intelligibility”. In: *The Journal of the Acoustical Society of America* 106.5 (1999), pp. 2719–2732.

[36] Zhi Zhu et al. “Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech”. In: *Acoustical Science and Technology* 39.6 (2018), pp. 379–386.

[37] L. Atlas, S. Greenberg, and H. Hermansky. “The Modulation Spectrum and Its Application to Speech Science and Technology”. In: *Interspeech2007, Tutorial*. 2007.

[38] Ingo Titze, Tobias Riede, and Ted Mau. “Predicting achievable fundamental frequency ranges in vocalization across species”. In: *PLoS computational biology* 12.6 (2016), e1004907.

[39] Krzysztof Tyburek. “Analysis of the Fundamental Frequency F0 of Oesophageal Speech in Patients Following Total Laryngectomy Surgery”. In: *Applied Sciences* 15.8 (2025), p. 4402.

[40] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. “A scale for the measurement of the psychological magnitude pitch”. In: *The journal of the acoustical society of america* 8.3 (1937), pp. 185–190.

[41] Roy D Patterson et al. “An efficient auditory filterbank based on the gammatone function”. In: *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*. Vol. 2. 7. 1987.

[42] Malcolm Slaney et al. “An efficient implementation of the Patterson-Holdsworth auditory filter bank”. In: *Apple Computer, Perception Group, Tech. Rep* 35.8 (1993).

[43] Zulfiqar Ali and Muhammad Talha. “Innovative method for unsupervised voice activity detection and classification of audio segments”. In: *Ieee Access* 6 (2018), pp. 15494–15504.

[44] Zixing Zhang et al. “Deep learning for environmentally robust speech recognition: An overview of recent developments”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.5 (2018), pp. 1–28.

[45] Khalid Zaman et al. “A survey of audio classification using deep learning”. In: *IEEE access* 11 (2023), pp. 106620–106649.

[46] Shawn Hershey et al. “CNN architectures for large-scale audio classification”. In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.

[47] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. “Rethinking CNN models for audio classification”. In: *arXiv preprint arXiv:2007.11154* (2020).

[48] Michele Scarpiniti et al. “Deep recurrent neural networks for audio classification in construction sites”. In: *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 810–814.

[49] Kalyanaswamy Banuroopa and D Shanmuga Priyaa. “MFCC based hybrid fingerprinting method for audio classification through LSTM”. In: *International Journal of Nonlinear Analysis and Applications* 12.Special Issue (2021), pp. 2125–2136.

[50] Yixiao Zhang et al. “Spectrogram transformers for audio classification”. In: *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE. 2022, pp. 1–6.

[51] Michele Scarpiniti, Raffaele Parisi, and Yong-Cheol Lee. “A Scalogram-based CNN approach for audio classification in construction sites”. In: *Applied Sciences* 14.1 (2023), p. 90.

[52] Keunwoo Choi et al. “Convolutional recurrent neural networks for music classification”. In: *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 2392–2396.

[53] Fatih Demir, Daban Abdulsalam Abdullah, and Abdulkadir Sengur. “A new deep CNN model for environmental sound classification”. In: *IEEE access* 8 (2020), pp. 66529–66537.

[54] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[55] Farheen Ramzan et al. “A deep learning approach for automated diagnosis and multi-class classification of Alzheimer’s disease stages using resting-state fMRI and residual neural networks”. In: *Journal of medical systems* 44.2 (2020), p. 37.

[56] Chao Yang et al. “ResNet based on multi-feature attention mechanism for sound classification in noisy environments”. In: *Sustainability* 15.14 (2023), p. 10762.

[57] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[58] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[59] Shinnosuke Takamichi et al. “JVS corpus: free Japanese multi-speaker voice corpus”. In: *arXiv preprint arXiv:1908.06248* (2019).

[60] R Plomp. “The role of modulation in hearing”. In: *HEARING—Physiological Bases and Psychophysics: Proceedings of the 6th International Symposium on Hearing, Bad Nauheim, Germany, April 5–9, 1983*. Springer. 1983, pp. 270–276.