Master's Thesis

# Study on Deep-fake Speech Detection Based on Spectro-temporal Modulation Analysis

Haowei Cheng

Supervisor: Masashi Unoki

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

September 2023

# Abstract

Deep-fake speech refers to the use of artificial intelligence and machine learning techniques to manipulate and generate synthetic speech that imitates a person's voice or speech patterns. Just like deep-fake videos, which manipulate visual content to create realistic but fake videos, deep-fake speech involves altering or creating speech content to make it appear as if someone is saying something they didn't say. The process of creating deep-fake speech typically involves training a neural network on a large dataset of speech samples from the target speaker. This enables the AI model to learn the unique nuances, intonations, and speech patterns of the individual. Once the model is trained, it can generate new speech that sounds like it is coming from that specific person.

While the technology can have positive applications in speech synthesis for individuals with speech impairments or for voice-over work in entertainment, it also raises concerns about the potential for misuse and deception. Deep-fake speech can be used maliciously to create fake speech recordings of individuals, leading to issues such as misinformation, identity fraud, or the spread of false rumors. As technology advances, it becomes increasingly important to develop methods to detect and counteract the harmful effects of deep-fake speech.

Deep-fake speech detection is a critical research area aiming to develop effective techniques for accurately identifying fake speech generated using advanced deep-learning methods. The increase in deep-fake speech poses significant risks, including the potential for misinformation, fraud, and social engineering attacks. Detecting and mitigating the spread of deep-fake speech is essential for maintaining the integrity of speech-based communication and ensuring the trustworthiness of speech-based applications in various domains.

While humans can relatively easily distinguish between genuine and fake speech due to the remarkable capabilities of the human auditory system, machines face significant challenges in achieving the same level of discrimination. One major obstacle lies in effectively separating speech content from human vocal system information. Common features used in traditional speech processing, such as Mel-frequency cepstral coefficients (MFCC) and Gammatone cepstral coefficients (GTCC) struggle to handle this issue, leading to difficulties for neural networks in learning the discriminative differences between genuine and fake speech.

To address this fundamental challenge, our research delves into the

concept of spectro-temporal modulation (STM) representations in genuine and fake speech. These STM representations simulate the complex auditory perception process in the human auditory system, capturing both spectral and temporal modulations in speech signals. By incorporating STM representations, we enable our deep-fake speech detection system to exploit the dynamic characteristics of speech signals and effectively distinguish between genuine and fake speech.

Our proposed approach involves fitting the STM representations into a light convolutional neural network bidirectional long short-term memory (LCNN-BiLSTM) model for classification. The LCNN-BiLSTM model effectively captures temporal dependencies and long-range patterns in the STM representations, thereby enhancing the effectiveness of deep-fake speech detection capabilities.

To evaluate the performance of our approach, we conducted extensive experiments on benchmark datasets, including the Automatic Speaker Verification and Spoofing Countermeasures Challenge 2019 (ASVspoof2019) and the Audio Deep synthesis Detection Challenge 2023 (ADD2023). The results demonstrated the effectiveness of spectro-temporal modulation representations in distinguishing genuine and deep-fake speech. Achieving an equal-error rate of 8.33% on ASVspoof2019 and 42.10% on ADD2023, our method showed the potential of STM representations in deep-fake speech detection.

In conclusion, our research contributes to the ongoing efforts to combat the proliferation of deep-fake speech. By leveraging the insights from human auditory perception and developing novel techniques using STM representations, we offer a promising solution to address the challenges posed by artificial speech generation. The proposed deep-fake speech detection system holds considerable practical importance, safeguarding individuals from deception and ensuring the authenticity and credibility of speech-based interactions in today's digitalized world.

While the focus of this study evaluates the performance against baseline models, conducting comparisons with other existing deep-fake speech detection systems could provide a more comprehensive assessment of the proposed method's effectiveness. Future work will involve further investigation of the specific physics-based acoustic features that can be accurately captured and represented by the STM representation. Moreover, it is essential to expand the evaluation to encompass other state-of-the-art methods and to assess the system's performance across a diverse range of datasets.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction

## 1.1 Research background

With the rapid advancement of deep learning technology, our society has witnessed the emergence of various applications that have greatly enhanced our daily lives. For example, audio-books have become more accessible and engaging through the use of deep learning algorithms, allowing people to enjoy literature in a new and immersive way [1]. Intelligent speech robots have also gained popularity, providing assistance and companionship to individuals in various settings [2]. Moreover, the transformative power of deep learning technology has proven invaluable in aiding individuals afflicted by throat diseases and other medical conditions, enabling them to regain their voices. Through voice synthesis techniques, these individuals can regain their ability to communicate, improving their quality of life. In addition, it can also promote the development of new industries in entertainment, such as voice simulation of virtual characters in film, television, games and other productions, and customized personalized voices for self-media and other content creators to be used in communication platforms.

However, the increasing advancement of deep learning technology has brought about a new challenge: the widespread production of deceptive synthetic speech, known as deep-fake speech. When harnessed by malicious actors, this technology can inflict severe repercussions on multiple facets of society, spanning livelihoods, politics, economies, and social stability. Deep-fake speech involves the creation of artificial voices that closely resemble the speech patterns of specific individuals, often achieved through the misuse of deep learning techniques. The consequences of deep-fake speech are far-reaching and pose significant threats to societal stability and individual security. It has the potential to deceive both human listeners and automated speaker verification (ASV) systems, leading to activities such as identity theft, fraud, and the dissemination of false information.

Malicious attacks on deep-fake speech can target two main groups: human auditory systems and machine auditory systems. Attacks on human

hearing, particularly the increasing use of deep-fake speech manipulation, have captured worldwide attention. This technology allows the synthesis of a target speaker's voice, enabling the manipulation of individuals to make false statements and incite violence or engage in fraudulent activities. These actions can undermine state security, national property, and social stability. One significant concern is the possibility of deep-fake speech altering the statements of public figures, such as politicians. The spread of fabricated remarks attributed to political leaders can harm public trust, social unity, and the democratic system [3]. In an era where information greatly influences people's opinions and choices, the demand for reliable methods to detect deep-fake speech is increasingly critical. Additionally, voice forgery poses challenges to judicial forensics. On the other hand, attacks on machine hearing systems specifically ASV are also a concern. In the era of artificial intelligence, voice-based identity authentication plays a crucial role in securing access to intelligent devices and network transactions. Maliciously manipulating speaker verification systems through forged voices allows unauthorized access to a user's voiceprint, compromising their security and enabling control over their devices and accounts. This poses a significant threat to voice-based security access control. Therefore, it is imperative to propose an effective method for detecting deep-fake speech.

## 1.2 Research issues

In recent years, in order to effectively defend against the above-mentioned harms caused by maliciously abused fake speech to human beings and machines, several challenges have been organized by scholars to advance the field of deep-fake speech detection. These challenges serve as platforms to encourage the development of robust strategies to combat deep-fake speech. One notable global challenge is the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof) [4]. The challenge aims to promote research and innovation in detecting and mitigating spoofing attempts in speaker verification systems. Another significant challenge is the Audio Deep Synthesis Detection (ADD) challenge [5], which focuses on detecting deep-fake speech in realistic scenarios. By providing a benchmark dataset and evaluation metrics, this challenge encourages the development of effective methods for deep-fake speech detection in real-life settings.

Despite the availability of various challenges and proposed methods for deep-fake speech detection, accurately distinguishing between genuine speech and fake speech remains a challenging task for machines. The main difficulty arises from the inherent differences in the characteristics of genuine speech

and machine-generated fake speech. Genuine speech exhibits not only linguistic content but also reflects human vocal system activity, including unique characteristics such as glottal vibration. In contrast, machine-generated fake speech lacks these human-like characteristics, making it challenging for machines to accurately differentiate between them. Commonly used features struggle to effectively capture the distinct patterns associated with genuine speech and human vocal system activity, resulting in insufficient discriminative information for training neural networks.

To achieve effective detection of deep-fake speech, it is crucial to successfully separate the components of speech signals. By effectively separating the linguistic content from the characteristics related to human vocal system activity, it becomes possible to capture the unique patterns and features associated with genuine speech. This separation can be achieved through advanced techniques, such as the utilization of spectro-temporal modulation (STM) representations, which simulate the auditory perception process in the human auditory system. By integrating STM representations into deep-fake speech detection systems, an effective method for distinguishing between genuine and fake speech can be achieved.

## 1.3 Research motivation

To address the issues in deep-fake speech detection, we drew inspiration from the remarkable capabilities of the human auditory mechanism. Extensive research has shown that the human auditory cortex possesses dynamic and adaptive properties that enable us to effectively differentiate between speech produced by humans and machine-generated speech [6]. Building upon this understanding, a study [7] discovered that auditory cortex neurons possess the capability to transform spectrograms, resulting in STM content. As a consequence, this breakthrough paved the way for the STM's development—an innovative multi-scale representation utilized for speech analysis, which has demonstrated its efficacy in explaining various psychoacoustic phenomena [8]. Another study [9] proposed an STM-based method for audio classification inspired by human auditory mechanisms. By utilizing an auditory model to capture relevant features, these methods have demonstrated their effectiveness in classification. Therefore, it is reasonable to consider that STM representation has the potential to discriminate deep-fake speech and enhance detection accuracy.

## 1.4 Research purpose

The primary research objective of this study is to develop effective techniques for deep-fake speech detection. By achieving this objective, the ultimate goal is to mitigate the negative impact of maliciously produced or disseminated fake speech in various real-life scenarios.

To achieve the research objective, the study investigates and analyzes the role of feature expressions and spectro-temporal modulation representations in both genuine and fake speech, simulating the human auditory perception process. These representations offer valuable insights into the dynamic characteristics of speech signals, providing crucial information for differentiating between genuine and deep-fake speech. Leveraging this knowledge, the research aims to design a deep-fake speech detection model that effectively exploits spectro-temporal modulation representations to make accurate and reliable classifications.

Furthermore, the study aims to contribute theoretically by offering valuable insights into the underlying mechanisms of human auditory perception and the unique features that distinguish genuine speech from deep-fake speech. By exploring the capabilities of cochlear and auditory cortex perception in recognizing deep-fake speech, the study sheds light on the challenges and opportunities in developing reliable detection systems.

## 1.5 Organization of thesis

The thesis is organized as illustrated in Figure 2.2. The organization along with its details can be described as follows:

Chapter 1 introduces the background of the thesis topic. It provides a brief overview of deep-fake speech and the potential harm it can cause. The importance of addressing the challenges associated with deep-fake speech is emphasized.

Chapter 2 reviews the types of deep-fake speech, the related works in deep-fake speech detection, the concept of human auditory mechanism, and spectro-temporal modulation.

Chapter 3 describes the details of spectro-temporal modulation and investigates the role of feature expressions, as well as the simulation process of implementing spectro-temporal modulation.

Chapter 4 presents a method for detecting deep-fake speech, which encompasses the framework, feature extraction, and details of the identification model.

Chapter 5 shows the datasets and evaluation metrics used in this thesis, along with comparison experiments that utilize these common measurements to assess the results. Subsequently, the obtained results serve as the foundation for a comprehensive general discussion.

Chapter 6 summarizes the whole work in the master's program, consisting of the research contribution and the remaining works.

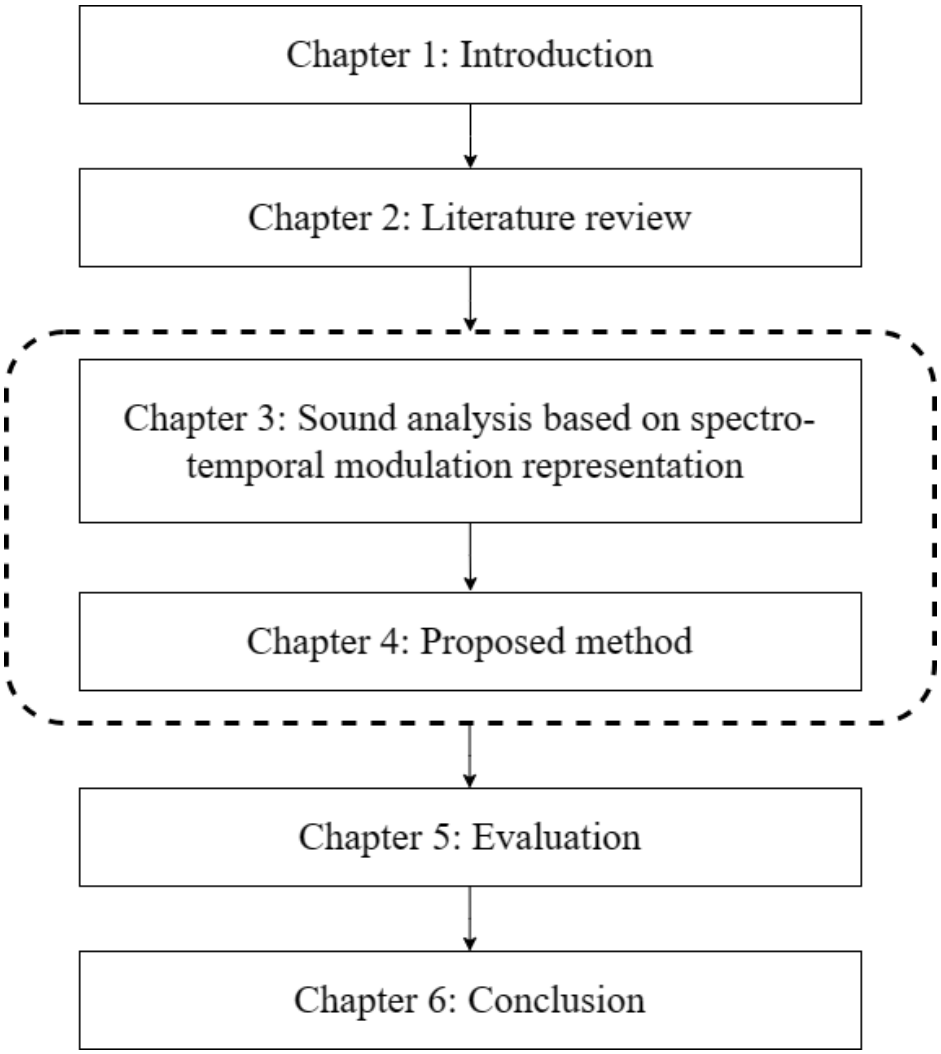


Figure 1.1: Organization of the thesis

# Chapter 2

# Literature Review

## 2.1 Deep-fake speech

### 2.1.1 Types of deep-fake speech

Deep-fake speech can be categorized into three distinct types, as shown in Figure 2.1.

1. **Replay-based:**
   Replay-based deep-fake speech refers to malicious techniques used to replicate or reproduce recorded voices of individuals. These deceptive works aim to create realistic imitations of the interlocutor's voice, often for fraudulent purposes [10]. Two distinct categories of fake speech can be identified: far-field detection and cut-and-paste detection. Far-field detection involves the attacker utilizing a hands-free phone to play a recorded voice of the victim as a test segment [11]. On the other hand, cut-and-paste attacks entail generating a requested sentence using a text-dependent system [12]. In response to replay-based attacks, text-dependent speaker verification methods have been suggested [10] [13]. These methods leverage the distinct traits of the speaker's voice to authenticate the speech. Additionally, an emerging strategy for detecting end-to-end replay attacks involves the use of convolutional neural networks, capable of learning intricate patterns and features to discern manipulated or synthesized voices [14].
2. **Synthetic-based:**
   This thesis centers on synthetic-based fake speech. The structure of this subsection is illustrated in Figure 2.2.
   Speech synthesis generates the target speaker's voice from the specified linguistic text, realizing text-to-speech (TTS), which converts written text into real-time, natural-sounding speech [15]. A common configuration for speech synthesis involves two main components: the analysis of front-end text and the generation of speech waveforms on the back-end, as shown in Figure 2.3. Text analysis generates the corresponding

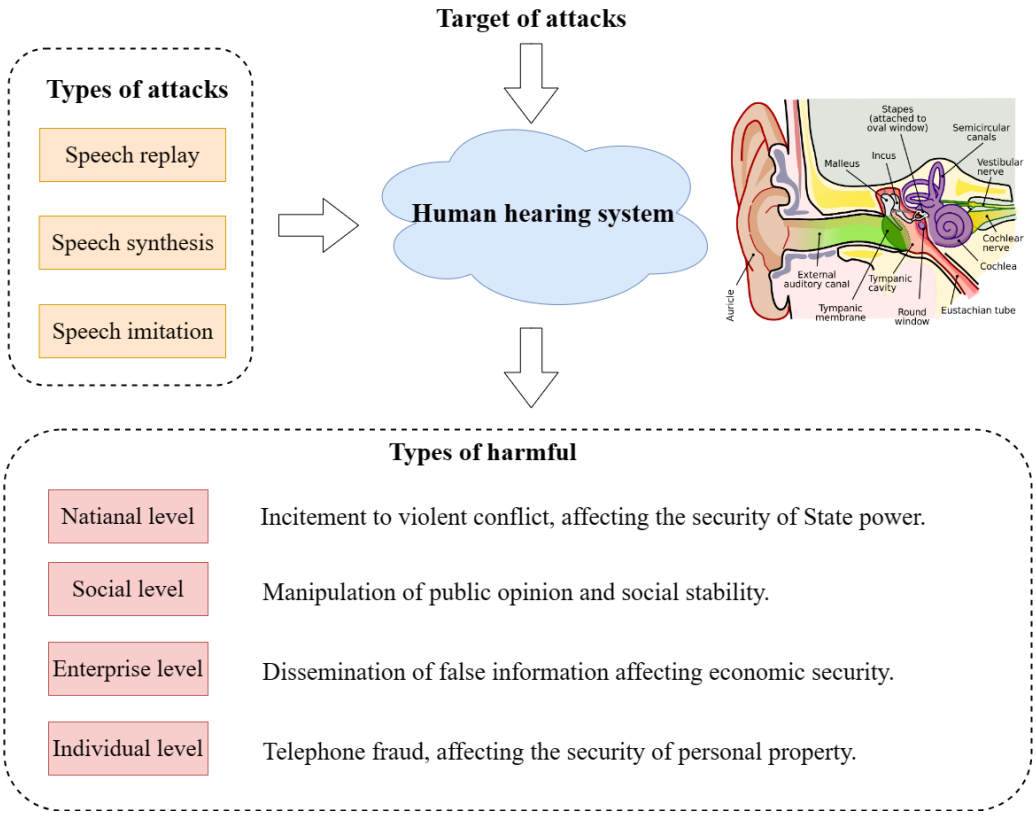


Figure 2.1: Types and hazards of deep-fake speech

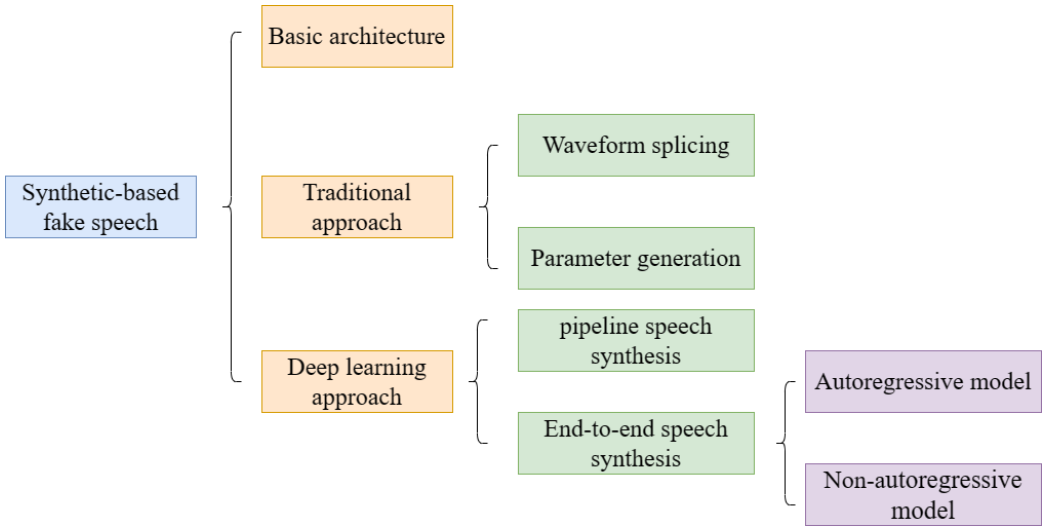


Figure 2.2: Structure of the synthetic-based fake speech subsection

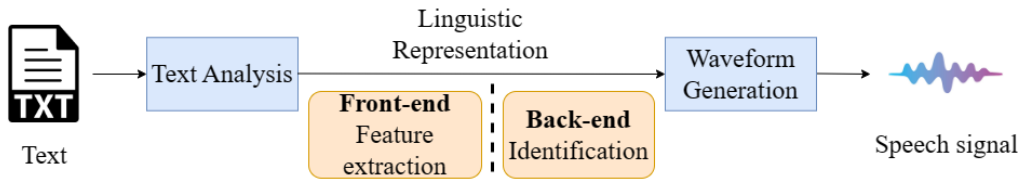


Figure 2.3: Basic architecture of synthetic-based approach

phoneme sequences, duration prediction, and other information from the input text through normalization, word segmentation, lexical annotation, and other steps. Speech waveform generation synthesizes the target speaker's voice waveform according to linguistic specifications generated by text analysis. The progress of deep learning has led to a gradual shift in speech synthesis, transitioning from traditional methods to deep learning-based speech synthesis. Currently, with the advancement of deep learning, the incorporation of the end-to-end speech synthesis mechanism into the technology has been a gradual process, i.e., connecting the text analysis and waveform generation process, directly inputting text or annotated characters, and outputting speech waveforms. This subsection will introduce representative traditional speech synthesis work and the latest deep learning speech synthesis work respectively.

**Traditional approach:**

Traditional speech synthesis mainly includes the waveform splicing method and the parameter generation method. Waveform splicing speech synthesis splices speech units from natural speech data according to certain rules to synthesize speech that is highly similar and natural to the target speaker, including corpus collection, acoustic unit selection, splicing and forgery, and other steps. Simple methods of waveform splicing speech synthesis use editing software to directly modify the audio signal by cutting, inserting, copying, and pasting operations, i.e., copy-paste tampering. More complex splicing methods will adjust and control the rhythm of each splicing unit in order to obtain more natural and smooth synthesized speech, representative works include the PSOLA technique for base-step superposition [16] and unit selection system using the Hidden Markov Model (HMM) to limit rhythmic parameters of target unit [17]. In recent years, waveform spliced speech synthesis mostly adopts deep learning techniques, such as Google proposed a system in 2017 that utilizes a sequence-to-sequence LSTM self-encoder for real-time unit selection. [18]. Waveform spliced speech synthesis is suitable for some specific domains such as weather forecasting, time reporting, financial services, etc. The method uses real speech segments, which maximizes the preservation of speech sound quality and allows for the synthesis of highly naturalistic speech. However, it requires a large amount of target speaker corpus and is not stable for text synthesis in different domains, which can be easily recognized by humans or machines. Parameter generation-based speech synthesis predicts acoustic parameters through acoustic modeling and synthesizes the target speaker's speech from the acoustic parameters

through a vocoder. The traditional representative works of parametric generation speech synthesis techniques include HMM-based statistical parameter synthesis methods [17] and DNN-based parameter synthesis methods [19]. Parametric speech synthesis can output stable and smooth speech, but the synthesized speech is usually not natural enough due to the defects of parametric synthesizers and the loss of statistical modeling, such as insufficient smoothing of the generation parameters and insufficient accuracy of HMM modeling.

**Deep learning approach:**

Deep learning technology has greatly influenced speech synthesis in recent years, leading to its predominant adoption of deep learning methods. These methods primarily fall into two categories: pipeline speech synthesis and end-to-end speech synthesis.

**Pipeline speech synthesis** as a whole can be categorized into three distinct components that encompass text analysis, acoustic modeling, and vocoder. In the text analysis component, each phoneme undergoes rhyme prediction and duration prediction based on the input text, the acoustic model establishes the connection between text features and acoustic features, and maps to acoustic features based on the output of the text analysis via DNN: the vocoder module realizes the conversion of acoustic parameters to speech waveforms. In 2017, Baidu Artificial Intelligence Laboratory used a neural network model to replace the submodule of the traditional parametric speech synthesis, combined with the improved WaveNet acoustic coder, which is the most efficient way to synthesize speech [20]. In 2018, Baidu AI Lab further proposed Deep Voice3, a fully convolutional speech synthesis system that employs an attention-based approach, which uses an encoder to convert text into high-level feature representations, and an autoregressive decoder to generate Mel-scale spectrograms, while introducing a non-autoregressive-post-processing-converter [21]. The network functions as a converter, utilizing the decoder's hidden state to predict vocoder parameters, thereby enhancing the speed of speech synthesis. Pipelined speech synthesis harnesses the potent learning capabilities of deep learning to address the limitations of traditional statistical construction mode speech synthesis to some extent, but the accumulation of errors between multiple modules and the need for costly text annotation as well as the mandatory alignment of textual and acoustic features, the above problems limit its speech synthesis effectiveness.

**The end-to-end speech synthesis** system realizes direct text input or phonetic characters and output of audio waveforms, which greatly

reduces the complexity of the speech synthesis system construction, and reduces manual intervention in the synthesis process and the need for linguistics-related background knowledge. The reduction of modules in the end-to-end system also effectively avoids the accumulation of errors caused by the multi-stage modeling of traditional methods and achieves a significant improvement in the performance of speech synthesis. The present deep learning-based end-to-end speech synthesis systems can be classified into two categories: autoregressive and non-autoregressive models.

Autoregressive speech synthesis is based on a sequence-to-sequence generation system, which can achieve the current optimal speech synthesis effect, but the synthesis speed is slow. In 2016, Google proposed the WaveNet speech synthesis algorithm, which generates the next sampling point based on the current moment sampling point by expanding a causal convolutional network [22]. The model directly models the raw speech data, avoiding the loss of sound quality that results when a vocoder parameterizes the speech. However, this system cannot realize the direct conversion of input text or labeled notes to output speech. In 2017, Wang et al. proposed Tacotron, the first end-to-end speech synthesis system [23]. The model introduces an attention mechanism, inputs text or annotated characters, outputs a linear spectrogram, and generates speech waveforms by the Griffin-Lim algorithm. All feature models in Tacotron can be self-learning and tuned, and it is easy to add language, timbre, emotion, and other constraints, but the model is complex, with poor error correction and human intervention, and the sound quality is not as good as that of the speech synthesized by the WaveNet-based synthesizer. In 2018, Tacotron2, an improved system of Tacotron, simplified the structure of the front-end model and generated the Mel spectrogram based on the text input, and then synthesized speech waveforms by an improved WaveNet vocoder to synthesize speech waveforms and build a complete speech synthesis system [24]. The system produces synthesized speech that is close to the human voice. In 2021, Weiss et al. extend the Tacotron system by adding normalized streams to the autoregressive decoder and propose the end-to-end speech synthesis system Wave-Tacotron, which does not generate intermediate features and therefore does not require a vocoder, allowing for end-to-end speech synthesis of text to waveforms [25].

Non-autoregressive speech synthesis relies on a fully parallel network structure, enabling the generation of complete speech through a single feed-forward computation. This significant advancement enhances the

speed of speech synthesis, offers better controllability, and achieves speech synthesis quality comparable to that of the autoregressive model. In 2019, Wang et al. proposed a real-time parametric speech synthesis model based on source filtering [26]. Based on the training criterion of spectral distance and phase distance, the model generates sinusoidal excitation signals for a given input acoustic feature by means of a source module, fundamental F0, and harmonic additive noise model, and then converts the excitation signals and spectral features into speech waveforms by means of cascaded dilation convolution and Long-Short Term Memory (LSTM) network. In 2019, Ren et al. proposed FastSpeech, a Transformer-based speech synthesis system that extracts attentional alignment from an encoder-decoder-based teacher model to predict phoneme durations, extends the source phoneme sequences to match the length of the target phoneme sequences through length adjustment and generates Mel spectrograms in parallel. This system is comparable to the autoregressive model in terms of speech quality, while greatly improving the speed of speech generation [27].In 2021, Elias et al. expanded the Tacotron system by incorporating a residual encoder derived from the Variational Auto-Encoder (VAE) and introduced an end-to-end speech synthesis system known as Parallel Tacotron [28]. The system incorporates VAE and selective spectral loss to enhance the naturalness of speech synthesis, while utilizing LCNN to implement the self-attention mechanism, thereby improving generation efficiency. Remarkably, this model achieves substantial improvements in speech generation speed while maintaining speech quality on par with Tacotron2.

3. **Imitation-based**

   Imitation is a technique used to transform the original speech of one speaker (referred to as the "original") to resemble that of another speaker (referred to as the "target") [29]. Through the use of an algorithm based on imitation, the spoken signal undergoes processing and modification to closely replicate the style, intonation, or prosody of the target voice while retaining the linguistic information of the original speech, resulting in a convincing imitation [30]. This process involves adjusting the acoustic characteristics of the speech signal to align with the target speaker, effectively converting the voice from one individual to another.

   The method is occasionally confused with the synthetic-based approach, as the generation process lacks a clear distinction between the two. Both techniques involve modifying the acoustic-spectral and style attributes of speech signals. However, in the context of imitation-

based speech, the original input and resulting output text usually remain unchanged. The primary focus lies in adjusting the sentence's delivery to match the unique characteristics of the target speaker. This approach aims to achieve a voice transformation while retaining the original linguistic content of the sentence [31].

### 2.1.2 Speech forgery challenges

Mainstream speech forgery competitions include Blizzard Challenge, Voice Conversion Challenge (VCC), etc. Some international conferences on acoustics also organize voice forgery challenges, such as the Multi-Speaker Multi-style Voice Cloning Contest (M2VoC) organized by International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

The Blizzard Challenge, an annual international speech synthesis competition co-sponsored by Carnegie Mellon University and Nagoya Institute of Technology, Japan, it was first organized in 2005 and has been held for 16 years. The competition aims to build an open and unified evaluation platform for speech synthesis technology and promote the rapid development of speech synthesis technology. Participants are required to assemble a set of prescribed test phrases based on official data, and The organizer will assess the synthesized speech's fidelity through auditory examination during the test.

The Voice Conversion Challenge (VCC) is a biennial international voice conversion competition jointly sponsored by several universities, including Nagoya University and the University of Science and Technology of China. It was first organized in 2016 and has been held three times (2016, 2018, and 2020). The competition promotes the development of Voice Conversion (VC) technology by releasing open datasets and organizing competitive evaluations for outstanding problems and challenges facing VC technology. Participants are required to realize speech conversion (e.g., semi-parallel conversion, cross-language conversion, etc.) for specific tasks based on the officially released specified datasets, and the quality of the converted speech will be evaluated by the organizer, considering both its naturalness and speaker similarity.

The Multi-speaker Multi-style Tone Drop Competition (M2VOC) is one of the signal processing challenges of ICASSP 2021. It addresses the current limitations of multispeaker and multi-style speech forgery and provides a common dataset and testbed to facilitate the development of speech cloning. M2VoC 2021 has two tracks, the few samples track and the very few samples track. Participants are required to generate sentences and short paragraphs based on speech samples, and the organizer evaluates the clones for speaker similarity, voice quality, style/expression, and pronunciation accuracy.

## 2.2 Deep-fake speech detection

### 2.2.1 Current common challenges

**ASVspoof challenge:** The realm of fake speech detection hosts its most prominent and comprehensive challenge known as the biennial Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof challenge). This event is organized and initiated by esteemed institutions like the University of Edinburgh (UK), EURECOM (France), NEC (Japan), and the University of Eastern Finland (Finland). Its core purpose revolves around fostering the advancement of ASV fake speech detection techniques through the dissemination of open datasets and the facilitation of competitive evaluations. The ASVspoof challenge was first held in 2015 and has been held four times (2015, 2017, 2019, 2021). The ASVspoof 2015 challenge addresses both synthetic and conversion forgery attacks, while the ASVspoof 2017 challenge targets replay attacks. In contrast, the ASVspoof 2019 challenge addresses multiple synthesis/conversion forgery attacks and replay attacks across LA and PA subtasks. Lastly, the ASVspoof 2021 challenge comprises three subtasks: LA scenario, PA scenario, and deep-fake speech detection without ASVs. Compared with the previous three competitions, ASVspoof 2021 has three improvements: first, the challenge of LA scenario is extended to explore detection algorithms robust to channel variations; second, the subtask of PA scenario is more realistic by using audio recorded in real physical environments; and lastly, it is the first time that this competition explores deep-fake detection of speech with the purpose of spoofing human beings.

**ADD challenge:** However, there has been a recognition that many real-life scenarios have not been adequately covered in the ASVspoof challenge. To address this gap, the Audio Deep synthesis Detection Challenge (ADD) was motivated. In 2022, the inaugural ADD challenge featured three subtasks: detecting low-quality fake audio, detecting partially fake audio, and audio fake game. These subtasks collectively aim to comprehensively evaluate the capabilities of ASV systems in identifying various types of voice spoofing attacks, covering both low-quality and partially manipulated audio samples. The inclusion of these tracks in the ADD challenge aims to advance research in the field and encourage the creation of more robust and efficient solutions to counter speech-spoofing attacks. Additionally, there is growing interest in going beyond the limitations of binary classification (genuine/fake) and instead focusing on identifying and localizing specific manipulated intervals within the partially fake speech. Moreover, there is a need to pinpoint the source responsible for generating the fake audio. In order to encourage researchers worldwide to develop new and innovative technologies to address

these challenges, ADD 2023 was proposed. This challenge aims to accelerate and promote research in the detection and analysis of deep-fake utterances.

### 2.2.2 Methods for detecting deep-fake speech

The basic idea of deep-fake speech detection is to find the feature differences between genuine speech and fake speech. The typical fake speech detection system is generally composed of two parts: front-end and back-end. The front-end extracts distinguishing features by analyzing the speech signal, and the back-end judges whether the speech is genuine or fake by classification. Traditional detection systems use manual features designed by experts for discriminative features in the front-end, while the back-end directly uses Gaussian Mixture Model (GMN) or Support Vector Machine (SVM) for classification decisions. In recent years, deep learning-based systems have gradually become mainstream systems. The front-end extracts speech features from input neural networks, while the back-end learns advanced representations of features through neural networks and then performs classification judgments. At present, some end-to-end systems have emerged that can directly use the original audio waveform as network input, learn advanced feature representations, and make decisions. The architecture of a generic fake speech detection system is shown in Figure 2.4

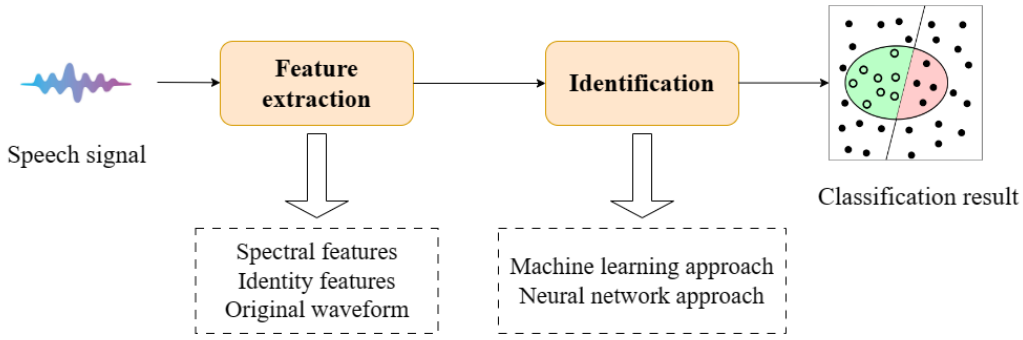


Figure 2.4: The architecture of a generic fake speech detection system.

1. **Feature extraction:** The front-end of fake speech detection system primarily concentrates on extracting distinctive features from fake speech that can be used for identification by the back-end. This involves constructing features based on the absence of spectral and temporal details in fake speech, enabling effective differentiation from genuine

speech. Traditional detection methods mostly use carefully designed manual features. With the popularization of data-driven deep learning methods, front-end features gradually develop towards relatively low-level features such as spectra. Figure 2.5 summarizes the commonly used features in current fake speech detection, which can be mainly divided into three categories: spectral features, identity features, and original waveforms, with spectral features being the most widely used. Numerous features have been developed, drawing inspiration from the human auditory mechanism. These features can be categorized as follows:

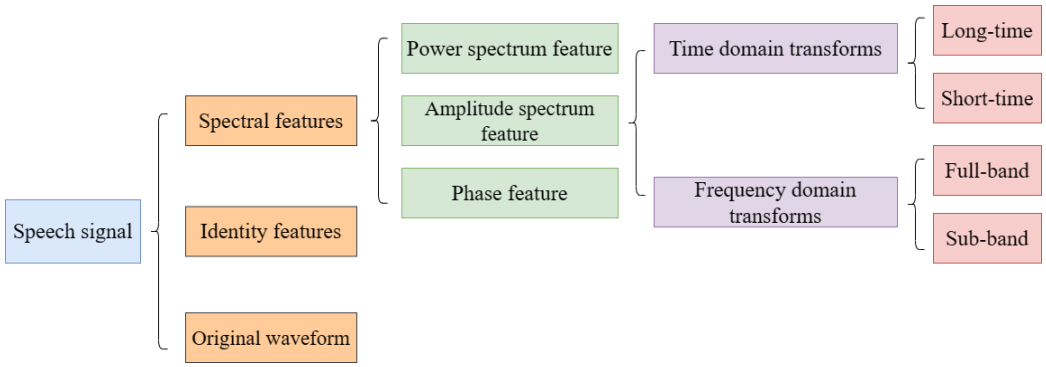


Figure 2.5: Commonly used features for fake speech detection

**The short-term power spectrum feature** describes the variation of signal power with frequency. Commonly used include Mel Frequency Cepstral Coefficients (MFCC), Rectangular Filter Cepstral Coefficients (RFCC), Linear Frequency Cepstral Coefficients (LFCC), Inverted Mel Frequency Cepstral Coefficients (IMFCC), and Linear Prediction Cepstral Coefficients (LPCC). Due to the inability to effectively simulate the temporal characteristics of forged speech, the high-order coefficients of cepstrum and the first-order and second-order dynamic differential coefficients are both beneficial for detecting fake speech [32].
**The amplitude spectrum features** currently used for fake speech detection include Log Magnitude Spectrum (LMS) and Residual Log Magnitude Spectrum (RLMS). LMS contains details of speech signals, such as Formants, pitch and harmonics in spoken vowels. The RLMS is extracted from the residual waveform of Linear predictive coding (LPC), which contains spectrum details such as harmonics. Compared with LMS, RLMS eliminates the Formant effect [33].

Due to the neglect of phase information during waveform reconstruction by the vocoder, **short-time phase feature** are one of the effective features for detecting fake speech. The phase spectrum obtained from the Fourier transform has phase distortion, so it needs to be processed to obtain stable and effective phase features. Common phase features include group delay (GD), modified group delay (MGD), the baseband phase difference (BPD), etc [33].

**The long-time transformation features** extracted from speech signals are effective for detecting fake speech. For example, the Cochlear Filter Instantaneous phase and frequency Coefficients with Instantaneous Frequency (CFCCIFs) [34] and the Constant Q Cepstral Coefficients (CQCC) [35] based on the long-term constant Q transform, which performed best in the ASVspoof 2015 challenge, they are all based on the long-time transform.

2. **Identification:** In a fake speech detection system, the back-end is responsible for processing and classifying the features extracted from the front-end to determine speech authenticity. Traditionally, feature-based machine learning methods, like Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) based classifiers, have been commonly utilized in the back-end to directly classify and screen the manual input features. GMM employs multiple Gaussian distribution functions in a linear combination to fit any distribution. However, with the rise of deep learning, the latest fake speech detection systems primarily rely on the feature representation capabilities of deep neural networks (DNNs) and classification network-assisted neural networks in the back-end. These systems harness the feature learning ability of neural networks to obtain more sophisticated feature representations from the input features of the front-end before conducting the classification.

   Currently, most research in this field is focused on specific attack types or is based on fixed datasets, making it challenging for a single system to efficiently detect multiple types of forgery attacks (such as Text-to-Speech (TTS), Voice Spoofing (VS), TTS-VC hybrid, etc.) and unknown attacks that were not encountered during training. Since it is difficult to anticipate the specific type of attack in practice, the latest research efforts are directed toward improving the generalization of detection. The goal is to design a universal detection system capable of handling different types of forgery attacks and resisting noise interference in various channel environments. Among the different architectures used for the back-end, the CNN architecture is widely adopted in the implementation of these systems. Variants like light Convolutional neural network (LCNN), deep residual network (ResNet), squeeze-and-

excitation networks (SENet), etc., are commonly used. In some cases, recurrent neural network (RNN) architectures like Gated Recurrent Units (GRU) are introduced to capture sequence context information. The primary focus in enhancing the discriminative performance of classifiers is to learn more advanced and robust feature representations. The aim is to ensure that feature distances between similar samples are kept as close as possible while maintaining a considerable separation between feature distances of dissimilar samples.

ResNet was originally introduced to address the problem of degradation in deep networks and the vanishing gradient problem. In fake speech detection networks based on CNN, when the network becomes deep, it encounters the vanishing gradient problem. This problem hinders the lower layers from receiving useful update information during training, which makes it difficult for the network to learn more advanced and differentiated feature representations. ResNet effectively solves this issue by using skip connections as shortcuts, reducing the training parameters of deep networks and allowing faster propagation of parameter updates to lower layers during training. As a result, ResNet has become one of the most widely used networks in fake speech detection. Several studies have utilized ResNet for fake speech detection. In 2019, Alzantot et al. proposed a deep ResNet-based detection scheme that fused three different front-end features (MFCC, spectrogram, CQCC) [36]. In 2020, Li et al. introduced a detection scheme based on Res2Net, which divided features within the same block into multiple channel groups, allowing different groups to undergo different scale transformations and incorporating residual-like connections between groups. This structure enabled the learning of multi-scale feature representations, improving the system's generalization against unknown attacks [37]. Additionally, in 2020, Parasu et al. proposed a Light ResNet structure, which demonstrated good universality, effectively crossing databases and attack types [38]. This simplified version of ResNet reduced parameters to prevent overfitting while maintaining a lightweight network. Experimental results showed that this model outperformed the CQCC-GMM baseline system of the ASVspoof2019 Challenge in cross-dataset detection.

LCNN, proposed by Wu et al. in 2018, it was initially applied in facial recognition and has proven to be effective in fake speech detection [39]. The best system of the ASVspoof2017 Challenge and the best single system of the ASVspoof2019 Challenge LA scene were based on LCNN. LCNN can efficiently process large-scale data with numerous noise labels while reducing computing costs and storage space. Its

core innovation lies in the introduction of the Max Feature Map (MFM) Activation function after each convolution layer. The MFM activation produces more compact feature maps compared to the ReLU Activation function, enabling feature selection and dimension reduction simultaneously. Experiments have shown that MFM activation discards noise effects, such as environmental noise and signal distortion while preserving core information and enhancing the learning ability of lower-level network features. In 2019, Gomez Alanis proposed a lightweight convolutional gated recurrent neural network (LC-GRNN) based on GRU, which combined LCNN and RNN [40]. LC-GRNN combined the strengths of LCNN and RNN to extract discriminative features at various levels and learn contextual features. In 2020, Wu et al. introduced a CNN-based Genuinization model for fitting genuine speech distribution [41]. This model received manual feature inputs extracted from the front-end and generated advanced feature representations more conducive to distinguishing speech authenticity. The model focused on fitting the distribution of real speech, which was considered easier than the diverse distribution of fake speech. The feature normalization model amplified the differences between true and false speech. Its structure resembled that of an Autoencoder. In 2021, Kuak et al. proposed the detection system ResMax, which combined the MFM Activation function and ResNet's residual structure within LCNN. ResMax was a single model with fewer parameters but exhibited good detection performance [42].

The Squeezing Excitation Network (SENet) was introduced by Hu et al. in 2018, focusing on demonstrating the interdependence between modeling channels through squeezing excitation operations [43]. The Squeeze Excitation (SE) module in the network first compresses the feature map to obtain channel-level global features and then stimulates these global features to learn the relationships between each channel. By assigning different influence weights to different channels, the model's ability to focus on the most relevant channel information for forgery detection is improved. In 2019, Lai et al. proposed a detection system called ASSERT, which integrated SENet and ResNet along with three other systems: mean standard deviation ResNet and expanded ResNet [44]. The combination of SENet and network architectures like ResNet led to enhanced detection results. In 2020, Li et al. introduced a detection scheme based on Res2Net, which integrated a squeezing excitation module into the Res2Net network, further improving the detection performance [37]. In 2021, Hemavathi et al. used blind source separation (BSS) technology based on non-negative Matrix de-

composition to decompose synthetic speech into real speech and artifact components [45]. They then employed a CNN-based classifier to classify the target speech. Although this paper did not compare with the most advanced detection methods, it presented a new idea in the detection field. Similarly, in 2021, Chen et al. proposed a spoof print system for counterfeit speech detection, which deviated from the traditional detection system framework [46]. This system resembling the ASV system consisted of two stages: registration and verification. In the registration stage, the speaker watermark model was learned based on real speech samples of the target speaker. In the verification stage, the authenticity of the test sample was determined by calculating the cosine similarity between the test sample and the speaker watermark. In the same year, Luo et al. introduced a detection network based on a capsule network (CapsNet) and modified the Dynamic routing algorithm of the original network [47]. This modification forced the network to learn to forge artifacts in the voice, thereby enhancing the system's generalization ability.

### 2.2.3 Evaluation metrics

Equal Error Rate (EER) and tandem detection cost function (t-DCF) serve as prevalent performance metrics in the evaluation of fake speech detection systems.

EER is the error rate when False Accept Rate (FAR) and False Rejection Rate (FRR) are equal. Fake speech detection involves categorizing speech as either genuine or fake. An error acceptance occurs when the detection system misclassifies fake speech as genuine speech. On the other hand, an error rejection happens when the system misclassifies real speech as fake speech. Given the detection score and threshold of the detection system. The EER is defined as follows. Let $P_{fa}(\theta)$ and $P_{miss}(\theta)$ denote the false alarm and miss rates at threshold $\theta$.

$$P_{fa}(\theta) = \frac{\text{fake samples with score} > \theta}{\text{total fake samples}} \tag{2.1}$$

$$P_{miss}(\theta) = \frac{\text{genuine samples with score} < \theta}{\text{total genuine samples}} \tag{2.2}$$

Therefore, $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are monitonically decreasing and increasing functions of $\theta$, respectively. The EER corresponds to the threshold $\theta_{EER}$ at which the two detection error rates are equal, i.e. $EER = P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER})$. A smaller EER value indicates the superior performance of the fake speech detection system.

t-DCF is an evaluation indicator introduced in the ASVspoof 2019 challenge. In practical scenarios, the fake speech detection system plays an auxiliary role in decision-making for the ASV system. This indicator is not employed for the isolated evaluation of the fake speech detection system. Instead, it reflects the collective influence of fake speech and the system on the actual performance of the ASV system in real-world scenarios. t-DCF draws on minimum risk Bayesian decision-making for system reliability evaluation. In real-world scenarios, ASV systems often face legitimate users, temporarily impersonated illegal users, and attackers trying to manipulate ASV decisions for malicious purposes. This indicator takes into account the potential consequences of misjudgments in various situations. The calculation processing is as follows.

$$\min\ t - DCF = \min\left\{\beta P_{miss}(\theta) + P_{fa}(\theta)\right\} \tag{2.3}$$

$P_{fa}(\theta)$ and $P_{miss}(\theta)$ are the error acceptance rate and error rejection rate of the fake speech detection system when the threshold is $\theta$. And the coefficient $\beta$ depends on the actual priority of fake attacks, misjudgment costs, and the detection performance of the ASV system. A smaller t-DCF value indicates the better generalization performance of the fake speech detection system.

### 2.2.4 Current limitations

Despite many features are employed in deep-fake speech detection tasks, it is difficult for machines to distinguish them accurately. On the other hand, humans can distinguish genuine and fake speech through our sense of hearing. Taking inspiration from the human auditory mechanism, our approach involves exploring feature representations that not only capture the speech content but also the subtle cues associated with human vocal system activity.

## 2.3 Human auditory mechanism

The effectiveness of features related to the human auditory mechanism in detecting deep-fake speech stems from the unique characteristics and adaptability of our auditory perception. With its heightened sensitivity and distinctive perception abilities, the human auditory system has evolved to adeptly process and analyze intricate auditory stimuli, encompassing speech. This natural adaptation enables the differentiation between genuine and fake speech by capturing and discerning subtle features within speech signals.

Here are some possible reasons why the human ear can distinguish between true and false speech:

1. Human beings grow up in language and speech environments, receiving speech training that provides us with rich experience and memory for genuine speech perception and patterns. This extensive training helps us distinguish between normal, natural speech patterns, and potential anomalies.
2. Humans develop unique perceptual models for the voices and pronunciation habits of different speakers through experiential contact and auditory memory. This cognitive model allows us to detect features that deviate from our familiar sound models, aiding in identifying potential inconsistencies in speech.
3. The human ear exhibits a keen ability to perceive multiple features in speech signals, such as pitch, formant, and sound quality, which reflect the sound source, vocal tract characteristics, and pronunciation habits during speech production. Analyzing these features enables us to identify subtle differences in sound and distinguish between true and false speech.
4. When people understand and interpret speech, they not only rely on individual sound features but also integrate contextual information, including grammar, vocabulary, sentence structure, and related nonverbal cues such as facial expressions and body language. This contextual information plays a crucial role in helping us judge the authenticity and consistency of speech.

The human auditory mechanism is critical in distinguishing between genuine and fake speech. It can analyze speech sounds by detecting changes in both the spectral and temporal domains, allowing us to understand speech even in challenging acoustic environments. For instance, a study conducted at the University of Geneva1 discovered that the auditory cortex amplifies different aspects of sounds based on the task at hand. Voice-specific information is prioritized for voice recognition tasks, while other aspects of the sound are amplified for other tasks. These findings highlight the complex and dynamic nature of human auditory processing, enabling us to effectively differentiate between genuine and fake speech. [48]

# Chapter 3

# Sound analysis based on spectro-temporal modulation representation

## 3.1 Concept of spectro-temporal modulation

Temporal modulation refers to changes in modulations over time in the spectrogram, while spectral modulation represents variations along the frequency axis. Spectro-temporal modulation (STM) combines both temporal and spectral modulations simultaneously, providing a comprehensive representation of the dynamic characteristics of a signal. In the field of auditory psychophysics and neuroscience, the auditory model is divided into two essential stages: transforming the acoustic signal into an auditory spectrogram and analyzing this spectrogram to estimate spectral and temporal modulation content using specialized filters that respond to specific modulations [49, 50].

In the initial stage, the acoustic signal undergoes the transformation into an auditory spectrogram, which serves as an internal neural representation. This spectrogram encapsulates the distribution of energy across various frequency bands over time. The second stage analyzes the auditory spectrogram and extracts information related to spectral and temporal modulations. This analysis is achieved using specialized filters that are sensitive to specific modulation rates and frequency ranges [51–53]. By separating different cues and characteristics associated with distinct auditory percepts, this stage resembles the adaptive and masking properties of the human auditory system. It enables vital information to be perceived even in noisy environments.

Incorporating STM analysis provides a more comprehensive understanding of human perception and can reveal meaningful characteristics in the speech signal that aid in the detection of deep-fake speech. The STM representation is obtained through a series of steps. Initially, the input signal is decomposed into frequency components using filterbanks, which divide the speech signal into distinct frequency bands. Following this, the power envelope is computed

through the process of squaring the output of the filterbank. Finally, a two-dimensional spectral analysis is performed on this power envelope to derive the STM spectrogram. This spectrogram represents the dynamic variations present in the speech signal across different spectral and temporal scales. The process of STM calculation processing is shown in Figure 3.1

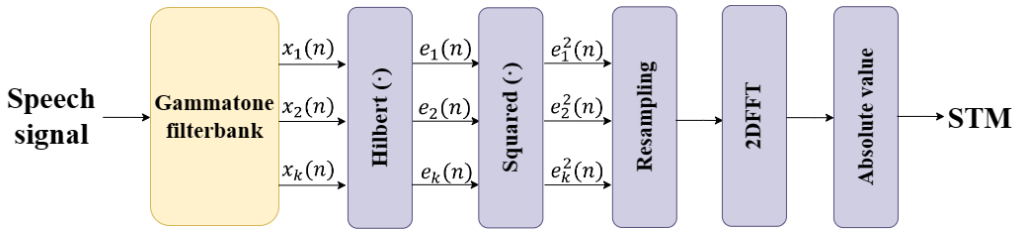


Figure 3.1: Block diagram of STM calculation

## 3.2 Investigation the role of feature expressions

In order to investigate the role of feature expressions, three filterbanks were employed to implement the STM independently. The Mel filterbank (Mel FB) and Gammatone filterbank (ERB FB) are both widely utilized filterbanks in the realm of speech signal processing [54].

### 3.2.1 Mel filterbank

The Mel filterbank (Mel FB) is constructed based on the Mel scale, which is a perceptual scale of pitches judged by listeners to have equal distances from one another [55]. This perceptual aspect, associated with musical melodies, serves as the correlate of repetition rate.

The Mel FB comprises a set of evenly spaced triangular-shaped filters positioned along the Mel scale. Each filter is centered at a specific Mel frequency and has a bandwidth determined by the adjacent filters. The filters' central frequencies are aligned with critical bands, representing specific regions of the auditory system that respond to distinct frequency ranges. The width of each triangular filter is typically defined by the Mel frequency difference between the center frequency and the frequencies where the filter response drops to half power.

The triangular shape of the filters in the Mel FB is essential for approximating the perceptual frequency resolution of the human auditory system. The wider filters capture the lower-frequency components, while the narrower filters focus on the higher-frequency components. By using a set of overlapping triangular filters, the Mel FB effectively covers the entire audible frequency range and provides a representation of the signal in a perceptually relevant manner.

In the context of deep-fake speech detection, the Mel FB is often applied to extract Mel-frequency cepstral coefficients (MFCCs) from speech signals. MFCCs are derived by taking the logarithm of the energy within each Mel filterbank channel and then applying the discrete cosine transform (DCT) to obtain a compact representation of the spectral information. The feature extraction process, utilizing the Mel FB and MFCCs, has found extensive application in speech recognition and related tasks, as it captures important perceptual information while reducing the dimensionality of the signal. The formula to convert a linear frequency ($f$) to the Mel scale ($m$) is as follows:

$$m = 2595 \log_{10} \left( 1 + \frac{f}{700} \right), \tag{3.1}$$

Figure 3.2 and Figure 3.3 provides visualizations of the spectrogram representations using Mel FB.

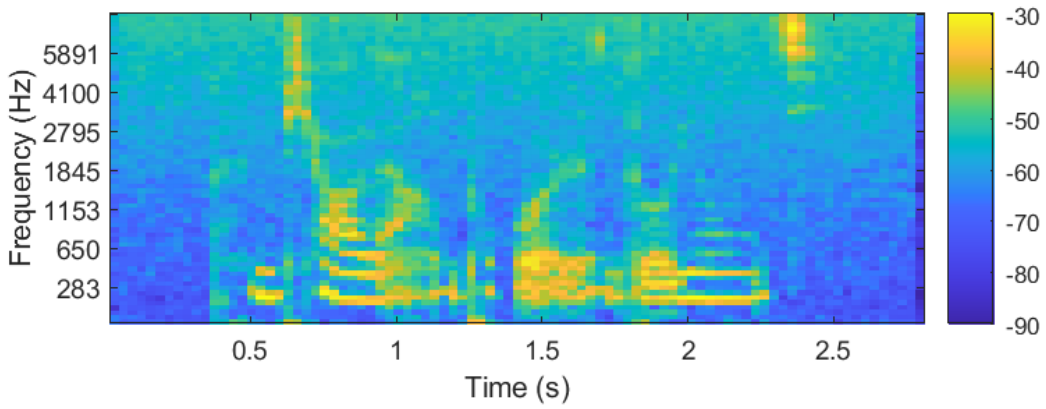


Figure 3.2: Spectrogram of genuine speech signal using Mel FB

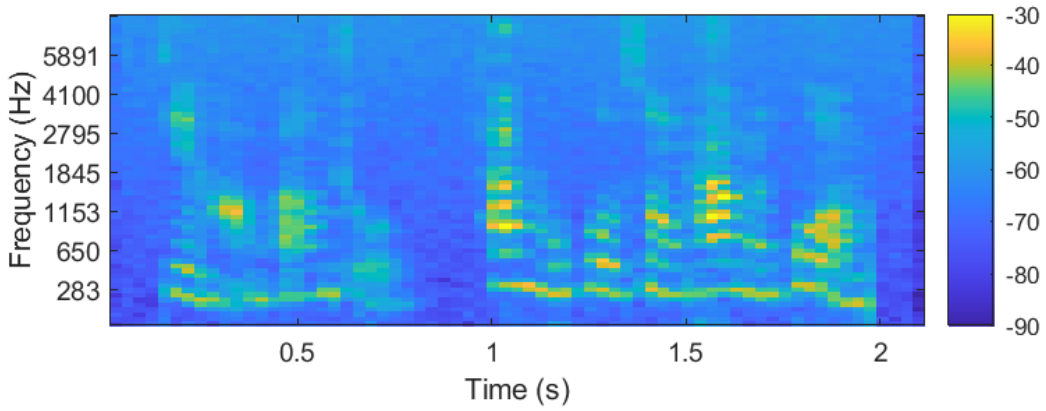


Figure 3.3: Spectrogram of fake speech signal using Mel FB

### 3.2.2 Constant bandwidth filterbank

In contrast, Constant Bandwidth filterbanks (CBW FB) stand out for their fixed bandwidth for each filter, irrespective of their center frequency. This characteristic makes CBW FB particularly useful in certain applications where a consistent bandwidth is preferred across all frequency channels. The construction of CBW FB involves the utilization of a consistent bandwidth parameter, allowing for the computation of center frequencies and channel count based on the provided lower and upper frequency limits.

To analyze the differences between genuine and fake speech signals as illustrated in spectrograms. Figure 3.4 and Figure 3.5 provide visualizations of the spectrogram representations using CBW FB.

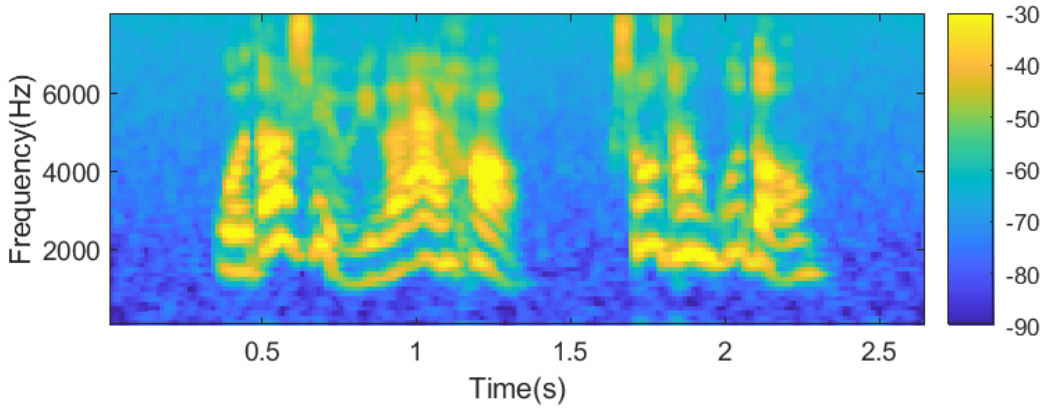


Figure 3.4: Spectrogram of genuine speech signal using CBW FB

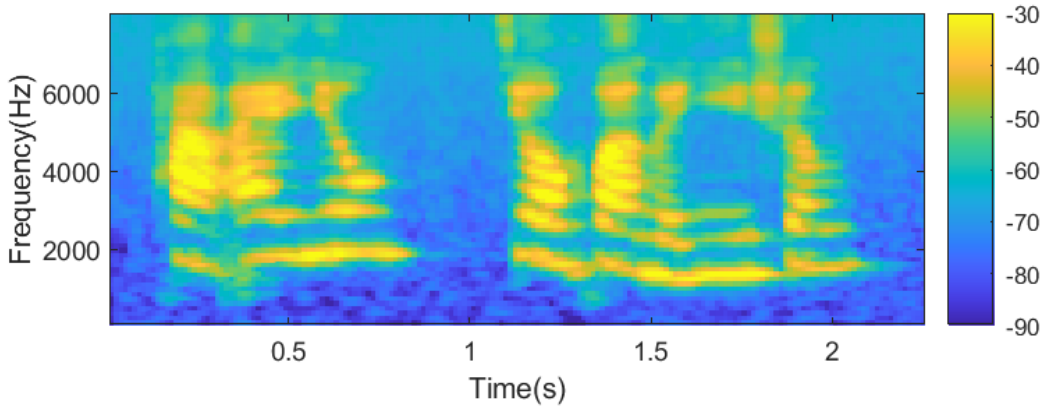


Figure 3.5: Spectrogram of fake speech signal using CBW FB

### 3.2.3 Gammatone filterbank

The ERB FB is designed to accurately model the response characteristics of the cochlea in the human auditory system [56, 57]. It utilizes filters based on the Gammatone function, derived from a combination of complex exponential functions and low-pass filters. These filters adeptly capture both the shape of cochlear filters and the frequency selectivity inherent in the auditory system. As a result, the filterbank enhances the representation of low-frequency components with narrow bandwidths and reduces the presence of high-frequency components with wider bandwidths, as shown in Figure 3.6. The integration of the ERB scale further enhances the accuracy by approximating the frequency resolution of the human auditory system. This integration allows the ERB FB to better capture the spectral characteristics of auditory signals and align with human auditory perception [58].

In the ERB FB, the center frequencies are based on the specified upper and lower frequency limits and the number of channels. These center frequencies are proportional to the corresponding bandwidths of the filters [59]. The output obtained from the ERB FB is as follows:

$$g_k(t) = \mathrm{A}t^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k)t) \cos(2\pi f_k t), \tag{3.2}$$

The amplitude term represented by the Gamma distribution is denoted as $At^{(n-1)} \exp(-2\pi b_f \mathrm{ERB}(f_k t))$, where $A$, $n$, and $b_f$ represent the amplitude, filter order, and bandwidth of the filter, respectively. We apply the fourth order Gammatone. The formula to convert a linear frequency ($f$) to the ERB scale is as follows:

$$\mathrm{ERB} = 24.7(4.37 f_k + 1), \tag{3.3}$$

where $f_k$ is the $k$-th center frequency (in kHz) of filterbank.

Figure 3.7 and Figure 3.8 provide visualizations of the spectrogram representations using ERB FB.

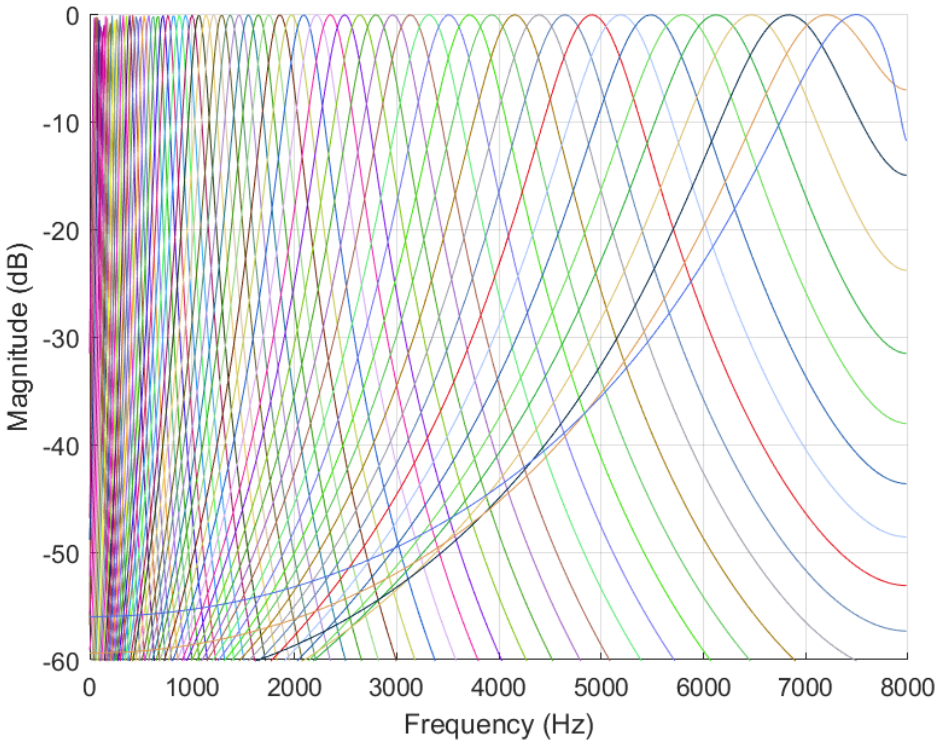


Figure 3.6: Frequency response of ERB FB

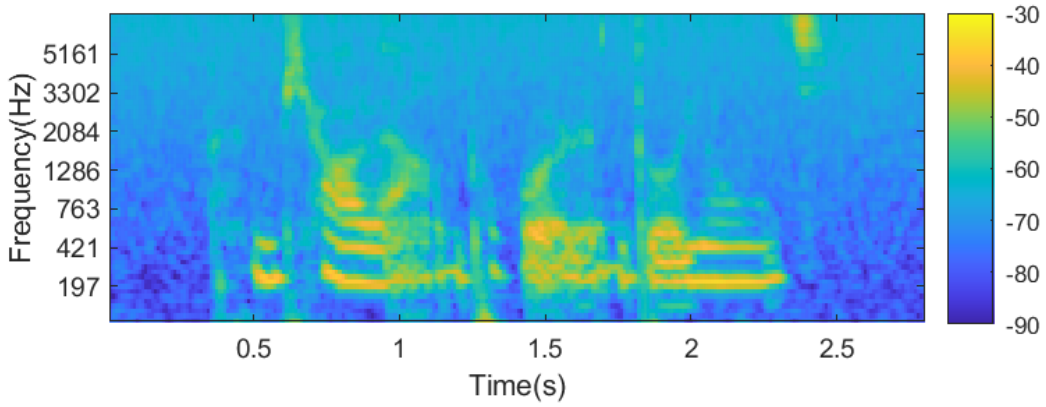


Figure 3.7: Spectrogram of genuine speech signal using ERB FB

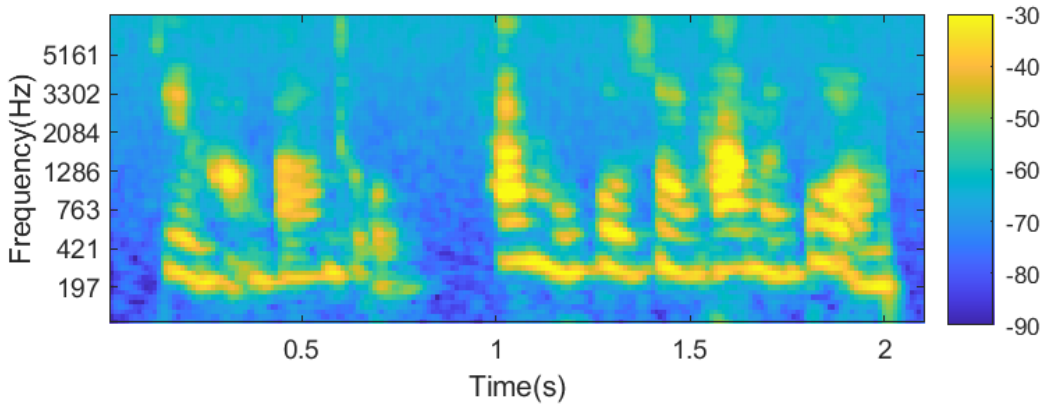


Figure 3.8: Spectrogram of fake speech signal using ERB FB

## 3.3 Procedure of STM implementation

First, the speech signal $s(t)$ undergoes an initial filtering process using a bank of filters. The output of the $k$-th channel is expressed as follows:

$$y_k(t) = g_k(t) * \mathrm{s}(t), \tag{3.4}$$

where * represents the convolution, $g_k(t)$ is impulse response of the $k$-th channel of filterbank.

Next, the power envelope is obtained through the application of the Hilbert transform, followed by squaring the signal. Additionally, LPF represents a low-pass filter with a cut-off frequency of 64Hz.

$$e_k^2(t) = \mathrm{LPF}\left[|\mathrm{y}_k(t) + j\mathrm{Hilbert}(y_k(t))|^2\right], \tag{3.5}$$

Finally, STM representation can be obtained by applying a two-dimensional Fourier transform to squared envelope $e_k^2(t)$, as shown in Eq. (3.6). It is important to note that the result of the two-dimensional Fourier transform is typically a matrix comprising complex numbers, where each element consists of both real and imaginary parts. To obtain the STM representation utilized in this study, the absolute value of the result is taken.

$$\mathrm{STM} = \mathrm{2DFFT}(\log e_k^2(t)). \tag{3.6}$$

where 2DFFT represents a two-dimensional fast Fourier transform.

The STM representations of genuine and fake speech signals using Mel FB, CBW FB, and ERB FB are shown in Figure 3.9 to 3.12.

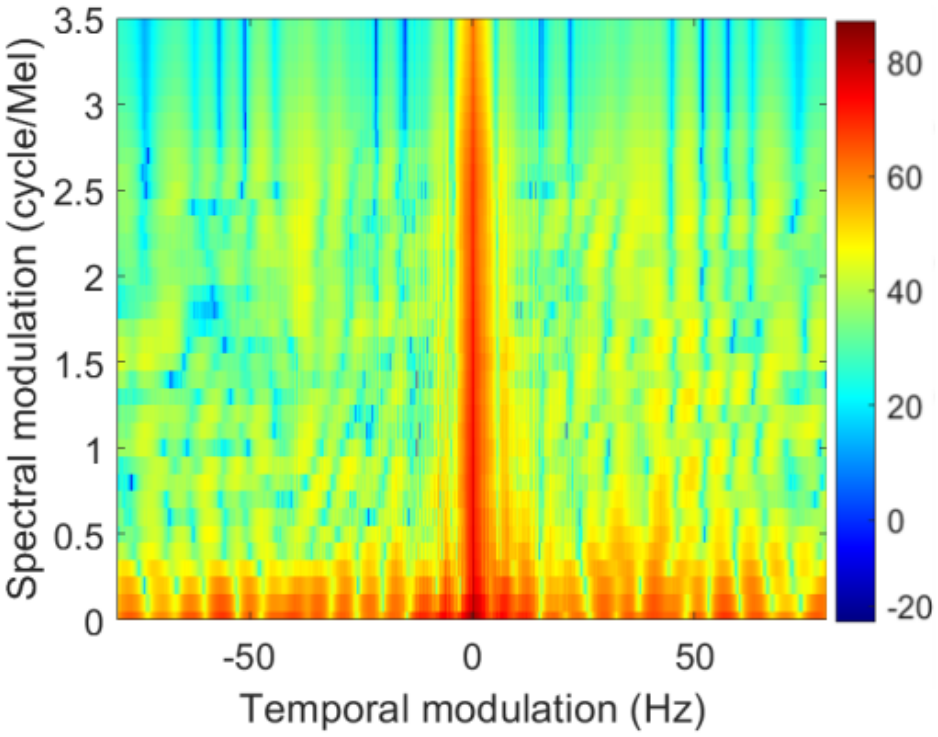


Figure 3.9: STM of genuine speech signal using Mel FB

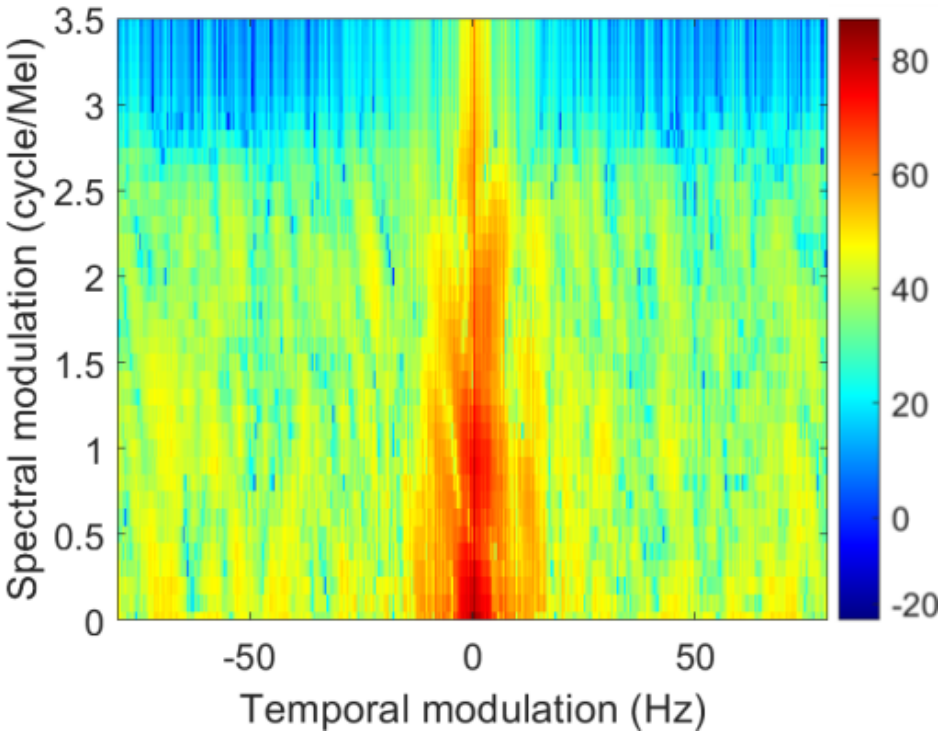


Figure 3.10: STM of fake speech signal using Mel FB

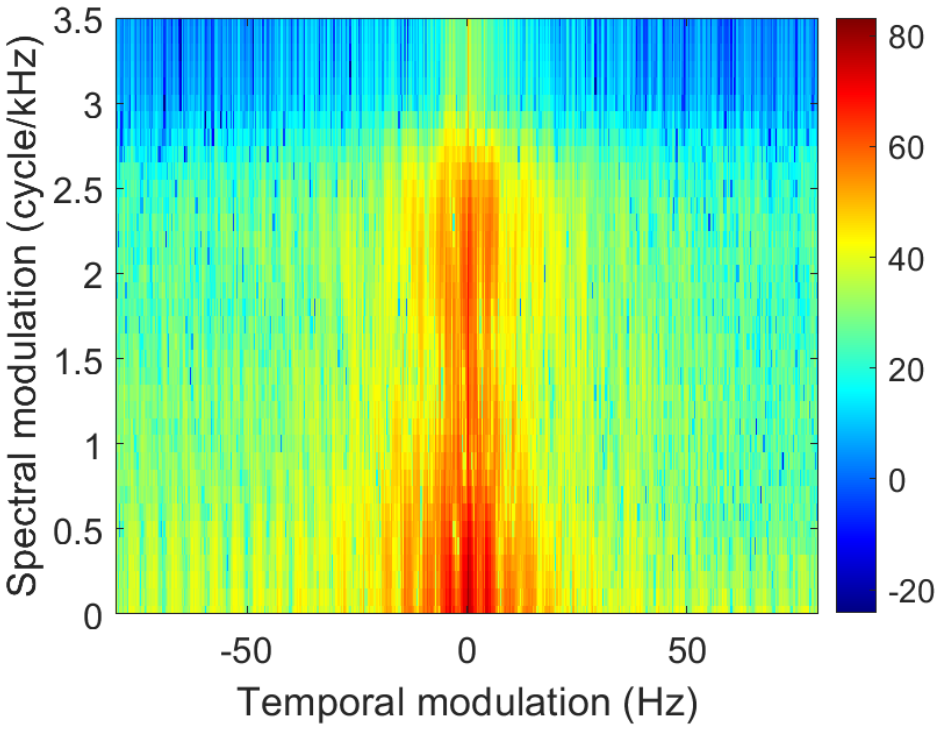


Figure 3.11: STM of genuine speech signal using CBW FB

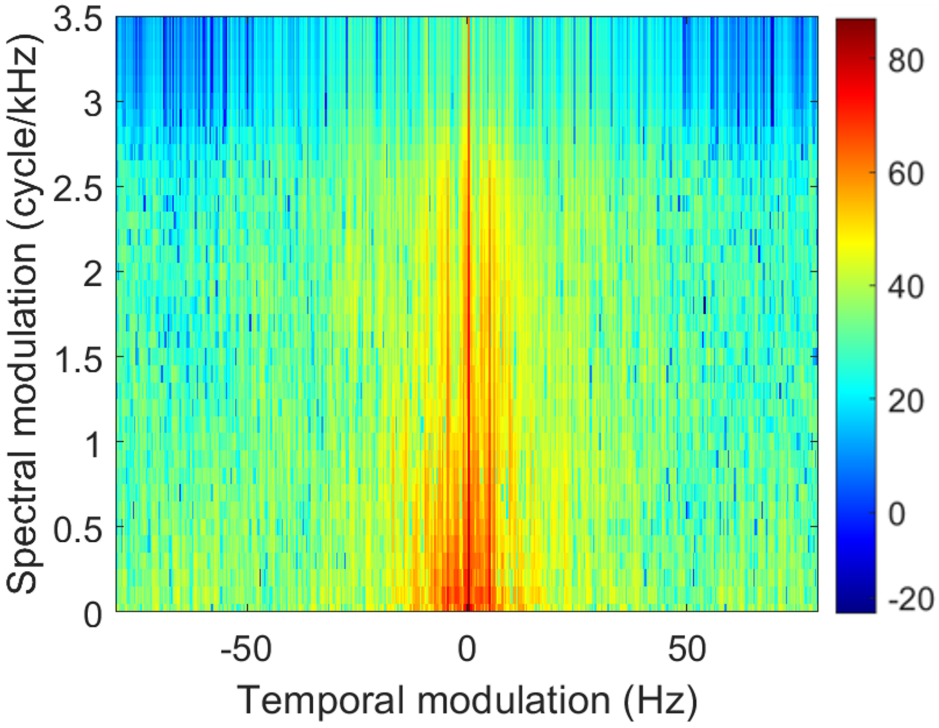


Figure 3.12: STM of fake speech signal using CBW FB

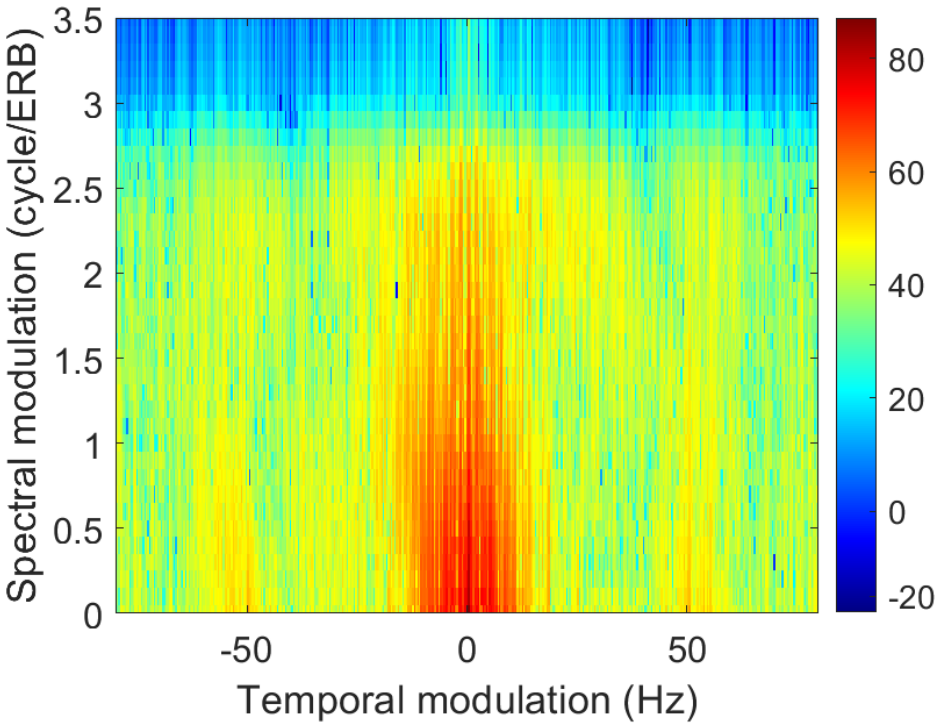


Figure 3.13: STM of genuine speech signal using ERB FB

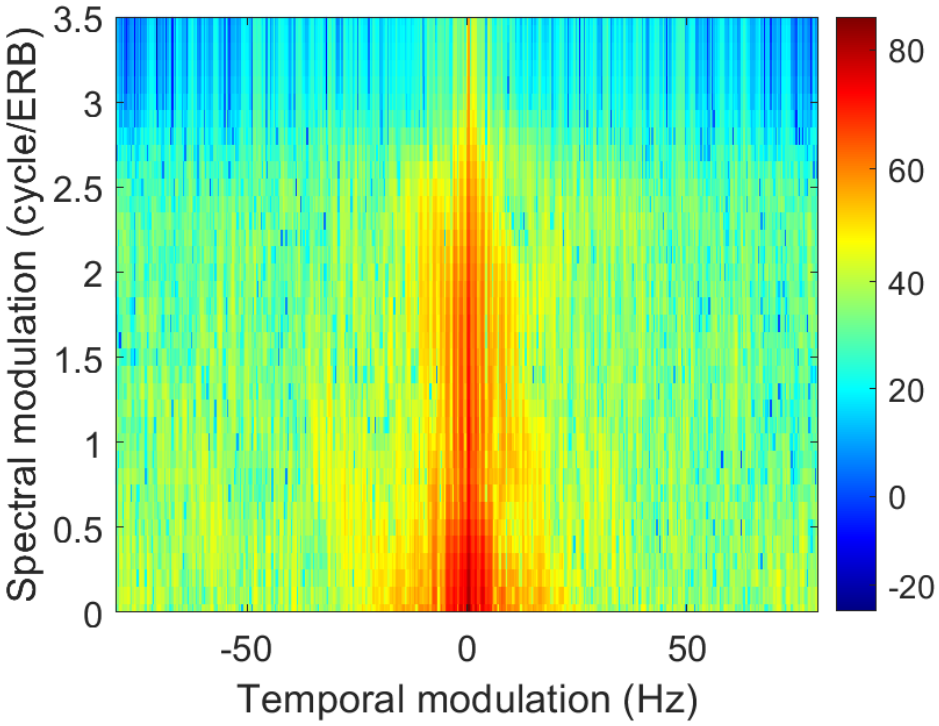


Figure 3.14: STM of fake speech signal using ERB FB

# Chapter 4

# Proposed Method

## 4.1 Framework

In this section, we first extract STM from original speech signals by employing the Gammatone filterbank, Hilbert transform, and fast Fourier transform. In order to enable neural networks to better capture feature information in STM representations, then present a deep-fake speech detection model that utilizes an LCNN and BiLSTM. LCNN is a convolutional neural network variant that is purposefully developed to strike a balance between computational complexity and performance. The strength of LCNN lies in its utilization of a max-out activation strategy, which enhances feature map activations. This characteristic enables faster training and inference times while minimizing the impact on overall performance. The block diagram of the proposed method is shown in Figure 4.1

## 4.2 Feature extraction

Feature extraction holds an essential role in speech detection tasks as it enables the distinction between different speech signals. Its primary function is to transform speech signals into a stream of feature vector coefficients that contain only the necessary information for identifying a particular utterance. Each speech signal possesses distinct attributes that are embedded within spoken words. These attributes can be extracted using various techniques, allowing them to be employed in speech recognition tasks. The process of feature extraction involves extracting relevant characteristics from speech signals and capturing essential acoustic and linguistic properties. These extracted features serve as representations of the underlying speech signals and carry vital information for subsequent analysis and classification. By employing appropriate feature extraction techniques, we can uncover specific attributes within speech signals that are relevant to the identification and discrimination of different utterances. These extracted features serve as valuable inputs for subsequent speech recognition systems, enabling accurate

classification and interpretation of speech.

To obtain the STM representation from the original signal, a series of steps is followed, the block diagram as shown at the bottom part in Figure4.1.

Initially, the input signal undergoes decomposition into frequency components using the Gammatone filterbank, the formula is shown in Eq.(3.2). The frequency range was set from 50 Hz to 8000 Hz, utilizing 64 channels. This frequency range was chosen due to the typical perception of speech signals by human hearing. The lower limit of 50 Hz captures the fundamental frequency component of speech, while the upper limit of 8000 Hz includes high-frequency resonances and harmonics. Setting the channel number to 64 aims to enhance spectral information and improve sound resolution. With 64 channels, a finer frequency division is achieved, enabling precise capture of speech characteristics across various frequency ranges. The increased channel count provides more frequency detail, leading to a more accurate representation of spectral features and capturing a wider range of speech features. However, it's important to consider the computational and memory costs associated with higher channel numbers, as they can impact real-time performance and computational efficiency. Thus, a careful balance was struck by selecting 64 channels. To accommodate the high-resolution STM representations, the TM domain underwent resampling at a rate of 1000 Hz, resulting in an STM representation size of [64, 1000]. This process separates the speech signal into different frequency bands, capturing the spectral information contained within the signal. The resulting frequency components represent the distribution of energy across different frequency ranges.

Next, the squared magnitudes of these frequency components are computed, effectively extracting the power envelope of the signal. The power envelope represents the overall energy variations in each frequency band over time. To further refine the power envelope, a low-pass filter is applied, smoothing out rapid fluctuations and emphasizing the overall temporal characteristics of the signal. The formula is shown in Eq.(3.5).

In the final step, a two-dimensional spectral analysis is performed on the power envelope. This analysis captures the interactions between spectral and temporal modulations in speech signals. By considering both the variations in frequency components and their corresponding temporal dynamics, the STM spectrogram is derived. The STM spectrogram provides a comprehensive representation of the dynamic variations present in the speech signal across different spectral and temporal scales. The formula is shown in Eq.(3.6).

By incorporating the STM representation into the feature extraction process, we can effectively capture the unique patterns and temporal dynamics that distinguish genuine speech from deep-fake speech. The STM representa-

tion offers valuable insight into the underlying processes of human auditory perception and allows us to leverage the dynamic and adaptive properties of the human auditory system for improved detection accuracy.

## 4.3 Identification

In order to extract useful information from STM, deep learning approaches such as BiLSTM have been widely employed. It can effectively model the temporal dependencies in the STM, which are critical for deep-fake speech detection tasks. Specifically, BiLSTM has a "memory" mechanism that allows it to keep track of past information and use it to inform the current prediction. During the task using BiLSTM, the speech feature sequences are individually fed into the hidden layers of both the forward LSTM (LSTM_F) and the backward LSTM (LSTM_B). This process generates two feature vectors that encapsulate the forward and backward information of the speech. Subsequently, the output vectors from these two layers are combined, forming a merged vector that is passed through two fully connected layers. Finally, the classification is performed by applying a sigmoid activation function to compute the score. This is particularly useful for distinguishing between genuine and fake speech, as speech signals often contain long-term dependencies. The dimensions of the BiLSTM layers are set to match the output dimensions of the LCNN. To optimize the model parameters, a binary cross entropy (BCE) objective function is utilized. The BCE objective function is defined as follows:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{4.1}$$

In the equation, $N$ represents the total number of samples in the dataset, while $y_i$ and $\hat{y}_i$ represent the ground truth and predicted output probability of the $i$-th training sample, respectively.

The LCNN-BiLSTM model is trained using the labels, the batch size of all data is 64, and the epoch number is 30. The optimization process employed an Adam optimizer with a learning rate of 0.0001. Validation was performed using the development dataset, and the model achieving the lowest EER score was considered the best-performing one.

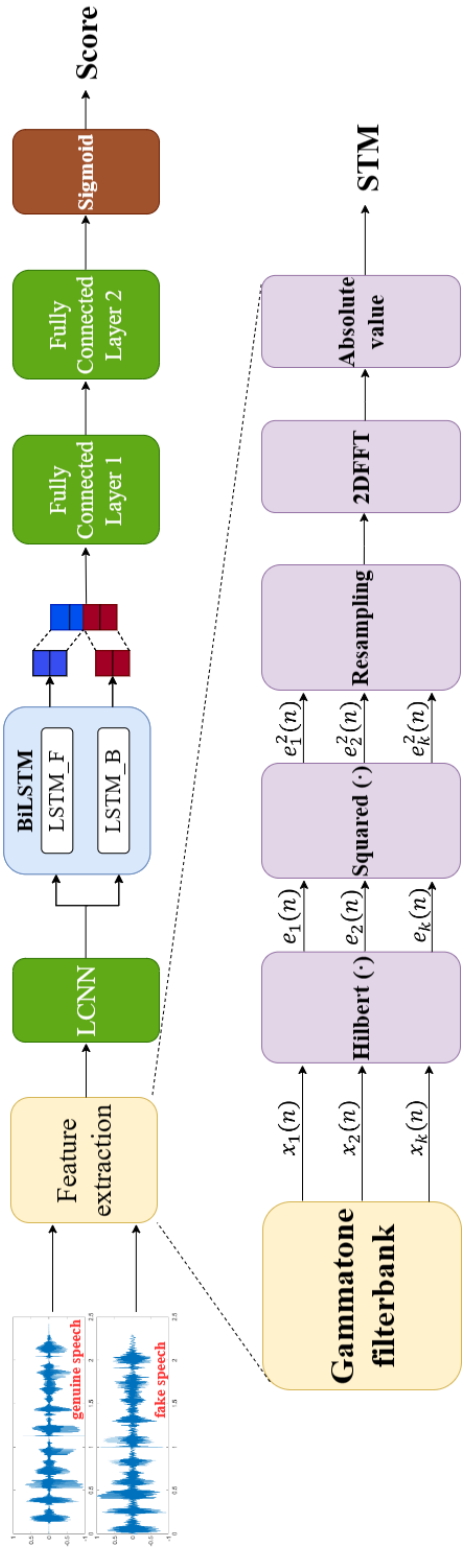


Figure 4.1: Block diagram of the proposed method

# Chapter 5

# Evaluation

## 5.1 Datasets

In this study, two datasets were employed. The Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof2019) pioneered the comprehensive treatment of all major attack types, including text-to-speech, voice conversion, and replay spoofing attacks, effectively covering real-world voice spoofing scenarios [60]. Another dataset used is the Audio Deep synthesis Detection Challenge (ADD2023) [61].

The ADD2023 dataset (as shown in Table 5.2) consists of Mandarin speeches with neutral emotions. The training and development sets have high signal-to-noise ratios (SNR), while evaluation sets have low SNR with various real-world background noises. The evaluation set lacks publicly accessible labels provided by the organizers, therefore the final scores are required to be submitted to the ADD2023 challenge's website CODALAB for online evaluation.

The ASVspoof2019 dataset is partitioned into three subsets: the training set, development set, and evaluation set, as presented in Table 5.1. Notably, the evaluation dataset in ASVspoof2019 includes provided labels, facilitating local evaluation without the need to submit results for external assessment. The proposed method's performance was assessed using the Equal Error Rate (EER). By comparing the results obtained from the two datasets, the objective was to prove the generalization capability and reliability of the proposed methods. This was achieved by utilizing a diverse and challenging collection of samples provided by the two datasets.

## 5.2 Evaluation metrics

In order to conduct a meaningful comparison between the results obtained by the proposed method and those from ASVspoof2019 and ADD2023, a consistent evaluation indicator EER was used. EER is a widely utilized metric for evaluating the performance of binary classification tasks, partic-

Table 5.1: Statistic for datasets of the ASVspoof2019 (Durations with three values denoted with minimum/average/maximum).

| Dataset | Number of utterances | | | Duration (sec) |
|---|---|---|---|---|
| | Genuine | Fake | Total | |
| Training | 2,580 | 24,072 | 26,625 | 0.65/3.42/13.19 |
| Development | 2,548 | 22,296 | 24,844 | 0.69/3.49/16.51 |
| Evaluation | 7,355 | 63,882 | 71,237 | 0.47/3.14/16.55 |

Table 5.2: Statistic for datasets of the ADD2023 (Durations with three values denoted with minimum/average/maximum).

| Dataset | Number of utterances | | | Duration (sec) |
|---|---|---|---|---|
| | Genuine | Fake | Total | |
| Training | 3,012 | 24,072 | 27,084 | 0.86/3.15/60.01 |
| Development | 2,307 | 26,017 | 28,324 | 0.86/3.16/60.01 |
| Evaluation | - | - | 111,977 | 0.35/5.51/217.49 |

ularly in voiceprint recognition and biometrics domains. It represents the percentage value of misclassifications, with lower values indicating better performance. EER is a straightforward and easily interpretable metric, making it convenient for comparing different classifiers or adjusting classifier thresholds. Additionally, EER is unaffected by sample imbalance and does not require equal sample sizes for genuine and fake categories. Hence, it can be applied to datasets with varying category proportions.

## 5.3 Comparison experiments

In addition to our proposed method, we conducted a comparative analysis by re-implementing three well-known features: MFCC, LFCC, and GTCC. This allowed us to assess the performance of our approach against these established methods [62–65].

Figure 5.1 illustrates the diagram for the classic features. In the feature extraction stage, the input signal undergoes initial pre-processing, including windowing with a window length of 25 ms and a step length of 10 ms. The window type used is Hamming. Subsequently, a fast Fourier transform (FFT) is applied to the windowed signal with 512 points. This yields the

Mel spectrum, Linear spectrum, and Gammatone spectrum after passing through the respective Mel FB, CBW FB, and ERB FB. Finally, the resulting spectrum is subjected to logarithmic transformation and discrete Cosine transform (DCT) to obtain the MFCC, LFCC, and GTCC. The back-end classifier used here is LCNN-BiLSTM, which shares the same architecture as the STM experiment. The Mel FB, CBW FB, and ERB FB were implemented with consistent parameters, including the frequency range and channel numbers as described in Section 4.2. The frequency range was set from 50 Hz to 8000 Hz, using 64 channels. Similarly, the parameters for the LCNN-BiLSTM model were set following the details in Section 4.3. The batch size for all data was set to 64, and the model was trained for 30 epochs. The model was optimized using an Adam optimizer with a learning rate of 0.0001.

## 5.4 Experiment results

To identify effective features for deep-fake speech detection, we conducted an analysis of the STM representation among genuine and fake speech. Subsequently, we applied the STM and common features to the LCNN-BiLSTM model and conducted comparative experiments for assessment.

The experiments were conducted using two datasets: ASVspoof2019 (Table 5.1) and ADD2023 (Table 5.2). In ASVspoof2019, the baseline model utilized LFCC and GMM, achieving an Equal Error Rate (EER) of 18.89%. In comparison, our STM-based approach with the ERB FB achieved a significantly improved EER of 8.33%, representing a performance gain of 10.56%. Notably, the STM (ERB FB) outperformed both STM (Mel FB) and STM (CBW FB) in terms of performance. These results, as shown in Table 5.3, underscore the effectiveness of integrating STM and the advantages of using the ERB FB for deep-fake speech detection. The improved performance can be attributed to the STM's ability to capture fine-grained temporal and spectral details, enabling more accurate discrimination between genuine and fake speech samples.

In the evaluation of the ADD2023 dataset (Table 5.4), We conducted a performance comparison between our proposed method and the baseline published by the organizer, which employed LFCC-LCNN and achieved an EER of 70.37%. In our experiments, we first applied Mel FB, CBW FB, and ERB FB as input features to the classifier, resulting in EERs of 77.61%, 83.37%, and 73.34%, respectively. Subsequently, we conducted further comparison experiments and found that the classic features outperformed the filterbanks. Additionally, the STM representations performed better than

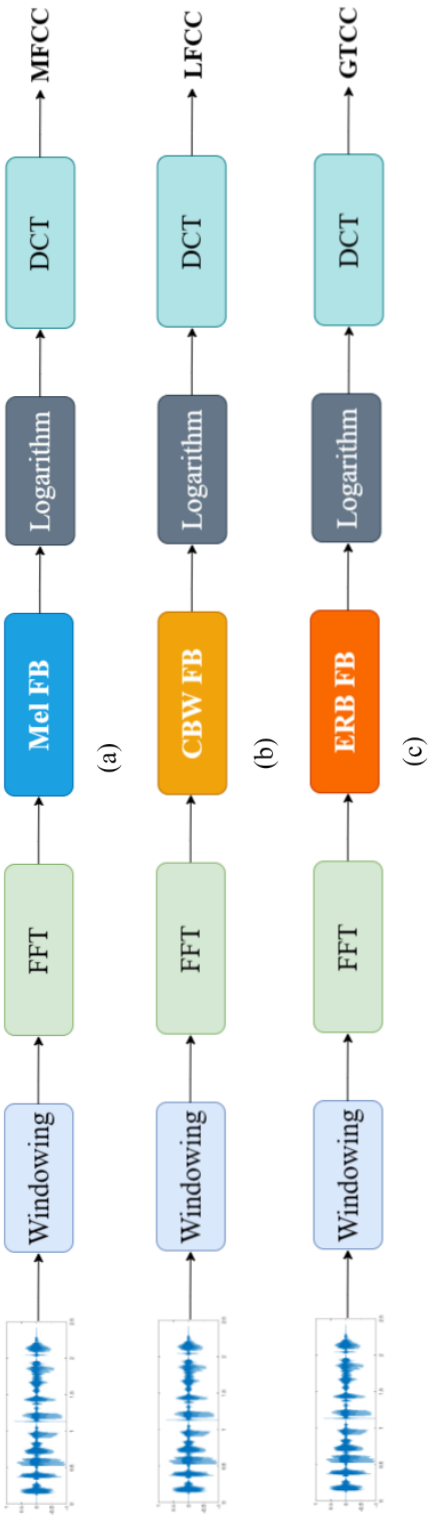


Figure 5.1: Flow process diagram of classic features: (a) MFCC, (b) LFCC, (c) GTCC

the classic features. Specifically, STM based on ERB FB achieved an EER of 42.10%, surpassing not only the baseline but also common features like MFCC (53.36%) and GTCC (63.69%). These results are shown in Table 5.4, which indicates that STMs provide crucial information for the detection of genuine and fake speech.

Table 5.3: Comparative results using the ASVspoof2019 dataset

| Methods | Equal Error Rate (%) | |
|---|---|---|
| | Development set | Evaluation set |
| STM (Mel FB) | 0.04 | 9.79 |
| STM (CBW FB) | 0.09 | 13.46 |
| **STM (ERB FB)** | **0.02** | **8.33** |

Table 5.4: Comparative results using the ADD2023 dataset

| Method | Equal Error Rate (%) | |
|---|---|---|
| | Development set | Evaluation set |
| Mel FB | 0.26 | 77.61 |
| CBW FB | 0.31 | 83.37 |
| ERB FB | 0.23 | 73.34 |
| MFCC | 0.14 | 53.36 |
| LFCC | 0.19 | 66.52 |
| GTCC | 0.21 | 63.69 |
| STM (Mel FB) | 0.14 | 47.65 |
| STM (CBW FB) | 0.26 | 55.55 |
| **STM (ERB FB)** | **0.09** | **42.10** |

## 5.5 General discussion

Our method has demonstrated remarkable success in deep-fake speech detection, achieving superior results. However, it is essential to delve into the underlying principles and address any remaining issues. The STM employed in our approach can be likened to a cepstrum, capturing subtle cues of human vocal system activity and speech information in a two-dimensional representation.

In the case of genuine speech, the STM representation exhibits a distinct pattern where the vocal system activity information tends to concentrate near the origin, while speech information spreads around it. This pattern

is a result of common physiological characteristics shared by humans during speech production, leading to consistent patterns in the STM representation. On the other hand, the STM representation of fake speech lacks this characteristic pattern observed in genuine speech. Machine-generated speech lacks the consistent patterns found in natural speech due to the absence of human-like physiological characteristics, leading to less regular waveforms in speech signals. Leveraging this distinction allows our method to effectively distinguish between fake and genuine speech.

While our approach has yielded promising outcomes, there are areas where further improvement is possible. One potential research direction involves refining the STM representation and optimizing its parameters to enhance its discrimination power between genuine and fake speech. Fine-tuning the STM parameters could potentially lead to even better performance, allowing the method to capture and highlight more relevant features in speech signals.

Additionally, it is crucial to assess the generalization capabilities of our method across different datasets and potential adversarial attacks. Evaluating the method's performance on new and diverse datasets will provide insights into its adaptability to various real-world scenarios. Furthermore, testing the method against adversarial attacks is vital to ensure its robustness and reliability in the face of deliberate attempts to deceive the detection system.

# Chapter 6

# Conclusion

## 6.1 Summary

This study presented in this work centers around the detection of deep-fake speech, which has become a significant concern in the era of advanced AI technology. Deep-fake speech refers to the manipulation of audio recordings to create deceptive or fabricated content that imitates human speech convincingly. Detecting such manipulations is crucial for maintaining the integrity of voice-based systems and ensuring the authenticity of speech information.

The proposed method in this research leverages the concept of spectro-temporal modulation (STM) to extract meaningful features from speech signals. STM combines both spectral and temporal modulations simultaneously, offering a comprehensive representation of the dynamic characteristics of a signal. The process involves transforming the acoustic signal into an auditory spectrogram and then analyzing this spectrogram to estimate spectral and temporal modulation content using specialized filters. Incorporating STM analysis allows for a deeper understanding of human perception and reveals meaningful characteristics in the speech signal, aiding in the detection of deep-fake speech.

To investigate the role of feature expressions, three different filterbanks, namely Mel filterbank (Mel FB), Gammatone filterbank (ERB FB), and Constant bandwidth filterbank (CBW FB), were employed to implement the STM independently. The Mel FB is created based on the Mel scale, which closely matches the way the human ear perceives frequencies. The ERB FB accurately models the response characteristics of the human auditory system. The CBW FB, on the other hand, maintains a fixed bandwidth for each filter across the frequency range. The experiments demonstrated that STM representations using ERB FB outperformed the other two filterbanks, showcasing the significance of the chosen feature expression.

The proposed deep-fake speech detection model combines STM representations with a hybrid deep learning architecture consisting of a Locally Connected Convolutional Neural Network (LCNN) and Bidirectional Long

Short-Term Memory (BiLSTM) layers. LCNN, a variant of CNN, is specifically developed to balance computational complexity and performance, and its max feature map activation strategy ensures faster training and inference times. BiLSTM is adept at modeling temporal dependencies in STM representations, which is crucial for deep-fake speech detection.

The evaluation of the proposed method was conducted using two datasets: ASVspoof2019 and ADD2023. The results demonstrated the superior performance of the STM-based approach over traditional features, achieving significant improvements in EER on both datasets. The STM representations, with their ability to capture fine-grained temporal and spectral details, proved highly effective in distinguishing between genuine and fake speech samples.

While the focus of this study evaluates the performance against baseline models, conducting comparisons with other existing deep-fake speech detection systems could provide a more comprehensive assessment of the proposed method's effectiveness. Future research efforts might concentrate on refining the STM representation and fine-tuning its parameters to bolster the discriminatory capability between genuine and fake speech. Additionally, investigating the generalization capabilities of the method across different datasets and potential adversarial attacks would ensure its robustness in real-world scenarios.

## 6.2 Contribution

This research makes contributions towards understanding the ability of cochlear and auditory cortex perception to recognize deep-fake speech and addresses the challenges posed by malicious production and dissemination of such content in real-life scenarios. The key contributions are summarized as follows:

1. By introducing the concept of Spectro-Temporal Modulation (STM) representation, we gain valuable insights into how the human auditory system perceives and processes speech. The incorporation of STM in our deep-fake speech detection model allows us to mimic the dynamic characteristics of cochlear and auditory cortex perception, enhancing the system's ability to differentiate between genuine and deep-fake speech.
2. We propose an LCNN-BiLSTM model that leverages STM representations to achieve highly effective deep-fake speech detection. Our method outperforms traditional features and other filterbank configurations, showcasing its capability to identify maliciously generated speech with better accuracy and reliability.

3. The proliferation of deep-fake speech poses significant threats to voice-based applications, including voice assistants, voice authentication, and audio content verification. Our work provides a useful solution to mitigate these threats and reduce the negative impact of maliciously produced or disseminated deep-fake speech in real-life scenarios.
4. Our work provides theoretical support for the design and implementation of deep-fake speech detection systems. The use of STM representations, coupled with the LCNN-BiLSTM model, demonstrates the feasibility and efficacy of integrating auditory-inspired features for reliable and practical deep-fake speech detection.

## 6.3 Remaining works

The future work mainly focuses on the following points:

1. It will be crucial to delve deeper into exploring the specific physics-based acoustic features that can be effectively captured and represented by the STM representation. Understanding the underlying mechanisms and intricacies of acoustic signals in relation to STM will enable us to optimize the representation and enhance its discriminative capabilities even further. This comprehensive investigation can lead to the identification of key acoustic attributes that contribute significantly to the detection of genuine and fake speech, thus improving the overall accuracy of the STM-based system.
2. To ensure the practical applicability and generalization ability of the proposed approach, the system's performance should be evaluated on a broader range of datasets that encompass different languages, accents, and speech styles. This broader evaluation will help verify the effectiveness and robustness of the STM-based detection system in real-world scenarios and across diverse populations. By testing the system on varied datasets, we can assess its ability to handle the inherent variations in speech characteristics and adapt to different linguistic nuances, making it more reliable and versatile in practical applications.

# Acknowledgments

First and foremost, I express my deepest gratitude to my family for their unwavering companionship, support, and for teaching me the meaning of learning. Their encouragement has given me the confidence to complete my study.

Secondly, I would like to extend my heartfelt appreciation to Professor Masashi Unoki, who granted me the invaluable opportunity to come to Japan and study at JAIST. I will cherish this experience always. His selfless guidance not only provided me with valuable advice on academic research but also taught me how to approach scientific inquiries and problem-solving. The dedication and passion he demonstrates in his work will continue to inspire me throughout my life.

Thirdly, I am deeply grateful to Junior Associate Professor Teeradaj Racharak for his open-mindedness and active approach as a research scientist and software engineer. Under his guidance, I conducted my minor research, which greatly broadened my perspective.

Furthermore, I owe a debt of thanks to Assistant Professor Candy Olivia Mawalim, without whom my master's program would not have been possible. Her insightful suggestions and unwavering support during moments of confusion helped steer me in the right direction. Additionally, her valuable feedback greatly aided me in the process of writing my papers.

I would also like to express my appreciation to my seniors in the Unoki laboratory. Their generous assistance and support have been instrumental in overcoming various challenges I encountered in both my academic and personal life.

I wholeheartedly dedicate this work to all those who have shown care and support throughout my journey.

With heartfelt thanks,

Haowei Cheng
JAIST, Ishikawa

# References


[1] A. Chadha, V. Kumar, S. Kashyap, and M. Gupta, "Deepfake: An overview," in *Proc. Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, 2021, pp. 557–566.

[2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *Proc. International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, 2010, pp. 537–541.

[3] R. Naika, "An overview of automatic speaker verification system," in *Proc. Intelligent Computing and Information and Communication, ICICC 2017*, pp. 603–610.

[4] 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.

[5] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9216–9220.

[6] B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Adaptation of the human auditory cortex to changing background noise," *Nature communications*, vol. 10, no. 1, p. 2509, 2019.

[7] S. Shamma, "Encoding sound timbre in the auditory system," *IETE Journal of research*, vol. 49, no. 2-3, pp. 145–156, 2003.

[8] R. P. Carlyon and S. Shamma, "An account of monaural phase sensitivity," *Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 333–348, 2003.

[9] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2004, pp. I–601.

[10] Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? focusing on audio deepfake: A survey," *arXiv preprint arXiv:2111.14203*, 2021.

[11] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.

[12] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza, and G. Arce, "Deep4snet: deep learning for fake speech classification," *Expert Systems with Applications*, vol. 184, p. 115465, 2021.

[13] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Proc. Carnahan Conference on Security Technology*, 2011, pp. 1–8.

[14] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention." in *Proc. Interspeech*, 2018, pp. 681–685.

[15] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[16] D. Bigorgne, O. Boeffard, B. Cherbonnel, F. Emerard, D. Larreur, J. Le Saint-Milon, I. Metayer, C. Sorin, and S. White, "Multilingual psola text-to-speech system," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1993, pp. 187–190.

[17] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems," *PhD dissertation, Nagoya Institute of Technology*, 2002.

[18] V. Wan, Y. Agiomyrgiannakis, H. Silen, and J. Vit, "Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders." in *Proc. Interspeech*, 2017, pp. 1143–1147.

[19] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[20] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International conference on machine learning*. PMLR, 2017, pp. 195–204.

[21] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *proc. ICLR*, pp. 214–217, 2018.

[22] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[23] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, vol. 164, 2017.

[24] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[25] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5679–5683.

[26] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5916–5920.

[27] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[28] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5709–5713.

[29] Y. Rodríguez-Ortega, D. M. Ballesteros, and D. Renza, "A machine learning model to detect fake voice," in *Proc. Applied Informatics*. Springer, 2020, pp. 3–13.

[30] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *arXiv preprint arXiv:1903.12389*, 2019.

[31] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proc. Advances in Neural Information Processing Systems*, vol. 31, 2018.

[32] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Proc. International Speech Communication Association (ISCA)*, 2018.

[33] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 2119–2123.

[34] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. International Speech Communication Association (ISCA)*, 2015.

[35] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients." in *Odyssey*, vol. 2016, 2016, pp. 283–290.

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

[37] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with res2net architecture," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6354–6358.

[38] P. Parasu, J. Epps, K. Sriskandaraja, and G. Suthokumar, "Investigating light-resnet architecture for spoofing detection under mismatched conditions." in *Proc. Interspeech*, 2020, pp. 1111–1115.

[39] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *Proc. IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[40] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional gru-rnn deep feature extractor for asv spoofing detection," in *Proc. Interspeech*, vol. 2019, 2019, pp. 1068–1072.

[41] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6316–6320.

[42] I.-Y. Kwak, S. Kwag, J. Lee, J. H. Huh, C.-H. Lee, Y. Jeon, J. Hwang, and J. W. Yoon, "Resmax: Detecting voice spoofing attacks with residual network and max feature map," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4837–4844.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[44] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *arXiv preprint arXiv:1904.01120*, 2019.

[45] R. Hemavathi and R. Kumaraswamy, "Voice conversion spoofing detection by exploring artifacts estimates," *Multimedia Tools and Applications*, vol. 80, pp. 23 561–23 580, 2021.

[46] T. Chen and E. Khoury, "Spoofprint: a new paradigm for spoofing attacks detection," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 538–543.

[47] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6359–6363.

[48] U. de Genève, "How the brain distinguishes between voice and sound." in *ScienceDaily*, 2019.

[49] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 382–395, 1995.

[50] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[51] F. Samson, L. Mottron, B. Jemel, P. Belin, and V. Ciocca, "Can spectro-temporal complexity explain the autistic pattern of performance on auditory tasks?" *Journal of Autism and Developmental Disorders*, vol. 36, pp. 65–76, 2006.

[52] S. M. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature neuroscience*, vol. 8, no. 10, pp. 1371–1379, 2005.

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

[54] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[55] R. S. Chavan and G. S. Sable, "An overview of speech recognition using hmm," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 6, pp. 233–238, 2013.

[56] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Replay attack detection in automatic speaker verification using gammatone cepstral coefficients and resnet-based model," *Journal of Signal Processing*, vol. 26, no. 6, pp. 171–175, 2022.

[57] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Proc. IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[58] T. Irino and M. Unoki, "A time-varying, analysis/synthesis auditory filterbank using the gammachirp," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, 1998, pp. 3653–3656.

[59] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[60] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[61] 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.

[62] R. K. Bhukya and A. Raj, "Automatic speaker verification spoof detection and countermeasures using gaussian mixture model," in *Proc. IEEE Electrical, Electronics and Computer Engineering (UPCON)*, 2022, pp. 1–6.

[63] R. Li, M. Zhao, Z. Li, L. Li, and Q. Hong, "Anti-spoofing speaker verification system with multi-feature integration and multi-task learning." in *Proc. Interspeech*, 2019, pp. 1048–1052.

[64] M. Sahidullah, H. Delgado, M. Todisco, A. Nautsch, X. Wang, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, "Introduction to voice presentation attack detection and recent advances," *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pp. 339–385, 2023.

[65] S. Ravindran and K. Geetha, "An overview of spoof detection in asv systems," *ECS Transactions*, vol. 107, no. 1, p. 1963, 2022.

# Publications

[1] H. Cheng, C. O. Mawalim, K. Li, L. Wang, M. Unoki, "Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2023)*. (Submitted)

[2] H. Cheng, C. O. Mawalim, K. Li, M. Unoki, "Study on Deep-Fake Speech Detection Based on Spectro-Temporal Modulation Representation," in *Proc. Joint conference of Hokuriku chapters of Electrical and information Societies (JHES 2023)*. (Submitted)