Doctoral Dissertation

Advanced Feature Representation of Audio Signal Based on Spectral Temporal Modulation Analysis and Its Applications

Kai LI

Supervisor: Masashi UNOKI

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

March, 2024

Abstract

Audio, including speech, music, machine sounds, etc., surrounds us every day. In recent years, a lot of advanced pattern recognition technologies have been proposed by using different kinds of audio and labels, greatly facilitating our human lives. For example, human and machine-synthesized speech is collected and used to perform automatic speaker verification (ASV) and fake audio detection (FAD) systems, which can be applied in authentication and access control, justice, healthcare, speech security, etc. Furthermore, the audios of different kinds of machines from the factory are collected and utilized to construct machine anomalous sound detection (ASD) systems, which can provide continuous monitoring of machine status and optimize production efficiency. For these techniques, extracting acoustic features that can capture distinguishing information serves as the foundation for achieving accurate performance.

Many acoustic features have been proposed and applied in different audio-related tasks, such as the linear prediction coefficient (LPC), the Mel-frequency cepstral coefficient (MFCC), and the constant Q cepstral coefficient (CQCC). These features are designed for general information extraction and can be used for many different tasks. For example, the MFCC and LPC are widely used as acoustic features for both automatic speech recognition (ASR) and ASV. However, there are different kinds of information, such as linguistics, individuality, and emotion, encoded in the audio signal, and different tasks need to have task-specific information (TSI) to separate out different patterns. Traditional acoustic features, such as MFCC and LPC, have problems with weak task-specific discrimination, resulting in extracted features containing a lot of redundant information or important information being filtered or smoothed out.

Spectral temporal modulation (STM) analysis in the auditory system deals with both spectral and temporal modulation of audio to perceive auditory attributes related to audio production, such as the timbral and prosody. This property enables the human ear to easily perceive and discriminate a wide range of TSI in various acoustic scenarios. Inspired by this, this study aims to investigate advanced feature representations based on STM analysis to extract more TSI for audio detection and verification tasks.

To achieve this objective, both frequency and time domain analyses are conducted to explore the importance of spectral and temporal attributes in the representation of TSI. Then, STM representations derived based on the frequency and time analysis results are proposed for extracting more TSI. The frequency and time domain analyses can help verify the effectiveness of spectral/temporal modulation and direct better designing for the auditory models. The effectiveness of proposed feature representations is verified in ASV, FAD, and machine ASD, which cover speech, machine-synthesized speech, and non-speech. The complexity

of the human auditory mechanism makes it challenging to fully understand the audio signal processing process and determine which auditory model best mimics this process.

In frequency domain analysis, this research first investigates the importance of different frequency regions for ASV, FAD, and machine ASD tasks. Specifically, we proposed two data-driven-based methods, including the F-ratio and a frequency-wise attention structure combined with a ResNet, to quantify the non-linear combined effect of frequency components. We then designed a non-uniform subband processing strategy based on the quantification results for task-specific feature extraction. The quantification results from these three applications consistently indicated that TSI distributed in the frequency domain non-uniformly, which is quite different from traditional auditory scales, such as the equivalent rectangular bandwidth. The designed non-uniform filterbanks can capture more TSI and further improve the performance of these three applications.

In time domain analysis, this study investigates the timbre and prosody information differences in the voice represented using the jitter and shimmer features. In accordance with the statistical analysis results, the most promising prosody features were selected and incorporated with a machine-learning-based FAD system. In addition, two additional F_0 estimation methods, namely YIN and IRAPT, were utilized in place of the IRAPT algorithm when extracting features. Different weights were tested to determine the optimal combination between the Mel-spectrogram and shimmer features. The experimental results indicate that both the static and dynamic shimmer features of voice can provide complementary knowledge to the traditional spectrum-based systems in the FAD task.

The STM representations simulate the complex auditory perception process in the human auditory system, capturing both spectral and temporal modulations in speech signals. The effectiveness of STM has been evaluated in speech intelligibility prediction and speech emotion recognition. Usually, the Gammatone filterbank is utilized in the extraction processes of STM representations. According to the results of frequency domain analysis, the non-uniform filterbanks are more effective in extracting TSI. Therefore, this study simulated the human auditory mechanism using the STM derived from well-designed non-uniform filterbanks and evaluated in SV, FAD, and machine ASD tasks. The experimental results indicate that the STM representations can achieve competitive performance in the FAD and machine ASD tasks, which covering synthesized speech and non-speech signals. However, in the ASV task, the current results are inconsistent with our expectations. One possible reason may be the mismatch between the proposed representations and the i-vector approach used for ASV, which are designed for short-term feature extraction. More experiments regarding to the deep-learning architectures will be conducted in the future.

Keywords: Spectral temporal modulation, Human auditory perception, Fake audio detection, Automatic speaker verification, Machine anomalous sounds detection.

Acknowledgment

I want to thank several individuals who have been instrumental in my academic journey. This hard-earned doctorate would not have been possible without their support, encouragement, and guidance.

First and foremost, I extend my most profound appreciation to Professor Unoki Masashi for his constructive suggestions, encouragement, and the funding provided during my time in the laboratory. Your guidance has dramatically improved my logical thinking ability in recent years.

I also thank Professor Akagi Masato, whose expertise, patience, and rigorous attitude towards scientific research have shaped my research and enriched my understanding of our field. Your guidance has been invaluable, and I am truly fortunate to have had the opportunity to work under your supervision.

To Professor Xugang Lu, you are a teacher and a good friend to me. You always thought about the problems I encountered in my research very seriously and gave me patient guidance. I am grateful for your contributions to my academic development.

I really appreciate my family for their unwavering support, understanding, and love throughout this journey. Your encouragement sustained me during the most challenging moments, and I am grateful for your belief in my abilities.

My heartfelt thanks go to my colleagues and friends who have shared this academic journey with me. We discuss academics, imagine the future, and encourage each other. Your presence has made this experience not only educational but also enjoyable.

Lastly, I give appreciation to myself. The path to this doctorate has been filled with pain and stress from physiology and psychology. Finally, with my strong will and optimistic attitude, I realized my life goal at this stage. I am proud of the effort I have invested in reaching this milestone.

In closing, I sincerely appreciate the collective efforts and contributions that have propelled me towards this academic milestone. I look forward to continuing to apply the knowledge and skills I have gained to impact our field and beyond positively.

Contents

List of Figures

List of Tables

List of Symbols/Abbreviations

ADD Audio deepfake detection

AFP Amplitude and frequency perturbation

ASD Anomalous sound detection

ASR Automatic speech recognition

ASV Automatic speaker verification

Bi-LSTM Bi-directional long short-term memory

CQCCs Constant Q Cepstral Coefficients

DCT Discrete cosine transform

DFT Discrete Fourier transform

DNN Deep neural network

EER Equal Error Rate

ERB Equivalent rectangular bandwidth

FAD Fake audio detection

FFT Fast Fourier transform

GFB Gammatone filterbanks

GFCCs Gammatone filterbank cepstral coefficients

IC Inferior colliculus

IHCs Inner hair cells

LCNN Light convolution neural network

LPCs Linear prediction coefficients

LSTM Long short-term memory

MFCCs Mel-frequency cepstral coefficients

MSE Mean Squared Error

UNFB Non-uniform filterbanks

UNFCC Non-uniform filterbank cepstral coefficient

PLDA Probabilistic linear discriminant analysis

SER Speech emotion recognition

STFT Short-time Fourier transform

STM Spectral temporal modulation

TDNN Time delay neural network

TSI Task-specific information

UBM Universal background model

Chapter 1

Introduction

1.1 Research Background and Problems

The world is filled with diverse sounds, ranging from human speech and music to the various sounds produced by machines and others. These sounds play a significant role in our everyday experiences and contribute to the richness of our auditory environment. With the continuous progress of technology, these different types of audio are collected, labeled, classified, and applied to different application scenarios, significantly improving and simplifying our daily lives. For example, human and machine-synthesized speech is collected and used to perform automatic speaker verification (ASV) and fake audio detection (FAD) systems, which can be applied in authentication and access control, justice, healthcare, speech security, etc. Furthermore, the audios of different kinds of machines from the factory are collected and utilized to construct machine anomalous sound detection (ASD) systems, which can provide continuous monitoring of machine status and optimize production efficiency.

In general, as depicted in Fig. 1.1, the first step of different tasks is the extraction of acoustic features from the raw waveform. The extracted feature may then be used to calculate feature representations by using advanced transformation. Subsequently, machine learning or deep neural network (DNN) techniques are applied to map these features or representations into a latent space where different patterns become more distinguishable and separable. In this process, the front-end features/representations that can provide as much as possible discriminative information are crucial for the final performance of a specific task.

Many acoustic features have been proposed and applied. According to Tomi Kinnunen and Haizhou Li [4], these features can be categorized as short-term spectral features, voice source features, spectro-temporal features, prosodic features, and high-level features. Mel-frequency cepstral coefficients (MFCCs) is one of the

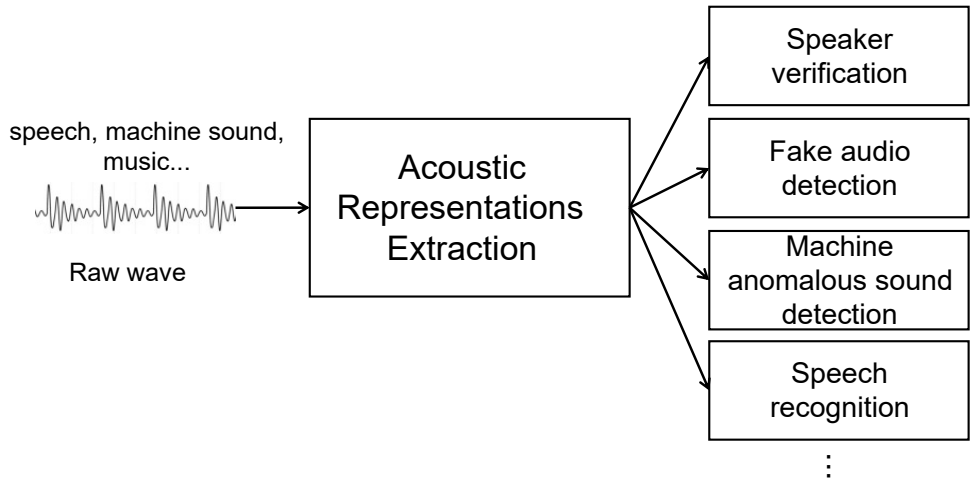

Figure 1.1: Some possible applications of acoustic features and their representations extracted from the raw wave.

most commonly used features in ASV, FAD, and machine ASD tasks [5]. MFCCs represent the short-term power spectrum of a sound signal and capture information about the spectral characteristics of speech. Another commonly used feature is the linear predictive coefficients (LPCs) [6]. It is based on the source-filter model of speech production, which assumes that speech is generated by filtering an excitation source (e.g., the vocal cord vibration) through the vocal tract filter. The CQCCs feature, which is derived from the constant-Q transform (CQT) and designed to capture the spectral characteristics of audio signals in a way that approximates human auditory perception, has been proved to be robust in the FAD task [7, 8]. More recently, researchers have preferred to input the Mel filterbank feature into a sophisticated DNN model to learn deep representations for ASV, FAD, and machine ASD tasks. More detailed reviews about feature extraction can be found in Chapter 2.2.

In summary, advanced feature extraction is crucial in many application fields, including human-computer interaction, speech security, industrial automation, and others. All of these technologies contribute to enabling more natural and intuitive interactions between humans and computers, improving the overall human-computer interaction experience, improving the overall security and stability of society and the economy, and accelerating the process of industrial development automation.

As we know, speech and machine sounds contain many different kinds of information. Speech, one of the most private forms of communication, contains a lot of personal information, such as the speaker's identity, age, gender, health, and emotional state. Machine sounds convey information about the types and condi-

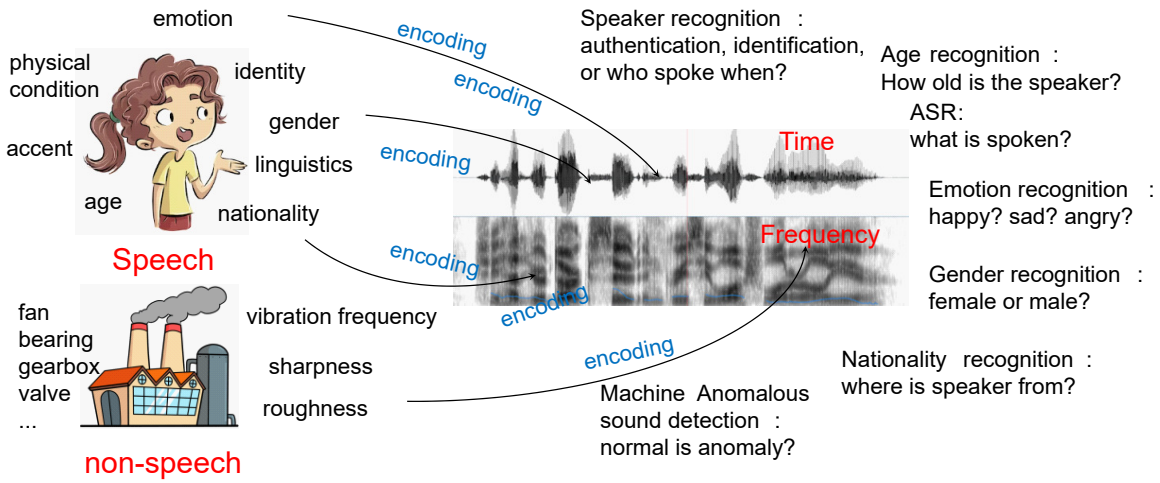

Figure 1.2: The explanation of task-specific information.

tions of a machine. As shown in Fig. 1.2, these different kinds of information are encoded in both frequency and time domain. Different discriminating information is needed when we perform different pattern recognition tasks. For example, a feature that provides enough similarities between the sentences from the same speaker while maintaining enough differences among different speakers is needed for the ASR task. A feature that can show a lot of differences between normal and anomalous sounds, while it is not interfered with by different machine types, is the need for machine ASD tasks. The task of FAD requires using acoustic features capable of distinguishing between genuine and fake speech while preventing potential interference from linguistic content, speech synthesis techniques, individual speakers, and other factors.

Based on literature reviews, most features are designed for general information extraction and can be used for many different tasks. For example, the MFCC and LPC features are widely used as acoustic features for automatic speech recognition (ASR) and ASV. However, as we explained above, there are different kinds of information, such as linguistics, individuality, and emotion, encoded in the audio signal, and different tasks need to have task-specific information (TSI) to separate different patterns. Traditional acoustic features, such as MFCCs and LPCs, have problems with weak task-specific discrimination, resulting in extracted features containing a lot of redundant information or important information being filtered or smoothed out. Feature representations that can capture more TSI are urgently needed.

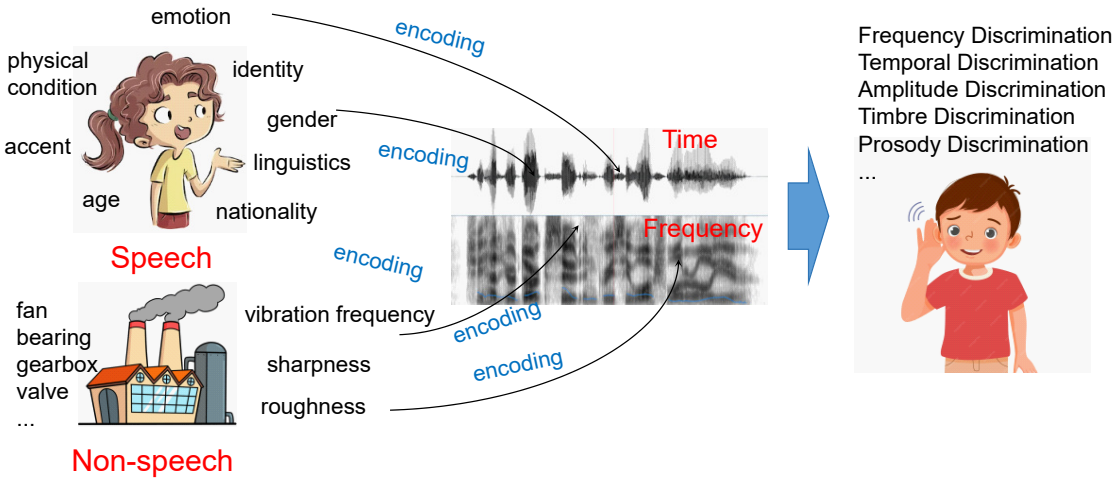

Figure 1.3: The statement of research motivation.

1.2 Research Motivation

Over billions of years of evolution, the human ear can efficiently extract TSI in different acoustic scenarios. Based on the literature, in the auditory system, sound signals are first analyzed by the cochlea and then transmitted to the auditory cortex for TSI perception. The cochlea is an important component of the auditory system. It can break down sound signals into multi-channel audio components along the basilar membrane. Inner hair cells (IHCs) detect movement of the basilar membrane and convert it into neuronal signals. From each transmitted signal, temporal amplitude envelope information is obtained. The temporal amplitude envelope information is propagated through the auditory nerve and cochlear nucleus to the inferior colliculus (IC) of the midbrain. There are a lot of studies have been conducted to reveal the complex mechanism of the human auditory system.

In 1971, Møller [9] found the specialized sensitivity of the mammalian auditory system to the amplitude modulation of narrowband signals. Subsequently, Suga [10] demonstrated the preservation of amplitude modulation information across various cochlear frequency channels in bats. Schreiner and Urbas [11, 12] further confirmed this neural representation of amplitude modulation, extending its presence to higher levels of mammalian auditory processing, including the auditory cortex. As a result, this sensitivity to amplitude modulation appears to be conserved across all levels of our auditory system. Also, some other researchers [13] not only confirmed the observability of these effects but also suggested that temporal envelope variations and amplitude modulation might be fundamental to the encoding mechanisms employed by mammalian auditory systems.

Recent physiological evidence has strengthened our understanding of these phe-

nomena. Kowalski et al. [14, 15] have conducted studies demonstrating that cells in the auditory cortex, which represents the highest processing stage in the primary auditory pathway, exhibit the greatest responsiveness to sounds that incorporate both spectral and temporal modulations. They used specially designed stimuli known as "ripples," characterized by dynamic broadband spectra that undergo amplitude modulation with drifting sinusoidal envelopes at varying speeds and spectral peak densities. Through the manipulation of ripple parameters and their correlation with neural responses, they were able to estimate the STM transfer functions of cortical cells, equivalent to their spectro-temporal receptive fields or impulse responses. Based on this data, they proposed that the auditory system effectively conducts a multi-scale spectro-temporal analysis, reencoding the acoustic spectrum in terms of its spectral and temporal modulations. The perceptual significance of these findings and formulations has been explored through psychoacoustic studies and applied in assessing speech intelligibility and communication channel fidelity.

There is also a lot of psycho-acoustic evidence to show the perceptual significance of signal modulations. For instance, Viemeister [16] conducted a comprehensive investigation into human perception of amplitude-modulated tones, highlighting it as a distinct window into auditory perception analysis. Houtgast [17] extended this understanding by demonstrating that the perception of amplitude modulation at one frequency can mask the perception of other nearby modulation frequencies. Further support for this notion comes from experiments conducted by Bacon and Grantham [18]. They concluded that, just as there are channels (known as critical bands or auditory filters) finely tuned for detecting spectral frequency, there are also specialized channels attuned to the detection of modulation frequency.

Recent psycho-acoustic studies have contributed further insights into human perception of modulation frequency. For instance, Ewert and Dau [19, 20, 21] have revealed dependencies between modulation frequency masking and carrier bandwidth. Chi et al. [20, 21] conducted experiments to measure human sensitivity to ripples with varying temporal modulation rates and spectral densities. An interesting discovery from these experiments is the striking alignment between the most sensitive range of modulations and the STM characteristics found in speech. This observation means that the preservation of the modulations inherent to speech could serve as an indicator of its intelligibility.

Various computational auditory models have been proposed to simulate the process of audio processing in our auditory system. One critical element of these models is the auditory filterbank. The auditory filterbank imitates the decomposition of time signals into time-frequency representations, much like the cochlear basilar membrane does in the human auditory system. Furthermore, the calcu-

lation of temporal envelopes from the output of different auditory filterbank can effectively simulate the transformation of mechanical signals into neural signals within the IHCs. Additionally, modulation filterbanks are integrated into these models to provide high-resolution temporal-modulation cues. These models closely replicate how the auditory system processes and interprets audio information. Recent studies have proved that STM contains loudness, timbre, and prosody information and can be widely applied in speech emotion recognition (SER) [22, 23], ASV [24], and ASR [25]. However, the complexity of the human auditory mechanism still makes it challenging to fully understand the process of audio signal processing and to determine which auditory model best mimics this process.

1.3 Research Goals

Inspired by the human auditory mechanism, this study aims to investigate advanced feature representations based on STM analysis to extract more TSI for audio detection and verification tasks. In STM analysis, spectral modulation refers to variations in the spectral content of the signal over time, capturing information related to formants and spectral features. Temporal modulation is associated with the temporal characteristics of sound, such as timbre or prosody-related attributes. To reach this objective, frequency and time domain analyses are first conducted to explore the importance of spectral and temporal attributes in the representation of TSI. For the frequency analysis, the importance of different frequency regions for ASV, FAD, and machine ASD tasks is investigated. This subgoal aims to find out which frequency regions are more important for TSI extraction. For the time analysis, we investigate the prosody and timbre attributes, aiming to propose advanced features that can capture TSI information in the time domain. Then, STM representations derived based on the frequency and time analysis results are proposed for extracting more TSI. Finally, all proposed feature representations are considered to solve the real-world problem, such as the ASV, FAD, and machine ASD.

1.4 Challenges

There are several challenges in the modeling, extraction, and application processes of feature representations under the inspiration of human auditory mechanisms.

(1) Modeling and extracting of feature representations

The human auditory system is very complex. Current research has not been able to clearly elucidate the mechanism of auditory signal processing. Therefore, mimicking these intricate processes in computational models can be

difficult. For example, in mimicking the frequency selectivity of the cochlea basilar membrane, many psychoacoustical experiments have been conducted to find a hearing scale that best matches the human ear. However, the distinguishing information of different tasks is not concentrated in the low-frequency region only. Focusing too much on the lower frequencies may smooth/filter out some important information. A method that can clarify/quantify where discriminative information is encoded in the frequency should be further considered. Another challenge is controlling for relatively low computational complexity and feature dimensions. This directly affects the popularity of the proposed features, which is just as important as the effectiveness of the proposed features.

(2) Application of extracted feature representations in different tasks

Recently, many advanced models have been proposed and applied in ASV, FAD, and machine ASD tasks. Most of these architectures are designed for short-term feature learning. However, according to the literature review, most of these computational models are based on a larger temporal context, resulting in fewer training vectors compared with short-term features. This problem makes the combination of proposed feature representations with commonly used deep-learning architectures challenging. In addition, the STM analysis-based feature representations have different frame rates from conventional features, making combining different features challenging at the frame level.

1.5 Organization of Thesis

As shown in Fig. 1.4, this thesis consists of seven chapters. Apart from the introduction chapter, the remaining chapters are organized as follows.

Chapter 2 presents a comprehensive literature review related to this study. It begins by introducing the human auditory system and various simulation methods employed in this context. Subsequently, it delves into the discussion of commonly used acoustic features, encompassing short-term spectral features, prosodic features, timbral features, and spectral-temporal modulation features. Lastly, Chapter 2 reviews the definitions and machine learning-based methodologies for ASV, FAD, and machine ASD tasks.

Chapter 3 describes the central concept and framework for extracting advanced feature representation. The overview of the proposed methods is introduced in the first section. Frequency and time domain feature analysis and extraction methods are introduced in the second and third sections, respectively.

Last, the spectral temporal modulation analysis method proposed in this thesis is introduced.

Chapter 4 introduces the application of the proposed method in the ASV task. This chapter contains the application of proposed features in an I-vector-based ASV system, the analysis results of frequency importance, experimental data and matrix, model setting, results and discussion, and summary.

Chapter 5 introduces the application of the proposed method in the FAD task. This chapter contains the application of spectral-temporal representations in an LCNN-based FAD system, experimental data, evaluation matrix, experimental settings, the analysis of differences between genuine and fake speech using temporal features, results discussion, and summary.

Chapter 6 introduces the application of the proposed method in the machine ASD task. This chapter contains the application of spectral-temporal representations in an autoencoder-based machine ASD system, experimental data, matrixes, model settings, the statistic analysis of frequency importance for different machines, results discussion, and summary.

Chapter 7 contains a summary, contributions, and the future works of this study.

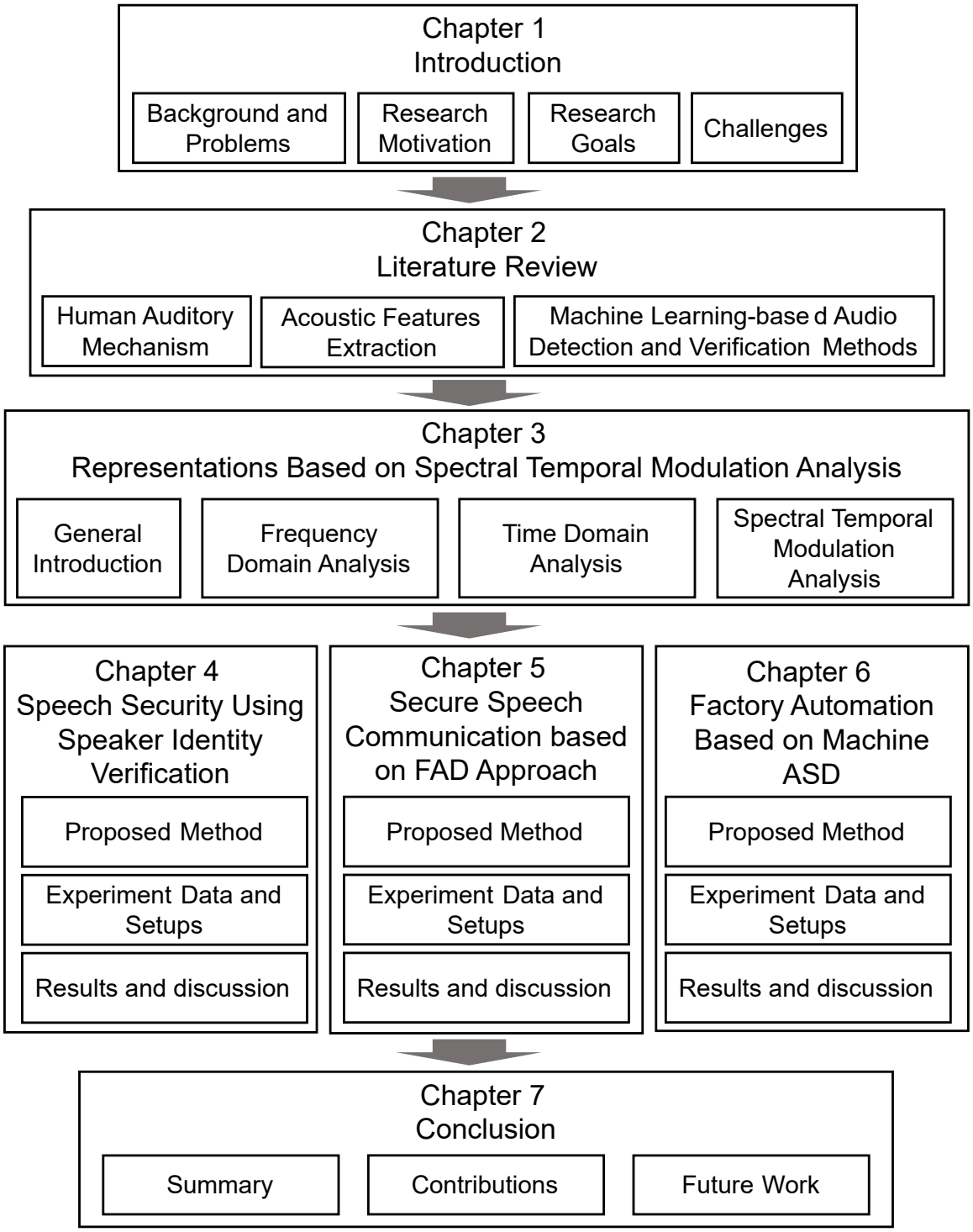

Figure 1.4: Organization of this dissertation

Chapter 2

Literature Review

2.1 Human Auditory Mechanism

The human auditory system is the complex biological system responsible for perceiving and processing sound. It includes several key components that enable humans to hear and understand auditory information. In this chapter, we provide a brief overview of the anatomy of the human auditory system and the process by which sound signals are conveyed to the brain. Then, computational auditory models that are popularly used in the simulation of the human auditory system are introduced.

2.1.1 Brief Introduction of Human Auditory System

Figure 2.1 illustrates the peripheral anatomy of the human auditory system. Sound waves initially enter the auditory system through the pinna, which serves as the external part of the ear. This structurally unique and intricate component plays a pivotal role by spectrally modifying incoming sound based on its angle of incidence. It is worth noting that beyond the pinna's contribution, other aspects of the human head and torso also contribute to the intricate process of sound modification, collectively enhancing our remarkable ability to pinpoint the source of a given sound.

As sound progresses from the pinna, it proceeds through the ear canal, a narrow tube-like structure that extends approximately 2.5 centimeters in length. The ear canal is not merely a passageway; it adds its own complexity to the auditory process. This is because the ear canal, similar to a tuning fork, possesses resonant properties, causing it to act as a highly effective band-pass filter. In this role, it plays an instrumental part in influencing the frequencies and characteristics of the sound waves as they make their way towards the eardrum, setting the stage for further processing within the auditory system.

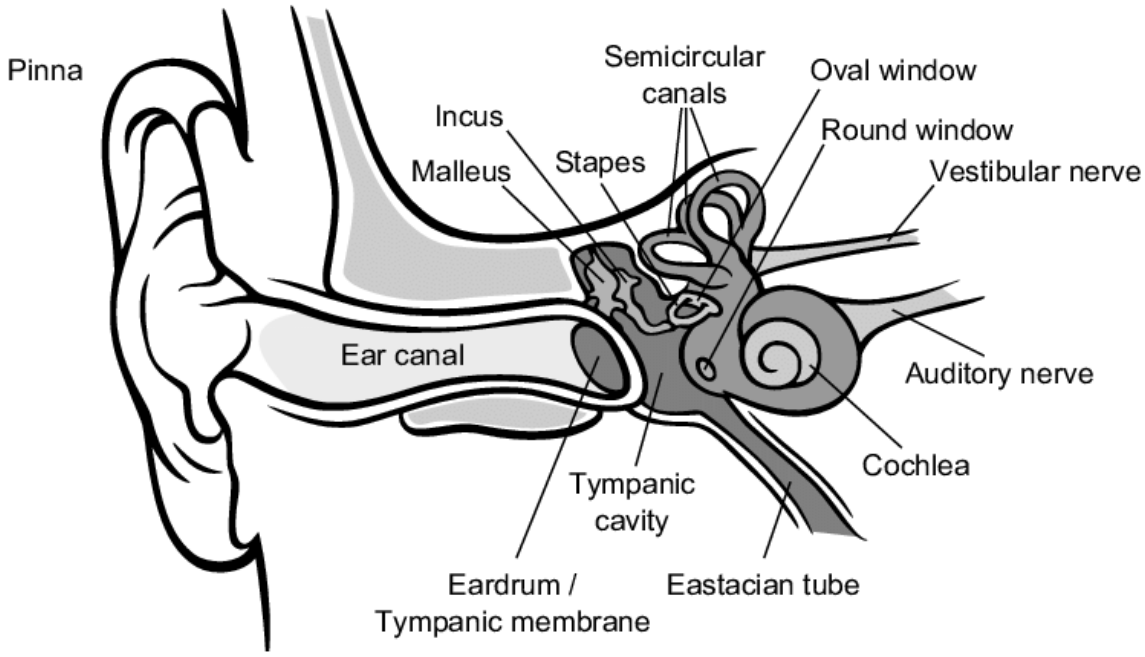

Figure 2.1: A simplified representation of the human auditory system (original graphics from [1])

The eardrum, a delicate membrane, responds to fluctuations in pressure within the ear canal. Connected to this membrane are three small bones: the malleus, incus, and stapes. These ossicles work in tandem to transmit the pressure changes from the ear canal to the cochlea through a minute structure known as the oval window (refer to Fig. 2.1). Remarkably, due to the size differential between the eardrum and the oval window, pressure is significantly amplified, typically by a factor of approximately 20. This amplification process is critical to the ear's mechanism for transforming sound waves into auditory signals [26].

The cochlea is a remarkable structure within the ear, resembling a coiled, snail-shell-like tube. It measures approximately 3.5 centimeters in length, with an average diameter of about 2 millimeters. Notably, its thickness is greater at the base, closer to the oval window, and gradually tapers towards the apex. Within the cochlea, a vital component known as the basilar membrane resides. This membrane initiates vibrations when subjected to pressure fluctuations within the cochlear fluid. What is interesting about the basilar membrane is its changing diameter along its length, which increases as it extends toward the cochlear apex. As a result of this gradient, the local mass density of the basilar membrane progressively rises toward the cochlear end. This crucial variation in mass density bestows upon each portion of the membrane a unique sensitivity to a particular frequency

range. At the cochlear base, the membrane is most responsive to higher frequencies, whereas lower frequencies stimulate segments of the membrane positioned closer to the apex. When a sine wave of a specific frequency propagates through the cochlea, it activates a distinct region of the basilar membrane characterized by a finite width. Consequently, the basilar membrane can be conceptualized as a series of auditory bandpass filters, each with its specific bandwidth, called critical bands. This intricate arrangement allows the cochlea to perform a sophisticated frequency analysis of incoming sounds, playing a pivotal role in our ability to perceive various frequencies and tones.

The vibrations of the basilar membrane are transformed into electrical signals through a process involving inner hair cells, which are positioned between the basilar membrane and another membrane called the tectorial membrane. As these membranes oscillate, minuscule hair-like structures attached to the inner hair cells undergo lateral movement relative to each other. These movements cause the stereocilia, or the hair-like structures, to bend, generating a mechanical force that results in the release of chemical neurotransmitters by the inner hair cell. This release of neurotransmitters triggers electrical activity, manifesting as neural spikes, in the neurons that are connected to the hair cell. Notably, the magnitude of basilar membrane displacement corresponds to the quantity of neurotransmitters released, thus influencing the extent of electrical activity generated. It is worth noting that the hair cells are responsive only to displacements in the upward direction; no neurotransmitter is released when the basilar membrane moves toward the center of the cochlea. This phenomenon leads to phase locking, where neural firing becomes synchronized with a specific phase of the basilar membrane's motion. Consequently, the timing of neural firing is directly linked to the period of the input signal, allowing for precise encoding of sound frequencies and their temporal characteristics.

Ultimately, the signals generated by the inner hair cells are conveyed through the auditory nerve, which consists of a bundle of nerve fibers. This nerve network comprises approximately 30,000 individual fibers, most of which are connected to the inner hair cells, as documented in [26]. Remarkably, each inner hair cell is linked to around 20 nerve fibers. Since each hair cell is situated at a distinct location along the basilar membrane, these fibers are finely tuned to different frequencies. This intricate arrangement ensures that the auditory nerve efficiently encodes and transmits a wide range of frequencies, contributing to our ability to perceive the full spectrum of sounds.

In the absence of external sound stimuli, many nerve fibers within the auditory system exhibit a continuous baseline of neural activity, referred to as spontaneous activity. When a sound with a consistent intensity is introduced, the neurons display a sudden surge in activity, followed by a gradual return to a stable, ongo-

ing level of activity. Similarly, when the sound ceases, a level of activity, albeit lower than the spontaneous baseline, persists for a certain duration. This response pattern to sound onsets and offsets is commonly referred to as adaptation. Furthermore, the magnitude of the steady-state neural activity is directly influenced by the sound's intensity. Higher sound levels correspond to higher levels of sustained neural activity. However, it is important to note that there is a limit to the amount of neural activity that can be generated. The neural response begins to saturate at very high sound levels, meaning that further increases in sound intensity do not produce proportionally higher levels of neural activity. This saturation effect represents a physiological constraint within the auditory system, ensuring that the neural response remains within manageable bounds, even in the presence of extremely loud sounds.

In Fig. 2.1, a separate set of nerve fibers, known as the vestibular nerve, is depicted. These nerve fibers play a crucial role in transmitting positional information from the semicircular canals. The semicircular canals function as the body's balance or equilibrium organ, contributing to our sense of spatial orientation and stability. The auditory and vestibular nerve converge to create a singular nerve known as the vestibulocochlear nerve. This combined nerve serves as the conduit for transmitting all signals from the auditory and vestibular systems to the brain. Within the brain, these signals undergo intricate processing through various neural pathways and structures, ultimately culminating in the perceptual experience of sound and the sense of balance and spatial orientation provided by the vestibular system.

2.1.2 Simulation of Human Auditory System

Numerous computational auditory models have been developed to describe the intricate signal processing that occurs in the ears. These models aim to simulate and understand the various stages of auditory signal processing, from the reception of sound waves to the neural encoding of auditory information. Computational auditory models have found valuable applications in the fields of SER, ASR, and speech quality evaluation. These models extract relevant acoustic features from speech signals that capture the nuances of task-specific information. Different auditory models are utilized to simulate specific stages of auditory signal processing, providing a framework for feature extraction tailored to the characteristics of human auditory perception. This section briefly introduces three critical operations commonly employed in the simulation of the human auditory system: auditory filterbank, temporal envelope extraction, and modulation filterbank.

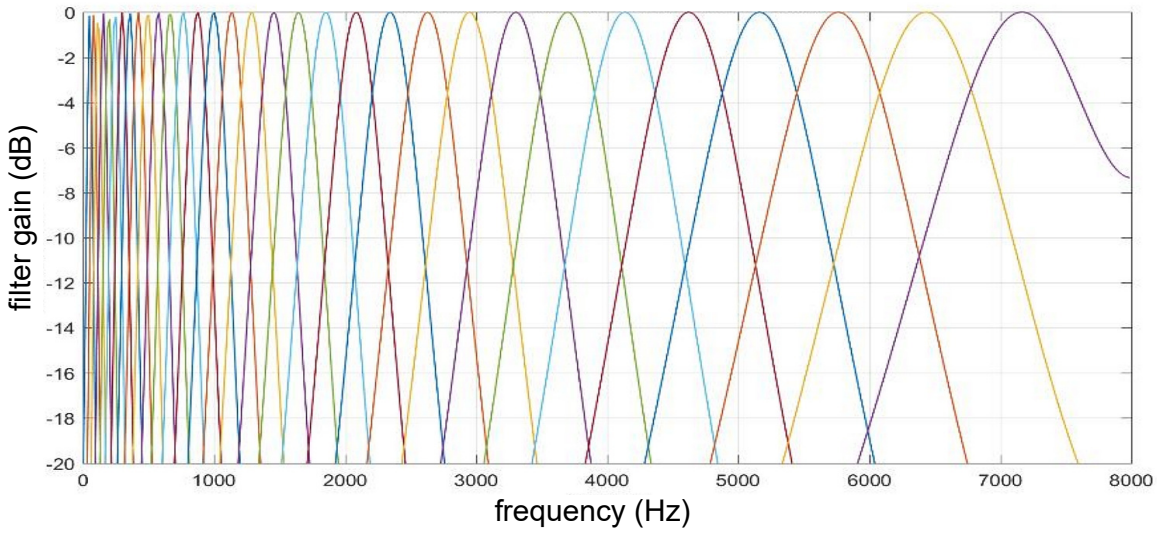

Figure 2.2: Frequency response of Gammatone filterbank

Auditory Filterbank

The auditory filterbank is a crucial component in auditory signal processing, primarily designed to simulate the time-frequency decomposition within the human ear's cochlear basilar membrane. The auditory filterbank consists of filters, each with a specific frequency response that mirrors the sensitivity of different regions along the basilar membrane. These filters divide the incoming acoustic signal into frequency bands, akin to how the cochlea separates sounds into distinct frequency channels. This decomposition is critical for capturing the spectral content of the signal as perceived by the human auditory system.

Two frequently employed cochlear models are Lyon's cochlear model [27] and the auditory filterbank model based on equivalent rectangular bandwidth (ERB) [28]. These models serve as simulations of the cochlea, aiding in the analysis and processing of auditory signals. Lyon's model utilizes a cascade of Gammatone filters to emulate the human cochlea's spectral processing, while Moore's model employs a set of filters designed to approximate the ERB of the cochlea, offering valuable tools for various audio processing applications.

Gammatone filterbanks (GFB) [29] and Gammachirp filterbanks [30] are commonly employed auditory filterbanks. The concept of Gammatone responses was initially proposed in 1972 to describe revcor (reverse correlation) functions measured in the cochlear nucleus of cats [31]. It has since become an important tool in auditory signal processing and auditory modeling, used to mimic the frequency analysis performed by the human auditory system and to capture essential auditory

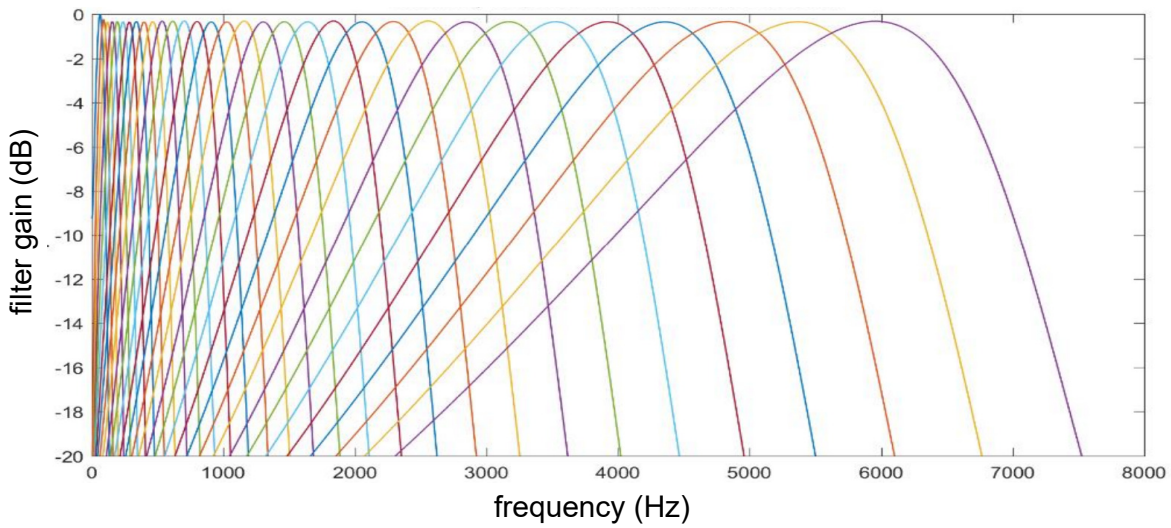

Figure 2.3: Frequency response of Gammachirp filterbank

characteristics. The impulse response of a Gammatone filter is characterized by the convolution of a Gamma distribution and a sinusoidal tone. Each Gammatone filter's bandwidth is quantified using the psychoacoustic measure known as ERB_n. This measure represents the width of the auditory filter at various positions along the cochlea, aligning with human auditory perception. Fig. 2.2 visually represents the frequency responses generated by the Gammatone filterbank.

In contrast, the Gammachirp filter, when compared to the Gammatone filter, exhibits asymmetric and nonlinear characteristics closely resembling auditory filter shapes. As shown in Fig. 2.3, the frequency responses of Gammachirp filters display pronounced asymmetry and feature a sharp decline on the high-frequency side relative to the center frequency. This pattern corresponds well with auditory filter shapes derived from masking data, making Gammachirp filters valuable for capturing auditory signal characteristics in various audio processing applications.

Both Gammatone and Gammachirp filters simulate the basilar membrane's behavior in auditory signal processing, each with its unique strengths and limitations. Gammatone filters offer the advantage of higher calculation efficiency compared to Gammachirp filters. However, Gammachirp filters excel in their ability to simulate the asymmetric and level-dependent characteristics of the auditory filterbank.

Temporal Envelope Extraction

The basilar membrane within the cochlea exhibits a remarkable characteristic: its frequency selectivity varies logarithmically along its length. This logarithmic

mapping means that different positions along the basilar membrane are sensitive to specific ranges of frequencies, mirroring the human auditory system's ability to analyze a wide range of frequencies in a highly organized manner. The inner hair cells are embedded within the basilar membrane, sensory receptors crucial for auditory processing. These hair cells are situated on the lower surface of the basilar membrane and become stimulated when the basilar membrane moves upward. They play a pivotal role in the transduction of mechanical motion into neural signals.

One essential component of this transduction process is extracting a temporal envelope. The temporal envelope is employed to simulate the signal transduction process performed by the inner hair cells. This envelope captures the variations in signal amplitude over time and is essential for tasks like speech perception and auditory analysis.

Traditionally, the temporal envelope from each frequency band is extracted using methods such as half-wave or full-wave rectification and low-pass filtering. However, a more recent approach involves the use of the Hilbert transform as an alternative method for extracting the temporal envelope. Unlike rectification, which can introduce distortions in the frequency components within the modulation domain, the Hilbert transform offers a distinct advantage by providing a clear separation between the signal's temporal envelope and its fine structure [32]. As a result, the Hilbert transform has been favored for temporal envelope extraction in this study, ensuring a more accurate representation of the envelope's characteristics.

Modulation Filterbank

Both physiological and psychological evidence strongly suggest the existence and importance of a modulation filterbank within the auditory system. From a physiological perspective, the processing of amplitude modulation frequencies is primarily carried out in the higher processing stages of the auditory system [33]. This temporal periodicity code is believed to undergo translation into a frequency-selective rate-based representation between the cochlear nucleus (CN) and the IC.

Furthermore, within the IC, there is evidence of a periodotopic organization of neurons that are specifically tuned to different modulation frequencies. These neurons are arranged in a manner nearly orthogonal to the tonotopic organization of neurons, which are tuned to specific acoustic frequencies. Physiological studies have demonstrated that the IC plays a crucial role in processing high-resolution temporal information, particularly by tuning to specific modulation frequencies [4].

Recent psychoacoustic experiments further underscore the significance of temporal modulation in speech perception and comprehension. A modulation filter-

bank has been introduced to effectively analyze envelope fluctuations of stimuli within each peripheral auditory filter. This filterbank enables the extraction of high-resolution temporal modulation cues in the frequency domain, enhancing our understanding of the role of temporal modulation in auditory perception and the processing of complex sounds.

2.2 Acoustic Features Extraction

This section reviews commonly used acoustic features related to our research, including short-term spectral features, prosodic features, timbral features, and spectral temporal modulation features.

2.2.1 Short-term Spectral Features

Breaking down a signal into short frames of about 20–30 milliseconds is a common practice in various signal-processing applications, including speech processing, audio analysis, and even some image-processing tasks. The specific frame duration (e.g., 20–30 ms) is often chosen based on the requirements of the application and the trade-off between temporal resolution and frequency analysis. Short frames strike a balance between capturing fine-grained temporal information and maintaining the stationarity assumption, making them suitable for a wide range of signal-processing tasks. The spectral feature can be extracted for each short-term frame.

Next, the pre-emphasis and windowing operations are often used to process each short-time frame further. Pre-emphasis involves applying a high-pass filter to the signal to emphasize higher-frequency components, reducing the impact of low-frequency noise and enhancing the spectral characteristics of the signal. This process helps improve the signal-to-noise ratio and can make subsequent analysis, such as speech recognition, more effective. Windowing, often done using a Hamming window or other window functions, segments the signal into overlapping frames, providing temporal localization for analysis [34, 35, 36]. The Hamming window reduces spectral leakage in the Fourier analysis and minimizes the abrupt discontinuities at frame boundaries. Together, pre-emphasis and windowing enhance the effectiveness of signal processing techniques by preparing the input data with appropriate spectral and temporal characteristics, ultimately leading to better feature extraction, analysis, and classification results.

The discrete Fourier transform (DFT) is a fundamental tool in signal processing for analyzing the frequency content of a signal [36]. The fast Fourier transform (FFT) is a highly efficient algorithm for computing the DFT and is widely used due to its computational speed. In practice, the magnitude spectrum of the FFT

is often retained while discarding the phase information. It is commonly believed that in many signal processing applications, the human perception system is more sensitive to the magnitude (amplitude) of the frequency components than to the phase. This belief has led to a focus on the magnitude spectrum, which contains essential information about the energy distribution across different frequency components. However, there has been some research that challenges the conventional wisdom of discarding phase information. Paliwal and Alsteris [37] provided evidence that phase information could carry valuable information for certain tasks. Additionally, Hedge et al. [38] proposed techniques that utilize phase information in signal processing applications.

The spectral envelope represents the global shape of the magnitude spectrum and is particularly informative about the resonance properties of the vocal tract, which are unique among different speakers. To capture the spectral envelope, a common approach is to employ a bank of bandpass filters. Each filter is designed to pass a specific range of frequencies while attenuating others. The output of each filter represents the energy within its respective frequency band. The spectral envelope modeling often involves allocating more filters with narrower bandwidths in the lower frequency range, such as the Mel filterbank and Gammatone filterbank. This is motivated by psychoacoustic studies, which have shown that human perception is more sensitive to changes in spectral shape at lower frequencies. Therefore, allocating more resources (i.e., more filters) to capture fine details in the spectral envelope where they matter most is beneficial. Once the signals from the bandpass filters are obtained, they are typically integrated or smoothed to estimate the energy in each frequency band. This integration process helps create a representation of the spectral envelope robust to noise and variations. The subband energy features have been popularly applied in different applications.

In practice, the dimensionality of the spectral envelope representation may be reduced further to obtain a compact feature vector. Techniques like principal component analysis (PCA) [39] or linear discriminant analysis [40] may be applied to select the most discriminative components of the spectral envelope for tasks like speaker recognition and emotion recognition. Additionally, followed by discrete cosine transform [35], the MFCCs [41], inverse MFCC [42], Gammatone filterbank cepstral coefficients (GFCCs) [43] are also popularly used.

Linear Prediction [44] is a powerful technique in signal processing, particularly in speech and audio analysis. It offers an intuitive interpretation in both the time and frequency domains. In the time domain, it models the correlation between adjacent samples in a signal, which can be viewed as a form of autoregressive modeling. In the frequency domain, LP models the signal as an all-pole filter, which corresponds to the resonance structure of the signal. This provides a meaningful and interpretable representation of the underlying signal dynamics.

LP estimates a set of predictor coefficients that describe the linear prediction model. These coefficients provide valuable information about the signal's characteristics, such as spectral properties and resonance frequencies. However, they are often transformed into other feature representations to enhance their robustness and reduce correlation. LPCCs [45] are derived from LP coefficients by applying the cepstral analysis, which models the spectral envelope of the signal in a manner similar to MFCCs. LPCCs are often used in speech and audio processing tasks like speech recognition. LSFs [45] represent LP coefficients that offer improved numerical stability and less correlation between features. They are frequently used in speech coding and synthesis. PLP [46] coefficients are another feature representation derived from LP coefficients. They are designed to mimic the human auditory system's perception of sound and are widely used in speech processing and recognition. The transformation of LP coefficients into these derived features allows for more robust and informative representations that are better suited for various applications.

2.2.2 Prosodic Features

Prosody plays a critical role in speech and speaker recognition because it encapsulates various non-segmental aspects of speech that convey important information about the speaker's identity, emotions, and communication style [47]. Prosody encompasses features extending over long speech segments, including syllables, words, and even entire utterances. These features capture speech's rhythm, melody, and intonation patterns, which can be highly distinctive for different speakers.

Prosody contains valuable cues for speaker recognition. [47, 48]. Differences in speaking style, language background, and emotional expression are often reflected in prosodic patterns. For example, individuals may have unique speaking rates, intonation contours, and stress patterns contributing to their distinct prosodic fingerprints. One of the challenges in text-independent speaker recognition is effectively modeling and extracting prosodic information. This includes capturing instantaneous prosodic features (e.g., pitch at a specific moment) and long-term prosodic patterns (e.g., speaking rate over an utterance). Balancing these two aspects is crucial for accurately characterizing speaker differences. To enhance the robustness of speaker recognition, prosodic features are often combined with other information sources, such as spectral features and phonetic information. This multimodal approach leverages the complementary nature of different feature types to improve recognition accuracy.

Fundamental frequency (F_o), also known as pitch, is one of the most crucial prosodic parameters in speech analysis, and it plays a significant role in various speech-related tasks [49]. The mean value of F_o is closely linked to the physiological

characteristics of the speaker, particularly the size and tension of the vocal folds within the larynx [50]. A larger larynx with longer vocal folds tends to produce a lower-pitched voice, resulting in a lower mean F_o. Conversely, a smaller larynx with shorter vocal folds leads to a higher-pitched voice, resulting in a higher mean F_o. Therefore, the mean F0 can be considered an acoustic correlate of larynx size. For example, male speakers, on average, have larger larynxes and longer vocal folds than female speakers, leading to lower mean F_o values in male voices compared to female voices. In addition to physiological factors, F_o also reflects learned characteristics related to the speaker's manner of speaking. The temporal variations in pitch, known as pitch contours or intonation patterns, are highly informative about how a speaker articulates words and conveys meaning. These variations are influenced by factors such as language, dialect, regional accent, and speaking style. For example, different languages and dialects may exhibit distinct pitch patterns for questions, statements, and exclamations. Additionally, individual speakers may have idiosyncratic pitch patterns that contribute to their unique speaking style.

Combining F0-related features with spectral features has proven to be effective, particularly in challenging and noisy conditions. F_o represents the rate at which the vocal folds vibrate during speech production and is closely related to the perceived pitch of the voice. F_o-related features, such as pitch contours and statistics, provide valuable information about a speaker's unique vocal characteristics [51, 52].

Duration-related prosodic features [53], such as pause statistics and phoneme or phone duration, capture temporal aspects of speech. Differences in speaking rate, hesitation patterns, and rhythm can be indicative of speaker-specific traits. Incorporating duration features can complement other prosodic and spectral features. Speaking rate refers to the speed at which a speaker articulates words or syllables. It can vary widely between speakers and maybe a distinctive characteristic for speaker recognition. Features related to speaking rate, such as syllable rate or phoneme rate, can be informative for identifying speakers.

2.2.3 Timbral Features

Timbre plays a key role in music and audio perception, conveying emotional and perceptual information vital in various audio information processing fields. Timbre is a multifaceted set of auditory attributes that collectively define the quality or character of a sound. Typically, timbre encompasses various spectral and harmonic features that provide distinct characteristics to a sound [54]. This unique quality allows us to differentiate between sounds, even when they share identical pitch and loudness levels. For example, when both a guitar and a flute play the same note with equivalent amplitude, each instrument produces a sound with its own unmistakable tone color or timbral identity [55].

The timbral characteristics belong to psychoacoustic attributes, and each corresponds to a specific sensation when one listens to a song [56]. Psychoacoustics is a field dedicated to exploring the intricate relationship between acoustics and psychology, specifically how humans perceive and interpret sound. It is important to notice that the scores generated by the algorithm do not replicate these characteristics accurately due to the inherent subjectivity associated with human sensory experiences.

Numerous studies have delved into modeling timbre and developing objective metrics for each timbral attribute [57]. An influential example comes from the University of Surrey, which formulated timbral models within the framework of the Audio Commons project. These models have gained widespread acceptance and are extensively employed in psychoacoustic research [58]. The algorithm underlying these models draws on various literature sources that describe exemplary computational models and incorporates findings from subjective experiments. Additionally, these models prove valuable for statistical analysis, as the calculated metric can serve as an indicator, enabling researchers to assess and analyze the timbral attributes of sounds quantitatively. In accordance with references [59, 60, 61, 62], this section provides a brief introduction to several key timbral attributes, which include sharpness, roughness, hardness, and brightness. More information, such as the boominess and depth, can be found in [62, 63].

Sharpness is a metric associated with the perception of sharp or shrill sensations. Sharpness exhibits an increase in magnitude when the center frequency is shifted to a higher range. Based on this, Zwicker introduced the concept of acum. One acum is defined as a unit of narrow-band noise centered at 1,000 Hz with a loudness level of 60-phon [61]. Subsequently, a sharpness model was developed, and it is calculated as follows:

$$S = 0.11 \frac{\int_0^{24Bark} N^{'}(x) g(x) x dx}{\int_0^{24Bark} N^{'} dx} \tag{2.1}$$

In the formula, S represents sharpness, $N^{'}(t)$ denotes the loudness density within the critical-band rate t. Loudness, an intrinsic attribute of auditory sensation, is quantified in phons. $g(t)$ refer to the weighting factor for S at that particular rate. Based on psycho-acoustic experiments, the weighting factor is defined as 1.0 for the frequency range up to 3,000 Hz, rapidly increasing to 4.0 for higher frequencies.

Roughness, as described in [61], is a perceptual sensation arising from relatively rapid amplitude modulation changes. In this context, the change is not associated with fluctuations in loudness but rather pertains to alterations in sound quality [59]. The score provided by the roughness extractor serves as a representation of how the amplitudes interact to simulate the sensation of roughness.

Calculating the apparent roughness of an audio signal involves several methods, with this thesis primarily reviewing the approach introduced in [64, 62]. Firstly, the audio signal is segmented into frames of 50 milliseconds each. These frames are then subjected to windowing using a Hanning window and subsequently zero-padded to the nearest power of two following each frame. Subsequently, an FFT is applied to each frame, and the magnitudes of the frequency components for all frames are normalized. This normalization ensures that the maximum magnitude across all frequencies and frames equals 1.0, facilitating consistent comparisons and measurements.

Following the FFT and normalization processing, a peak-picking algorithm is employed on each frame to identify the peaks present within the frequency spectrum. For each pair of peaks detected within a frame, the roughness is calculated as follows:

$$R = 0.5L^{0.1}M^{3.11}N \tag{2.2}$$

with:

$$L = A_{min} * A_{max}, \tag{2.3}$$

$$M = \frac{2A_{min}}{A_{min} + A_{max}}, \tag{2.4}$$

$$N = e^{(-3.5g(f_{max}-f_{min}))} - e^{(-5.75g(f_{max}-f_{min}))}, \tag{2.5}$$

$$g = \frac{0.24}{0.0207f_{min} + 18.96}, \tag{2.6}$$

where R refer to the roughness, A_{max} and A_{min} represent the maximum and minimum magnitudes of the peak pairs, while f_{max} and f_{min} correspond to the maximum and minimum frequencies of the two peaks, respectively. The cumulative roughness for a frame is calculated by summing the roughness values of all the roughness pairs within that frame. The overall roughness for an entire audio file is determined by calculating the mean of the roughness values across all frames.

Hardness in sound perception is essentially a blend of two factors: loudness and harshness. When a sound is described as harsh, it means that there is an imbalance in the audio core of the sound. The audio core, in this context, refers to

the frequency range between 2kHz and 5kHz, which corresponds to the range where the human ear is most sensitive. Therefore, hardness is a metric to measure how pleasant or soothing a sound is by evaluating the harmony between its loudness and frequency range, as perceived by the human ear. In essence, it quantifies how well the balance between these two attributes aligns with human auditory preferences.

The timbral attribute of hardness has garnered significant attention in research, as highlighted by Pearce et al. in 2016 [65]. Notably, [66] conducted a study suggesting that the initial phase or onset of a sound plays a pivotal role in shaping our perception of hardness. Solomon [67] made a contribution by identifying the perceptual dimension of hardness/softness as a fundamental psychological dimension of timbre. This work posited a hypothesis that this attribute may be linked to rhythmic distinctions among stimuli, thus suggesting a potential connection between timbral qualities and rhythmic characteristics. Additionally, Freed's research in 1990 [68] introduced a model for perceiving mallet hardness in the context of single percussive sounds. This model considered four key acoustic correlates: Spectral Mean Level (This represents a form of the long-term average spectrum, which provides insights into the overall spectral content of a sound.), Spectral Level Slope (Similar to cepstrum analysis, this attribute pertains to the rate of change in spectral levels across frequencies, shedding light on the sound's spectral characteristics.), Spectral Centroid Mean (This is a measure of the mean spectral centroid over time, quantifying the center of mass of the sound's frequency distribution on the bark scale.), and Spectral Centroid (This attribute calculates the time-weighted mean of the spectral centroid, providing a dynamic perspective on the sound's centroid location over time.). [62] developed three key metrics to represent hardness, including attack time, attack gradient, and spectral centroid of attack. Subsequently, a linear regression model was employed to predict the perceived hardness based on these parameters. Interested readers may refer to [66] for further details.

Brightness, a timbral attribute of considerable research interest, has been explored in depth. Some studies have revealed that the spectral centroid is a metric strongly correlated with the perception of brightness [69, 70]. However, alternative research suggests that an even more effective predictor of brightness is the ratio of high-frequency components to the total energy in the sound signal [71, 72]. Recently, Pearce's recent work surveyed existing models and introduced a new model for brightness that incorporates both a spectral centroid variant and a spectral energy ratio, offering a comprehensive approach to capturing the perception of brightness in sound [66]. The calculation processes are introduced in the following paragraph.

Firstly, an audio signal is divided into small frames and converted to the fre-

quency domain using FFT. A sample-by-sample half-octave smoothing technique is then used to smooth the magnitude frequency response for each audio frame. This smoothed response is subsequently employed to derive two metrics, including the frequency-limited spectral centroid, which specifically considers frequencies above 3 kHz (SC_{3k}) and ratio.

$$SC_{3k} = \frac{\sum_{n(3kHz)}^{n(Nyquist)} f(n)x(n)}{\sum_{n(3kHz)}^{n(Nyquist)} x(n)}, \tag{2.7}$$

$$\text{Ratio} = \frac{\sum_{\text{n(3kHz)}}^{\text{n(Nyquist)}} \text{x(n)}}{\sum_{\text{n(20Hz)}}^{\text{n(Nyquist)}} \text{x(n)}}, \tag{2.8}$$

Where $n(s)$ represents the bin number corresponding to frequency s, $f(n)$ denotes the frequency associated with the nth bin, $x(n)$ represents the magnitude of the nth bin, $n(Nyquist)$ is the bin number corresponding to the Nyquist frequency. Finally, a linear regression is employed to derive results related to brightness.

$$B = 25.8699 + 64.0127((log_{10}(Ratio) + 0.44log_{10}(SC_{3k})). \tag{2.9}$$

Recently, timbre features have been analyzed and applied for music emotion classification [73, 74, 75] and audio classification (machine sound) [76].

2.2.4 Spectral Temporal Modulation Features

The spectral and temporal signal details, such as formant transitions temporal amplitude variations, intensity, duration, and pitch variations, are rich sources of task-specific information in speech and audio processing. Properly extracting and analyzing these features can significantly improve the accuracy and robustness of various speech and audio analysis systems.

Numerous temporal or spectral modulation features have been introduced and put into practice. For example, in the FAD task, Wu et al. [77] suggested utilizing magnitude and phase modulation features for detecting synthesized speech, enhancing the security of speaker verification systems. As shown in Fig. 2.4, the modulation spectrum extraction procedure involves taking the STFT of the speech signal and computing the magnitude and phase spectrum. The modulation spectrum is then obtained by taking the STFT of the logarithm of the magnitude spectrum. Finally, the temporal modulation features are extracted from the modulation spectrum using PCA to reduce the dimensionality of the feature vector. This study [77] shows that the fusion of phase modulation features and phase spectrum features achieves the best detection performance.

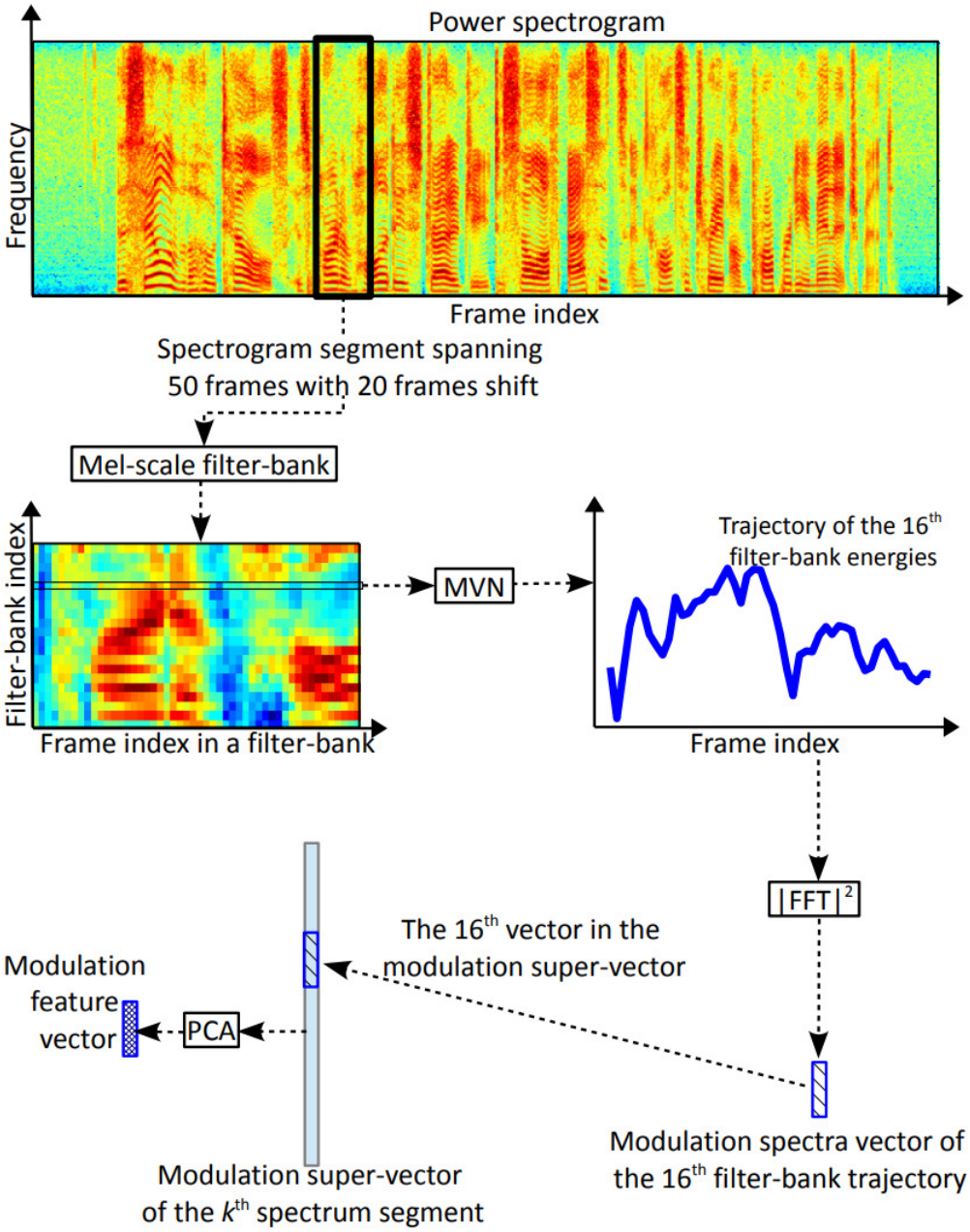

Figure 2.4: Illustration of temporal modulation feature extraction from power spectrogram.

In the emotion recognition field, Zhu et al. [23] believe that the temporal modulation clues derived from the temporal envelope provide a lot of important information related to the perception of vocal emotion. Based on this, they inves-

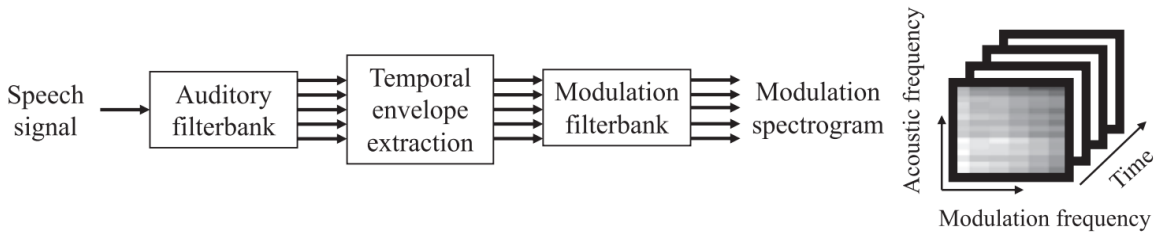

Figure 2.5: Schematic diagram of the calculation of modulation spectrogram.

tigate the effectiveness of modulation spectral features with consideration of the concept from the human auditory mechanism. These features were used to perceive data collected from vocal-motion recognition experiments. Fig. 2.5 shows the specific extraction process of modulation spectrograms. In Fig. 2.5, the auditory filterbank, such as the Gammatone and Gammachirp filterbanks introduced in Section 2.1.2, is firstly used to filter the emotional speech signal into subbands. The Hilbert transformation is then utilized to calculate the temporal envelope. The modulation filterbanks were finally used to decompose temporal envelopes into different modulation frequency bands. The results show that the modulation spectrograms are quite different among emotions, especially sadness and joy. It means the modulation spectral features are crucial for vocal emotion perception.

The aforementioned finding can be further verified in [22] and [78]. These two studies focus on applying temporal modulation features in DNN-based emotion recognition systems. They also believe that the human auditory system can easily perceive the variation in the time domain of different kinds of emotions. However, combining these informative features with sophisticated NN architectures should be deeply considered. As shown in Fig. 2.6, [22] proposed a novel method by combining the 3D convolutions and attention-based sliding recurrent neural networks (ASRNNs) to extract more helpful information. The evaluation results in two commonly used emotion datasets show that the proposed method (ASRNNs) can effectively recognize speech emotions based on temporal modulation cues.

Additionally, inspired by the multi-resolution analysis characteristics of human brains, [78] proposed a novel temporal modulation feature, named the multi-resolution modulation-filtered cochleagram (MMCG), to capture the variations of arousal and valence values of short-term frames in a long sequence. As shown in Fig. 2.7, four modulation-filtered cochleagrams at varying resolutions are combined to create the MMCG. Furthermore, a parallel long short-term memory (LSTM) architecture was designed to mimic the multi-temporal dependencies of the MMCG. Extensive trials conducted on the RECOLA and SEWA datasets show that, out of all the analyzed features, MMCG offers the best recognition

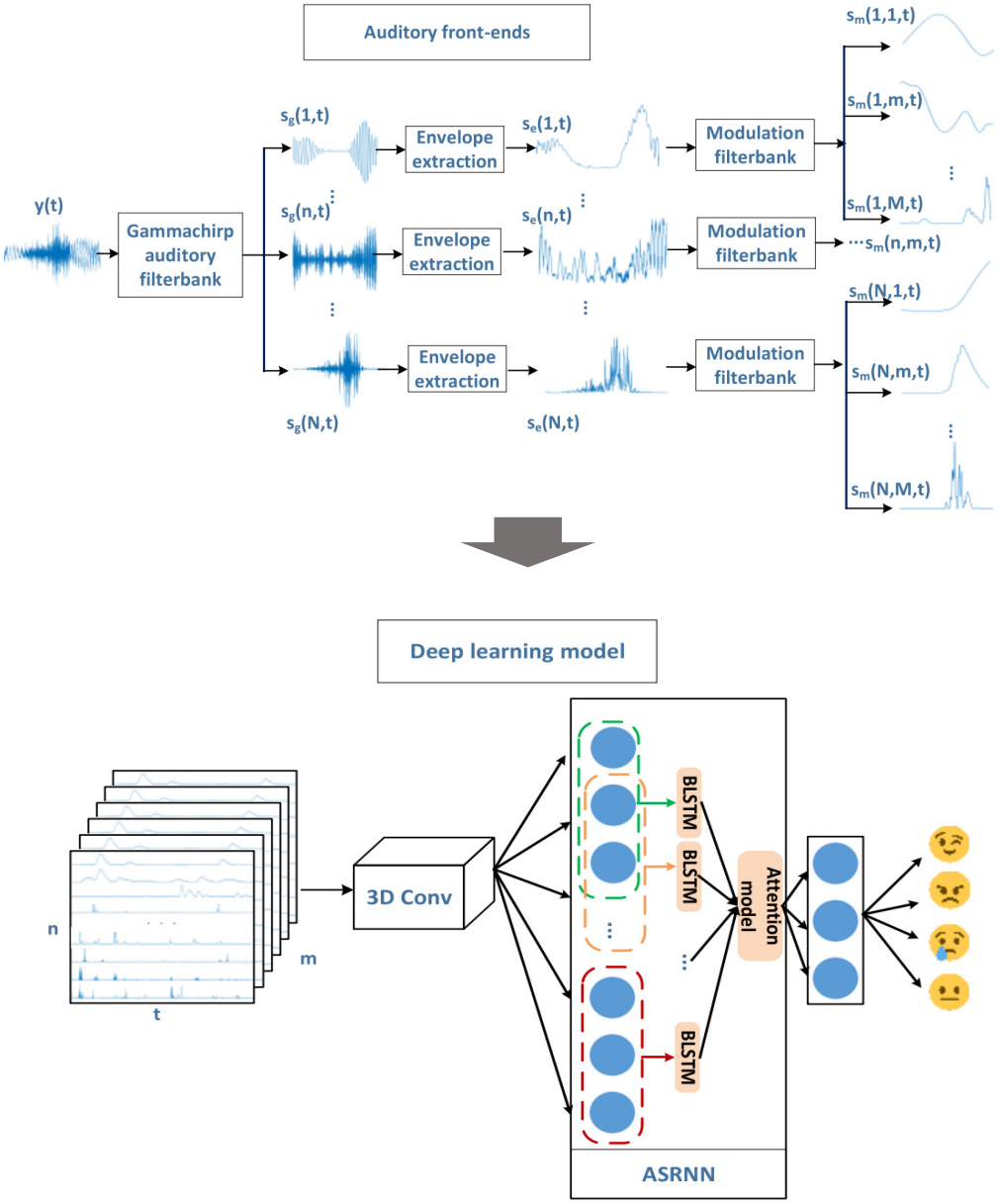

Figure 2.6: Speech emotion recognition system with auditory front-ends.

performance in both datasets.

STM features can also be used for discriminating different speakers [79]. For example, as shown in Fig. 2.8, Kinnunen et al. proposed to use the joint acoustic-modulation frequency feature for speaker recognition [80]. The extraction proce-

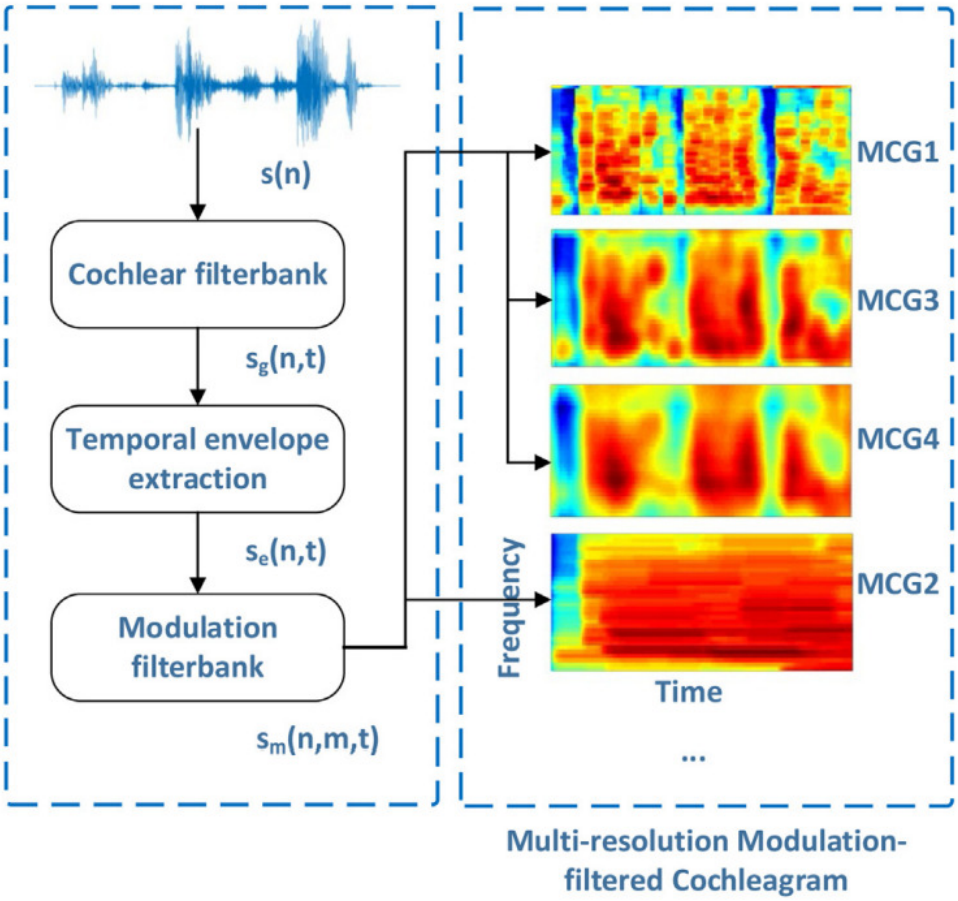

Figure 2.7: The extraction of MMCG features with different resolutions.

dure consists of three steps. The spectrogram is first calculated using conventional short-term Fourier analysis. Next is another short-term Fourier analysis along the DFT output amplitude envelopes. Lastly, these short-term modulation spectra are subjected to a time-average. By using this feature with the conventional static and dynamic cepstra feature, a slight improvement was achieved. In order to extract more effective speaker discriminative information and decrease the feature dimension, a low-dimensional feature that captures the shape of the modulation spectra was also proposed by Kinnunen [81]. As shown in Fig. 2.9, they believed that the proposed modulation representation could provide more information related to speaking rate and other stylistic attributes, hence improving the performance of the speaker recognition system.

More applications of modulation representations can also be found in speech intelligibility [20, 21] and speech recognition [25, 82].

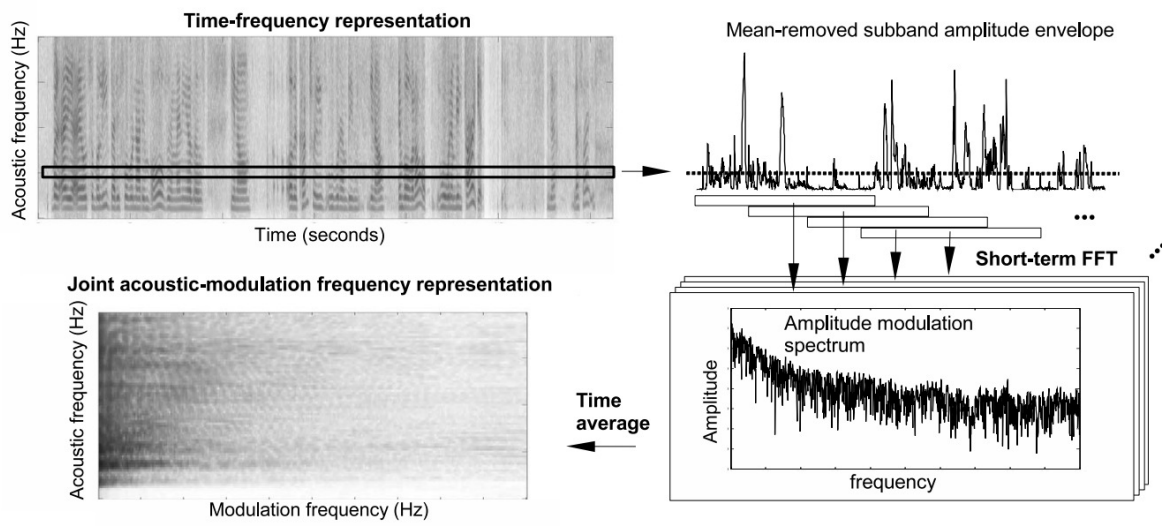

Figure 2.8: Computation of joint acoustic-modulation frequency representation.

2.3 Machine Learning-based Audio Detection and Verification Methods

2.3.1 Speaker Verification (SV)

ASV aims to confirm a speaker's stated identity by comparing tested and registered speech. Usually, for both registration and test speech, a low-dimensional feature rich in speaker individuality is extracted first. These features are then mapped to deep embedding representations and used to calculate the verification score using some comparison criterion. There are text-dependent and text-independent SV variations. Text-dependent SV requires the speech content to be fixed to a specific phrase. This thesis focuses on proposing advanced feature representations for text-independent SV. A detailed flow diagram of ASV is depicted in Fig. 2.11.

SV is commonly used for security and access control purposes. It is employed in various applications, including telephone banking, voice-controlled devices, secure facility access, and other scenarios where voice authentication is necessary to confirm a person's identity. For example, banks and financial institutions use SV as an additional layer of security for phone-based customer authentication, allowing customers to access their accounts, check balances, or perform transactions over the phone. Law enforcement agencies use SV in forensic investigations to match voice recordings with known suspects or to identify potential witnesses. Also, SV may be used in voice-activated in-car systems to authenticate drivers and provide personalized settings and access to features.

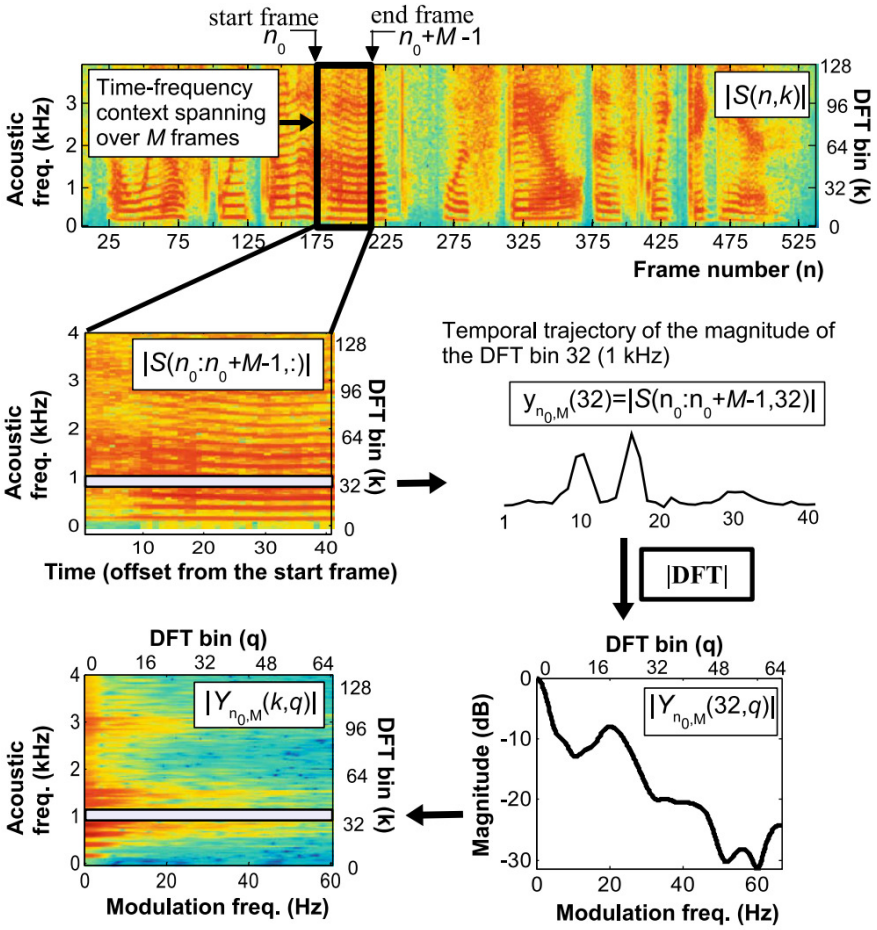

Figure 2.9: The extraction of modulation spectrogram features for speaker verification.

The extraction of speaker discriminative information is commonly based on the i-vector system [83, 84]. Then, the probabilistic linear discriminant analysis(PLDA) backend is employed for scoring. This method can be further improved by replacing the universal background model (UBM) with DNN models [85, 86], improving the ability of phonetic modeling of the UBM. However, an i-vector system's components are often trained on complimentary subtasks; they are not jointly optimized for verification purposes.

In 2014, Variani et al. [87] introduced a DNN-based approach to obtain more speaker-related discriminative information for text-dependent speaker recognition. This method involved extracting features from the activation function of the last hidden layer, known as d-vectors, which served as sentence-level representations.

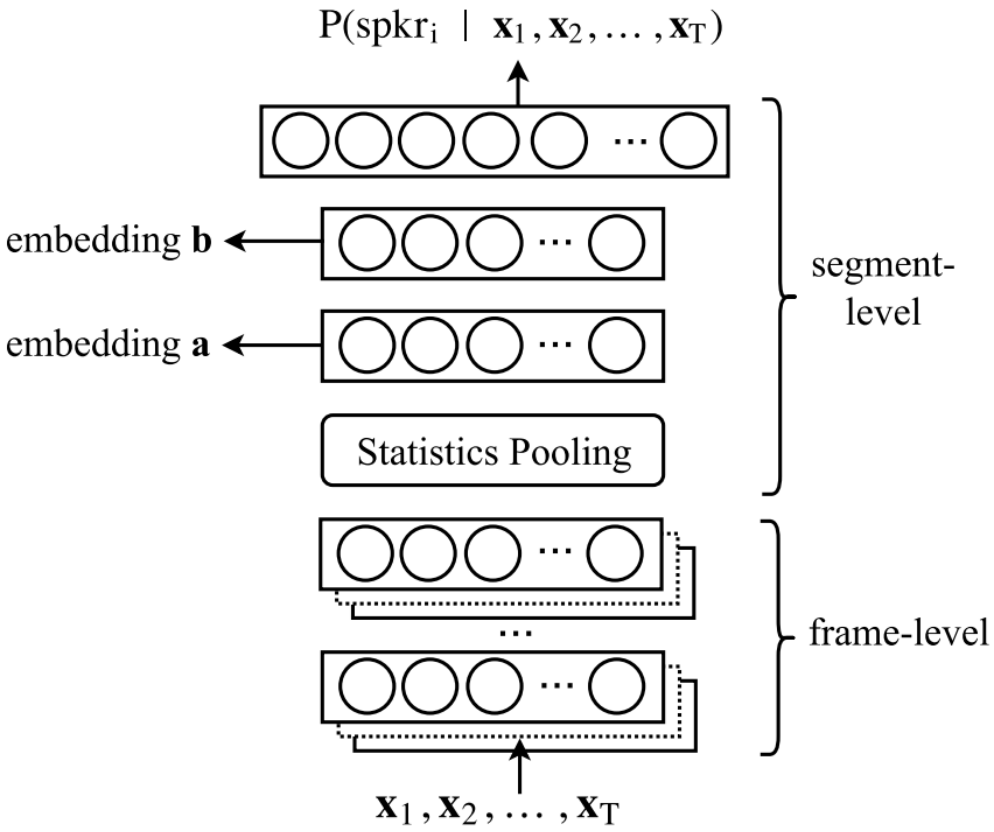

Figure 2.10: Architecture of the x-vector extractor proposed by David Snyder [2].

Subsequently, similarity scores were assessed by calculating the cosine distance between enrollment and test utterances. In 2017, David Snyder [2] proposed a different set of speaker embeddings referred to as x-vectors, which were derived after the statistics pooling layer. The complete architecture of the x-vector extractor is illustrated in Fig. 2.10.

As shown in Fig. 2.10, the first five layers incorporated a time delay neural network (TDNN) to extract frame-level features. Subsequently, the statistics pooling layer computed the mean and standard deviation of the TDNN output, thereby generating a segment-level representation. The x-vector has served as a fundamental baseline in numerous competitions. More recently, more advanced DNN-based systems have emerged, leading to substantial improvements in ASV performance due to their robust nonlinear mapping and learning capabilities. However, it is worth noting that the performance of a DNN-based SV system is significantly dependent on the effectiveness of the front-end acoustic features.

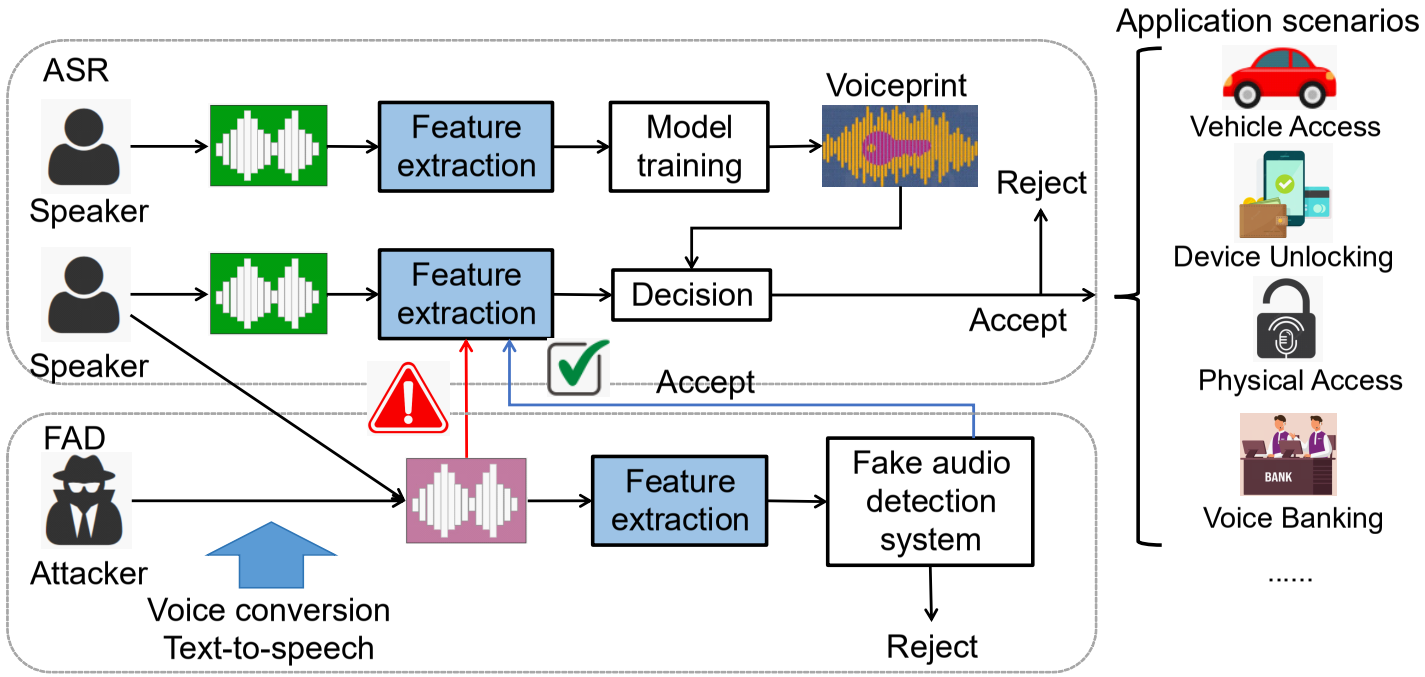

Figure 2.11: Flow chart of ASV and FAD tasks.

2.3.2 Fake Audio Detection (FAD)

With the advancement of deep learning and generative models, generating high-quality synthetic audio has become increasingly easier. This includes speech synthesis, voice cloning, and audio deepfakes. As a result, distinguishing between real and fake audio has become a significant challenge. FAD, also known as Audio Deepfake Detection (ADD), is the process of identifying and verifying whether an audio recording has been manipulated, synthesized, or altered with the intent to deceive or mislead. The goal of this technology is to distinguish between genuine, unaltered audio recordings and those that have been created using various audio deepfake techniques. The importance of FAD techniques can be found in Fig. 2.11.

The ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) [88, 89] and ADD [90] challenges are well-known competitions in the fields of FAD. The ASVspoof challenge is a series of competitions designed to address the vulnerabilities and security concerns in ASV systems. The ADD challenge focuses on detecting audio deepfakes, which are manipulated or synthesized audio recordings generated using deep learning techniques. Both the ASVspoof and ADD challenges provide a platform for researchers and data scientists to collaborate, benchmark their models, and advance the capabilities of audio deepfake detection systems. The latter is more challenging due to more advanced synthesis techniques being added.

In recent years, a lot of advancements have been made in the FAD field. The architecture, including a light convolution neural network (LCNN) followed by two bi-directional long short-term memory (Bi-LSTM) units, a global average pooling layer, and a fully connected output layer, is popularly used as a baseline [91]. This is because LCNN and Bi-LSTM models are well-suited for sequential data processing tasks, a fundamental requirement in audio analysis. In the context of ADD, audio signals are inherently sequential in nature, as they involve variations in sound over time. LCNN and bi-LSTM can capture temporal dependencies and patterns within audio data, making them suitable for this task.

In addition, Subramani and Rao [92] proposed two lightweight convolutional network models, named EfficientCNN and RES-EfficientCNN, to detect synthetic audio. Their experimental results show that the proposed method can perform better with fewer parameters. To overcome the vanishing gradients problem during the training of very deep neural networks, Moustafa Alzantot [7] proposed to use a deep residual convolutional network to perform FAD. In this work, three different acoustic features, including the LFCC, log-magnitude STFT, and CQCC, are used to train three models separately. Finally, model fusion is conducted to improve the performance further. More recently, self-supervised methods have been introduced [93]. Self-supervised models are pre-trained on large-scale datasets, learning valuable representations from generic tasks. These learned representations can be

fine-tuned on the target ADD task. This transfer of knowledge can significantly boost the model's performance.

Since this study focuses more on acoustic feature extraction, more advanced deep-learning architectures will not be reviewed. Interested readers can refer to [94] for more details.

2.3.3 Machine Anomalous Sound Detection (ASD)

Machine ASD, also known as audio anomaly detection, is a field of research focused on automatically identifying unexpected or irregular sounds in audio data. This area has gained significant attention due to its relevance and importance in various applications and industries. For example, in industrial settings, the ability to detect abnormal sounds in machinery can prevent costly breakdowns and downtime. This leads to more efficient maintenance practices and cost savings. Additionally, Machine ASD can assist in remote patient monitoring and early disease detection. It enables healthcare professionals to identify health issues in patients through audio cues. This thesis focuses on the voice data of different machines, such as the fan, bearing, gearbox, etc., to detect factory machine anomalies.

A complete workflow of the machine ASD system is depicted in Fig. 2.12. It involves continuously collecting data from sensors embedded in machines, extracting features, detecting anomalies using machine learning algorithms, and informing the workers to take necessary measures. The research on ASD can be classified into two types of problems [95], i.e., supervised ASD and unsupervised ASD. In supervised ASD, recordings of anomalous events that need to be detected are available during the training phase. This means that the system is provided with labeled examples of both normal and anomalous sounds. In unsupervised ASD, there is a lack of recordings of the anomalous events during the training phase. The system is only provided with recordings of normal sounds, and no labeled examples of anomalous events are available. Generally, researchers focus on unsupervised ASD due to the difficulties in collecting anomalous events that can cover all kinds of anomalies.

In recent years, a lot of DNN-based methods have been proposed to improve the performance of unsupervised ASD. These method contain the architectures of CNN [96], WaveNet [97, 98], KNN [99], autoencoder (AE) [100], and even pretrained [101] models. Among these methods, AEs and their variations are the most popular in unsupervised ASD. There are three main reasons. First, AEs reduce the dimensionality of audio data by encoding it into a lower-dimensional representation (the "latent space") and then decoding it back to its original form. This dimensionality reduction can help capture essential features and minimize noise, which is crucial for ASD. Second, AEs are trained to minimize the difference between the input and the output (reconstruction) in an unsupervised manner. During in-

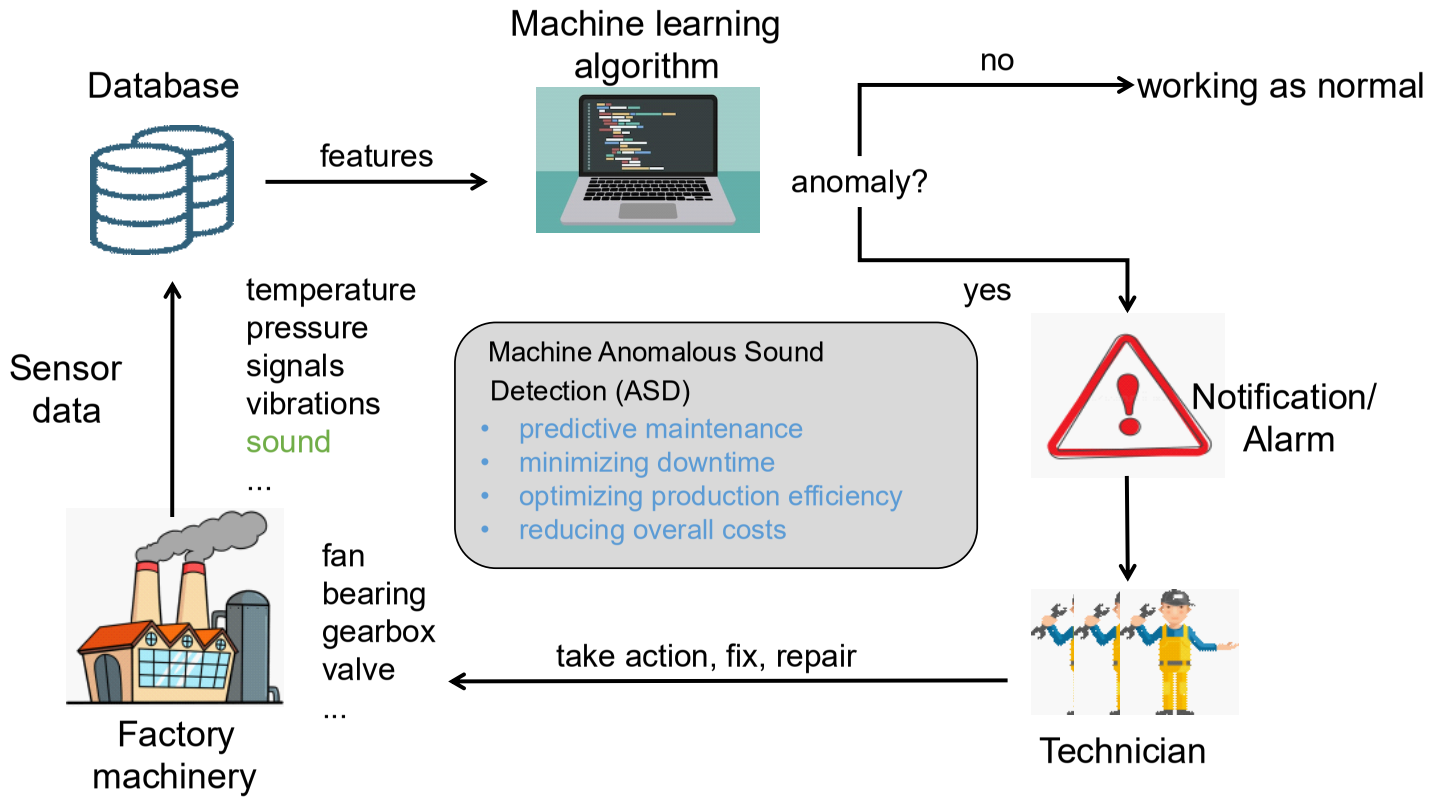

Figure 2.12: A brief introduction of machine anomalous sound detection.

ference, if the reconstruction error for a given audio segment is significantly high, it indicates an anomaly. This makes the use of anomalous sounds in the training stage not necessary. Last, AEs can be designed in various architectures, including convolutional AEs for capturing spatial patterns, recurrent AEs for modeling temporal dependencies, and more. This versatility allows them to adapt to different types of audio data. This thesis tests the effectiveness of our proposed feature representation for detecting machine anomalies using a simple AE-based approach. More specifics can be found in Chapter 6.

Chapter 3

Representations Based on Spectral Temporal Modulation Analysis

3.1 General Introduction of Proposed Features

As we explained above, the human auditory system can effectively perceive TSI in different acoustic scenarios. This ability depends on the special signal-processing mechanism of the human auditory system. A large number of studies have shown that STM analysis of temporal amplitude envelope is very important in the human auditory system, and the simulation of this process using computable models can be applied in several applications.

In STM analysis, spectral modulation refers to variations in the spectral content of a signal over time. It can capture information related to changes in the formants and other spectral attributes. Temporal modulation, on the other hand, is associated with variations in the temporal characteristics of sound, such as timbre and prosody-related attributes. These temporal modulation representations can reveal how sound changes over time and help in the perception of timbre and other temporal aspects. STM analysis deals with both spectral and temporal modulation of audio to perceive auditory attributes related to audio production.

Inspired by this, as shown in Fig. 3.1, this thesis proposed feature representations based on frequency domain analysis, time domain analysis, and STM analysis and introduced in Section 3.2, Section 3.3, and Section 3.4, respectively. These three sections correspond to three research questions to be studied: (Q1) Which frequency regions are more important for TSI extraction? (Q2) Can we represent TSI in the time domain? and (Q3): Can STM analysis obtain more TSI? The first and second research questions can help us confirm the importance of spectral

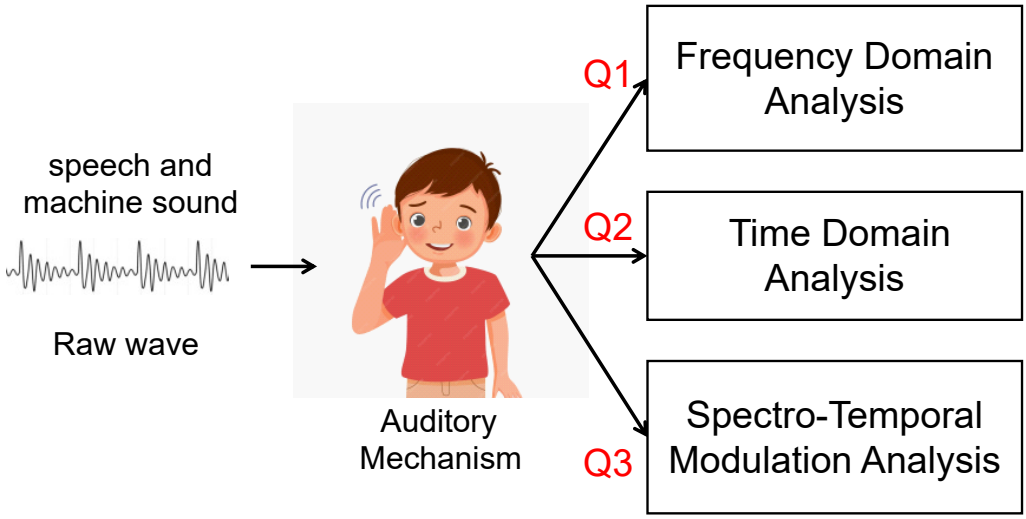

Figure 3.1: The statement of research philosophy. Q1, Q2, and Q3 correspond to three research questions described in the text content.

and temporal attributes in the representation of TSI, hence extracting more TSI in the STM analysis.

3.2 Frequency Domain Analysis

Influenced by the differences in the physiological structure of speech organs, the speech synthetic methods, and the physical characteristics of different machines, discriminative information used for ASV, FAD, and machine ASD tasks are distributed nun-uniformly in the frequency domain. For example, the glottis is an important articulator to modulate the air input from the lung. The vibration frequency of the glottis from a normal adult ranges between 60 and 400 Hz due to the differences of glottis length and stiffness among different speakers [102]; The nasal cavity is the largest side branch within the vocal tract. The nasal cavity with the sinuses demonstrates significant SDI from 1 kHz to 2 kHz when producing nasal and nasalized sounds [103, 104]; In addition, the distinguishing information tends to be concentrated in the high-frequency area when the machine is wearing out. Which frequency regions are more important for a specific task is still not clear.

In this section, two data-driven methods based on F-ratio and frequency-wise attentional neural networks are proposed to quantify the importance of different frequencies in ASV, FAD, and machine ASD tasks. The quantification results of each task are described in Chapter 4, Chapter 5, and Chapter 6.

3.2.1 Quantification of Frequency Importance Using F-ratio

The Fisher's ratio (F-ratio) is a statistical-based method and widely used to measure the discriminative ability of a feature for pattern recognition [105]. It has been used to evaluate the importance of different frequencies in speaker recognition [106], [107], emotion recognition [108], and replay attack detection [109]. The calculation of the F-ratio requires no training data and is comparatively straightforward and efficient.

For ASV, the frequency bands with more discriminative features should possess high inter-class variances and low intra-class variances among different speakers. Based on this, the F-ratio is defined as:

$$\text{F-ratio} = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_i^j - u_i)^2}, \tag{3.1}$$

where x_i^j is the acoustic feature variable (subband energy is used in this study) of the jth speech frame of speaker i with $j = 1, 2, ..., N$, and $i = 1, 2, ..., M$, and u_i and u are variables that represent the subband energy averages for speaker i and for all speakers, respectively, which are defined as:

$$u_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j; \qquad u = \frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N} x_i^j. \tag{3.2}$$

Equation (3.1) is the ratio between the inter-speaker variance and intra-speaker variance of speech power in a given frequency band. A larger value obtained in a frequency band means that more speaker information is encoded in that band. Similarly, the calculation of the F-ratio for FAD and machine ASD can use the same method introduced above but with different character meanings.

In Eq. (3.1), the discrimination measurement that uses F-ratio is based on a single-mode Gaussian distribution assumption of the subband power energy variable. It is possible that the distribution could be multi-mode with a mixture of distributions. In addition, the frequency importance is calculated in each frequency band independently. Therefore, it cannot reflect the nonlinear and joint relationship among different frequency bands. To solve this disadvantage, a DNN-based quantification method is proposed in the following section.

3.2.2 Quantification of Frequency Importance Using Frequency-wise Attentional Neural Network

DNN-based models have been successfully used for ASV. Due to the strong capacity in speaker discriminative feature extraction, the performance of ASV has

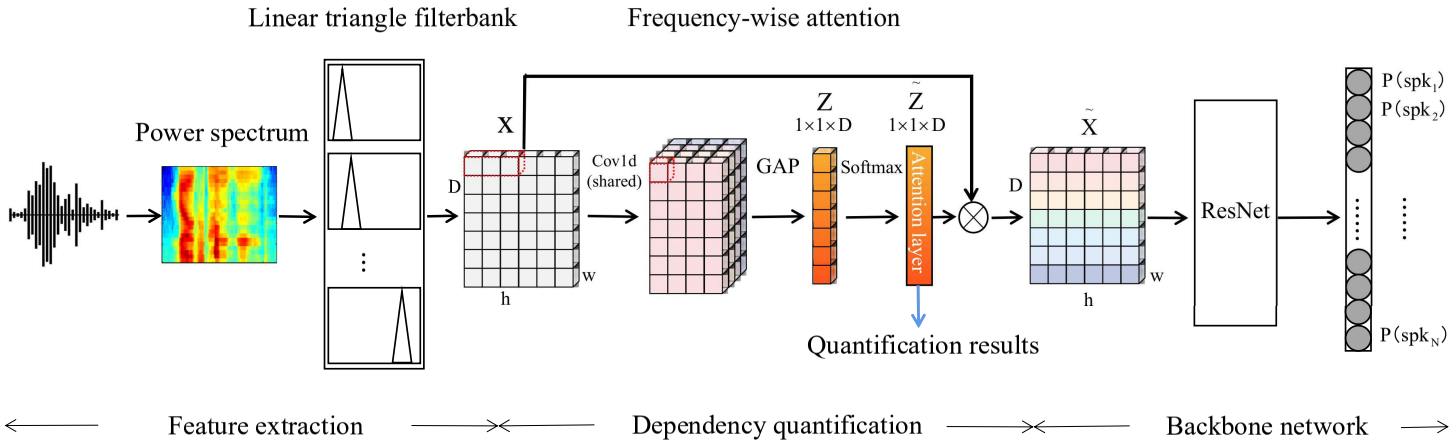

Figure 3.2: Proposed residual network architecture augmented with frequency-wise attention to learn dependencies between frequency components and speaker individuality.

been significantly improved. However, as a black box modeling method in DNN, it is difficult to understand which acoustic features are specifically relevant to speaker discrimination. Unlike most studies, we obtain information about which frequency components are important for speaker discrimination by explicitly inserting a frequency-wise attention module in a DNN-based speaker recognition task. Our method consists of a frequency-wise attention architecture and a simple ResNet, which is illustrated in Fig. 3.2, to learn the importance of each frequency band. Motivated by channel-wise attention in image recognition [110], we designed the frequency-wise attention module to map the input feature $\mathbf{X}$ to weighted feature $\widetilde{\mathbf{X}}$, and $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_D}]$, where $\mathbf{x}_i \in R^{w \times h}$ represents the i-th frequency band of input feature $\mathbf{X}$, h is the frame index, $w = 1$, D is the number of frequency components. Specifically, convolution operations are first carried out in $\mathbf{x}_i$ along the time axis using a shared one-dimensional convolution layer. We then apply global average pooling (GAP) for each channel to obtain the channel feature $\mathbf{Z}$:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_D], \tag{3.3}$$

where $\mathbf{z}_i$ represents the feature of the i-th frequency component. The importance of each frequency component $\widetilde{\mathbf{Z}}$ (attention layer) can be learned after using the softmax function. The weighted feature map is calculated as follows:

$$\widetilde{\mathbf{X}} = \widetilde{\mathbf{Z}} \otimes \mathbf{X}, \tag{3.4}$$

where $\otimes$ represents the outer product of vectors. The architecture of this ResNet-based SV system is shown in Fig. 3.3. Three one-dimensional convolutional layers combined with three residual blocks are used to generate a frame-level feature for utterance X_i. For the three convolutional layers, the kernel size is (5×5) and the number of channels varies from 64 to 256. For each residual block, two convolution layers with the same kernel size (3×3) and stride (1×1) and a rectified-linear-unit function are used in the back of the first convolution layer. After the average pooling layer, segment-level speaker embeddings can be extracted from a 1024-dimensional fully connected layer (FCL). Then, the embeddings are mapped into a number corresponding to that of speakers in the training data. Finally, we use the cross-entropy loss as an objective during the optimization of the entire network.

3.2.3 Extraction of Cepstral Features Using Data-Driven Non-uniform Filterbank

The previous two sections describe how to quantify the importance of different frequencies. By using the above method, we can get the importance weights for different frequencies. These weights can be used to design data-driven non-uniform

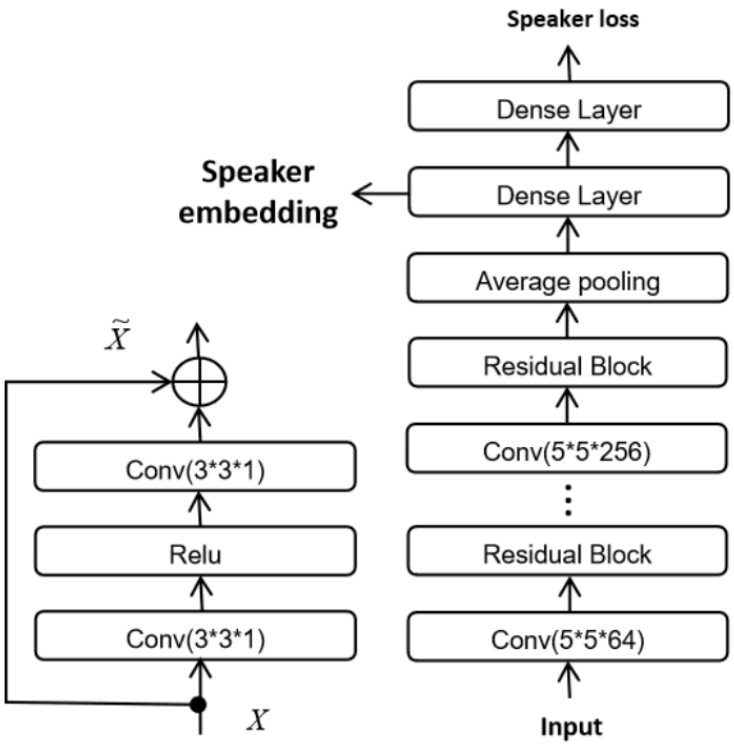

Figure 3.3: Architecture of residual network (ResNet).

filterbanks (UNF), which then be used to extract cepstral features, named non-uniform filterbank cepstral coefficient (NUFCC).

Specifically, to emphasize the importance of frequency regions with relatively high quantification scores, the distribution density of the triangular band-pass filters is assigned to be directly proportional to the average quantification score (Q_{score}). Q_{score} is calculated by:

$$Q_{score} = \frac{\sum_{i=1}^{N} \widetilde{\mathbf{Z}}_i}{N}, \tag{3.5}$$

where N is the number of utterances, $\widetilde{\mathbf{Z}}_i$ is the quantification score of ith utterance. The steps for designing an NUF are as follows:

- calculate the weight k based on the Q_{score}, $k = fs/(2 \times Sum(Q_{score}))$, where fs is the sampling rate,
- calculate the cumulative Sum of weighted Q_{score}, $Sum = Cumsum(k \times Q_{score})$,
- fit the curve of the mapping frequency from the linear scale to the adaptive scale by cubic spline interpolation,
- calculate the center frequency of the triangular band-pass filters $C(i)$ based on the fitting curve, and
- design a NUF with the same bandwidth.

The cepstral feature (NUFCC) is calculated by applying a discrete cosine transform (DCT) to the log filterbank outputs. DCT plays a significant role in compressing the information contained in the filterbank outputs and reducing the dimensionality of the feature representation. This operation is the same as the MFCC feature extraction.

3.3 Temporal Features Using Jitter and Shimmer

Prosody refers to the melodic and rhythmic aspects of speech, including variations in pitch, loudness, duration, and intonation. Jitter and shimmer are acoustic measures that provide information about the stability and irregularities in vocal fold vibration and intensity. These measures can be related to prosody because variations in vocal stability and irregularities can affect the melodic aspects of speech, such as pitch and loudness modulation. Previous studies have demonstrated the

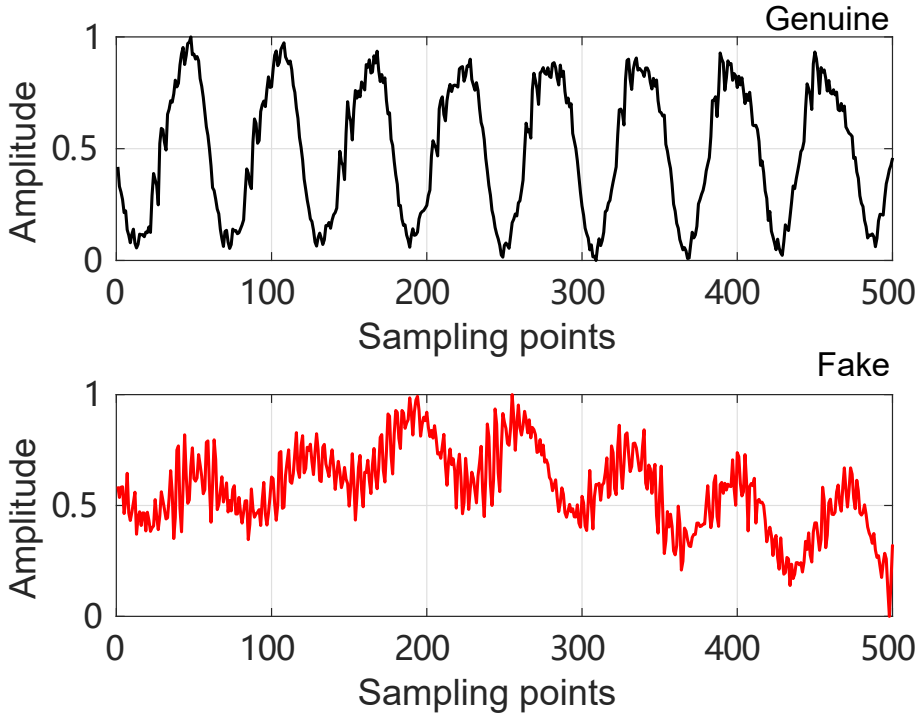

Figure 3.4: Comparison of differences among genuine and fake speech waveforms. These segments retain the same linguistic content (/i/). The sampling frequency used for the comparison is 16 kHz.

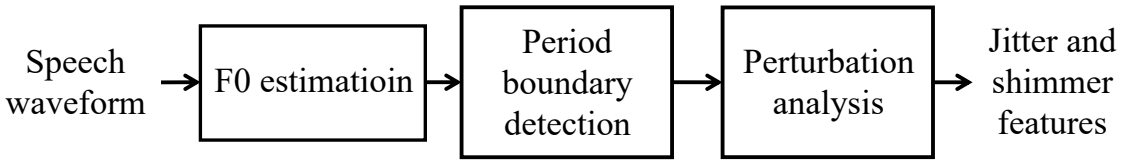

Figure 3.5: Extraction process of jitter and shimmer features.

efficacy of these features in characterizing voices with pathological prosody [111]. It is reasonable to regard jitter and shimmer as valuable features for distinguishing between genuine and fake speech.

Jitter and shimmer features represent variations in F_0 and amplitude of adjacent glottis periods, respectively. They reflect the characteristics of amplitude and frequency perturbation (AFP). To illustrate the disparity in AFP between genuine and fake speech, particularly under degraded speech quality, two segments of genuine and fake speech were chosen with identical linguistic content (/i/). These segments are depicted in Figure 3.4. To visualize the difference in AFP, an amplitude normalization operation was conducted, scaling the amplitudes to the range of [0,1]. From Figure 3.4, it is evident that the stability of the fake-speech segment notably decreased in terms of both amplitude and frequency/period.

As shown in Fig. 3.5, the extraction process of jitter and shimmer involves three essential steps [112]. First, the F_0 is estimated using general F_0 estimation methods. The estimated F_0 contour is used as a "reference" signal for further period detection. Therefore, the accuracy of F_o estimation directly affects the effectiveness of jitter and shimmer features. Second, the boundary of each fundamental period is detected using waveform matching with a phase constraint algorithm. Lastly, jitter and shimmer are calculated by considering several adjacent periods. In this section, we roughly introduce three state-of-the-art methods for estimating F_0: IRAPT [113], YIN [114], and SWIPE [115]. Subsequently, we provide the calculation method for the jitter and shimmer features.

3.3.1 F_o Estimation Algorithms

Choosing an appropriate F_0 estimation algorithm entails considering several trade-offs. These include the upper and lower bounds of the F_0 search range, time and frequency resolution, robustness, computational complexity, and delay.

IRAPT: The main target of the IRAPT algorithm [113] is to estimate the instantaneous pitch values accurately, particularly in scenarios where there are rapid frequency modulations or noisy conditions. The IRAPT algorithm utilizes a robust framework that is less sensitive to rapid frequency changes and noise. Although designed to be robust, the IRAPT algorithm may still be influenced by

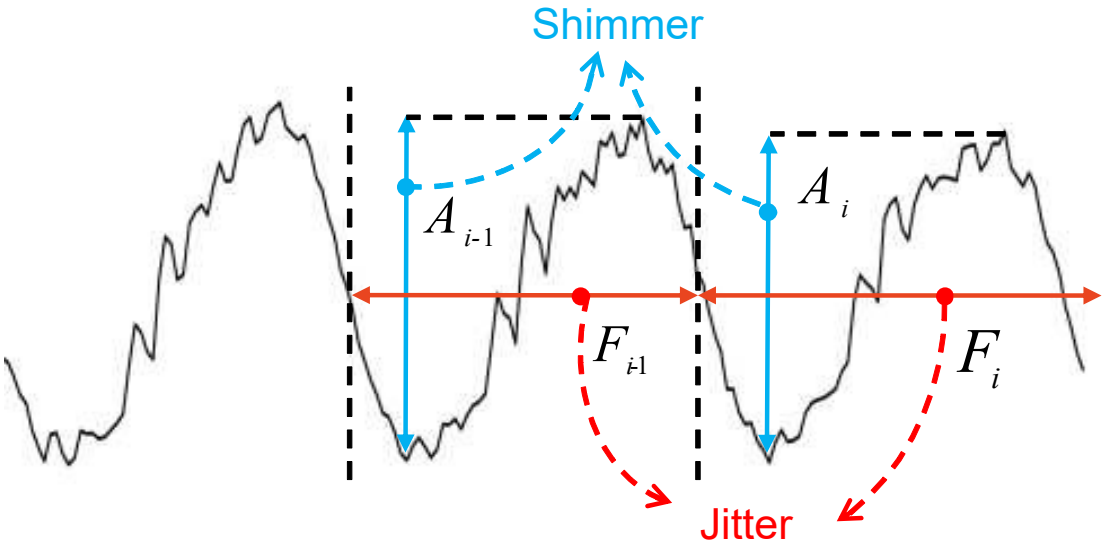

Figure 3.6: Schematic diagram of calculation of jitter and shimmer. A_i refer to the amplitude of the ith period, and F_i refer to the frequency of the ith period

certain artifacts or specific types of noise, which can affect the accuracy of pitch estimation. In addition, the algorithm may perform less accurately for extreme pitch ranges, where the instantaneous pitch values exhibit significant variations.

YIN: The YIN algorithm is based on the concept of the autocorrelation function and is particularly effective in handling non-periodic and noisy signals. The YIN algorithm provides an effective method for F_0 estimation, particularly in speech and music signals, with notable advantages in terms of computational efficiency and noise robustness. However, its applicability may be limited in certain scenarios with complex harmonic content or overlapping sounds.

SWIPE: The main theory behind the SWIPE algorithm is inspired by the sawtooth waveform, which is known for its periodic nature. The algorithm utilizes a comb-filtering approach to identify the F_0 by searching for the best match between the input signal and a series of synthetic sawtooth waveforms with varying periods. The key idea is to find the period that produces the highest correlation or similarity measure between the synthetic waveform and the signal being analyzed. The SWIPE algorithm offers a robust and efficient approach to F_0 estimation, particularly in scenarios involving speech and music signals.

3.3.2 The Definition of Jitter

Jitter, as depicted in Fig. 3.6, measures the variability of the fundamental period between consecutive periods, representing short-term variations rather than volun-

tary changes in F_0. In this paper, the jitter is utilized to provide some information related to the stability of speech-synthesis systems. Building upon the findings in [112], the following jitter features are considered:

1. average and continuous differences of jitter between consecutive periods ($AJ1/CJ1$);
2. relative average and continuous perturbation of jitter, which evaluates the smoothness of period duration over 3 adjacent periods ($AJ2/CJ2$) [116];
3. average and continuous period-perturbation quotient of jitter, which quantifies the pitch period variability over 5 consecutive periods ($AJ3/CJ3$);
4. average and continuous frequency perturbation quotient of jitter ($AJ4/CJ4$), which aims to eliminate the influence of frequency "drift" and provide a more accurate index of underlying jitter in 55 consecutive periods.

Let $F(i)$ represent the frequency of the ith fundamental period in an utterance. The parameters L_p, with $p = 2, 3, 4$, represent the number of consecutive periods used in calculating $AJ2/CJ2$, $AJ3/CJ3$, and $AJ4/CJ4$, respectively. Specifically, L_2 is set to 3, L_3 to 5, and L_4 to 55. N is the total number of fundamental periods. With these definitions, we can calculate $AJ1/CJ1$, $AJ2/CJ2$, $AJ3/CJ3$, and $AJ4/CJ4$ as follows:

$$AJ1 = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|F(i) - F(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{3.6}$$

$$CJ1 = \frac{|F(i) - F(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{3.7}$$

$$AJ_P = \frac{\frac{1}{N-L_p+1}\sum_{i=1+\frac{L_p-1}{2}}^{N-\frac{L_p-1}{2}}|F(i) - \widetilde{F(i)}|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{3.8}$$

$$CJ_P = \frac{|F(i) - \widetilde{F(i)}|}{\frac{1}{N}\sum_{i=1}^{N}F(i)} \times 100, \tag{3.9}$$

where

$$\widetilde{F(i)} = \frac{1}{L_p}\sum_{k=i-\frac{L_p-1}{2}}^{i+\frac{L_p-1}{2}}F(k). \tag{3.10}$$

3.3.3 The Definition of Shimmer

Shimmer, a measure of variation in expiratory flow during articulation, has been successfully utilized in previous studies [112]. The ADD2022 and ADD2023 databases exhibit frequent amplitude variations in fake audio. Therefore, in this paper, we explore the potential usefulness of shimmer as a feature in FAD. Five shimmer features are considered for analysis. The first feature, $AS1/CS1$, represents the average and continuous basic shimmer measure. It is defined as the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude [116]. To mitigate the influence of long-term changes in vocal intensity on $AS1/CS1$ and obtain a more effective representation of shimmer, we calculate four additional amplitude-perturbation quotients of shimmer. These are denoted as $AS2/CS2$, $AS3/CS3$, $AS4/CS4$, and $AS5/CS5$. The computation of these shimmer features follows a similar approach as that used for jitter. The calculations for the shimmer features are presented as follows:

$$AS1 = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|A(i)-A(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}A(i)} \times 100, \tag{3.11}$$

$$CS1 = \frac{|A(i)-A(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}A(i)} \times 100, \tag{3.12}$$

$$AS_P = \frac{\frac{1}{N-L_P+1}\sum_{i=1+\frac{L_P-1}{2}}^{N-\frac{L_P-1}{2}}|A(i)-\widetilde{A(i)}|}{\frac{1}{N}\sum_{i=1}^{N}A(i)} \times 100, \tag{3.13}$$

$$CS_P = \frac{|A(i)-\widetilde{A(i)}|}{\frac{1}{N}\sum_{i=1}^{N}A(i)} \times 100, \tag{3.14}$$

where

$$\widetilde{A(i)} = \frac{1}{L_P}\sum_{k=i-\frac{L_P-1}{2}}^{i+\frac{L_P-1}{2}}A(k). \tag{3.15}$$

A(i) is the amplitude of the ith period, L_p, with $p = 2, 3, 4, 5$, represent the number of consecutive periods used in calculating $AS2/CS2$, $AS3/CS3$, $AS4/CS4$ and $AS5/CS5$, respectively. Specifically, L_2 is set to 3, L_3 to 5, L_4 to 11, and L_5 to 55.

3.4 Spectral Temporal Modulation Representations

The calculation processes of STM representations are depicted in Fig. 3.7. Firstly, different filterbanks are used to decompose the input signal $x(t)$ into a series of frequency bands. In this step, we choose the MFB, GTFB and NUF. The calculation of NUF has been introduced in Section 3.2.3. The design of GFB is based on the ERB scale. The shape of a Gammatone filter has been visualized in Fig. 2.2. The center frequencies and bandwidth of Gammatone filters can be derived based on the number of filterbanks and ERB expression. Therefore, the GFB is also named ERB filterbank. The thesis set the number of channel as 80, the frequency range from 60 to 7600Hz. The impulse response of k-th Gammatone filter can be represented as:

$$g_k(t) = At^{(n1)}exp(2\pi b_f ERB(f_k)t)cos(2f_k t), \tag{3.16}$$

with

$$ERB = 24.7(4.37f_k + 1), \tag{3.17}$$

where A refers to the amplitude, n is the order of GFB, which is set as 4 in our experiments, b_f represents the bandwidth of the k-th filter, and f_k is the center frequency if k-th filter. The output of the k-th channel is expressed as:

$$y_k(t) = g_k(t) * x(t), \tag{3.18}$$

where $*$ is the convolution.

Secondly, the Hilbert transform and squared operations are implemented to compute the power envelope of each frequency band. Following is the down-sampling step, which aims to convert the modulation sampling frequency to an audible range and decrease the feature dimension. The Hilbert transform that we used is a standard function from MATLAB. Let's define LPF as a low-pass filter with a cut-off frequency of 160Hz. The output of LPF is:

$$e_k^2(t) = LPF[|Hilbert(y_k(t))|^2], \tag{3.19}$$

Lastly, a two-dimensional FFT is performed to transfer the resampling power envelope to the STM representations. The absolute value is calculated as the final representation.

$$STM = |2DFFT(e_k^2(t))|, \tag{3.20}$$

The STM representations represent the dynamic variations present in the speech signal across different spectral and temporal scales.

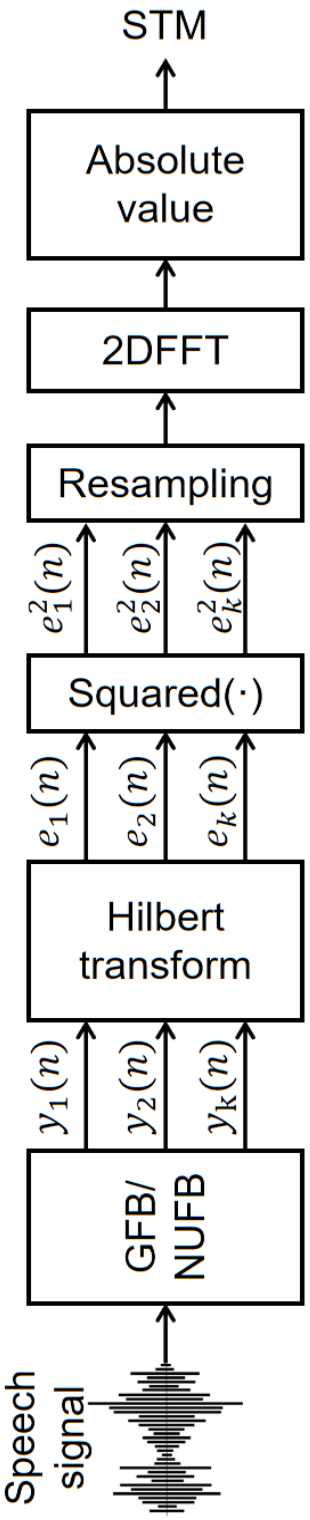

Figure 3.7: The calculation procedure of STM representations.

Chapter 4

Speech Security Using Speaker Identity Verification

The diversity of speech organs (e.g., the glottis [102], nasal cavity [117, 103, 104], piriform fossa cavities [118, 119], and vocal tract length [120]) non-uniformly provides speaker-dependent information to different frequency components in the acoustic spectrum. We believe that quantifying the effect of frequency components on the speaker identity can help us to understand the relationship between speakers' physiological structure and acoustic speech signals and extract more speaker discriminative information.

This Chapter is the application of proposed feature representations in the speaker identity verification task. The quantification methods have been introduced in Section 3.2. In this chapter, first, the i-vector-based ASV system is introduced. The quantification results of frequency importance for the ASV task are then introduced. Finally, the effectiveness of proposed NUFCC and STM representations are evaluated using the I-vector-based ASV system.

4.1 Application of Proposed Feature Representations to I-vector-based SV

I-vector is proposed by Dehak [121]. It combines the subspace of speaker difference and the subspace of channels together in the modeling stage, which reduces the limitation on the training corpus. Also, the calculation is simple, and the performance is stable. Given a speech, its Gauss mean supervector M can be divided into the following form:

$$M = m + T\omega \tag{4.1}$$

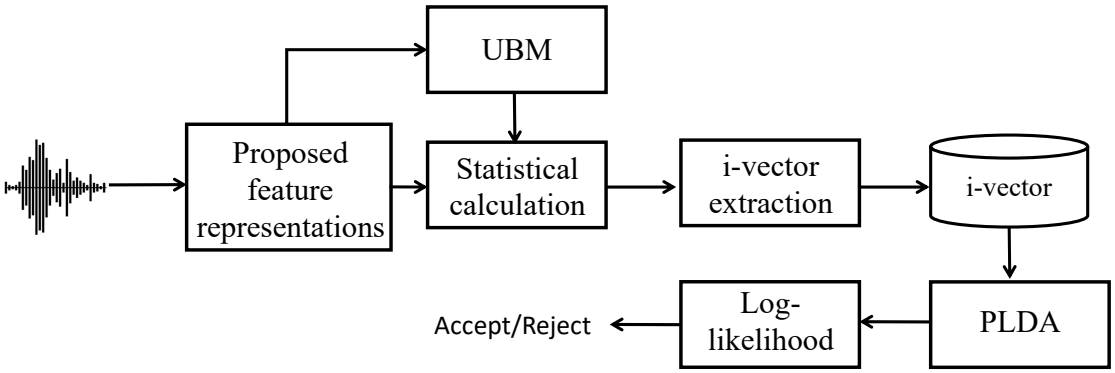

Figure 4.1: The flow diagram of i-vector-based ASV.

where m refer to UBM mean supervector, it has nothing to do with the specific speaker and the channel information, and M is the UBM mean supervector for the given speech, and T is the global differential space matrix. ω is the change factor containing the speaker and channel information, namely I-vector. In the training and feature mapping process of the i-vector model, sufficient statistics of each speech segment should be calculated (Baum-Welch statistics):

$$N_c^{(k)} = \sum_t \gamma_{c.t}^{(k)} \tag{4.2}$$

$$F_c^{(k)} = \sum_t \gamma_{c.t}^{(k)} O_t^{(k)} \tag{4.3}$$

$$S_c^{(k)} = \sum_t \gamma_{c.t}^{(k)} O_t^{(k)T} \tag{4.4}$$

where $N_c^{(k)}$, $F_c^{(k)}$ and $S_c^{(k)}$ represent the zero-order statistics, first-order statistics, and second-order statistics of speech segment K on the c-th GMM component, $O_t^{(k)}$ represents the acoustic characteristics of speech segment k at time index t, $\gamma_{c.t}^{(k)}$ represents the posterior probability of acoustic characteristics $O_t^{(k)}$ on the c-th GMM component.

$$\gamma_{c.t}^{(k)} = \frac{\widetilde{\omega}_c N_{UBM}(O_t^{(k)}; \mu_c, \sum_c)}{\sum_{c'=1}^{C} \widetilde{\omega}_{c\prime} N_{UBM}(O_t^{(k)}; \mu_{c\prime}, \sum_{c\prime})} \tag{4.5}$$

where c denotes the total number of mixed components, $\widetilde{\omega}_c$, μ_c and $\sum_c$ correspond to the weighted of the c_{th} Gauss components, mean and covariance respectively. The i-vector is able to cope with more massive data due to its significantly reduced computational load. At the same time, i-vector has an excellent performance in cross-channel situations. The flow diagram of i-vector-based ASV are depicted in

Table 4.1: Statistics of training and testing sets

Set	Speakers	Utterances
Train	70	6,999
Test	30	2,998
Total	100	9,997

Fig.4.1. In this thesis, different features are compared based on the i-vector ASV system.

4.2 Experiment data and matrix

The Japanese versatile speech corpus [122] consists of audios from 100 native Japanese speakers. The database was recorded in a clean environment at a 24-kHz sampling rate. To train our model, we selected a set of 9,997 sentences from 100 speakers to learn the nonlinear combined effect of the frequency components on speaker identity. All the sentences were downsampled from 24 kHz to 16 kHz. The average length of each utterance was 7.92 s, and the total length of the speech data was 21.86 hrs. In i-vector-based ASV, the same speech data introduced above was divided into training (70 speakers) and testing (30 speakers) sets. The details of the training and testing sets are summarized in Table 4.1.

4.3 Experimental setting

We used power spectrum (PS), subband power spectrum (SPS), and subband log power spectrum (SLPS) as front-end input of the proposed frequency-wise attentional neural network. The extraction of these three features was without frequency warping operations. The filterbank used for SPS and SLPS features extraction was a triangular band-pass filter with a linear frequency scale, and the dimension was set to 512.

In the frequency-wise attention module shown in Fig. 3.2, the kernel size of the one-dimensional convolution layer is (5×1), and the number of output channels is 64. The dimension of the attention layer corresponds to the number of frequency components that were set to 512. In the ResNet-based backbone module, three one-dimensional convolutional layers combined with three residual blocks are used to generate a frame-level feature for an utterance. For the three convolutional layers, the kernel size is (5×5) and the number of channels varies from 64 to 256. For each residual block, two convolution layers with the same kernel size (3×3) and stride (1×1) and a rectified-linear-unit function are used in the back of

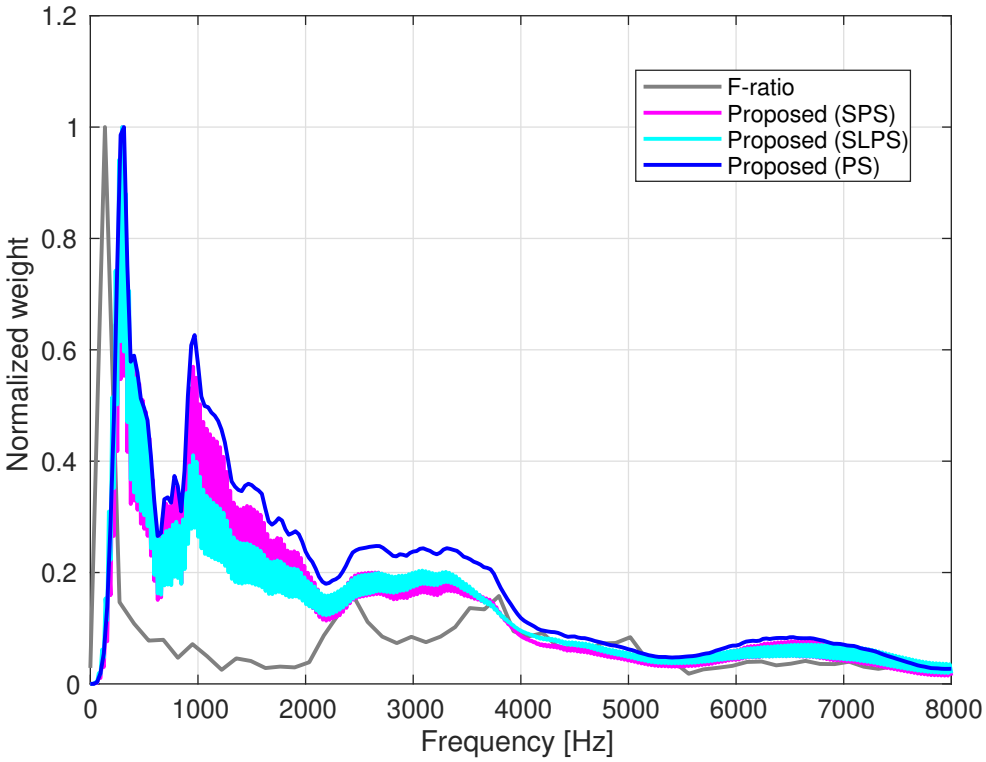

Figure 4.2: Comparison of quantification results from using F-ratio-based and proposed quantification methods. Three different features were used as front-end input of the proposed architecture.

the first convolution layer. After the average pooling layer, segment-level speaker embeddings can be extracted from a 1024-dimentional fully connected layer. Then, the embeddings are mapped into a number corresponding to that of speakers in the training data. In the training stage, 3-s utterances, including 300 frames, were randomly selected from each raw waveform, and 128 utterances were grouped as one batch fed into the NN. The number of training epochs is set to 30.

For i-vector-based ASV, the UBM and i-vector extractor were trained on the training set, and 30,000 test pairs, including half positive trials and half negative trials, were randomly generated from the testing set. The Gaussian mixture number of UBM was set to 128, and the dimension of the i-vector was set to 300. We used the equal error rate (EER) and minimum decision cost function (minDCF) with $P_{target} = 0.01$ as the evaluation metrics of the ASV [123].

4.4 Analysis of Frequency Importance for ASV

Figure 4.2 shows the speaker discriminative abilities of each frequency component quantified using the F-ratio-based quantification method and DNN-based quantification method. The comments in the parentheses refer to the front-end input feature types for the training of our method. We compare the two methods by showing all the results using normalization with values ranging from 0 to 1. The quantification results show the distribution of speaker discriminative information in the frequency domain was non-uniform and most of the discriminative information was concentrated in the low-frequency region. Using our method with different input features, we could obtain consistent results with peaks and valleys located in similar frequency regions on the curves. Moreover, the quantification results with PS as an input feature had fewer fluctuations than others.

Figure 4.3 illustrates the normalized plot of different frequency warping scales to compare our method with other methods. We can observe that the frequency warping based on data-driven methods has a high-frequency resolution in the low-frequency regions (below 400 Hz), which is discriminative information expected from the glottis based on knowledge from [102]. Compared to the F-ratio-based quantification method, the normalized scale from our method (red-solid curve) indicates a higher frequency resolution from 1 kHz to 2 kHz. Based on [103] and [104], this peak is possibly related to the antiresonance contributed by the nasal cavity and sinuses, which is another essential factor for speaker discrimination that the F-ratio method could not reveal.

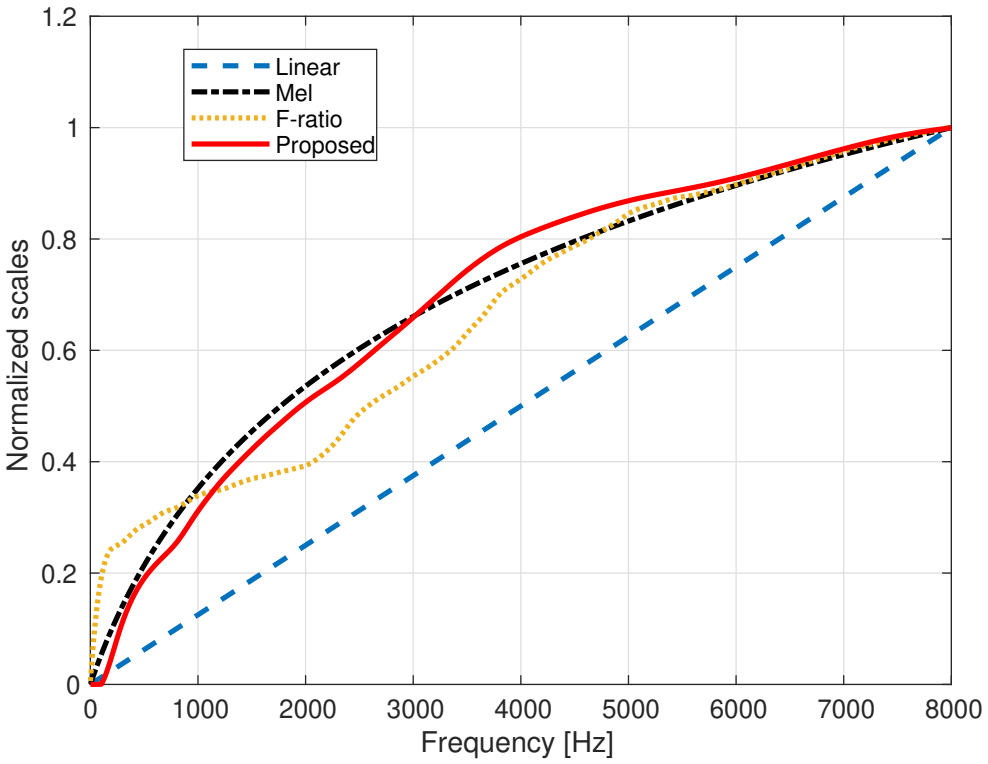

Figure 4.3: Frequency warping for linear, Mel, F-ratio, and proposed scale.

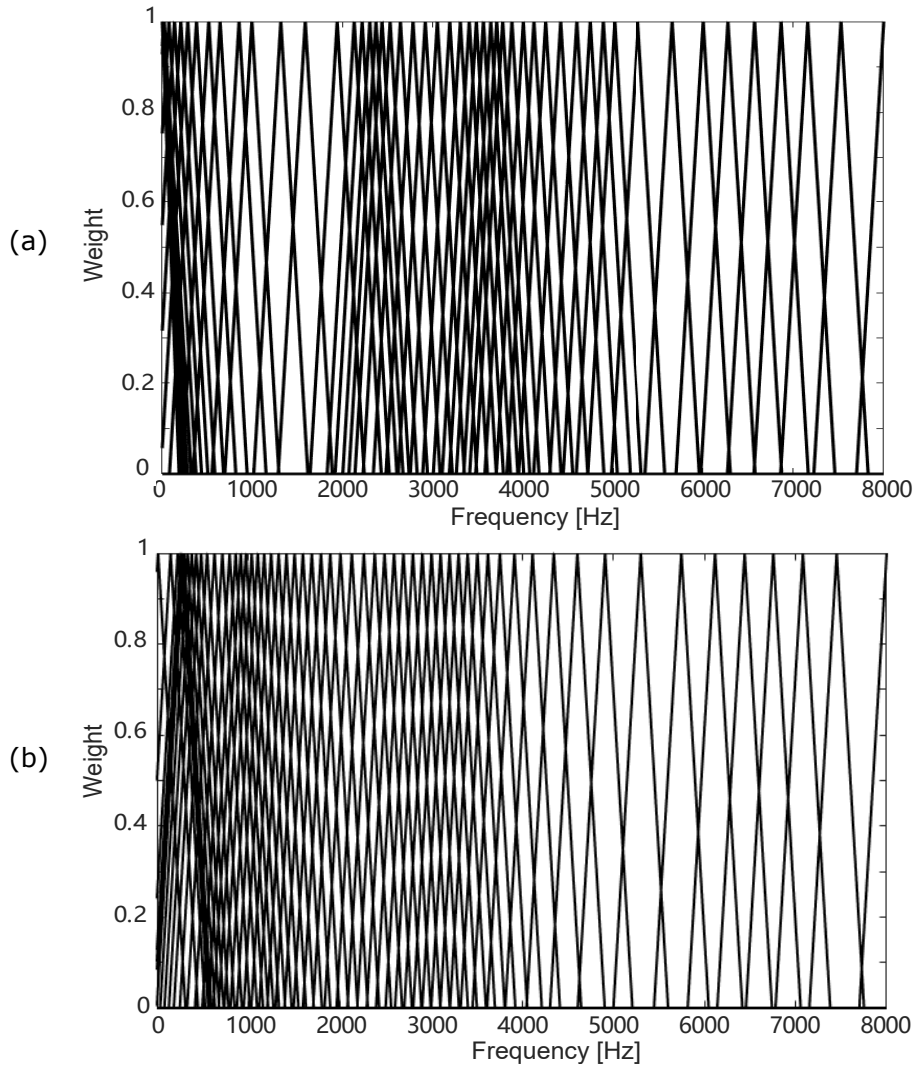

Figure 4.4: Comparison of NUFs designed with F-ratio-based method (a) and proposed method (b). Number of filters was 60, and bandwidth of each sub-band filter was fixed.

Table 4.2: Results of the proposed feature representations in i-vector-based ASV systems. The results are shown in terms of EER and minDCF based on a Japanese databases.

Acoustic feature	EER (%)	minDCF (0.01)
UFCC	3.092	0.417
MFCC	2.977	0.363
GTCC	1.908	0.204
F-ratio-based NUFCC	2.084	0.215
DNN-based NUFCC (SLPS)	1.800	0.219
DNN-based NUFCC (PS)	1.698	0.236
DNN-based NUFCC (SPS)	**1.597**	**0.206**
STM	9.47	0.680

4.5 Results and discussion

Based on different frequency warping scales, NUF can be designed using the steps described in Section 3.2.3. Two examples of the designed NUF are depicted in Fig. 4.4. The specially designed NUF can then be used to extract the cepstral features for ASV.

The ASV results in Table 5.3 indicate that acoustic feature extraction using an NUF can substantially improve speaker discrimination abilities. In addition, NUFCC extraction with our method can perform better than the F-ratio-based quantification method in both EER and minDCF. This indicates that the designed NUF can work well in the cepstral domain. This also indicates that the quantification results from using our method can capture more speaker discriminative factors, such as the relationships among different frequency bands. The NUFCC feature designed with our quantification method using SPS as input decreases the EER from 2.084% (F-ratio-based method) to 1.597, resulting in a relative improvement of 23.4%.

However, the STM representations derived from the NUF cannot further improve the performance according to our current experiments. This may because of the i-vector that we used initially designed for short-term features. Due to the broader temporal context of STM representations, there is generally a lower count of available training vectors, which will decrease the discriminative ability of the i-vector-based ASV system. More sophisticated DNN architectures will be designed to be more suitable for the proposed feature representations.

4.6 Summary

In this chapter, we quantified the nonlinear combined effect of frequency components on speaker identity. A frequency-wise attention structure combined with a ResNet was designed to learn the importance of different frequency bands by considering resonance and antiresonance. The quantification results with our method using three input features consistently indicated that SDI is non-uniformly distributed in the frequency domain and most of the discriminative information is concentrated in the low-frequency region. In addition, the quantification results from using our method indicated that the antiresonance frequency induced by the nasal cavity from 1 kHz to 2 kHz is another essential factor for speaker discrimination that the F-ratio method could not reveal. To further evaluate our findings, we designed a non-uniform subband processing strategy based on the quantification results using our method for speaker feature extraction and did ASV. Finally, compared with the NUFCC designed with the F-ratio-based method, the NUFCC designed with DNN-based method achieved 23.4% relative improvement in EER. These results also confirmed that further emphasizing the spectral structure around the antiresonance frequency region could enhance speaker discrimination. Finally, we try to use the proposed DNN-based NUF to conduct STM analysis. Unluckily, the current results by using the STM representations cannot achieve good performance in the ASV task. More attempts will be made in the future.

Chapter 5

Secure Speech Communication based on FAD Approach

Usually, the distinctions between genuine and fake speech stem from the challenging issue of unnaturalness in speech synthesis [124]. Moreover, the unnaturalness in synthesized speech is often caused by the limitations in capturing and reproducing rich and diverse prosody information. Prosody related to representations of non-linguistic information of voice is a key issue for solving the unnaturalness of synthesized speech. Therefore, exploring prosody differences between genuine and fake speech holds great promise in providing discriminative information for FAD.

This chapter applies proposed feature representations, especially the jitter and shimmer features, to the FAD task to build a secure speech communication environment. First, the combination of proposed feature representations and the light convolutional neural network (LCNN)-based FAD System is introduced. Considering the differences in feature dimension, AFP (jitter and shimmer) features are used as complementary features for conventional features. Two different architectures are designed for the evaluation of different features. The architecture shown in Fig. 5.1 is used for the evaluation of AFP features. While architecture shown in Fig. 5.2 is used for the evaluation of STM representations. In the parts of the result, we analyze the differences between genuine and fake speech using temporal features. Finally, the evaluation results in the ADD2022 and ADD2023 challenges are reported.

5.1 Application of Proposed Feature Representations in LCNN-based FAD System

Previous studies [125, 126] have demonstrated that a shallow network is sufficient for downstream tasks, including anti-spoofing tasks. Therefore, this thesis

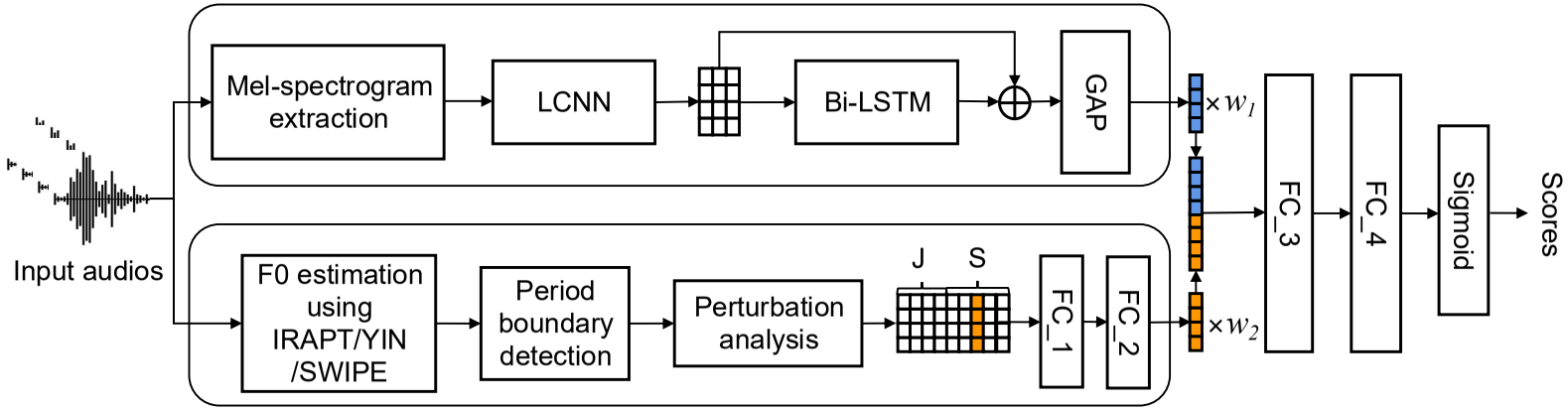

Figure 5.1: Proposed system used for the evaluation of AFP features. The jitter features, consist of $CJ1$, $CJ2$, $CJ3$, and $CJ4$, are denoted by J. The shimmer features, encompassing $CS1$, $CS2$, $CS3$, $CS4$, and $CS5$, are denoted by S.

chooses a LCNN [126] based architecture as the baseline system to evaluate the performance of all feature representations. As depicted in Fig. 5.2, this LCNN is accompanied by two Bi-LSTM, a global-average pooling layer, and two fully connected output layers [127]. The dimensions of the Bi-LSTM layers match the output dimensions of the LCNN. This specific architecture is commonly referred to as an LLGF network in the literature [91, 128].

Designing an FAD system that can combine the jitter and shimmer features with a conventional acoustic feature reasonably is also a challenging point. Fig. 5.1 illustrates the proposed FAD system, which combines jitter and shimmer features with a Mel-spectrogram. We adopt a late-fusion approach to add jitter and shimmer features to the baseline system. Specifically, the jitter and shimmer features are first extracted using the method introduced in Section 3.3. Next, these extracted features are utilized as input for two fully connected layers (FC_1 and FC_2). The resulting output from FC_2 is combined with the output of global average pooling (GAP) [129], employing distinct weights (w_1 and w_2). Different weights used here aim to regularize dynamic ranges of different features. This combined output is then passed into two additional fully connected layers (FC_3 and FC_4). Finally, to compute the score of each audio, a sigmoid function is employed. The detailed architecture of the LCNN-BLSTM model, including the kernel shape, the output shape of each layer, and the number of trainable parameters are listed in Table 5.1.

An objective function named binary cross entropy (BCE) is used for optimizing the model parameters. BCE is defined as:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N}[y_i \log P_\theta(\boldsymbol{x}_i) + (1 - y_i)\log(1 - P_\theta(\boldsymbol{x}_i))] \tag{5.1}$$

where N refers to the number of samples, θ denotes model parameters, y_i and $P_\theta(\boldsymbol{x}_i)$ are respectively the ground truth of the i-th training sample and its corresponding output probability from the model.

5.2 Experiment data and matrix

The datasets from the ADD2022 [90] and ADD2023 [130] challenges were selected to assess the effectiveness of the proposed method. These challenges aim to shape the future direction of detecting deep synthetic and manipulated audio in multimedia. In ADD2022, all tracks share the same training and development datasets, while an individual adaptation dataset is provided for fine-tuning and evaluation in each track. The ADD2023 comprises only the training and development datasets. For evaluation purposes, test datasets for ADD2022 and ADD2023 are available

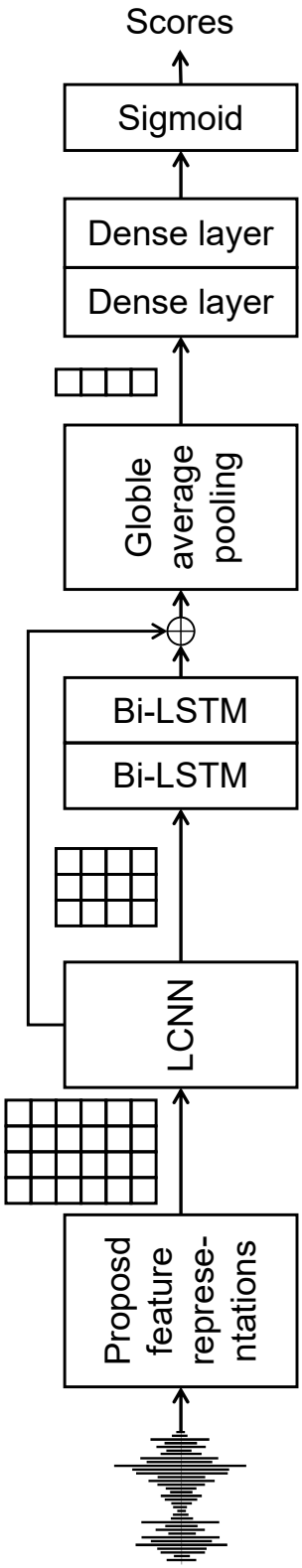

Figure 5.2: The LLGF system used for the evaluation of proposed feature representations.

Table 5.1: Architecture of LCNN-BLSTM-based deep classifier for FAD.

Type	Kernel Shape	Output Shape	Param
Feature_CS3	-	[64, 404]	-
Feature_Mel-spectrogram	-	[64, 80, 404]	-
Conv2d_0	[1, 64, 5, 5]	[64, 64, 404, 80]	1.66k
MaxFeatureMap2D_1	-	[64, 32, 404, 80]	-
MaxPool2d_2	-	[64, 32, 202, 40]	-
Conv2d_3	[32, 64, 1, 1]	[64, 64, 202, 40]	2.11k
MaxFeatureMap2D_4	-	[64, 32, 202, 40]	-
BatchNorm2d_5	-	[64, 32, 202, 40]	-
Conv2d_6	[32, 96, 3, 3]	[64, 96, 202, 40]	27.74k
MaxFeatureMap2D_7	-	[64, 48, 202, 40]	-
MaxPool2d_8	-	[64, 48, 101, 20]	-
BatchNorm2d_9	-	[64, 48, 101, 20]	-
Conv2d_10	[48, 96, 1, 1]	[64, 96, 101, 20]	4.704k
MaxFeatureMap2D_11	-	[64, 48, 101, 20]	-
BatchNorm2d_12	-	[64, 48, 101, 20]	-
Conv2d_13	[48, 128, 3, 3]	[64, 128, 101, 20]	55.42k
MaxFeatureMap2D_14	-	[64, 64, 101, 20]	-
MaxPool2d_15	-	[64, 64, 50, 10]	-
Conv2d_16	[64, 128, 1, 1]	[64, 128, 50, 10]	8.32k
MaxFeatureMap2D_17	-	[64, 64, 50, 10]	-
BatchNorm2d_18	-	[64, 64, 50, 10]	-
Conv2d_19	[64, 64, 3, 3]	[64, 64, 50, 10]	36.93k
MaxFeatureMap2D_20	-	[64, 32, 50, 10]	-
BatchNorm2d_21	-	[64, 32, 50, 10]	-
Conv2d_22	[32, 64, 1, 1]	[64, 64, 50, 10]	2.11k
MaxFeatureMap2D_23	-	[64, 32, 50, 10]	-
BatchNorm2d_24	-	[64, 32, 50, 10]	-
Conv2d_25	[32, 64, 3, 3]	[64, 64, 50, 10]	18.50k
MaxFeatureMap2D_26	-	[64, 32, 50, 10]	-
MaxPool2d_27	-	[64, 32, 25, 5]	-
Dropout_28	-	[64, 32, 25, 5]	-
LSTM_l_blstm	-	[25, 64, 160]	154.88k
LSTM_l_blstm	-	[25, 64, 160]	154.88k
GAP	-	[64,160]	-
FC_1	-	[64,256]	103.68k
FC_2	-	[64,160]	41.12k
Feature_Concat	-	[64, 564]	-
FC_3	-	[64, 128]	41.09k
FC_4	-	[64, 2]	258
Total	-	-	653.41k

Table 5.2: Statistics information for the training, development, adaptation, and test datasets of the ADD2022 and ADD2023 challenges. The duration values are presented in a format indicating the minimum, mean, and maximum durations.

Dataset		**Genuine**	**Fake**	**Total**	Duration (sec.)
ADD 2022	Training	3,012	24,072	27,084	0.86/3.15/60.01
	Development	2,307	21,295	223,602	0.86/3.16/60.01
	Adaptation	300	700	1,000	1.13/3.63/60.01
	Test	-	-	109,199	0.35/5.51/217.46
ADD 2023	Training	3,012	24,072	27,084	0.86/3.15/60.01
	Development	2,307	26,017	28324	0.86/3.16/60.01
	Test	-	-	11,8477	0.35/5.51/217.46

online. These datasets contain unseen audio samples obtained from various speech-synthesis systems. Notably, these samples present more real-life and challenging multimedia scenarios than those in the ASVspoof2021 challenge [88]. This paper uses the data from the track of low-quality FAD (LF) in ADD2022 and the track of audio fake game detection (FG-D) track in ADD2023. The difference between these two datasets is from the setting of the competition system, the FG-D track in the ADD2023 challenge includes two rounds of testing. The second round test is more difficult, so this paper considers the results from the second round only. Table 5.2 provides statistical information regarding these datasets.

The performance of the proposed FAD system was assessed using the equal error rate (EER), following the same evaluation method used in the ADD2022 and ADD2023 challenges.

5.3 Experimental setting

The Mel-spectrogram was extracted by using the $MelSpectrogram$ module in the $torchaudio.transforms$ library [131]. The parameters used in the STFT were configured as follows: a fast Fourier transform size of 1024, a window length of 512, and a hop length of 256. In cases where the audio duration is shorter than 4 seconds, zero padding is applied. The resulting Mel-spectrogram has dimensions of $[64 \times 80 \times 404]$, denoting the batch size, number of Mel filterbanks, and number of frames. The YIN and SWIPE algorithms, implemented through the $libf0$ toolbox [132], are used in this paper. The dimension of each continuous shimmer feature is $[64 \times 1 \times 404]$.

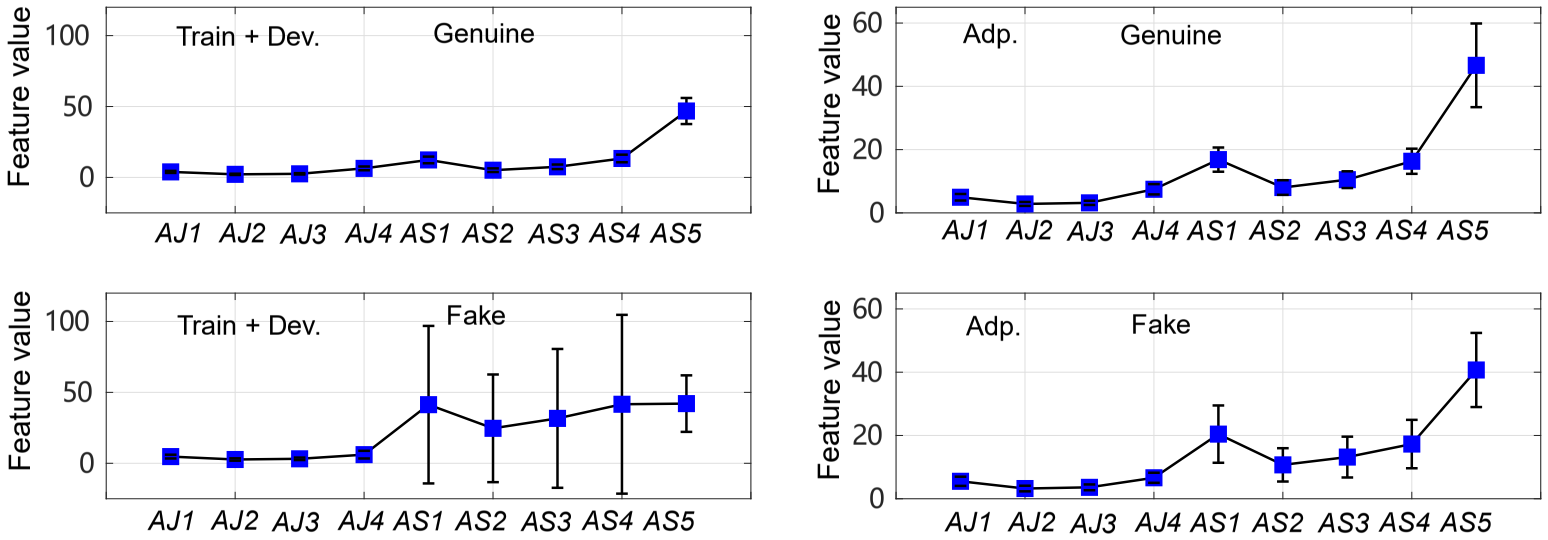

Figure 5.3: Statistical results using means and variances of averaged jitter and shimmer features in both Train + Dev. and Adp. datasets.

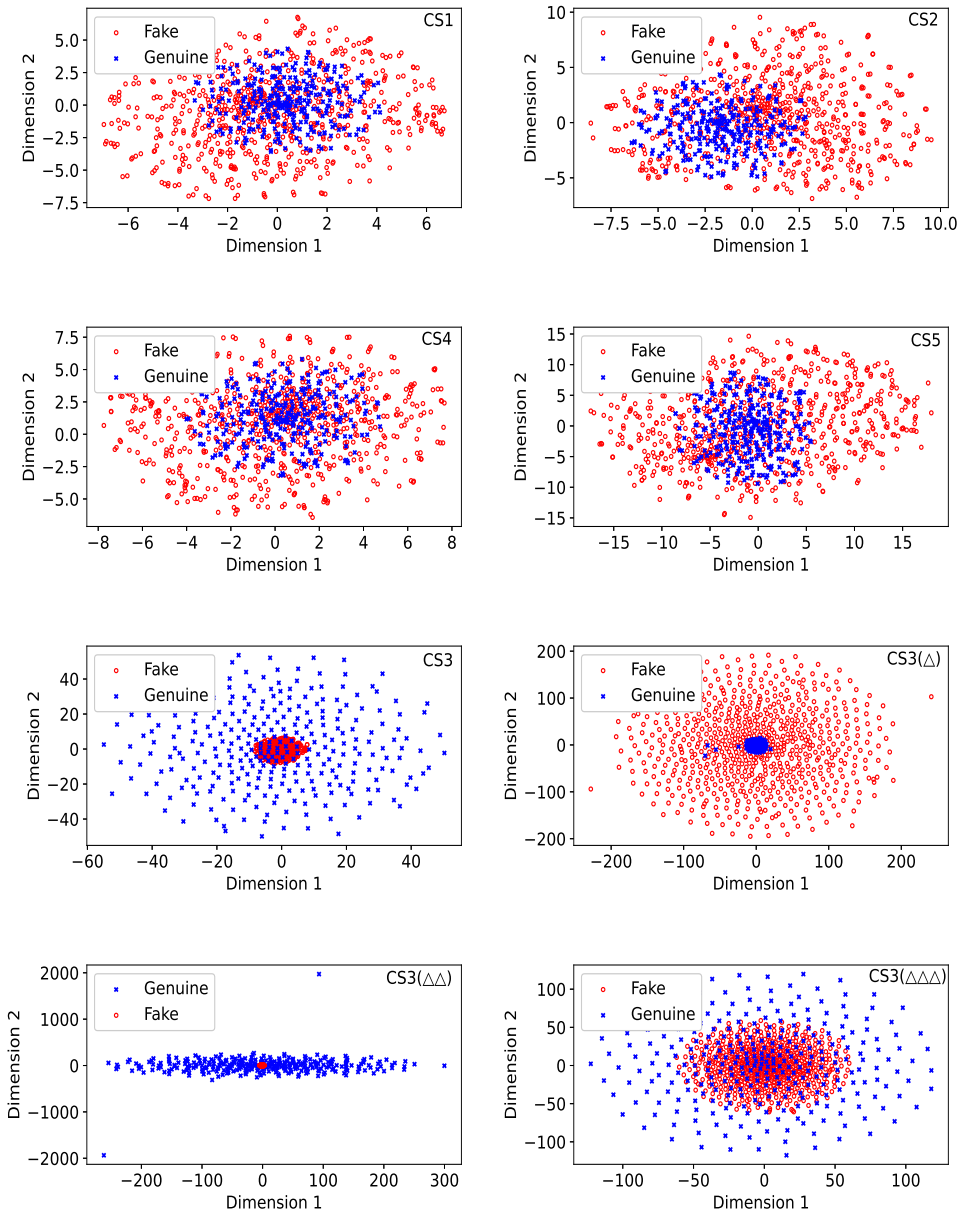

Figure 5.4: Comparison of discrimination of CS1, CS2, CS3, CS4, CS5, CS3 (△), CS3 (△△), and CS3 (△△△) for the ADD2022 adaptation set. The dimensions of these features were decreased to two and plotted by using the t-SNE toolkit [3].

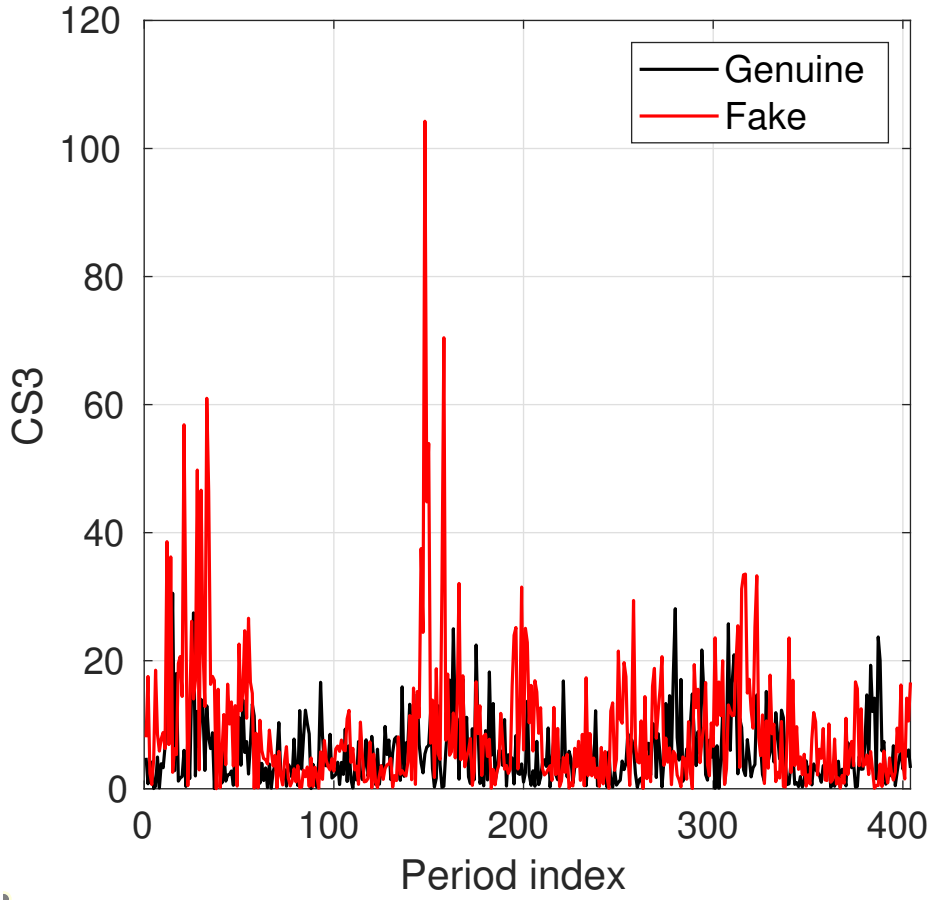

Figure 5.5: Comparison of CS3 features extracted from genuine and fake speech.

Data augmentation techniques (including the introduction of reverberation, babble, and music noise) were used during the extraction of the Mel-spectrogram feature to enhance the diversity of the training dataset, hence enhancing the robustness of the back-end classifier. Additionally, voice activity detection (VAD) [133] was utilized in the pre-processing stage to minimize disturbances caused by silence clips. The training process consisted of 30 epochs, and the model that yielded the best results was compiled using the Adam optimizer, with a learning rate of 0.0001.

5.4 Analysis of Differences Between Genuine and Fake Speech Using Temporal Features

This section examines the discrimination of various jitter and shimmer features by using statistical methods. The aim is to identify effective features for FAD. The most promising features are selected and combined with the Mel-spectrogram feature as the front-end input for an LCNN-BLSTM classifier.

The mean and variance of four averaged jitter features ($AJ1$, $AJ2$, $AJ3$, and $AJ4$) and five averaged shimmer features ($AS1$, $AS2$, $AS3$, $AS4$, and $AS5$) were analyzed in both the Train + Dev. (left column) and Adp. datasets (right column) as depicted in Fig. 5.3. The top graphs represent the results of genuine speech, while the bottom graphs illustrate the results of fake speech. Notably, the mean and variance of $AS1$, $AS2$, $AS3$, $AS4$, and $AS5$ exhibited an increase in the case of fake audio compared with genuine audio. This difference was more pronounced in the Train + Dev. dataset. It is also clear that amplitude perturbation in the fake audio is much more unstable. This amplitude-perturbation instability of fake audio could provide discriminative information for accomplishing FAD.

The continuous-shimmer features ($CS1$, $CS2$, $CS3$, $CS4$, and $CS5$) were calculated to capture continuous variations in the speech waveform, providing more discriminative information than the averaged shimmer features. PCA was used to visualize the discrimination capability of the continuous-shimmer features, reducing the dimensions to two and depicted in Fig. 5.4. Among these features, $CS3$ exhibited the fewest overlapping samples, indicating that it facilitated easier separation between genuine and fake speech. To enhance feature discrimination and incorporate dynamic variation characteristics, the first, second, and third derivatives of $CS3$ were considered ($CS3$ ($\triangle$), $CS3$ ($\triangle\triangle$), and $CS3$ ($\triangle\triangle\triangle$)), as depicted at the bottom of Fig. 5.4. It is evident from the figure that discrimination performance improved with the use of $CS3$ ($\triangle$) and further with $CS3$ ($\triangle\triangle$), with the best performance achieved by $CS3$ ($\triangle\triangle$). However, the discrimination performance of $CS3$ ($\triangle\triangle\triangle$) was lower than that of $CS3$ ($\triangle\triangle$). In general, $CS3$ and its dynamic features exhibit greater potential for successful FAD.

Fig. 5.5 illustrates the $CS3$ feature values for both genuine (blue) and fake (red) audio, providing an intuitive distinction between the two. The plot makes it evident that the CS3 feature exhibits more pronounced perturbation in fake audio than genuine audio.

5.5 Performance of Shimmer Features

This section is divided into two parts. The first part presents the results and discussion derived from ADD2022, demonstrating the efficacy of the shimmer features.

Table 5.3: FAD results (EER) in the adaptation (Adp.) and test sets of ADD2022 Challenge. Data augmentation and VAD are applied in the extraction of the Mel-spectrogram only.

Front-end Features	Data Augmentation (Mel-spectrogram)	VAD (Mel-spectrogram)	Results (EER %)	
			Adp. set	Test set
Mel-spectrogram	✗	✗	17.40	38.63
CS3	✗	✗	31.00	45.46
CS3 (△)	✗	✗	32.45	43.99
CS3 (△△)	✗	✗	32.00	42.49
CS3 (△△△)	✗	✗	37.00	45.83
Mel-spectrogram	✓	✗	3.31	33.47
Mel-spectrogram + CS3	✓	✗	4.31	32.48
Mel-spectrogram + CS3 (△)	✓	✗	3.90	31.95
Mel-spectrogram + CS3 (△△)	✓	✗	3.90	**31.50**
Mel-spectrogram + CS3 (△△△)	✓	✗	3.62	32.60
Mel-spectrogram + CS3 + CS3 (△△)	✓	✗	4.81	32.28
Mel-spectrogram	✓	✓	4.55	30.41
Mel-spectrogram + CS3 (△△)	✓	✓	3.31	**29.90**

Table 5.4: FAD results (EER) in the development (Dev.) and test set of ADD2023 Challenge. Different F_0 estimation methods, including IRAPT, YIN, and SWIPE, were utilized.

Front-end features	**Data augmentation**	**Loss**			**Dev. set (%)**			**Test set (%)**		
		IRAPT	YIN	SWIPE	IRAPT	YIN	SWIPE	IRAPT	YIN	SWIPE
Mel-spectrogram	✗		0.34			1.09			61.21	
Mel-spectrogram	✓		0.39			2.99			41.29	
Mel-spectrogram+CS3	✓	0.36	0.37	0.38	3.03	3.08	1.81	38.14	**36.63**	37.32
Mel-spectrogram+CS3(△)	✓	0.39	0.41	0.37	4.25	3.38	2.64	37.31	37.05	**36.70**
Mel-spectrogram+CS3(△△)	✓	0.37	0.39	0.38	2.95	3.46	2.21	39.98	**36.18**	41.24

Table 5.5: FAD results (EER) in the test set of ADD2023 Challenge using different combination weights between the Mel-spectrogram and CS3 △△ feature.

Front-end features	Data augmentation	Weight	Test set (%)
Mel-spectrogram+CS3(△△)	✓	4:1	38.79
Mel-spectrogram+CS3(△△)	✓	3:2	**35.77**
Mel-spectrogram+CS3(△△)	✓	1:1	36.18
Mel-spectrogram+CS3(△△)	✓	2:3	40.70
Mel-spectrogram+CS3(△△)	✓	1:4	40.14

The second part encompasses the results and discussion obtained from ADD2023, focusing on the utilization of various F_0 estimation methods.

5.5.1 Results and discussion in ADD2022

The discrimination performance of $CS3$ and its dynamic features ($CS3$ (△), $CS3$ (△△), and $CS3$ (△△△)), measured by the EER, is presented in Table 5.3 and categorized into three parts on the basis of the utilization of data augmentation and VAD methods. We focus on the results obtained from the test dataset only, as they exhibit the same trend as the Adp. dataset. The baseline system, which utilizes an LCNN-BLSTM model with a Mel-spectrogram as input, achieved an EER of 38.63%. However, the static and dynamic CS3 features yielded higher EERs than the Mel-spectrogram. It is important to note that the dimension of the Mel-spectrogram is 80 times larger than that of the shimmer features. A specific-designed classifier could further improve the performance of shimmer features.

By incorporating additional reverberation, noise, and music during Mel-spectrogram extraction, the EER was decreased to 33.47%. Combining the Mel-spectrogram with $CS3$, $CS3$ (△), $CS3$ (△△), and $CS3$ (△△△) further decreased the EER, which is consistent with the results of statistical analysis. The best result (31.50%) was achieved when combining the Mel-spectrogram with $CS3$ (△△), resulting in a 5.89% improvement in EER compared with using only the Mel-spectrogram (33.47%). However, combining the Mel-spectrogram with $CS3$ and $CS3$ (△△), which have the same distribution state (as shown in Fig. 5.4), led to a slight increase in EER (from 31.50% to 32.28%). Applying VAD to filter out interference information from silent clips decreased EER to 29.90%.

5.5.2 Results and discussion in ADD2023

Table 5.4 presents the experimental results conducted using three distinct methods for F_0 estimation. To ensure fairness, the losses of the selected epoch remain

Table 5.6: Evaluation results of STM representations. Different numbers and types of filterbank are used for the calculation of STM representations.

Front-end Features	Results (EER %)	
	Adp. set	Test set
LFB	0.24	11.02
MFB	3.34	10.82
GFB	1.18	9.68
NUF	2.20	9.33
STM (NUF,64)	7.61	9.06
STM (NUF,80)	6.81	9.80
STM (NUF,128)	7.30	9.54

nearly unchanged (around 0.37). Implementing data augmentation significantly reduces the EER from 61.21% to 41.29%. Moreover, utilizing $CS3$, $CS3$ ($\triangle$), and $CS3$ ($\triangle\triangle$) in conjunction with the Mel-spectrogram feature contributes to further reducing EER. These results validate the efficacy of utilizing pathological prosody information, specifically the shimmer features, for FAD.

Comparing the F_0 estimation algorithms, both the YIN and SWIPE methods improve the effectiveness of the shimmer features and exhibit lower EER values than the IRAPT algorithm. The reason for this may be that both YIN and SWIPE encompass a broader frequency search range and higher robustness for natural speech. The best result is achieved when extracting the $CS3$ ($\triangle\triangle$) feature with the YIN algorithm, resulting in an EER of 36.18%.

The exploration results of different combination weights between the Mel-spectrogram and CS3 ($\triangle\triangle$) features are presented in Table 5.5. The optimal result is achieved when the weight is set at a ratio of 3:2. This results in a significant improvement of 13.3% compared with using the Mel-spectrogram only, which yields a performance of 41.29%. This finding indicates that setting different combination weights can balance the effects of inconsistencies in the dynamic range of different features.

5.6 Performance of STM Representations

This section first tests the performance of different filterbanks, including the linear filterbank (LFB), MFB, GFB, and the proposed UNFB. Then, STM representations derived from NUF with channel numbers 64, 80, and 128 are tested. All results tested in the ASVspoof 2019 challenge are listed in Table 5.6. This thesis focus on the introduction of test set results since the test set contains more unseen data.

By comparing various filterbanks, the newly proposed DNN-based NUF demonstrates superior performance, yielding an EER of 9.33%. This outcome reaffirms that a data-driven NUF can extract more TSI. The utilization of such filterbanks for deriving STM representations should theoretically be more effective, and the performance of STM align with our expectations. The best performance is 9.06% when the channel number setting as 64.

5.7 Summary

This chapter applies the proposed feature representations in the FAD task. First, we investigate the prosody information differences in the voice represented by using the jitter and shimmer features for the FAD task. In accordance with the statistical analysis results, the most promising features were selected and incorporated with a DNN-based FAD system. To further enhance the performance of the proposed FAD system, two additional F_0 estimation methods, namely YIN and IRAPT, were utilized in place of the IRAPT algorithm when extracting features. Different weights were tested to find out the optimal combination between the Mel-spectrogram and shimmer features.

Statistical analysis results indicate prosody differences captured by the shimmer features, especially the CS3, can provide important information to distinguish between fake and genuine speech. This finding can be further verified by combining the static and dynamic CS3 features with the Mel-spectrogram and integrating them into the LCNN-BLSTM-based FAD system. The results obtained from the ADD2023 dataset indicate that utilizing YIN and SWIPE algorithms can further improve the performance of the FAD system due to the accuracy of F_0 detection and broader frequency search range. During the online test of ADD2022, EER decreased from 33.47 % to 31.50 % in the absence of VAD, namely an improvement of 5.89 %. During the online test of ADD2023, a combined weight of 3:2 resulted in a significant improvement. The EER decreased from 41.29 % to 35.77 %, achieving an improvement of 13.37 %.

To propose more effective STM representations, we compare the performance of different filterbanks in the ASVspoof challenge. The most potential filterbanks are used to derive the STM representations. Filterbanks with different channel numbers are also discussed. The overall results show that data-driven NUF outperforms commonly used filterbank, including the LFB, MFB, and GFB. Finally, the STM representation with channel number 64 achieved the best performance in the test set of ASVspoof 2019 challenge.

Chapter 6

Factory Automation Based on Machine ASD

ASD for machine condition monitoring enables workers to arrange maintenance work to fix machine problems in the earliest stages of anomaly, thus reducing maintenance costs and preventing damage. Developing advanced ASD systems is an important component of the fourth industrial revolution and has received increasing attention in recent years.

The log Mel spectrum (LMS) is widely used as an acoustic front-end in an NN-based ASD system [134, 135, 136]. The Mel filterbank (FB) is designed on the basis of the pitch perception of the human ear. It has a higher resolution in the low-frequency regions and a lower resolution in the high-frequency regions [41]. However, it can be argued that the human ear is not the most effective in detecting machine anomalies. Moreover, different types of machines have different vibration frequency regions depending on their physical property. Consequently, the discriminative information of sounds emitted from different types of machines may be encoded non-uniformly in the frequency domain. The Mel FB may filter out important information at the high-frequency regions, decreasing the performance of an ASD system. Therefore, quantifying the importance of the frequencies of different types of machines for ASD is necessary.

This chapter aims to apply the proposed quantification method of frequency importance, which is introduced in Chapter 3, in the machine ASD task. Additionally, STM representations are also applied to extract more TSI information. First, an autoencoder-based ASD architecture is introduced. All of our proposed features are tested using this architecture. The experiment data and matrix, experimental setting, and experimental results and discussion are described in Section 6.2,6.3, and 6.4, respectively.

6.1 Proposed Methods

The calculation of the F-ratio for machine m can be defined as

$$F_m = \frac{\frac{1}{2}\sum_c (u_{m,c} - u_m)^2}{\frac{1}{2N}\sum_c \sum_{i=1}^{N} (x_{m,c}^i - u_{m,c})^2}, \tag{6.1}$$

where $x_{m,c}^i$ is the sub-band energy of the i-th audio of class c with $i = 1, 2, ..., N$, $m \in \{\text{fan}, \text{gearbox}, \text{bearing}, \text{slider}, \text{toycar}, \text{toytrain}, \text{valve}\}$, and $c \in \{\text{normal}, \text{anomaly}\}$. The equations

$$u_{m,c} = \frac{1}{N}\sum_{i=1}^{N} x_{m,c}^i \qquad \text{and} \qquad u_m = \frac{1}{2N}\sum_c \sum_{i=1}^{N} x_{m,c}^i$$

are used to calculate the variables that represent the sub-band energy averages for class c and for all classes, respectively.

Equation (6.1) is the ratio between the inter-class variances and intra-class variances of speech power in a given frequency band. A larger value obtained in a frequency band means that more discriminative information is encoded in that band.

Based on the F-ratios, seven different filterbanks for seven types of machines are designed to extract more discriminative information. The log non-uniform spectrum (LNS) was extracted by replacing the filterbank used in the extraction processes of the log mel spectrum (LMS). The detailed process of our proposed ASD systems is shown in Fig. 5.1, the autoencoder (AE) model includes an encoder, bottleneck layer, and decoder modules. All modules consist of fully connected layers. The mean squared error (MSE) is used as the cost function to optimize the overall system. In the testing phase, audio with high reconstruction error was treated as anomalous sound.

6.2 Experiment data and matrix

We used the dataset provided by the DCASE2022 Challenge Task 2 [137, 138]. The dataset is comprised of normal and anomalous sounds produced by seven types of machines, i.e., fan, gearbox, bearing, slide, tor car, toy train, and valve. Sounds recorded from each type of machine are divided into six sections in accordance with the differences in machine configurations; sections 01, 02, and 03 are organized in the development dataset; sections 04, 05, and 06 are in the evaluation dataset. During the analysis step, we used {100 normal,100 anomaly}×3×7 clips of the development data to calculate F-ratios. During the training step, we used 1,000×3×7 clips of the development data to train the AE model. It is worth noting

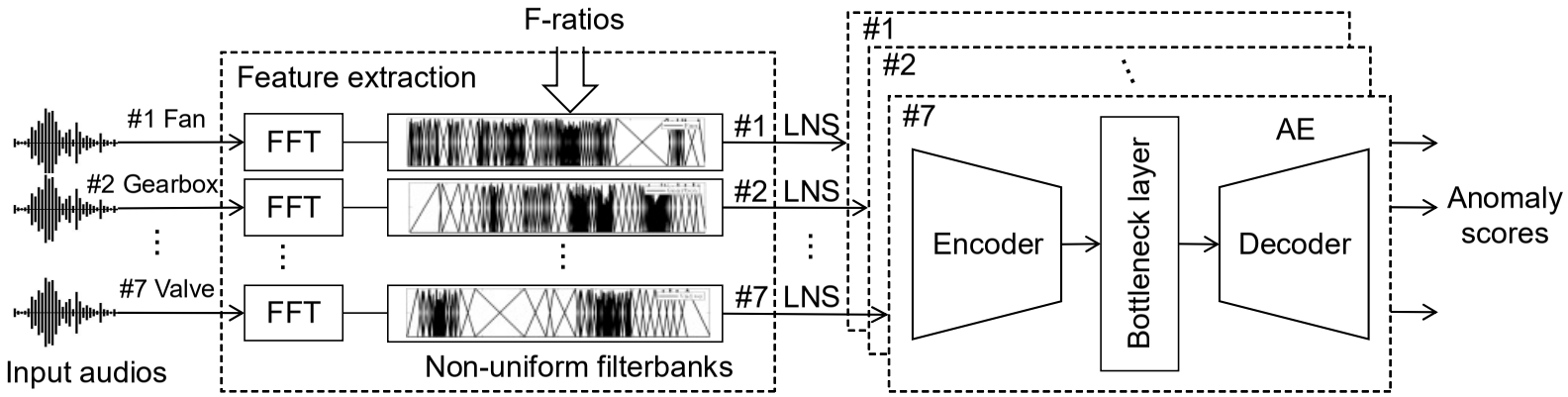

Figure 6.1: Systems using LNSs extracted by the proposed data-driven non-uniform FBs with AE-based detectors for machine ASD.

that the data used to calculate the F-ratio is not used for model training. During the evaluation step, we used 200×3×7 clips of test data in both the development and evaluation datasets to evaluate the effectiveness of the proposed method. The length of each clip was fixed to 10s.

The area under the curve (AUC) and partial-AUC (pAUC) for the receiver operating characteristic (ROC) curves were used to evaluate the proposed ASD method. The formulae of AUC and pAUC can be found in [134]. Generally, AUC and pAUC are the average sums of anomaly scores. However, the difference between pAUC and AUC is that pAUC is designed to focus on a low false-positive-rate portion of the ROC curve over a pre-specified range of interest $[0, p = 0.1]$. In practical situations, if an ASD system generates false alarms frequently (high false-positive rate), the system is not trustworthy. Therefore, using pAUC to encourage a high true-positive rate under low false-positive-rate conditions is essential.

The DCASE2022 challenge [139] provides the formulae to calculate AUC and pAUC for each machine type, section, and domain as follows:

$$AUC_{m,n,d} = \frac{1}{N_d^- N_n^+} \sum_{i=1}^{N_d^-} \sum_{l=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_l^+) - \mathcal{A}_\theta(x_i^-)) \,, \tag{6.2}$$

$$pAUC_{m,n} = \frac{1}{\lfloor pN_n^- \rfloor N_n^+} \sum_{i=1}^{\lfloor pN_n^- \rfloor} \sum_{l=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_l^+) - \mathcal{A}_\theta(x_i^-)) \,, \tag{6.3}$$

where $\mathcal{H}(x)$ is a sign function that returns 1 if $x > 0$ and 0 otherwise. $\mathcal{A}_\theta(\cdot)$ is the proposed ASD that produces an anomaly score for each input clip.

where m, n, and $d = \{\text{source}, \text{target}\}$ are a machine type, a section, and a domain, respectively. N_d^- is the number of normal test clips in the domain d in the section n in the machine type m, N_n^+ the number of anomalous test clips in the section n in the machine type m, and N_n^- the number of normal test clips in the section n in the machine type m. $\mathcal{H}(x)$ is a sign function that returns 1 if $x > 0$ and 0 otherwise. $\mathcal{A}_\theta(\cdot)$ is the proposed ASD that produces an anomaly score for each input clip. $\{x_i^-\}_{i=1}^{N_d^-}$ and $\{x_l^+\}_{l=1}^{N_n^+}$ are normal and anomalous test clips in domain d in section n in machine type m, respectively. The difference between pAUC and AUC is that pAUC is design to focus on a low false-positive-rate portion of the ROC curve over a pre-specified range of interest $[0, p = 0.1]$. In practical situations, if a ASD system generates false alarms frequently (high false-positive-rate), the system is not trustworthy. Therefore, using pAUC to encorage high true-positive-rate under low false-positive-rate conditions is essentially important.

6.3 Experimental setting

To extract the LMS, 10-s audio clips were first split into different frames with frame lengths of 64 ms and hop lengths of 32 ms. The Mel-spectrogram feature was then extracted with the following parameters: n_fft=1024, hop_length=512, $num\_filters$=128, and $power = 2.0$. We extracted the LNS using the same configuration but different FBs compared with the LMS. Five consecutive frames with a sliding window were concatenated into one feature vector with a dimension of 640 and fed into the detector. For example, we assume that the input signal is $X = \{X_t\}_{t=1}^{T}$ where $X_t \in R^M$, and M and T are the number of Mel-filters and time-frames, respectively. Then, the acoustic feature at t is obtained by concatenating consecutive frames of the feature as $\psi_t \in R^D$, where $D = P \times M$, $P = 5$, $M = 128$ and $D = 640$. The reconstruction error is calculated as:

$$E(X) = \frac{1}{DT} \sum_{t=1}^{T} \parallel \psi_t - r(\psi_t) \parallel_2^2, \tag{6.4}$$

where $r(\psi_t)$ is the vector reconstructed by the AE model, and $\| \cdot \|_2$ is $\mathcal{L}_2$ norm.

The AE model had four dense layers with 128 dimensions for the encoder/decoder and one bottleneck layer with 8 dimensions. We trained the model for 100 epochs using the Adam optimizer [140] with a learning rate of 0.0001 and batch size of 128.

6.4 Results and discussion of Spectral Features

The quantification results of discriminative information for ASD for each machine are shown in Fig. 6.2. We also illustrate the frequency-band importance of the Mel scale for comparison. This result can be understood as the derivative of the Mel scale. It is evident that the frequency importance in the Mel scale decreased with the increase in frequency. The quantification results using the machine-wise F-ratio indicate that the discriminative feature for ASD of each type of machine was encoded non-uniformly in the frequency domain. There are many discriminative features concentrated in the high-frequency regions, such as the gearbox and slider.

Based on the F-ratios, we designed the machine-wise ASD systems on accordance with the pipeline shown in Fig. 6.1, and carried out experiments to examine its effectiveness. The results of harmonic mean (HM) and arithmetic mean (AM) in both development and evaluation datasets are listed in Tables 6.1 and 6.2, respectively. All results are shown in percentages, and the improved results are highlighted in bold.

The proposed LNS generally improved the performance in both development and evaluation datasets for most of the machines. The LNS obtained the highest

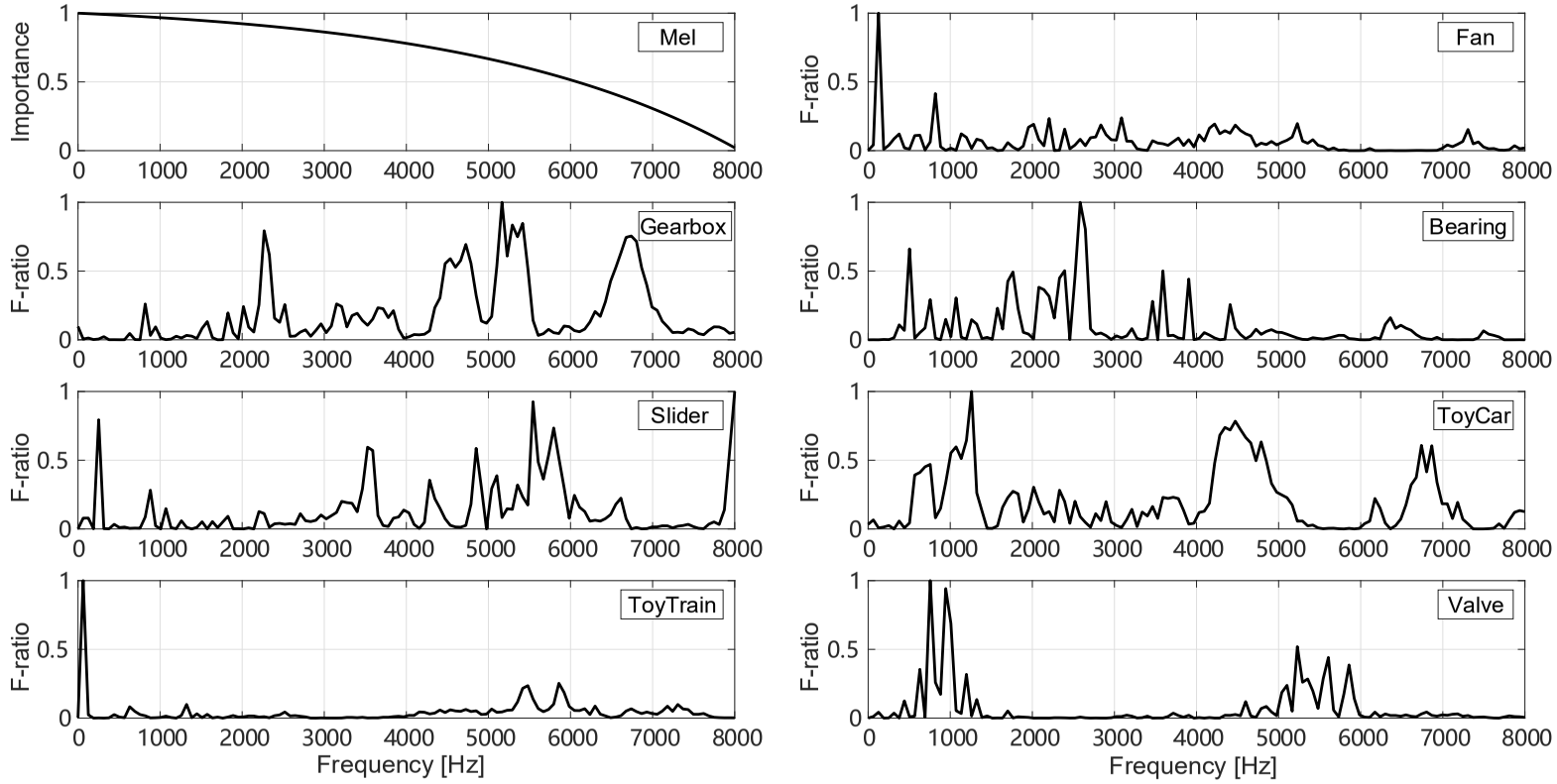

Figure 6.2: Frequency-band importance of Mel scale and quantified frequency-band importance using machine-wise F-ratio for each machine. All frequency-band importances were normalized from 0 to 1.

Table 6.1: Overall results by using the proposed (LNS) and baseline (LMS) features in terms of AUC (%) and pAUC (%) in the development dataset.

Machines	Sections	AUC		pAUC	
		LMS	LNS	LMS	LNS
Toy car	AM	64.22	**66.48**	53.25	**56.11**
	HM	63.01	**65.00**	53.23	**56.05**
Toy train	AM	51.20	**57.35**	50.49	**52.65**
	HM	49.55	**56.74**	50.48	**52.62**
Bearing	AM	56.68	**65.75**	50.93	**58.18**
	HM	55.65	**65.12**	50.86	**57.70**
Fan	AM	**64.45**	60.68	**58.39**	56.79
	HM	**63.14**	59.23	**57.93**	56.40
Gearbox	AM	65.54	**70.10**	59.00	**59.93**
	HM	65.28	**69.79**	58.74	**59.26**
Slider	AM	63.68	**69.95**	56.54	**62.49**
	HM	62.77	**68.98**	56.27	**62.28**
Valve	AM	50.59	**54.53**	50.33	**50.72**
	HM	50.38	**54.41**	50.29	**50.70**
Average	AM	59.48	**63.55**	54.13	**56.69**
	HM	55.05	**60.13**	53.76	**56.20**

Table 6.2: Overall results by using the proposed (LNS) and baseline (LMS) features in terms of AUC (%) and pAUC (%) in the evaluation dataset.

Machines	Sections	AUC		pAUC	
		LMS	LNS	LMS	LNS
Toy car	AM	59.20	**70.43**	56.91	**63.32**
	HM	57.07	**66.74**	56.46	**62.49**
Toy train	AM	44.73	**46.13**	**50.26**	49.09
	HM	**44.44**	43.53	**50.25**	49.06
Bearing	AM	44.79	**51.86**	50.23	**51.09**
	HM	43.20	**51.43**	50.17	**51.09**
Fan	AM	48.81	**50.90**	51.07	**51.46**
	HM	47.91	**50.54**	51.02	**51.43**
Gearbox	AM	51.63	**54.45**	50.40	**52.32**
	HM	50.40	**53.09**	50.40	**52.19**
Slider	AM	**49.16**	48.35	**50.61**	50.23
	HM	**44.52**	44.45	**50.56**	50.18
Valve	AM	45.60	**45.31**	49.65	**49.68**
	HM	45.15	**45.27**	49.62	**49.67**
Average	AM	49.13	**52.49**	51.31	**52.45**
	HM	47.14	**49.78**	51.13	**52.00**

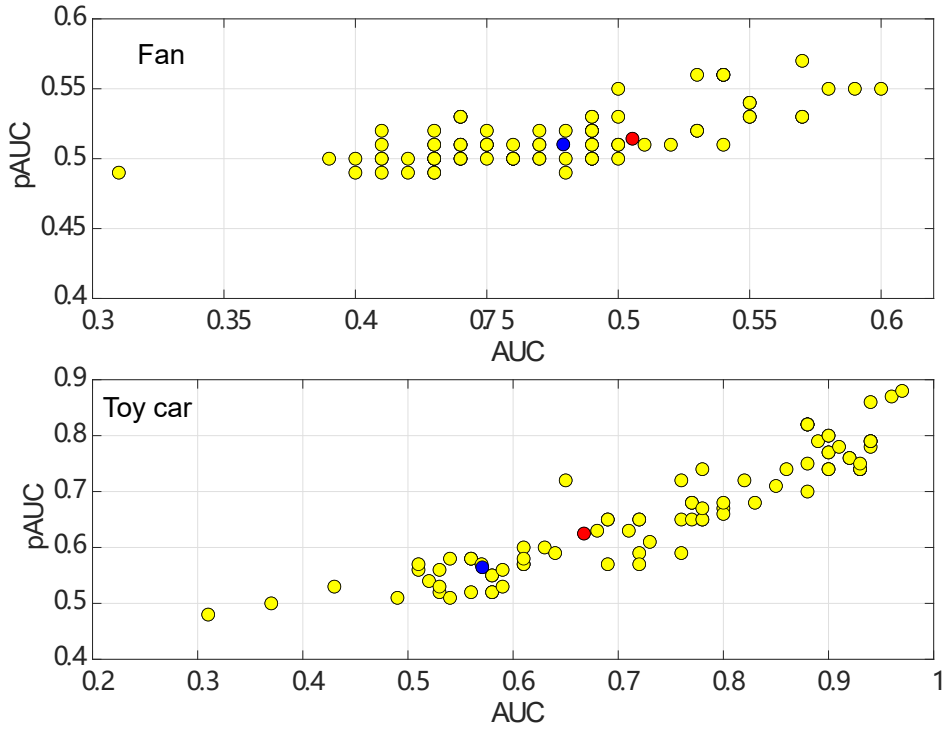

Figure 6.3: Results in the evaluation dataset of DCASE2022 Challenge Task 2 using sounds recorded from the fan and toy car. Blue and red dots correspond to the baseline and proposed systems, respectively. The higher AUC and pAUC, the better performance.

improvement in the bearing of the development dataset and in the toy car of the evaluation dataset, which are relative improvements of 17.02 and 16.94 % for HM, respectively. By using the proposed LNS, averaged HMs improved from 55.05 to 60.13% in the development dataset and from 47.14 to 49.78% in the evaluation dataset, achieving relative improvements of 9.22 and 5.60% respectively.

There are two exceptions. The proposed LNS provided better performance in the fan of the evaluation dataset, but the performance degraded in the development dataset. In contrast, the LNS performed better in the slider of the development dataset even when degradation occurred in the evaluation dataset. This could be because the frequency-band importance is calculated independently with Eq. (6.1), which makes it difficult to consider the combined effects of each frequency component for all machines with the F-ratio. A more suitable quantification method could further improve performance.

Figure 6.3 shows a comparison of results between our proposed method with 83 other methods for fan and toy car. Each dot corresponds to a different system proposed in the DCASE2022 Challenge Task 2. There was a significant improvement by replacing the LMS of the baseline (blue dot) with the LNS extracted with our quantification results (red dot). The other state-of-the-art methods appeared to have higher performances; however, they employ heavy deep NNs containing hundreds of millions of parameters and pre-trained models. It appears that our proposed LNS feature with a simple AE-based detector can achieve results competitive with systems composed of more sophisticated models and training data.

6.5 Results and discussion of STM Representations

In this section, we derive the temporal modulation (TM), spectral modulation (SM), and STM representations using the Gammatone (GFB), Mel (MFB), and the proposed data-driven non-uniform filterbanks (NUF) with channel number 80, respectively. The overall results of using the proposed feature representations in terms of AUC (%) and pAUC (%) in the development dataset are listed in Table 6.3. All results are shown in percentages, and the improved results are highlighted in bold.

The results indicate that incorporating TM representations can enhance the performance of time-frequency (TF) features. For instance, utilizing TM leads to an improvement in the AUC (target) metric for AM from 47.55 % to 49.62 %, and for HM, the enhancement is from 41.48 % to 45.21 %. Additionally, the results demonstrate that the inclusion of STM, which encompasses both SM and TM, further enhances accuracy. Particularly, the STM derived from the proposed NUF

Table 6.3: Overall results by using the proposed STM representations in terms of AUC (%) and pAUC (%) in the development dataset. TF: time-frequency feature, SM: spectral modulation representation, TM: temporal modulation representation, STM: spectral temporal modulation representation.

Filterbanks	Metrics	Arithmetic mean (AM)				Harmonic mean (HM)			
		TF	SM	TM	STM	TF	SM	TM	STM
GFB	AUC (source)	69.01	62.28	69.69	**70.00**	66.39	60.01	65.65	**67.00**
	AUC (target)	47.55	43.27	**49.62**	49.05	41.48	37.45	**45.21**	43.09
	pAUC	**54.10**	51.52	52.34	52.56	**53.70**	51.36	52.16	52.32
MFB	AUC (source)	71.85	57.41	**72.25**	71.69	68.68	56.00	**68.95**	68.70
	AUC (target)	48.36	46.42	49.80	**50.24**	44.11	43.65	44.41	**44.52**
	pAUC	**54.48**	51.68	52.59	52.62	**54.10**	51.55	52.45	52.47
NUF	AUC (source)	72.02	55.74	**72.57**	72.53	68.35	53.01	68.32	**68.91**
	AUC (target)	49.92	45.48	51.29	**52.19**	43.09	42.04	44.34	**44.79**
	pAUC	54.16	51.48	54.02	**54.64**	53.54	51.28	53.36	**54.04**

achieves a noteworthy 52.19 % in the AM of AUC (target), reflecting a significant 7.9 % improvement compared to the MFB feature (48.36 %).

6.6 Summary

This chapter applies the proposed spectral features and STM representation in the machine ASD task. First, we quantified the importance of different frequencies for the anomalous detection of seven types of machines using a data-driven statistical-based quantification method (machine-wise F-ratio). We found that the discriminative features of each machine were encoded non-uniformly in the frequency domain. To highlight such important frequencies, we designed non-uniform FBs that have high resolutions in the frequencies with high F-ratios and used them to extract the LNS. The correctness of quantification results and effectiveness of the proposed LNS were verified in the DCASE2022 Challenge Task 2 with a simple AE-based detector. Compared with the LMS, the LNS achieved a relative improvement of 9.22 and 5.60% in development and evaluation datasets in terms of averaged HM of AUC, respectively. Furthermore, the performance was further improved when we used the STM representations derived from the proposed NUF, and the TM contribute more effectiveness compare with the SM in the machine ASD task.

Chapter 7

Conclusion

7.1 Summary

Acoustic feature extraction from raw speech can be useful in many application fields, including human-computer interaction, speech security, industrial automation, and others. This study aims to extract advanced feature representations that can contain more task-specific discriminative information for automatic speaker verification (ASV), fake audio detection (FAD), and machine anomalous sound detection (ASD) tasks. These three applications cover data of speech, machine-synthesized speech, and non-speech. Therefore, the conclusion of this study has a certain generalization.

Inspired by the human auditory mechanism, frequency domain analysis, time domain analysis, and spectral temporal modulation (STM) analysis are implemented to achieve the research target. Both time and frequency domain analyses can help us confirm the importance of different auditory attributes and provide guidance for the designing of the STM analysis methods. This study believes that the STM analysis deals with both spectral and temporal modulation of audio to perceive auditory attributes related to audio production, hence providing more discriminative information for different tasks.

In the frequency domain analyses, this study investigates the importance of different frequency regions for ASV, FAD, and machine ASD tasks. The proposed methods aim to answer the question of which frequency regions are more important for task-specific information (TSI) extraction. The frequency-wise attention structure combined with a ResNet is proposed to quantify the nonlinear combined effect of frequency components. The experimental results from these three applications consistently indicate that TSI is non-uniformly distributed in the frequency domain, and more distinguish information can be extracted by highlighting such important frequencies. Moreover, the frequency modulation is related to the cep-

stral analyses. Therefore, the effectiveness of the proposed non-uniform filterbank cepstral coefficients (NUFCC) indicates that the spectral modulation contains a lot of discriminative information for different applications.

In the time domain analysis, the jitter and shimmer features which are related to the timbre and prosody information are investigated in the FAD task. Jitter and shimmer features reflect the characteristics of amplitude and frequency perturbation (AFP). We aim to answer the research question of whether can we represent TSI in the time domain. The experimental results indicate that both the static and dynamic shimmer features of voice can provide discriminative information and are complementary to the traditional spectrum-based systems in the FAD task. This finding can help us confirm the importance of temporal modulations.

In the STM analysis, we want to answer the research question of can STM representations obtain more TSI. The data-driven NUF designed from the frequency analysis stage is utilized to derive the STM representation, which is an advanced feature representation proposed in this study. By using the STM representations, variations in the spectral and temporal domains should be included. According to the current results from these three applications, the STM representations can achieve competitive performance in the FAD and machine ASD tasks, which covering synthesized speech and non-speech signals. However, in the ASV task, the current results are inconsistent with our initial expectations. It seems that the potential problem arises from the machine learning methods (i-vector) that we used initially designed for short-term features. Due to the broader temporal context of STM representations, there is generally a lower count of available training vectors. Moreover, various modulation frequencies carry distinct auditory information, and the significance of this information varies across different applications. Feeding all the information to the back-end classifier results in redundant data, as not all of it is equally crucial for different purposes. Consequently, a specific-designed DNN architecture may be required to deal with these problems.

In conclusion, spectral analysis, temporal analysis, and STM analysis are crucial in the human auditory system. More advanced and distinguished feature representations can be developed by considering these important auditory attributes. Current research still has plenty of room to improve in the application of STM representations.

7.2 Contributions

This study focuses on the extraction of feature representations with the inspiration of the human auditory mechanisms. The proposed methods can be used in different audio detection and verification tasks, such as ASV, FAD, and machine ASD. Therefore, this study contributes to society by improving the security and

reliability of digital speech communication systems. This study can also accelerate the process of factory automation, providing predictive maintenance, minimizing downtime, optimizing production efficiency, and reducing overall costs. Moreover, this study contributes to our broader understanding of how humans perceive speech from production, benefiting not only for security and technology but also for fields like biology, linguistics, psychology, and cognitive science.

7.3 Future Work

Applying STM representations in different acoustic scene analyses still has a lot of room for improvement. Future works will focus on the following issues to maximize the effectiveness of STM representations in different acoustic applications.

- Analyzing the robustness of STM representations in the presence of noise is a crucial aspect, especially for practical applications where real-world environments often introduce various forms of noise. Future work will conduct a comprehensive analysis of the types and levels of noise present in real-world scenarios. Understand the characteristics of noise in the application domain and how it interacts with STM representations.

- Designing DNN architectures to extract discriminative information from STM representations is necessary. The STM representations contain more long-term information and different feature channels have different resolutions. One possible way is to design a DNN to handle multiple resolutions. This can involve incorporating multiple parallel pathways or using architectures like U-Net, which are known for handling multiple resolutions effectively. Utilizing convolutional layers with different filter sizes to capture information at different scales is another possible way. In addition, incorporating attention mechanisms allows the model to focus on specific parts of the input sequence. Spatial and temporal attention mechanisms can be particularly useful for dealing with varying resolutions.

- Integrating STM representations with traditional acoustic features is a promising approach to leverage the strengths of different types of features. There are several feature fusion methods will be considered. Early fusion concatenates the STM representations and traditional features at the input level before feeding them into the model. This method combines information from both feature sets at the beginning of the processing pipeline. Late fusion trains separate models on STM and traditional features and combine their predictions at a later stage. This method allows the model to learn independent representations for each feature set and then combine them. In addition,

attention mechanisms are another commonly used feature fusion approach. It implements attention mechanisms to dynamically weigh the importance of STM and traditional features based on the context. Attention mechanisms allow the model to focus on relevant features for different inputs.

Bibliography

[1] J. van Dorp Schuitman, "Auditory modelling for assessing room acoustics," 2011.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[3] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[4] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2906–2919, 1997.

[5] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[6] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE transactions on information forensics and security*, vol. 15, pp. 1616–1629, 2019.

[7] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," *arXiv preprint arXiv:1907.00501*, 2019.

[8] L. Huang and C.-M. Pun, "Audio replay spoof attack detection using segment-based hybrid feature and densenet-lstm network," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 2567–2571.

[9] A. R. Møller, "Unit responses in the rat cochlear nucleus to tones of rapidly varying frequency and amplitude," *Acta Physiologica Scandinavica*, vol. 81, no. 4, pp. 540–556, 1971.

[10] N. Suga, "Analysis of information-bearing elements in complex sounds by auditory neurons of bats," *Audiology*, vol. 11, no. 1-2, pp. 58–72, 1972.

[11] C. E. Schreiner and J. V. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. i. the anterior auditory field (aaf)," *Hearing research*, vol. 21, no. 3, pp. 227–241, 1986.

[12] ——, "Representation of amplitude modulation in the auditory cortex of the cat. ii. comparison between cortical fields," *Hearing research*, vol. 32, no. 1, pp. 49–63, 1988.

[13] G. Langner, "Periodicity coding in the auditory system," *Hearing research*, vol. 60, no. 2, pp. 115–142, 1992.

[14] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. i. characteristics of single-unit responses to moving ripple spectra," *Journal of neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[15] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[16] N. F. Viemeister, "Temporal factors in audition: A systems analysis approach," in *Psychophysics and physiology of hearing*. Academic Press, 1977, pp. 419–428.

[17] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1676–1680, 1989.

[18] S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2575–2580, 1989.

[19] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1181–1196, 2000.

[20] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[21] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[22] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16 560–16 572, 2020.

[23] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants." in *INTERSPEECH*, 2016, pp. 262–266.

[24] ——, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 234–242, 2018.

[25] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 416–422, 2002.

[26] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[27] M. Slaney, *Lyon's cochlear model*. Citeseer, 1988, vol. 13.

[28] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[29] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[30] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2222–2232, 2006.

[31] P. Johannesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus," in *Symposium on Hearing Theory, 1972*. IPO, 1972.

[32] F.-G. Zeng, "Trends in cochlear implants," *Trends in amplification*, vol. 8, no. 1, pp. 1–34, 2004.

[33] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[34] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[35] J. R. Deller Jr, "Discrete-time processing of speech signals," in *Discrete-time processing of speech signals*, 1993, pp. 908–908.

[36] A. V. OPPENHEM, "Discrete-time signal processing," 1999.

[37] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[38] R. M. Hegde, H. A. Murthy, and G. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 1. IEEE, 2004, pp. I–517.

[39] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[40] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *Robust data mining*, pp. 27–33, 2013.

[41] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[42] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter," *International Journal of Signal Processing*, vol. 5, no. 1, pp. 11–19, 2009.

[43] G. K. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *arXiv preprint arXiv:1806.09010*, 2018.

[44] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[45] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[46] H. Hermansky, "Perceptual linear prediction (plp) analysis of speech. 87 (4): 1738–1752," 1990.

[47] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 4. IEEE, 2003, pp. IV–788.

[48] K. Bartkova, D. L. Gac, D. Charlet, and D. Jouvet, "Prosodic parameter for speaker identification," in *Seventh International Conference on Spoken Language Processing*, 2002.

[49] D. Gerhard *et al.*, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, SK, Canada, 2003.

[50] P. Rose, *Forensic speaker identification*. cRc Press, 2002.

[51] B. S. Atal, "Automatic speaker recognition based on pitch contours," *The Journal of the Acoustical Society of America*, vol. 52, no. 6B, pp. 1687–1697, 1972.

[52] T. Kinnunen and R. González-Hautamäki, "Long-term f0 modeling for text-independent speaker recognition," in *Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece*. Citeseer, 2005, pp. 567–570.

[53] J. Adell, A. Bonafonte, and D. Escudero, "Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech," *Procesamiento del lenguaje natural*, no. 35, pp. 277–283, 2005.

[54] S. McAdams and B. L. Giordano, "The perception of musical timbre," 2014.

[55] M. Scale, "Wikipedia the free encyclopedia," *Last modified on Oct*, vol. 13, 2009.

[56] A. Pooransingh and D. Dhoray, "Similarity analysis of modern genre music based on billboard hits," *IEEE Access*, vol. 9, pp. 144 916–144 926, 2021.

[57] K. Jensen, "The timbre model," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2238–2238, 2002.

[58] A. Pearce, T. Brookes, and R. Mason, "Timbral attributes for sound effect library searching," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[59] P. N. Vassilakis and K. Fitz, "Sra: A web-based research tool for spectral and roughness analysis of sound signals," in *Proceedings of the 4th Sound and Music Computing (SMC) Conference*, 2007, pp. 319–325.

[60] T. H. Pedersen, "The semantic space of sounds," *Delta*, 2008.

[61] E. Zwicker and H. Fastl, *Psycho-acoustics: Facts and models*. Springer Science & Business Media, 2013, vol. 22.

[62] A. Pearce, T. Brookes, and R. Mason, "First prototype of timbral characterisation tool for semantically annotating non-musical," *Audio Commons project deliverable D*, vol. 5, 2017.

[63] S. Hatano and T. Hashimoto, "Booming index as a measure for evaluating booming sensation," in *Proc. Inter-Noise*, vol. 233, 2000, pp. 1–5.

[64] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International conference on digital audio effects*, vol. 237. Bordeaux, 2007, p. 244.

[65] A. Pearce, T. Brookes, and R. Mason, "Hierarchical ontology of timbral semantic descriptors," *Audio Commons project deliverable D*, vol. 5, 2016.

[66] A. Pearce, *Perceived differences between microphones*. University of Surrey (United Kingdom), 2017.

[67] L. N. Solomon, "Search for physical correlates to psychological dimensions of sounds," *The Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 492–497, 1959.

[68] D. J. Freed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events," *The Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 311–322, 1990.

[69] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?" *Acta acustica united with acustica*, vol. 92, no. 5, pp. 820–825, 2006.

[70] E. Schubert, J. Wolfe, A. Tarnopolsky *et al.*, "Spectral centroid and timbre in complex, multiple instrumental textures," in *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*. sn, 2004, pp. 112–116.

[71] L. Olivier, "Mir in matlab (ii): A toolbox for musical feature extraction from audio," *ISMIR2007, Sep.*, 2007.

[72] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cognition & Emotion*, vol. 19, no. 5, pp. 633–653, 2005.

[73] S. McAdams, "Musical timbre perception," *The psychology of music*, pp. 35–67, 2013.

[74] M. M. Farbood and K. C. Price, "The contribution of timbre attributes to musical tension," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 419–427, 2017.

[75] W. Jiang, J. Liu, X. Zhang, S. Wang, and Y. Jiang, "Analysis and modeling of timbre perception features in musical sounds," *Applied Sciences*, vol. 10, no. 3, p. 789, 2020.

[76] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, 2023.

[77] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7234–7238.

[78] Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech," *Neural Networks*, vol. 140, pp. 261–273, 2021.

[79] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[80] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[81] T. Kinnunen, K.-A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification." in *Odyssey*, 2008, p. 30.

[82] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[83] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[84] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[85] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 92–97.

[86] P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, and M. J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition." in *Odyssey*, vol. 2014, 2014, pp. 293–298.

[87] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[88] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[89] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[90] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[91] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *arXiv preprint arXiv:2111.07725*, 2021.

[92] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5859–5866.

[93] Z. Jiang, H. Zhu, L. Peng, W. Ding, and Y. Ren, "Self-supervised spoofing audio detection scheme." in *INTERSPEECH*, 2020, pp. 4223–4227.

[94] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022.

[95] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," *arXiv preprint arXiv:2006.15321*, 2020.

[96] T. Inoue, P. Vinayavekhin, S. Morikuni, S. Wang, T. H. Trong, D. Wood, M. Tatsubori, and R. Tachibana, "Detection of anomalous sounds for machine condition monitoring using classification confidence." in *DCASE*, 2020, pp. 66–70.

[97] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on wavenet," in *Proc. 26th EUSIPCO*. IEEE, 2018, pp. 2494–2498.

[98] T. Komatsu, T. Hayashiy, R. Kondo, T. Todaz, and K. Takeday, "Scene-dependent anomalous acoustic-event detection based on conditional wavenet and i-vector," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 870–874.

[99] L. Marchegiani and I. Posner, "Leveraging the urban soundscape: Auditory perception for smart vehicles," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 6547–6554.

[100] K. Li, Q.-H. Nguyen, Y. Ota, and M. Unoki, "Unsupervised anomalous sound detection for machine condition monitoring using temporal modulation features on gammatone auditory filterbank." in *DCASE*, 2022.

[101] Y. Deng, J. Liu, and W.-Q. Zhang, "Improving unsupervised anomalous sound detection performance of autoencoder and its variant with pretrained deep belief network," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1–5.

[102] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460–475, 1976.

[103] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *J. Acousti. Soc. Am.*, vol. 96, no. 4, pp. 2088–2100, 1994.

[104] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *J. Acousti. Soc. Am.*, vol. 100, no. 5, pp. 3374–3383, 1996.

[105] A. R. Webb, K. D. Copsey, and G. Cawley, *Statistical pattern recognition*. Wiley Online Library, 2011, vol. 2.

[106] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.

[107] K. Li, X. Lu, M. Akagi, J. Dang, S. Li, and M. Unoki, "Relationship between speakers' physiological structure and acoustic speech signals: Data-driven study based on frequency-wise attentional neural network," in *Proc. EUSIPCO*. IEEE, 2022, pp. 379–383.

[108] Y. Zhou, Y. Sun, J. Li, J. Zhang, and Y. Yan, "Physiologically-inspired feature extraction for emotion recognition," in *Proc. Tenth Annual Conference of the International Speech Communication Association*. Citeseer, 2009.

[109] S. Hyon, J. Dang, H. Feng, H. Wang, and K. Honda, "Detection of speaker individual information using a phoneme effect suppression method," *Speech Communication*, vol. 57, pp. 87–100, 2014.

[110] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. CVPR*, 2017, pp. 5659–5667.

[111] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[112] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar als detection based on analysis of voice perturbation and vibrato," in *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2019, pp. 267–272.

[113] E. Azarov, M. Vashkevich, and A. Petrovsky, "Instantaneous pitch estimation based on rapt framework," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2787–2791.

[114] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[115] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[116] R. J. Baken, "Clinical measurement of speech and voice," *(No Title)*, 1987.

[117] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[118] J. Dang and K. Honda, "An improved vocal tract model of vowel production implementing piriform resonance and transvelar nasal coupling," in *Proc. ICSLP'96*, vol. 2. IEEE, 1996, pp. 965–968.

[119] ——, "Acoustic characteristics of the piriform fossa in models and humans," *J. Acousti. Soc. Am.*, vol. 101, no. 1, pp. 456–465, 1997.

[120] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2011, pp. 559–564.

[121] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[122] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[123] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[124] G. Pamisetty and K. Sri Rama Murty, "Prosody-tts: An end-to-end speech synthesis system with prosody control," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 361–384, 2023.

[125] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[126] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[127] K. Li, S. Li, X. Lu, M. Akagi, M. Liu, L. Zhang, C. Zeng, L. Wang, J. Dang, and M. Unoki, "Data augmentation using mcadams-coefficient-based speaker anonymization for fake audio detection," pp. 664–668, 2022.

[128] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," *arXiv preprint arXiv:2103.11326*, 2021.

[129] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[130] J. Yi, J. Tao, R. Fu, X. Yan, C. Wang, T. Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren *et al.*, "Add 2023: the second audio deepfake detection challenge," *arXiv preprint arXiv:2305.13774*, 2023.

[131] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[132] S. Rosenzweig, S. Schwär, and M. Müller, "Libf0: A python library for fundamental frequency estimation," in *Late Breaking Demos of the International Society for Music Information Retrieval Conference (ISMIR), Bengaluru, India*, 2022.

[133] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, 2018.

[134] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *arXiv preprint arXiv:2206.05876*, 2022.

[135] S. Kapka, "Id-conditioned auto-encoder for unsupervised anomaly detection," *arXiv preprint arXiv:2007.05314*, 2020.

[136] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. ICASSP*. IEEE, 2021, pp. 336–340.

[137] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *In arXiv e-prints: 2205.13879*, 2022.

[138] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proc. 6th DCASE*, Barcelona, Spain, November 2021, pp. 1–5.

[139] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[140] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Publications

Main Publications

International Journals

1. Kai Li, Xugang Lu, Masato Akagi, and Masashi Unoki, "Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection," *IEEE Access*, vol. 11, pp. 84689-84698, 2023,
DOI:10.1109/ACCESS.2023.3301616.

International Conference

1. Kai Li, Dung Kim Tran, Xugang Lu, Masato Akagi, and Masashi Unoki, "Data-driven Non-uniform Filterbanks Based on F-ratio for Machine Anomalous Sound Detection," In *the European Signal Processing Conference (EUSIPCO)*, pp. 201–205, 2023.

2. Kai Li, Yao Wang, Minh Le Nguyen, Masato Akagi, and Masashi Unoki. "Analysis of Amplitude and Frequency Perturbation in the Voice for Fake Audio Detection," In *the Asia-Pacific Signal and Information Processing Association (APSIPA)*, IEEE, pp.929–936, 2022.

3. Kai Li, Xugang Lu, Masato Akagi, Jianwu Dang, Sheng Li, Masashi Unoki. "Relationship Between Speaker Physiological Structure and Acoustic Speech Signal: Data-driven Study Based on Frequency-wise Attentional Neural Network," In *the European Signal Processing Conference (EUSIPCO)*, pp. 379–383, 2022.

4. Kai Li, Quoc-Huy Nguyen, Masashi Unoki. "Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Using Temporal Modulation Features on Gammatone Filterbank," In *the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Nancy, France, 2022.

Domestic Conference

1. Kai Li, Xugang Lu, Masato Akagi, Jianwu Dang, Sheng Li, Masashi Unoki. "Study on Relationship Between Speakers' Physiological Structure and Acoustic Speech Signals: Data-Driven Study Based on Frequency-Wise Attentional Neural Network," In *the Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report*, Okinawa, Japan, 2022.

Other Publications

International Journals

1. Yongwei Li, Jianhua Tao, and Kai Li. "Speech Emotion Recognition Based on Glottal Source and Vocal Tract Features," *Journal of Signal Processing*, Vol. 39 Issue 4, pp. 632-638, Apr. 2023, *DOI:10.16798/j.issn.1003-0530.2023.04.004*.

International Conference

1. Kai Li, Masato Akagi, Yibo Wu, and Jianwu Dang. "Segment-level Effects of Gender, Nationality and Emotion Information on Text-independent Speaker Verification," In *the INTERSPEECH*, pp. 2987–2991, 2020, *DOI:10.21437/Interspeech.2020-1700*.

2. Kai Li, Masashi Unoki, Jianwu Dang and Masato Akagi. "Study on Simultaneous Estimation of Glottal Source and Vocal Tract Parameters by ARMAX-LF Model for Speech Analysis/Synthesis," In *the Asia-Pacific Signal and Information Processing Association (APSIPA)*, IEEE, pp. 36–43, 2021.

3. Xiaohui Liu, Meng Liu, Lin Zhang, Linjuan Zhang, Chang Zeng, Kai Li, Nan Li, Kong Aik Lee, Longbiao Wang, Jianwu Dang. "Deep Spectro-temporal Artifacts Detection for Synthesized Audios," In *the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pp. 69–75, 2022.

4. Quoc-Huy Nguyen, Kai Li, Masashi Unoki. "Non-Intrusive Speech Assessment Method with Temporal Modulation Feature on Gammatone Filterbank," In *the INTERSPEECH*, pp. 4526–4530, 2022, *DOI:10.21437/Interspeech.2022-528*.

5. Kai Li, Sheng Li, Xugang Lu, Masato Akagi, Meng Liu, Zhang Lin, Chang Zeng, Qin Yang, Longbiao Wang, Masashi Unoki. "Data augmentation based

on speaker anonymization for fake audio detection," In *the INTERSPEECH*, pp. 664–668, 2022, *DOI:10.21437/Interspeech.2022-10088*.

Domestic Conference

1. Kai Li, Masashi Unoki and Masato Akagi. "Estimation of Glottal Source Parameters of the LF Model Using Feed-forward Neural Network," In *the Acoustical Society of Japan (ASJ) spring meeting*, online, 2022.

2. Kai Li, Masashi Unoki and Masato Akagi. "Estimation of Glottal Source Waveforms and Vocal Tract Shapes Based on ARMAX-LF Model," In *the Acoustical Society of Japan (ASJ) spring meeting*, online, 2021.

3. Kai Li, Masato Akagi, and Yibo Wu. "Segment-level Effects of Gender, Nationality and Emotion Information on Text-independent Speaker Verification," In *the Acoustical Society of Japan (ASJ) 2020 spring meeting*, pp:879-882, Saitama University, Saitama, March 2020.