Doctoral Dissertation

Human-Imitated Speech Detection Using Acoustic Features Related to Auditory Perception

Khalid ZAMAN

Supervisor: Masashi UNOKI

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

September, 2025

Abstract

Speech plays an important role in human communication, allowing individuals to express thoughts, emotions, and intentions. In digital communication, speech is frequently targeted by attacks that manipulate or forge audio to deceive systems and compromise speaker identity, posing serious challenges to speech authentication and privacy. Among these threats, AI-generated synthetic speech is a prominent concern, produced using text-to-speech (TTS), voice conversion (VC), and deep learning-based techniques. These forms of spoofing are commonly evaluated in challenges such as ASVspoof and Audio Deepfake Detection (ADD), which assess vulnerabilities in automatic speaker verification systems. Despite their realism, AI-generated speech typically exhibits detectable artifacts or a uniform robotic tone, allowing current detection systems to identify it more reliably. In contrast, human-imitated speech, which is produced organically by humans mimicking others, often retains natural acoustic characteristics, making it harder to detect for both human listeners and machines.

Addressing this critical threat, the study initially aims to introduce and assess detection approaches based on standard and auditory-based features with deep learning, using a custom human-imitated speech dataset tailored for machine learning analysis. To evaluate the effectiveness of these approaches, a spoof detection model was trained using standard acoustic features exclusively on ASVspoof 2019 LA synthetic speech and then evaluated on both ASVspoof 2019 LA synthetic speech and human-imitated speech. Although the model performed well in detecting synthetic speech, its accuracy declined significantly when tested on human-imitated speech, revealing a clear generalization gap. However, when trained directly on the proposed human-imitated dataset, the same model and standard features achieved improved detection performance for imitated speech. To further examine this issue, auditory-based acoustic features derived from gammatone and gammachirp filterbanks, which are designed to closely mimic the filtering characteristics of the human inner ear, were evaluated to assess whether they could better capture relevant acoustic cues and improve detection performance. Furthermore, the experiments showed that these auditory-inspired features were more effective than standard acoustic representations in capturing discriminative variations. Although these auditory-based features performed better than standard acoustic representations, they were still not sufficient to achieve reliable detection. These findings underscored that effectively addressing the challenge of human-imitated speech will require acoustic features that are closely related to auditory perception.

Based on these considerations, this research proposes a two-phase framework: the first phase examines human listener tests through listening experiments to evaluate how accurately humans can distinguish human-imitated speech, while

the second phase applies machine-based analysis of acoustic features related to auditory perception combined with machine learning techniques to distinguish between genuine and imitated speech. In the first phase, a human listening test was conducted as part of a human-centered three-phase approach to evaluate the participants' ability to distinguish accurately between genuine and imitated speech. For this evaluation, a representative subset of samples was selected from the human-imitated speech dataset proposed in this study, which was specifically designed to be compatible with machine learning frameworks. In the test, listeners were asked to classify each sample as genuine or imitated and their performance was measured by the percentage of correct responses. Building on insights from the human listening test, the second phase of the study proposes two sets of acoustic features related to auditory perception: eight timbral features, including boominess, depth, brightness, warmth, etc., and spectro-temporal modulation (STM) representations.

In the human listening test, the participants performed well, indicating that with sufficient training and exposure, listeners can effectively distinguish between genuine and imitated speech. Similarly, machine-based experiments using timbral features and STM representations, which are closely related to auditory perception, demonstrated promising discriminative capacity compared to standard acoustic features and reflected trends similar to human evaluation results.

In conclusion, this study highlights the limitations of current spoof detection systems in handling human-imitated speech and demonstrates that even auditory-based features alone are not sufficient for reliable detection. To address this challenge, it begins with human listening tests to evaluate how accurately listeners can distinguish human-imitated speech and proposes an auditory perception-based detection framework supported by new benchmark datasets.

Keywords: Human-imitated speech, Acoustic features, Auditory perception, Timbral features, Auditory models, Machine learning, Deep learning.

Acknowledgment

This doctoral journey has not been merely an academic endeavor; it has been a crucible of transformation. It required every ounce of endurance, discipline, and belief I could summon. It has tested the limits of my mind and the resilience of my spirit. Now, at the end of this journey, I stand with a heart brimming with profound gratitude for those who shaped this journey with their wisdom, sacrifice, and unwavering support.

My highest and most reverent tribute goes to my supervisor, Professor Masashi Unoki, a towering figure in my academic life, whose influence has left an indelible mark on my identity. He has been much more than a mentor; he has been a pillar of strength, a beacon of insight, and a relentless force of encouragement. His intellectual brilliance is matched only by his generosity of spirit. In moments when doubt clouded my path, his calm resolve and precise guidance cut through the chaos. He demanded rigor, nurtured originality, and nurtured in me the courage to challenge conventions. Through his example, I learned not just how to conduct research, but how to think deeply, act ethically, and strive relentlessly for excellence. Studying under his tutelage has been one of the greatest honors of my life. His legacy will forever echo on the trajectory of my career and the convictions that I carry forward.

To my beloved family, your love has been my sanctuary. Your sacrifices, often invisible but immense, built the foundation on which this achievement is based. Your faith in me was a lifeline when mine faltered. Every late night, every battle with exhaustion, every quiet tear, was softened by your steady presence and silent strength. This success is a shared triumph, carved from your devotion and sustained by your unwavering belief.

To my dearest friends and fellow travelers in this search: thank you for the camaraderie, the fierce conversations, the sleepless nights filled with both frustration and laughter. You turned isolation into belonging, deadlines into milestones, and pressure into passion. Your presence made this journey bearable and, at times, even beautiful.

And in the end, I proffer an unvarnished obeisance to the transmogrified self I now embody. I give thanks for the obstinate endurance in the face of capitulation's beguiling overtures, for the relentless ascent beneath the Sisyphean weight of despair, and for clinging, however evanescently, to a glimmer of transcendence beyond the maelstrom of immediacy. This crucible exacted the totality of my essence, and I surrendered it without remit. In return, it conferred apperception, adamantine resilience, and a profound, imperturbable certitude.

As I consign this chapter to conclusion, I emerge not merely adorned with a doctoral title but irrevocably transfigured, my soul annealed in the crucible of adversity and fortified through relentless tribulation. I advance bearing not only

the weight of knowledge but the gravitas of purpose: to serve with unyielding integrity, to lead with unwavering compassion, and to pursue without cessation the uncharted realms that lie beyond the periphery of the known.

Contents

List of Figures

List of Tables

List of Symbols/Abbreviations

TTS Text-to-speech

ADD Audio deepfake detection

ASR Automatic speech recognition

ASV Automatic speaker verification

SER Speech emotion recognition

SVM Support vector machine

Bi-LSTM Bi-directional long short-term memory

DCT Discrete cosine transform

DFT Discrete Fourier transform

DNN Deep neural network

EER Equal error rate

ERB Equivalent rectangular bandwidth

FAD Fake audio detection

FFT Fast Fourier transform

GTFB Gammatone filterbank

GCFB Gammachirp filterbank

GFCCs Gammatone filterbank cepstral coefficients

GCCCs Gammachirp filterbank cepstral coefficients

STM Spectro-temporal modulation

IC Inferior colliculus

IHCs Inner hair cells

OHCs Outer hair cells

LCNN Light convolution neural network

RNNs Recurrent neural networks

LSTM Long short-term memory

MFCCs Mel-frequency cepstral coefficients

STFT Short-time Fourier transform

Chapter 1

Introduction

1.1 Research Background and Problems

Speech plays a crucial role in human communication, allowing individuals to express thoughts, emotions, intentions, and experiences. In digital communication, speech is frequently targeted by speech-related attacks (as illustrated in Fig. 1.1), which manipulate or forge audio data to deceive systems and impersonate speaker identity, posing serious challenges to speech authentication and privacy [4], [7–9]. Among these threats, AI-generated synthetic speech is a prominent concern, where techniques like text-to-speech (TTS), voice conversion (VC), and deep learning-based synthesis are used to generate highly realistic speech. These methods are often used in challenges like ASVspoof [10–14], audio deepfake detection (ADD) [15, 16] and related spoof studies [17–19] to assess the vulnerabilities of automatic speaker verification systems. Despite their sophistication, AI-generated speech often leaves behind detectable artifacts or sounds somewhat uniform and robotic, tending to have a more consistent and predictable quality, making it easier to identify using current detection systems.

Replay-based speech attacks are a type of audio deepfake where a previously recorded voice of a target speaker is maliciously reused to deceive voice authentication systems [20, 21]. These attacks are broadly categorized into two types: far-field replay and cut-and-paste attacks. In far-field replay, an attacker plays the victim's voice through a loudspeaker, which is then re-recorded by a distant microphone, introducing distortions related to room acoustics and playback devices. Cut-and-paste attacks, on the other hand, involve stitching together short audio segments from existing recordings to synthesize specific phrases required by text-dependent verification systems [21, 22]. Replay audio typically carries several acoustic artifacts, such as reverberation, background noise, device distortions, and spectral anomalies, especially in high-frequency bands due to the multiple stages

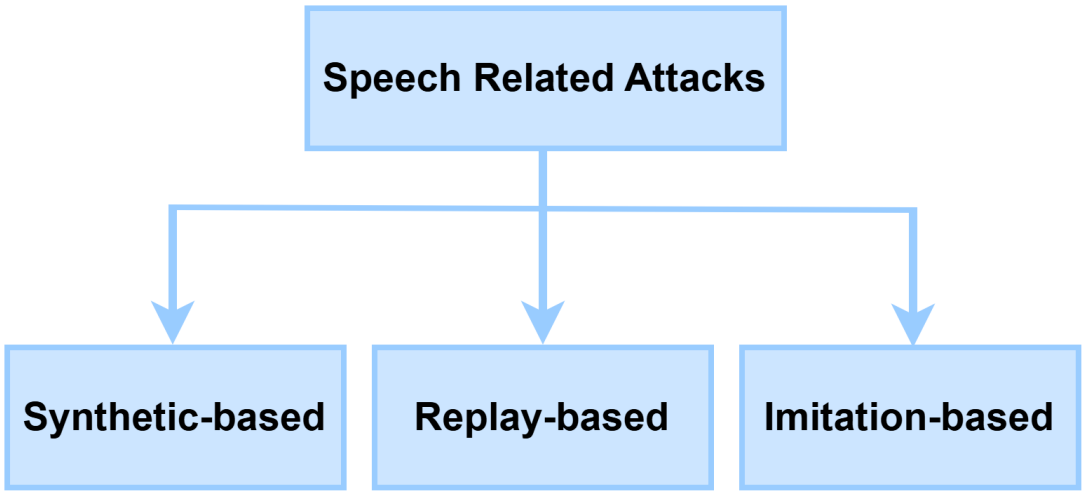


Figure 1.1: Speech-related attacks.

of recording and playback. These signal degradations and unnatural consistencies in timing and prosody make replayed audio distinguishable from live human speech. Current deepfake detection models, particularly those based on acoustic features and deep convolutional neural networks, are effective in identifying such artifacts [12, 20].

Unlike AI-generated speech and replay-based, human-imitated speech poses a much greater challenge for detection systems. Human-imitated speech, which is mimicked or reproduced by humans [21], closely resembles genuine speech in its naturalness as shown in Fig. 1.2. It replicates the natural characteristics of genuine speech such as pitch, rhythm, and timbre with such accuracy that it becomes difficult for both human listeners and machines to detect [4], [21]. Additionally, the shortage of datasets specifically designed for imitation-based detection further complicates this challenge [4].

This challenge becomes especially significant when considering human listening tests, which have been widely studied and applied in various detection tasks to improve performance by focusing on how humans perceive speech. Human listening is highly effective in analyzing and distinguishing subtle nuances in speech, making it a valuable tool for detecting differences between genuine and imitated speech. Understanding and further leveraging the strengths of human listening in this context is crucial for improving imitation detection systems.

Human listening tests have been the focus of numerous studies investigating how listeners process and interpret sound in complex auditory environments. For example, research has shown that auditory experiences such as exposure to quarter tones in classical Arab music can enhance pitch discrimination [23], while linguistic

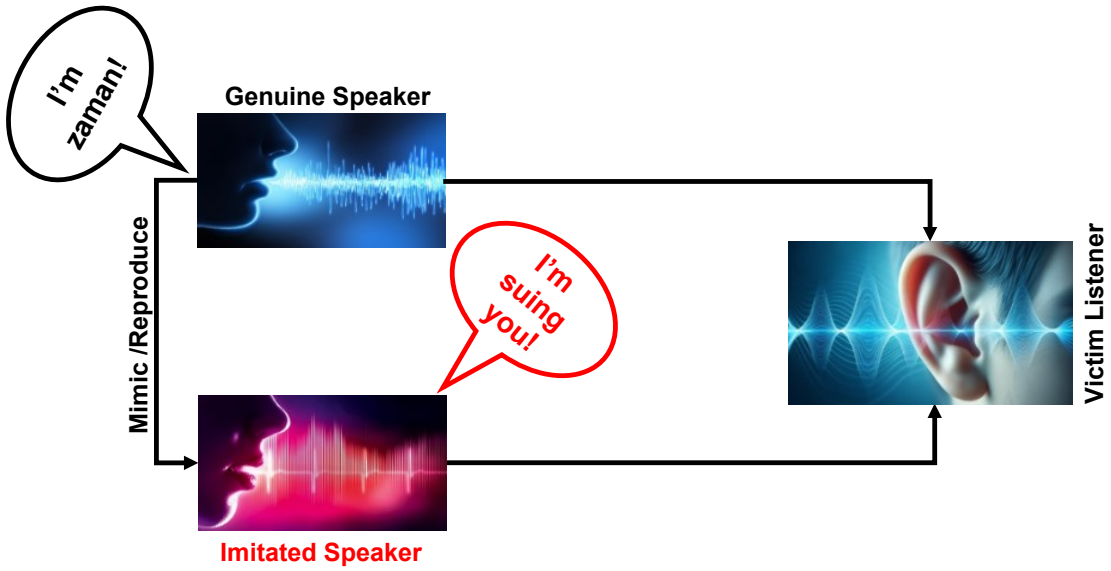


Figure 1.2: Example of imitated speech-based attack .

proficiency influences masked speech perception [24]. Studies like Zetterholm's [25] have examined voice imitation using acoustic analyses and listening evaluations to assess how pitch, formants, and speech tempo affect the recognition and perceived quality of imitated voices. Madureira [26] highlighted the role of multimodal perception by analyzing how vocal and visual cues together contribute to recognizing imitated voices.

Other research has applied listening tests in clinical and applied contexts, including studies on noise annoyance [27], binaural source separation [28], and speech intelligibility for hearing-impaired individuals [29]. Additional work has investigated human reaction times to environmental sounds [30], temporal pitch perception in cochlear implant users [31], and comparisons between human and machine speech recognition [32]. Research on emotional perception in speech, such as studies on Mandarin Chinese using auditory and visual cues [33], and investigations into how temporal cues affect speaker individuality and vocal emotion in noise-vocoded speech [34], further contribute to understanding the role of listening tests in exploring perceptual judgments.

Several studies have proposed imitation-based tasks, with a primary focus on vocal imitation without utilizing actual speech data. For instance, studies such as [35–38] concentrated on vocal sound imitation using standard acoustic features and machine learning approaches. Additionally, [39] employed Qur'anic audio clips (cantillations) from both authentic reciters and imitators, extracted MFCC-based acoustic features, and applied classifiers including SVM, XGBoost, CNN, and BiLSTM to identify reciters and distinguish imitations. Similarly, [7] utilized

the Imitation-based and Deep Voice datasets, converting speech into histogram images as features, and applied a custom 2D-CNN model enhanced with dropout and data augmentation for fake speech detection. Furthermore, [40] used a dataset comprising drum sound imitations by musicians, extracting auditory image-based and acoustic features (e.g., temporal, spectral, MFCC) and applying linear mixed-effects models to analyze perceptual similarity between the imitations and the original percussion sounds. More details about these studies can be found in Chapter 2.

Previous studies have primarily focused on vocal sound imitation which differs from actual imitated speech as well as on cantillation and music imitation. While these areas are related, they do not directly address the complexity of human-imitated speech. This highlights a significant gap in the development of advanced datasets compatible with machine learning analysis, as well as methodologies specifically designed to detect human-imitated speech. Addressing this gap requires combining evidence from human listening tests and acoustic features related to auditory perception with machine learning and deep learning techniques to build more effective detection systems.

1.2 Research Motivation

In recent years, speech-based biometric systems have seen rapid adoption across a wide range of domains, including finance, telecommunications, smart home devices, and personal authentication. These systems offer convenience and hands-free interaction, but they are also vulnerable to security threats that exploit the very modality they rely on: the human voice. Among the various forms of spoofing attacks, AI-generated and replay-based audio have received significant attention, with substantial progress made in detecting such manipulations using deep learning and acoustic feature analysis. However, human-imitated speech, where a person deliberately attempts to mimic another individual's speech, remains a considerably underexplored yet highly dangerous form of attack.

The primary motivation behind this research stems from the realization that human-imitated speech is inherently more deceptive than its synthetic or replayed counterparts. Unlike AI-generated speech, which often contains robotic artifacts, or replayed speech, which suffers from background noise and reverberation, human-imitated speech maintains the naturalness and fluidity of genuine vocal expression. Skilled impersonators can replicate essential vocal attributes such as pitch, tone, rhythm, and emotion with high fidelity. As a result, current detection systems largely trained on synthetic and replayed spoofing examples are not equipped to identify these subtle variations. This mismatch between threat and detection capability forms a critical research gap that must be addressed.

From a security perspective, the implications are profound. If an attacker can successfully imitate a target's speech in real-time or during a speech authentication session, they may gain unauthorized access to sensitive systems. This is especially concerning in contexts where speech is used as a standalone biometric or as part of multi-factor authentication. Without reliable mechanisms to detect such attacks, systems are left exposed to impersonation-based fraud, which may be difficult to trace due to the lack of artificial artifacts or acoustic distortions.

The scientific motivation for this research lies in the complex and often ambiguous nature of human-imitated speech, which presents a significant challenge for both human listeners and machine-based detection systems. It raises an important question: can computational models be developed to detect imitated speech that even trained human ears struggle to recognize? This challenge presents a valuable opportunity to study how humans judge imitated speech and to evaluate acoustic features inspired by auditory perception within computational frameworks. Human listeners play a critical role in fine-grained speech discrimination, enabling the perception of subtle cues such as emotional undertones, prosodic variations, and speaker-specific characteristics, even under noisy or ambiguous conditions. By modeling key aspects of human listening, such as spectral resolution, temporal integration, and perceptual thresholds, researchers can develop auditory-based features that capture information often overlooked by conventional acoustic features like MFCCs or spectrograms.

Another strong motivation for this study is the lack of publicly available datasets and standardized evaluation frameworks that focus specifically on imitated speech. Most existing datasets target synthetic or replayed speech, while those dealing with imitation are either small in size, limited in language diversity, or focused on non-speech vocal sounds like music or drum imitation. This scarcity of resources not only limits the development of robust detection systems but also hampers comparative research in this domain. There is an urgent need to develop or expand datasets that include both genuine and imitated speech, ideally covering different languages, speakers, and imitation styles, to support the training and evaluation of detection systems.

In summary, this research is motivated by the urgent need to address the emerging threat posed by human-imitated speech, a high-risk and low-resource challenge for voice-based security systems. It is driven by the difficulty of detecting naturally produced vocal imitations that closely resemble genuine speech. This work also reflects the opportunity to leverage insights from human listening to improve machine learning models, and it supports the broader goal of developing robust, generalizable approaches for speech integrity and secure human-machine interaction.

1.3 Research Goals

The goal of this research is to develop a robust and interpretable framework for detecting human-imitated speech by combining insights from human listening tests with machine-based analysis of features related to auditory perception, using a custom dataset of genuine and imitated speech designed for both human and machine evaluation.

The study begins by evaluating whether existing spoofing countermeasure models, trained on synthetic speech, can generalize to naturally produced imitation attacks. Performance comparisons between synthetic and human-imitated speech will highlight the generalization gap and expose the limitations of current systems.

In parallel, auditory-based acoustic features, such as gammatone and gammachirp filterbanks, will be examined. While these features simulate human cochlear processing and outperform standard representations, early results suggest they are insufficient for reliable detection, motivating the need for more perceptually aligned approaches.

To address these gaps, the research proposes a two-phase framework. In the first phase, human listening tests will be conducted using a subset of the proposed dataset to evaluate participants' ability to distinguish between genuine and imitated speech. The goal is to uncover how accurately humans can discriminate between the two.

Building on the findings from the listening tests, the second phase will apply acoustic features related to auditory perception including timbral features and STM representations within traditional machine learning models to assess their effectiveness in detecting imitation.

By integrating insights from human listening tests and machine-based analysis, this research aims to develop a scalable, interpretable, and perception-informed framework for detecting human-imitated speech.

1.4 Challenges

Detecting human-imitated speech presents unique and complex challenges. Unlike synthetic or replayed speech, which often contains obvious artifacts, human-imitated speech retains natural vocal characteristics such as pitch, prosody, rhythm, and timbre, making it sound highly similar to genuine speech. This close resemblance makes detection difficult not only for machine-based systems but also for human listeners. Even with the ability to recognize fine-grained differences in speech, humans still face important limitations. These limitations become more pronounced when large volumes of imitated speech must be evaluated. In such cases, phumans often experience fatigue, are influenced by personal biases, and

make inconsistent decisions over time, reducing the reliability of their judgments. This challenge is especially significant in real-world applications, where scalable, objective, and consistent solutions are essential.

Moreover, although insights from human listening have inspired the development of perceptually motivated features, accurately modeling the underlying mechanisms in computational systems such as frequency selectivity, spectral resolution, temporal resolution, and perceptual thresholds remains a considerable challenge due to the biological complexity of the auditory system. Another major barrier is the lack of dedicated, machine learning-compatible datasets containing diverse and well-labeled examples of genuine and imitated speech. Without such datasets, training and benchmarking reliable models become difficult. Additionally, existing spoof detection systems, which are primarily trained on synthetic data, struggle to generalize to human-imitated speech, as they rely heavily on detecting artifacts that are absent in naturally produced imitations.

These challenges highlight the urgent need for: (1) scalable detection systems capable of distinguishing human-imitated speech from genuine speech; (2) the development and integration of perceptually related acoustic features that reflect important perceptual characteristics and can be combined with deep learning; and (3) the creation of high-quality, imitation-specific datasets to support effective training, evaluation, and benchmarking of future detection models.

1.5 Organization of Thesis

An overview of the thesis structure is provided in Fig. 1.3, which outlines the seven chapters. With the introduction presented first, the subsequent chapters are organized as follows.

Chapter 2 provides a comprehensive review of the literature relevant to this study. It begins by describing the structure and function of the human auditory system, followed by computational approaches for modeling auditory perception. The chapter then introduces deep learning methods for audio detection, with a focus on architectures and techniques commonly used in the field. Finally, it addresses fake speech detection, including methods specifically targeting human-imitated speech.

Chapter 3 presents the core framework for detecting human-imitated speech using acoustic feature representations and deep learning methods. It begins by introducing the conceptual basis and motivation for this work, followed by a description of the human-imitated speech dataset developed for the study. The chapter then discusses findings from a human listening test designed to examine how people distinguish genuine from imitated speech. Next, it focuses on acoustic features inspired by auditory perception, such as hardness, depth, brightness,

roughness, warmth, sharpness, boominess, and reverberation, and how these perceptual attributes are quantified. Subsequently, the chapter introduces features based on auditory models, including those extracted using the gammatone filterbank, the gammachirp filter, and other standard methods. Finally, it presents a deep learning framework developed to evaluate the effectiveness of these features in detecting human-imitated speech.

Chapter 4 presents the data collection process and evaluation metrics used in this study. It provides a detailed overview of the datasets, including the scenario, collection process, and composition of the data. It also describes the evaluation metrics used to assess model performance, including confusion matrix, accuracy, Equal Error Rate (EER), F1-score, and the D-prime measure, which offers a perceptual evaluation of discriminability.

Chapter 5 presents a method for detecting human-imitated speech using auditory-inspired acoustic features in combination with deep learning models. It begins by formulating the problem and introducing the proposed approach, which includes feature extraction using the gammatone and gammachirp filterbank, as well as the deep learning architectures applied. The chapter outlines the experimental setup and provides a comparative evaluation of different training and testing strategies using both standard and auditory-based features. Results are presented across multiple scenarios, including training and testing on ASVspoof 2019 LA and the imitated speech dataset. The effectiveness of auditory-based features is discussed in depth, demonstrating their potential advantages over conventional approaches. The chapter concludes with a summary of key findings.

Chapter 6 explores the ability of human listeners to detect human-imitated speech through subjective evaluation. It begins by detailing the experimental methodology, including the stimuli, apparatus, participant information, listening conditions, and evaluation procedures. The chapter then presents the results of the listening tests and provides a thorough discussion of their implications. A machine-based study is also introduced to evaluate how computational models can detect differences between genuine and imitated speech using features related to auditory perception. Finally, the chapter offers a general discussion of the findings and concludes with a summary, emphasizing the influence of perceptual limitations and the potential of perceptually inspired approaches in imitated speech detection.

Chapter 7 summarizes the overall findings and underscores the major contributions of this research, bringing the thesis to a conclusion. It provides a concise summary of the work conducted throughout the thesis, including the investigation of human-imitated speech detection using auditory-inspired features and deep learning models. The chapter outlines the primary contributions in terms of methodological innovation, dataset development, and evaluation insights, and discusses potential directions for future work, including dataset expansion, im-

provements in perceptually motivated feature design, and further integration of auditory perception in machine learning frameworks.

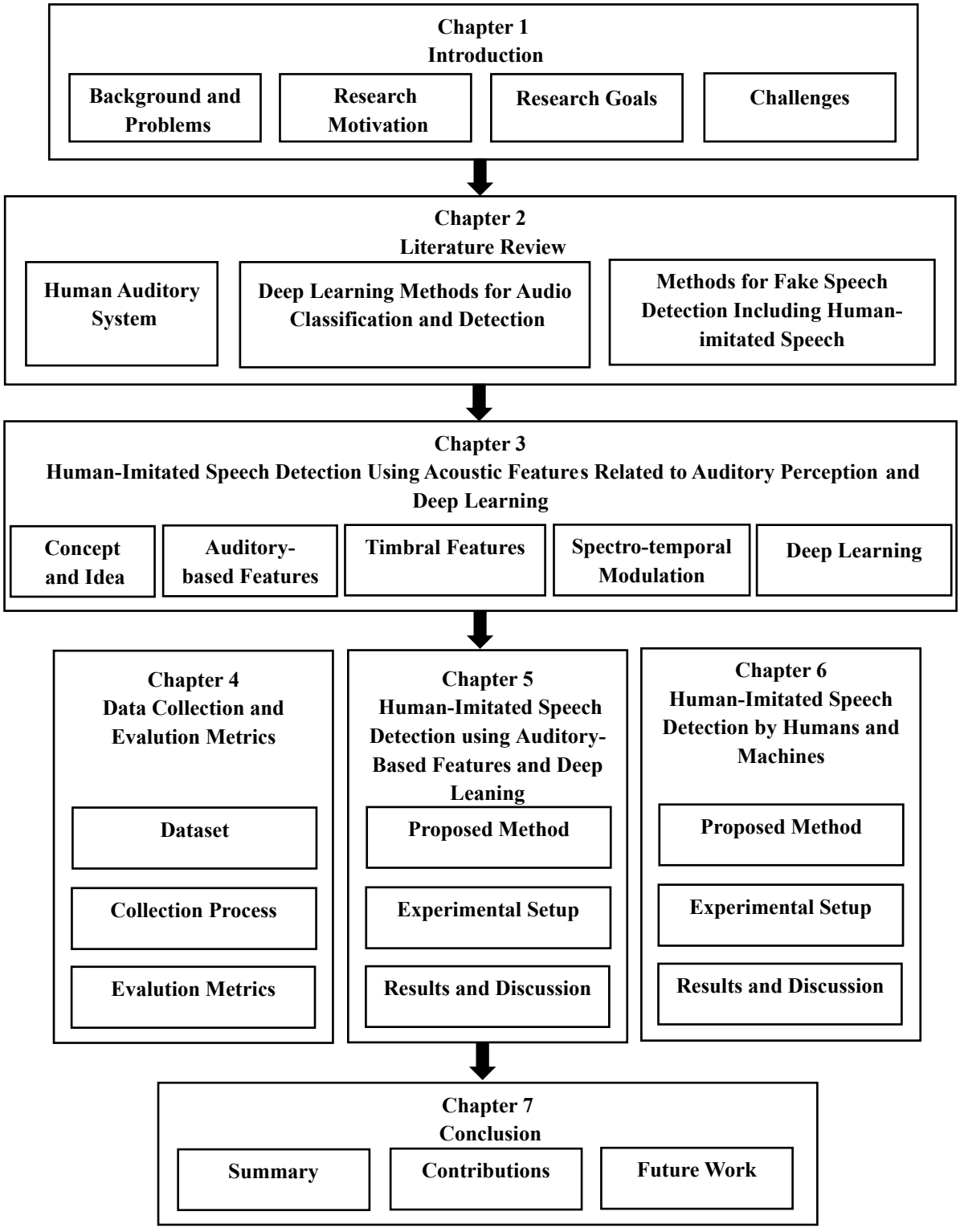


Figure 1.3: Organization of this dissertation.

Chapter 2

Literature Review

2.1 Human Auditory System

The human auditory system is a complex biological mechanism that enables the perception and processing of sound, extending from the outer ear through to the inner ear. It comprises several essential components that allow humans to detect, analyze, and interpret acoustic signals. This chapter provides a concise overview of the anatomy of the auditory system and the pathways through which sound signals are transmitted to the brain. It then examines widely used computational models that simulate the processes of the human auditory system.

2.1.1 Structure and Function of the Human Auditory System

The peripheral anatomy of the human auditory system is illustrated in Fig. 2.1. Sound waves enter the ear through the pinna, the external part of the ear. This structurally unique and intricate component spectrally modifies incoming sounds depending on their angle or direction of incidence. In addition to the pinna, other parts of the head and torso also shape the sound, contributing to the complex process of acoustic modification and enhancing the ability to localize sound sources. As the sound travels past the pinna, it enters the ear canal, a narrow tube-like structure approximately 2.5 cm long. The ear canal is not simply a conduit; it adds further complexity to auditory perception. Much like a tuning fork, it exhibits resonant properties, acting as a highly effective band-pass filter. By selectively amplifying certain frequencies, it influences the characteristics of sound waves before they reach the eardrum, thereby preparing them for subsequent processing within the auditory system.

The eardrum, or tympanic membrane, is a thin, flexible membrane that vibrates

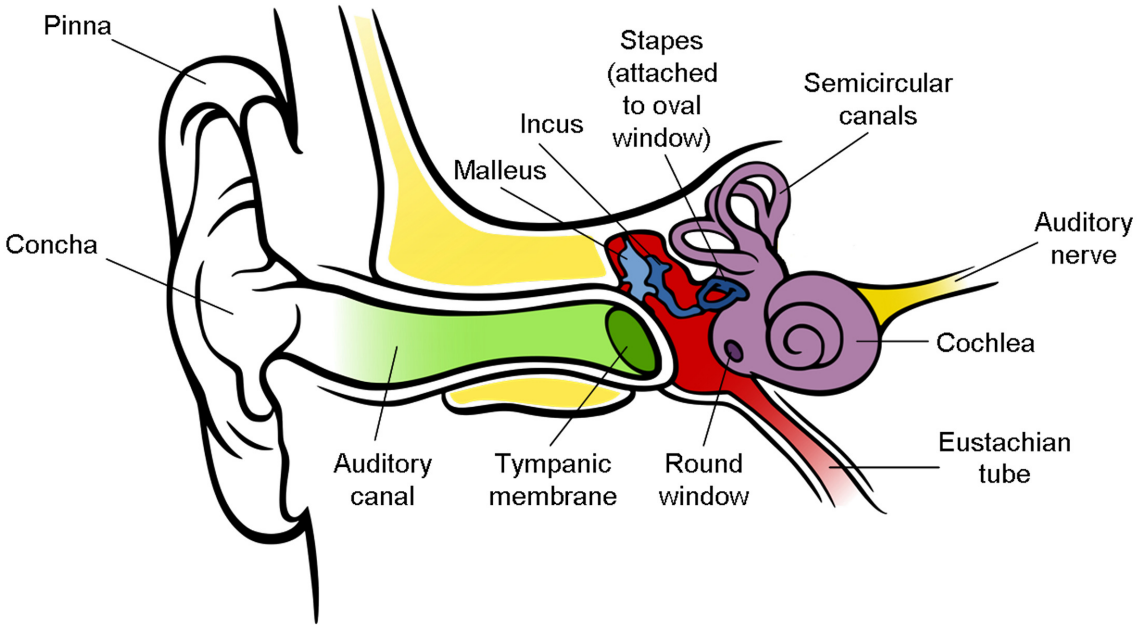


Figure 2.1: A schematic diagram of the peripheral auditory system. Image adapted from Brockmann [1, 2].

in response to pressure fluctuations generated by sound waves traveling through the ear canal. These vibrations are conveyed to three small bones in the middle ear, collectively known as the ossicles: the malleus (hammer), incus (anvil), and stapes (stirrup). The malleus attaches directly to the eardrum and transmits its motion to the incus, which in turn passes the vibrations to the stapes. The stapes then delivers these mechanical vibrations to the inner ear via a small membrane-covered opening called the oval window (see Fig. 2.1).

The oval window connects to the cochlea, a fluid-filled, spiral-shaped structure essential for auditory signal processing. Because the oval window has a much smaller surface area than the eardrum, this difference creates a substantial amplification of pressure approximately a factor of 20, as noted by Plack et al. [41]. This amplification is critical for efficiently transferring sound energy from the air-filled middle ear to the fluid-filled inner ear, ensuring that the vibrations are strong enough to stimulate the sensory cells within the cochlea.

The cochlea is a slender, fluid-filled, spiral-shaped structure within the inner ear, measuring approximately 3.5 cm in length and having an average diameter of about 2 mm. Structurally, it is wider at the base, near the oval window, and its diameter gradually narrows toward the apex. Due to its coiled configuration, it resembles the shape of a snail shell, which is where it gets its name. Inside the cochlea lies a crucial structure known as the basilar membrane, which plays

a fundamental role in auditory frequency analysis. When pressure changes are introduced into the cochlear fluid, typically via the vibrations transmitted through the stapes at the oval window, the basilar membrane begins to vibrate in response. The physical properties of the basilar membrane vary along its length: it becomes wider and more flexible toward the apex, while it is narrower and stiffer at the base. Consequently, the local mass and stiffness of the membrane differ along its length, making each segment selectively responsive to specific frequency ranges. High-frequency sounds cause maximal vibration near the base of the cochlea, whereas low-frequency sounds travel further and peak near the apex. When a pure tone, such as a sine wave, enters the cochlea, it excites a particular region of the basilar membrane with a finite spatial spread. This spatial activation allows the membrane to function as a bank of overlapping band-pass filters, each tuned to a different frequency band. These frequency-selective filters are referred to as critical bands, and they play a vital role in how we perceive pitch, loudness, and timbre. The arrangement of critical bands along the basilar membrane forms the foundation for tonotopic organization, a key principle in auditory neuroscience, whereby different frequencies are mapped to specific locations within the cochlea.

Outer hair cells(OHCs), located on the basilar membrane within the organ of Corti, play a critical role in modulating the mechanical response of the cochlea. Arranged in three rows along the length of the cochlea, these sensory cells exhibit a unique property known as electromotility, whereby they change length in response to electrical stimulation. This motion allows them to amplify sound-induced vibrations of the basilar membrane, particularly at low sound intensities. By enhancing the sensitivity of the cochlea and sharpening its frequency selectivity, outer hair cells contribute significantly to the fine-tuning and nonlinear characteristics of auditory signal processing.

The movements of the basilar membrane are converted into neural signals by specialized sensory cells known as inner hair cells, which are located between the basilar membrane and a second overlying structure called the tectorial membrane. As these membranes move relative to one another in response to sound-induced vibrations, the stereocilia, tiny hair-like projections on top of the inner hair cells, are deflected sideways. This mechanical displacement triggers the release of neurotransmitters from the inner hair cells, which subsequently generates electrical activity (neural spikes) in the auditory nerve fibers connected to them. The magnitude of the basilar membrane's displacement influences the amount of neurotransmitter released: larger displacements cause more neurotransmitter to be released, leading to a higher rate of neural firing. Importantly, inner hair cells respond only to upward movements of the basilar membrane; no neurotransmitter is released during downward displacements (i.e., when the membrane moves toward the center of the cochlea). This directional sensitivity leads to phase locking, where neural

firing is synchronized to a specific phase of the basilar membrane's oscillation. As a result, the timing of neural spikes reflects the periodicity of the incoming sound signal, providing crucial temporal information for auditory perception.

Finally, the signals generated by the inner hair cells are transmitted to the brain via the auditory nerve, which is a bundle of approximately 30,000 nerve fibers [Plack, 2005]. The majority of these fibers are connected to the inner hair cells, with each inner hair cell being linked to around 20 individual nerve fibers. Because each inner hair cell is positioned at a unique location along the basilar membrane, the associated nerve fibers are tuned to specific frequencies, corresponding to the region of the membrane they serve. This arrangement forms the basis of the tonotopic organization in the auditory pathway, allowing the brain to interpret and distinguish different sound frequencies.

Even in the absence of sound, most auditory nerve fibers exhibit a baseline level of electrical activity, known as spontaneous activity. When a sound with a constant intensity begins, the neurons produce a sharp initial spike in firing rate, which then gradually decreases to a steady-state level. When the sound ends, the firing rate often drops below the spontaneous level for a short duration before returning to baseline. This pattern of neural response to the onset and termination (or "offset," as it is commonly referred to in psychoacoustics) of a stimulus is known as adaptation. As the sound level increases, the steady-state firing rate also increases. However, this response has a limit: at very high sound levels, the neural firing rate reaches saturation, beyond which further increases in sound level no longer produce additional neural activity.

A second set of nerve fibers, known as the vestibular nerve, is shown in Fig. 2.1. These fibers are responsible for transmitting positional information from the semicircular canals, which serve as the body's balance organs. The auditory nerve and the vestibular nerve come together to form the vestibulocochlear nerve, which carries all signals from the auditory system to the brain. There, the information is processed in various ways, ultimately leading to the perception of sound.

2.1.2 Modeling the Human Auditory System

Numerous computational auditory models have been developed to replicate the complex signal transformations that occur within the human auditory system. These models are designed to simulate and analyze the various stages of auditory processing, from the reception of sound waves to the neural encoding of auditory information. Computational auditory models play an important role in applications such as speech emotion recognition (SER), automatic speech recognition (ASR), and speech quality evaluation. By extracting relevant acoustic features that capture task-specific nuances in speech signals, these models help improve the accuracy and robustness of auditory analysis. Different auditory models focus

on simulating specific stages of the auditory pathway, providing a framework for feature extraction that closely aligns with the characteristics of human auditory perception.

Auditory Filterbank

The auditory filterbank is a fundamental component in computational auditory models, designed to emulate the time-frequency decomposition that occurs along the cochlear basilar membrane in the human auditory system. Each filter in the bank simulates the frequency tuning of a specific region of the basilar membrane, decomposing the input acoustic signal into multiple frequency bands. This decomposition mirrors the biological process by which the cochlea separates complex sounds into distinct frequency components and plays a critical role in extracting perceptually relevant information from speech and audio signals.

Two widely used models for simulating cochlear filtering are Lyon's cochlear model [42] and the Equivalent Rectangular Bandwidth (ERB)-based filterbank [43]. Lyon's model employs a cascade of gammatone filters to simulate the nonlinear spectral dynamics of the cochlea, while the ERB-based model utilizes filters spaced according to the ERB scale, which approximates the human auditory system's frequency resolution.

Gammatone filterbank (GTFB) [44] and gammachirp filterbank (GCFB) [45] are widely used auditory filterbank in signal processing and auditory modeling. The concept of gammatone responses was first introduced in 1972 to describe revcor (reverse correlation) functions observed in the cochlear nucleus of cats [46]. Since then, gammatone filters have become fundamental tools for simulating the frequency analysis performed by the human auditory system and for capturing key auditory characteristics. The impulse response of a gammatone filter is defined as the product of a gamma distribution envelope and a sinusoidal carrier, modeling human auditory filtering. The bandwidth of each gammatone filter is described by the psychoacoustic measure ERB_N, where "N" denotes normal hearing. This measure represents the equivalent rectangular bandwidth of auditory filters at different locations along the cochlea, aligning with human auditory perception. An illustration of the frequency responses produced by the gammatone filterbank is shown in Fig. 2.2.

Compared to the gammatone filter, the gammachirp filter exhibits asymmetric and nonlinear properties that more accurately reflect the shapes of auditory filters. As shown in Fig. 2.3, the frequency responses of gammachirp filters display pronounced asymmetry, characterized by a steep decline on the high-frequency side relative to the center frequency. This behavior closely corresponds to auditory filter shapes observed in masking experiments, making gammachirp filters especially effective for capturing auditory signal characteristics in a wide range of

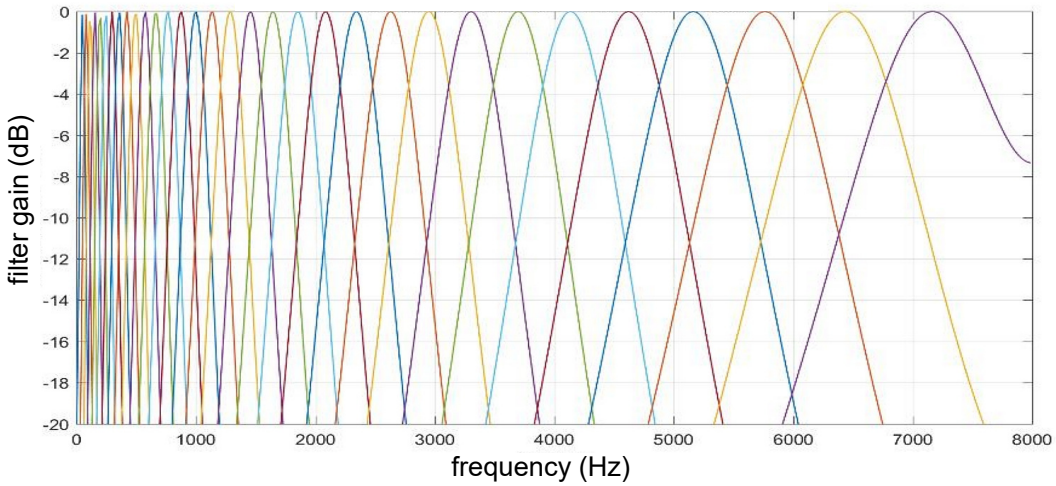


Figure 2.2: Gammatone filterbank frequency response.

audio processing applications. The exponential term $e^{c \cdot \theta(f)}$ introduces a frequency-dependent phase and magnitude modulation, yielding a sharper roll-off on the high-frequency side of the center frequency, consistent with auditory filter shapes derived from masking experiments. In addition, a bandwidth normalization coefficient is also applied for perceptual alignment.

While both gammatone and gammachirp filters simulate the spectral processing of the basilar membrane, their use depends on the application. Gammatone filters are computationally efficient and suitable for large-scale or real-time systems, whereas gammachirp filters offer superior physiological fidelity, making them ideal for applications requiring detailed modeling of auditory perception.

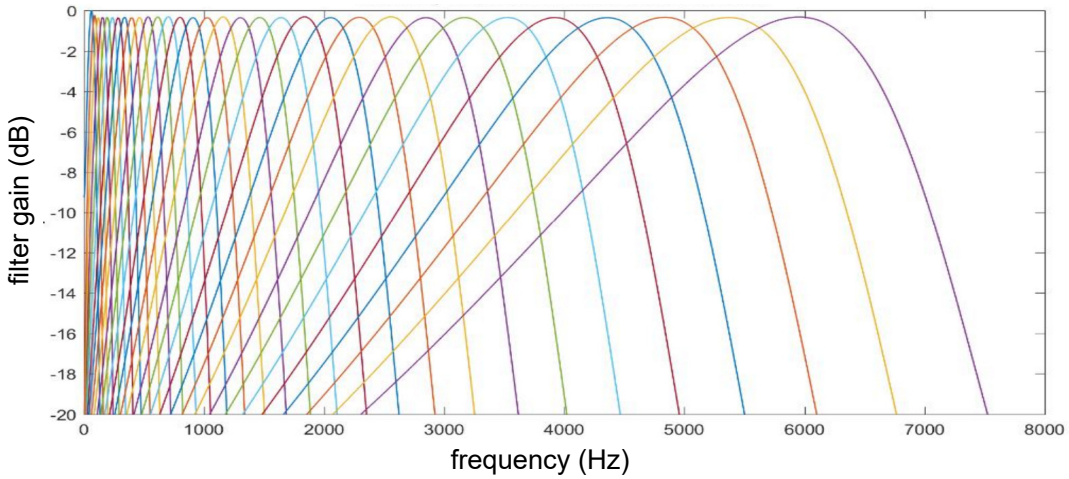


Figure 2.3: Gammachirp filterbank frequency response.

Modulation Filterbank

Both physiological and psychoacoustic evidence highlight the importance of modulation filterbanks in the auditory system. Physiologically, amplitude modulation is primarily processed in higher auditory stages, such as between the cochlear nucleus (CN) and the inferior colliculus (IC), where temporal periodicity is transformed into a frequency-selective rate-based representation [47].

Within the IC, neurons exhibit a periodotopic organization, tuned to specific modulation frequencies and arranged nearly orthogonally to the tonotopic (frequency-based) organization. This suggests the IC's critical role in resolving fine temporal structures by selectively responding to modulation frequencies [48].

Psychoacoustic studies further emphasize the relevance of temporal modulation in speech perception. A modulation filterbank has been proposed to analyze envelope fluctuations within each auditory filter, enabling the extraction of high-resolution temporal modulation cues and advancing our understanding of complex sound processing in the auditory system.

Spectro-temporal Modulation

Spectral and temporal characteristics in speech signals, such as formant transitions, fluctuations, amplitude, duration, intensity, and pitch, carry rich information essential for audio analysis. Capturing and modeling these aspects effectively can substantially improve the robustness and accuracy of speech processing systems.

Various approaches have been proposed to represent temporal and spectral modulations. For example, Wu et al. [49] introduced magnitude and phase mod-

ulation features to detect synthetic speech, demonstrating that combining these cues improves speaker verification security. Their method involves transforming the speech signal into a modulation spectrum by applying the short-time Fourier transform (STFT) to the log-magnitude spectrum and then using principal component analysis (PCA) to reduce dimensionality and extract relevant features.

Recent work has also highlighted the advantages of STM representations for detecting deepfake speech [50]. STM provides a multiscale view of how energy fluctuates across time and frequency, offering richer information than conventional cepstral features. This representation has been shown to simulate aspects of human auditory processing and improve discrimination between genuine and fake speech, particularly by capturing subtle cues related to vocal system activity that are often missing in synthesized signals.

In addition, Li et al. [51] demonstrated that combining machine-specific non-uniform filterbanks with spectro-temporal modulation analysis can significantly improve anomaly detection in machine sounds. By designing filterbanks that emphasize discriminative frequency regions and extracting STM representations inspired by auditory cortex processing, their approach achieved superior performance compared to traditional filterbanks and spectral representations.

2.2 Deep Learning Methods for Audio Detection

Deep learning offers powerful methods for classifying audio signals across a range of categories, including speech, music, and environmental sounds. These models excel at identifying intricate patterns within audio data and can achieve high accuracy when trained on large, diverse datasets. Before deep learning models can be applied, audio signals must be transformed into suitable representations. Common techniques such as spectrograms, Mel-frequency cepstral coefficients (MFCCs), linear predictive coding, and wavelet decomposition are used to extract informative features from the raw signals. These feature representations are then provided as input to deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models to enable effective audio classification. Zaman et al. conducted an extensive survey of modern deep learning approaches for a variety of audio classification tasks [3], as shown in Fig. 2.4. Their review highlights five major categories of deep neural network architectures commonly used in audio analysis: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, Transformers, and Hybrid Models. CNNs, in particular, are highly effective for distinguishing among speech, music, and environmental sounds, and are widely applied to tasks such as speech recognition, speaker identification, and emotion detection. RNNs, with their strength in capturing temporal patterns, are commonly utilized for audio

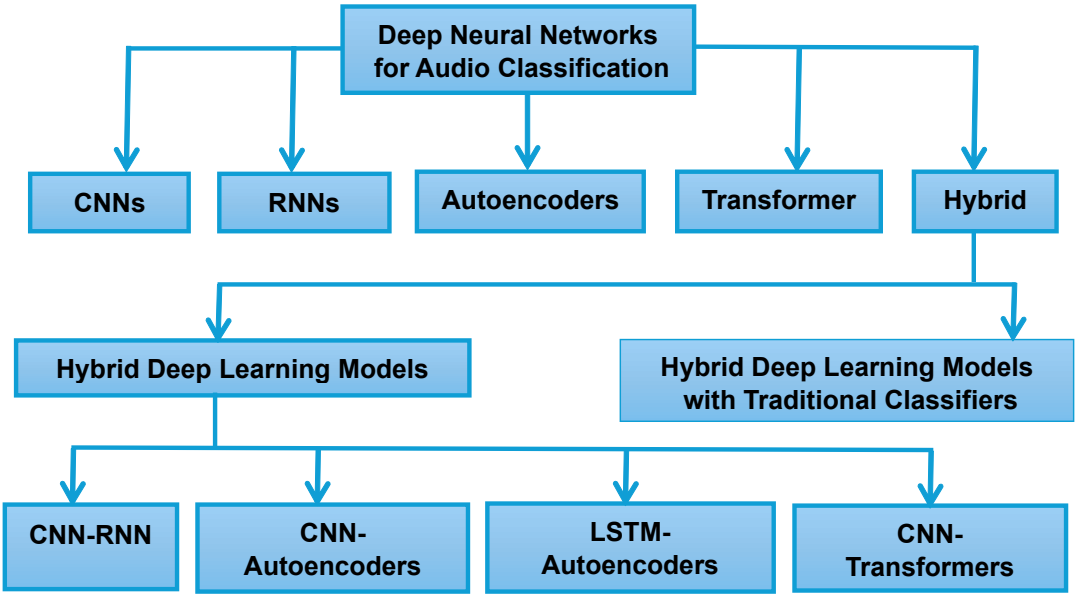


Figure 2.4: Audio classification using deep learning. Orignal image adapted from Zaman and Unoki (2023, 2024) [3, 4].

segmentation and classification tasks. Autoencoders serve as another approach, enabling unsupervised learning of meaningful audio representations before classification. Transformers have also gained traction for audio analysis due to their capacity to capture both temporal and spectral features. Lastly, hybrid models combine different deep learning architectures, such as CNN-RNN integrations, or blend deep learning with traditional classifiers, for example, pairing CNNs with Support Vector Machines, to improve classification performance.

Transformers have emerged as the leading architecture for audio detection tasks, consistently surpassing other deep learning models thanks to their remarkable ability to capture complex, long-range dependencies in sequential audio data. This advantage stems from their self-attention mechanisms [52], which enable flexible modeling of contextual relationships, and their capacity for parallel processing [53]. In contrast, traditional sequential models like RNNs and LSTMs often face limitations in effectively learning long-term dependencies. This simultaneous processing enables transformers to learn complex temporal and spectral patterns, offering a holistic understanding of audio context. Additionally, their scalability and compatibility with large datasets, combined with the increasing availability of pre-trained models, make them highly suitable for transfer learning across diverse audio classification tasks.

Transformers, initially designed for natural language processing, have achieved outstanding performance in a wide range of audio tasks, including sound event de-

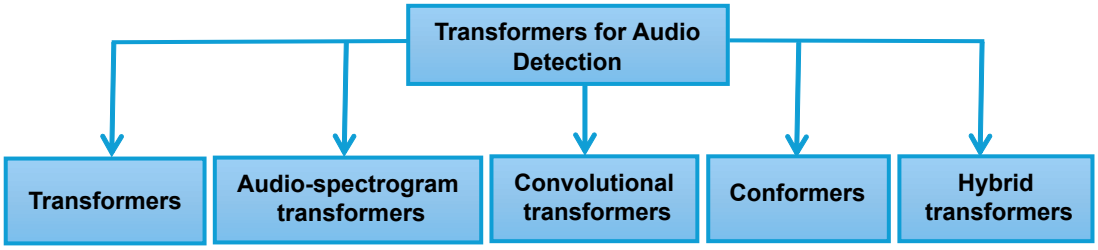


Figure 2.5: Audio detection using transformers .Image adapted from Zaman *et al* (2024) [5].

tection, speech recognition, and deepfake audio detection. Zaman *et al.* (2024) [5] present a comprehensive overview of recent transformer-based architectures and foundational principles. They detail preprocessing techniques that convert raw audio waveforms into spectrograms, enabling effective input representation for transformers, and examine the internal structure of transformer encoders, highlighting the role of self-attention and feedforward networks in modeling both temporal and frequency-based dependencies. As illustrated in Fig. 2.5, these architectures can be extended with task-specific output heads, which allow for the accurate mapping of learned audio representations to specific detection tasks, such as identifying acoustic events, recognizing speech patterns, or detecting manipulated audio content.

2.3 Fake Speech Detection including Human-imitated Speech

There is many methods used for fake speech detection, which are based on feature extraction and deep learning.

With the rapid progress of deep learning and generative technologies, creating high-quality synthetic audio such as speech synthesis, voice cloning, and audio deepfakes has become more accessible than ever. This growing accessibility poses a serious challenge in distinguishing genuine audio from fake. ADD, also referred to as FAD, is dedicated to identifying whether an audio recording has been manipulated, synthesized, or altered with deceptive intent. Its primary goal is to differentiate authentic, unmodified recordings from those generated through advanced deepfake techniques.

Prominent initiatives in this domain include the ASVspoof (Automatic Speaker Verification Spoofing and Countermeasures) challenges [54, 55] and the ADD challenge [15], both of which are widely recognized benchmarks in fake audio detection. The ASVspoof challenge addresses security vulnerabilities in Automatic Speaker Verification (ASV) systems, focusing on detecting spoofing attacks that target

voice authentication. In contrast, the ADD challenge is specifically designed to detect deepfake audio created using state-of-the-art generative models. These competitions provide a critical platform for researchers to benchmark detection methods, exchange ideas, and drive advancements in the field. Notably, the ADD challenge poses greater difficulty due to the use of highly sophisticated and evolving synthesis techniques.

In recent years, FAD research has made notable progress, driven by the development of increasingly sophisticated deep learning architectures. A common baseline model combines a lightweight convolutional neural network (LCNN) with two bidirectional long short-term memory (Bi-LSTM) layers, followed by a global average pooling layer and a fully connected output layer [56]. This configuration is popular due to the complementary strengths of LCNNs and Bi-LSTMs—LCNNs excel at extracting local features from audio spectrograms, while Bi-LSTMs effectively model temporal dependencies inherent in sequential audio data, making the architecture particularly suitable for ADD.

In addition to baseline models, several alternative architectures have been introduced to improve both efficiency and performance. Subramani and Rao [57] proposed two compact convolutional networks, EfficientCNN and RES-EfficientCNN, which achieved strong detection accuracy with low parameter counts, making them well-suited for real-time deployment. To address issues such as vanishing gradients in deeper networks, Alzantot [58] employed a deep residual convolutional network for FAD. This approach involved training separate models on different acoustic representations—Linear Frequency Cepstral Coefficients (LFCC), log-magnitude Short-Time Fourier Transform (STFT), and Constant Q Cepstral Coefficients (CQCC) and combining their outputs through model fusion to enhance robustness.

More recently, self-supervised learning (SSL) has emerged as a promising approach [59]. By pretraining models on large-scale unlabeled audio datasets, SSL makes it possible to learn rich and transferable feature representations. These pretrained models can then be fine-tuned for ADD tasks, significantly improving detection performance by leveraging knowledge acquired from broader audio contexts.

Unlike AI-generated speech and replay-based, human-imitated speech poses a much greater challenge for detection systems. Human-imitated speech, which is mimicked or reproduced by humans [21], closely resembles genuine speech in its naturalness. Replicates the natural characteristics of genuine speech, such as pitch, rhythm, and timbre, with such precision that it becomes difficult for both human listeners and machines to detect [4], [21]. Additionally, the shortage of datasets specifically designed for imitation-based detection further complicates this challenge [4]. Several studies have proposed imitation-based tasks, with a primary

focus on vocal imitation without utilizing actual speech data. For instance, studies such as [35–38] concentrated on vocal sound imitation using standard acoustic features and machine learning approaches. Additionally, Lataifeh et al. [39] used a dataset of 15,810 audio clips from 30 authentic Qur'anic reciters and 397 clips from 12 skilled imitators. Mel-frequency cepstral coefficients (MFCCs) were extracted to represent the speech acoustics. Several models were evaluated, including Support Vector Machines, Random Forests, and a GMM-UBM baseline. In addition, two deep learning architectures a Convolutional Neural CNN and a BiLSTM were implemented to examine their performance in both speaker identification and distinguishing authentic from imitated cantillations. The limitations of this study include its narrow focus on Quranic recitations, which may not generalize well to broader speech or language contexts due to the unique and highly stylized nature of cantillation. The dataset, while comprehensive within its domain, lacks diversity in content, speaker demographics, and speaking styles, limiting the applicability of the findings to more general speaker recognition tasks. Furthermore, the deep learning models showed a slight performance drop when evaluated on unseen Surahs, indicating sensitivity to text variations and a potential challenge in achieving robust text-independent recognition. Additionally, the imitation dataset, although valuable, is relatively small compared to the authentic dataset, which could lead to imbalance issues during training and evaluation.

Mehrabi et al. [40] studied perceptual similarity between vocal imitations and real percussion sounds using a dataset of 30 drum sounds and 420 vocal imitations from 14 musicians. Sixty-three listeners provided 11,340 similarity ratings via a MUSHRA-style interface. Similarity was modeled using linear mixed-effects regression with MFCCs, temporal descriptors, and PHG auditory image features, with PHG performing best. Limitations include the narrow focus on percussion sounds, lack of deep learning methods, limited gender diversity among imitators, and variability in listener focus. Vocal imitations, shaped by speech-based articulation, may not fully reflect the acoustic richness of drum sounds, limiting generalization to speech-related tasks.

Rodríguez-Ortega *et al.* [60] proposed a logistic regression model to detect fake voice recordings generated through signal processing-based voice imitation, specifically using wavelet coefficient reordering. The dataset included 1,086 original recordings from 43 speakers across five languages, and 10,000 forged samples. Entropy features were manually extracted per second and across the full signal, resulting in 11 features per sample. The study is limited by its use of handcrafted entropy features, lack of deep learning-based representation learning, and evaluation restricted to algorithmically-generated imitation rather than human or AI-synthesized speech. As such, its generalizability to other spoofing methods and natural speech variability remains uncertain.

Kim *et al.* [35] introduced the Vocal Imitation Set, a large-scale dataset of 11,242 human vocal imitations covering 302 sound event classes based on the AudioSet ontology. The dataset includes 5,601 listener-vetted imitations and 2,985 original sound recordings from Freesound. Each class includes an average of 10 original recordings and 19 imitations, enabling fine-grained query-by-vocal imitation (QBV) retrieval. The authors also evaluated TL-IMINET, a Siamese-style convolutional neural network trained on language and sound classification tasks, to measure its effectiveness in fine-grained QBV search. The study is limited by its exclusive focus on human-generated vocal imitations of non-speech environmental sounds, making it less relevant for applications involving voice or speech spoofing detection.

Ballesteros et al. [7] developed Deep4SNet, a convolutional neural network (CNN)-based system aimed at identifying fake voice recordings. The approach targeted two types of synthesized audio: Deep Voice, representing AI-generated speech, and an imitation technique that applied wavelet coefficient reordering for signal processing-based voice transformation. To enable the model to treat the problem as an image classification task, audio signals were converted into histogram representations. The dataset used for training comprised 720 authentic and 720 imitation-based fake samples, along with 76 recordings created with Deep Voice. The study is limited by its reliance on fake samples generated through a specific algorithmic imitation technique, which may not reflect the variability and naturalness of human-performed voice imitation. Additionally, the histogram-based representation may not fully capture dynamic vocal features relevant to detecting more subtle or expressive forms of imitation. As a result, generalizability to real-world human-imitation attacks remains uncertain.

Zetterholm [25] investigated voice imitation by analyzing 22 imitations of 9 Swedish male public figures performed by three impersonators (two professionals and one amateur). The study used both perceptual listening tests and acoustic analysis of speech production features such as fundamental frequency, formant frequencies, articulation rate, and /s/ spectral characteristics. The study is limited by a small dataset with only three impersonators and 22 imitations, reducing speaker and imitation diversity. It relied on subjective listening tests and a limited set of speech production features (e.g., F0, formants, articulation rate) analyzed descriptively. No statistical or computational modeling was used, limiting objectivity and generalizability.

Table 2.1: Summary of imitation detection studies: features, models, datasets, and limitations.

Study	Features	Model	Dataset	Limitation
[39]	MFCC	SVM, LR, DT, RF, XGBoost, GMM-UBM, CNN, BiLSTM	15,810 real, 397 imitation	Quranic-only data; class imbalance; limited generalization to broader speech tasks.
[40]	PHG, MFCC, Temporal	Linear Mixed-Effects Regression	420 vocal imitations	Percussion-only focus; no DL; limited generalization to speech.
[60]	Entropy features	Logistic Regression	1,086 original, 10,000 forged (wavelet-based)	Limited real-world generalization to speech produced by human imitators, as the fake samples were generated using signal processing techniques; no DL.
[35]	Spectrogram embeddings	Siamese CNN (TL-IMINET)	11,242 human imitations	Human-only, non-speech sounds (e.g., clapping, cheers); lacks speech-focused evaluation.
[7]	Histogram images	Deep4SNet (CNN)	720 original, 720 imitation, 76 Deep Voice	Algorithmic imitation only; ignores human imitation; histogram may miss vocal cues.
[25]	F0, formants	None	22 imitations voice by 3 impersonators	Small dataset; subjective listening and basic speech production features; no modeling; only descriptive analysis.

Chapter 3

Human-Imitated Speech Detection Using Acoustic Features Related to Auditory Perception and Deep Learning Methods

3.1 Concept and Idea for Detecting Human-Imitated Speech

This dissertation aims to propose a framework based on the idea that, *To detect human-imitated speech, a machine itself imitates how humans perceive.* Unlike AI-generated synthetic speech which often contains detectable artifacts human-imitated speech is naturally produced and often closely resembles genuine speech, making it especially difficult to distinguish. It preserves perceptual characteristics such as vocal texture, which refers to how the voice is experienced by listeners in terms of qualities. These qualities make it especially challenging to distinguish from genuine speech, both for humans and machines.

Despite this, human-imitated speech remains an underexplored problem in the field of speech-based security. Existing research has largely focused on detecting synthetic speech, and in rare cases where human-imitated speech is addressed, deep learning methods are applied without perceptual grounding, treating the problem as a standard classification task. This 'black-box' approach ignores how humans naturally perceive genuine speech, resulting in models that fail to capture the nuanced features unique to imitated speech. In light of this consideration, initial experiments were conducted to assess whether auditory-based features de-

rived from gammatone and gammachirp filterbanks designed to simulate cochlear filtering in the human auditory system could offer more perceptually relevant representations. Although these features demonstrated improvements over conventional acoustic representations and helped capture certain discriminative patterns, they were ultimately insufficient to achieve reliable detection.

To fill this gap, this dissertation proposes a framework for detecting imitated speech based on human listening tests and acoustic features related to auditory perception, starting with the fundamental question: *How accurately can human listeners distinguish between genuine and imitated speech?* This question was addressed through a human study conducted on a carefully selected subset of data, where participants were asked to classify samples as genuine or imitation. The primary objective of this study was to examine human judgment in this task, focusing on how listeners perceive and evaluate speech without prior exposure to imitation samples.

Building on these insights, the study further explores the question: *How effectively can auditory perceptual features be used to detect human-imitated speech by machine/deep learning techniques like human listeners?* These features, grounded in how humans perceive voice quality, were evaluated through a machine-based analysis to assess their potential for computational modeling of perceptual characteristics. Specifically, this phase focuses on two complementary sets of acoustic features related to auditory perception: timbral features, including boominess, depth, brightness, and warmth, which capture dimensions of voice quality central to how listeners experience and assess authenticity; and spectro-temporal modulation (STM) representations, which capture how energy varies jointly over time and frequency by decomposing the spectrogram into temporal and spectral modulation content. Inspired by how the auditory cortex processes complex sounds, STM provides a multiscale representation that helps detect subtle differences in vocal texture and speech dynamics. Both feature sets are closely related to auditory perception, offering complementary perspectives on how humans analyze and interpret speech signals.

This philosophy reinforces the central belief that effective detection of human-imitated speech begins with understanding how accurately humans can perceive it. Rather than treating machine learning as an isolated solution, this work integrates human listening tests and perceptually inspired acoustic features into the core of the detection framework. By combining human evaluation in small-scale experiments with machine-based analysis of perceptually motivated features, the framework ensures that the machine's decision-making process reflects how accurately humans naturally distinguish between genuine and imitated speech. In doing so, this dissertation advances a human-informed approach that goes beyond black-box modeling to deliver scalable and interpretable solutions for reliably de-

tecting imitated speech.

Figure 3.1 illustrates the overall concept, which bridges subjective and objective detection pathways by combining acoustic features related to auditory perception with machine and deep learning techniques. Further details of this approach are presented in Chapter 5 (see Figures 5.1 and 5.2) and Chapter 6 (see Figures 6.2, 6.8, 6.10).

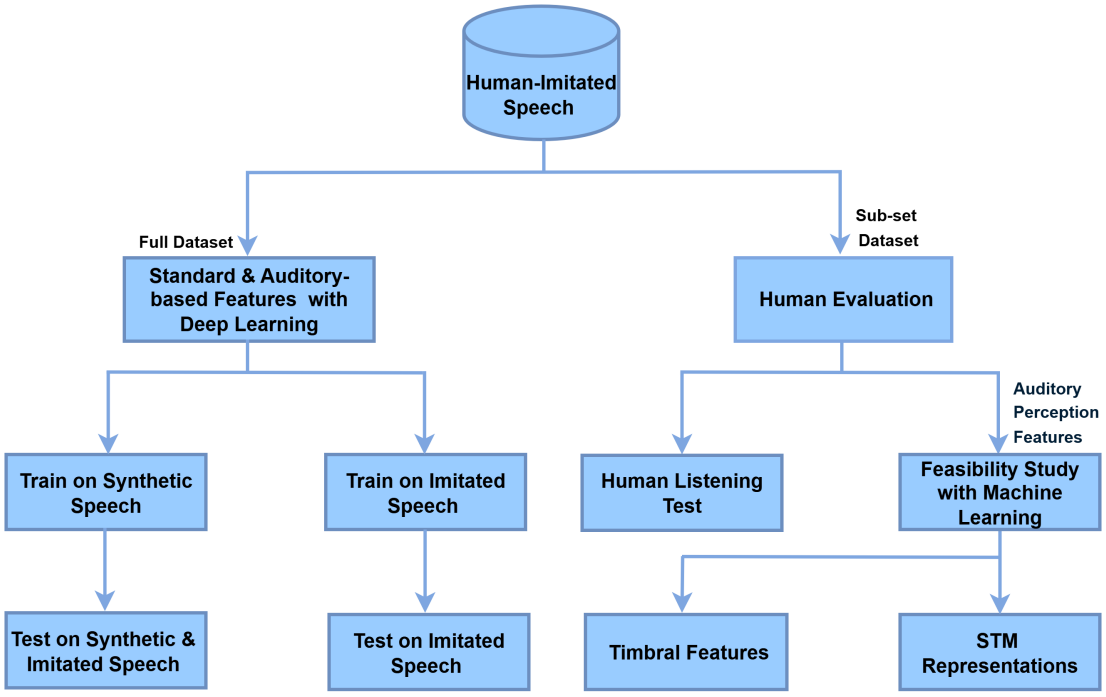


Figure 3.1: Overall concept: Human listening tests and acoustic features related to auditory perception for detecting human-imitated speech.

3.2 Human-imitated Speech Dataset

In this dissertation, a new human-imitated speech dataset is proposed, specifically created to address the limitations of existing spoofing datasets and it is presented in detail in Chapter 4. Most current datasets, such as those used in ASVspoof challenges, focus on synthetic speech generated by machines. As a result, existing spoofing detection systems are mainly trained to detect synthetic speech, which is relatively easier to identify due to its robotic and repetitive patterns. However, human-imitated speech is more natural and realistic, making it more challenging to detect. The proposed dataset fills this gap by providing genuine and human-

imitated speech samples, enabling both human listening tests and machine learning analysis to develop and evaluate systems that can better detect this type of attack.

3.3 Human Listening Studies

To scientifically validate the detection of human-imitated speech, it is essential to involve human listeners in the evaluation process. Human studies play a critical role in discerning subtle acoustic variations that may be overlooked by automated systems. Therefore, they are not only supportive but foundational to the credibility of this research.

In this study, after the creation of the speech dataset, a human listening test was conducted, as thoroughly discussed in Chapter 6. Initially, participants were trained using genuine speech samples from the target speakers to familiarize them with the speakers' natural vocal characteristics. This training phase was especially important, as it ensured that all participants developed a similar and high level of familiarity with the target speakers before the final evaluation. Such consistency was critical for reducing variability in judgments, avoiding bias, and enabling fair comparisons across listeners.

Following training, a preliminary test was conducted to assess each participant's ability to correctly identify the target speakers. Only those who met a predetermined accuracy threshold in identifying the target speakers were selected to proceed to the final evaluation phase.

During this evaluation, participants were asked to differentiate between genuine and imitated speech samples, including imitations specifically crafted to closely mimic the target speakers. This setup allowed us to directly assess the differences between genuine and imitated speech from the listener's perspective.

Notably, the evaluation included challenging cases, such as high-quality imitated speech, to rigorously test the limits of human judgment. The details of the experimental procedure and results are provided in Chapter 6 (see Fig. 6.2).

These human studies are indispensable, as they offer a benchmark for evaluating machine learning models. Understanding how humans identify and classify imitated speech provides crucial insights that can guide the selection of acoustic features and inform model design.

3.4 Acoustic Features Related to Auditory-based Features

This dissertation explores auditory-based features such as the gammatone filterbank (GTFB) and gammachirp filterbank (GCFB) in the context of various deep

learning models which is elaborated upon in Chapter 5.

This study primarily explored gammatone filterbank (GTFB) and gammachirp filterbank (GCFB) features, which closely model human auditory processing, for detecting imitated speech. This study follows Slaney's Auditory Toolbox[1] and the study by Nguyen et al. [61]. To benchmark their effectiveness, the mel-spectrogram, mel-frequency cepstral coefficient (MFCC), and linear-frequency cepstral coefficient (LFCC) were also used as standard features for evaluating a spoofing-countermeasure system [56], [62–65]. Additionally, the gammatone cepstral coefficient (GTCC) and gammachirp cepstral coefficient (GCCC) were introduced for comparative analysis.

3.4.1 Feature Extraction with Gammatone Filterbank

The GTFB is widely used in auditory modeling and speech processing, as it closely mimics the frequency response of the human cochlea [66]. The frequency response of a gammatone filter describes how it reacts to different frequency components of an input signal. It behaves as a complex bandpass filter centered at a specific frequency, with its bandwidth and order determining how selectively it responds and how quickly it decays. Mathematically, the frequency response is defined as:

$$H_{GT}(f) = \frac{a}{(j2\pi(f - f_c) + 2\pi b)^n} \tag{3.1}$$

where $H_{GT}(f)$ is the frequency response of the gammatone filter at frequency f, centered at f_c where the filter has peak sensitivity. The term $2\pi b$ controls the bandwidth and damping, while $j2\pi(f - f_c)$ shifts the filter's response to be centered at f_c. The exponent n (typically 4) determines how steeply the response rolls off from the center frequency, and a is a gain normalization constant. The resulting gammatone spectrogram undergoes log-power transformation by applying a logarithmic function to enhance perceptual relevance. In addition to computing the gammatone cepstral coefficients (GTCCs) from the gammatone filterbank energies, a discrete cosine transform (DCT) is applied to the logarithm of these energies. The DCT serves two main purposes: (1) it de-correlates the filterbank outputs, making the features more statistically independent, and (2) it emphasizes the spectral envelope by concentrating most of the energy into the lower-order coefficients. This transformation results in a compact and robust representation of the speech signal's spectral characteristics. The resulting GTCC representations are shown in Fig. 3.2.

[1] `https://engineering.purdue.edu/~malcolm/interval/1998-010/`

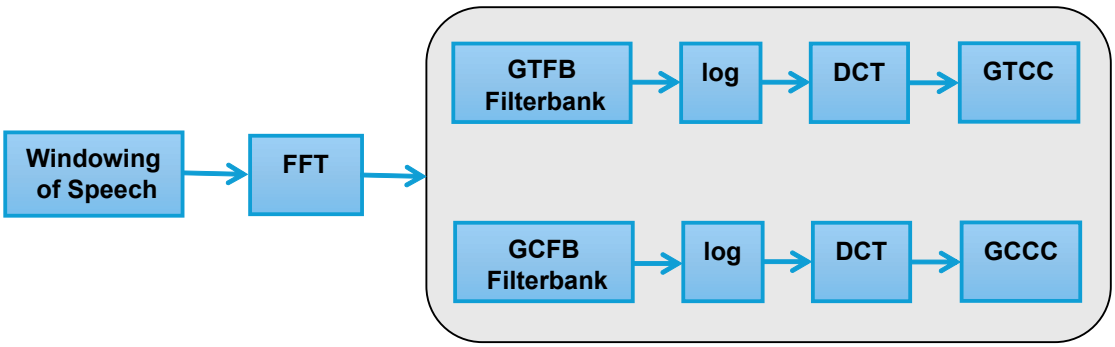


Figure 3.2: Block diagram to derive GTCCs and GCCCs using DCT.

3.4.2 Feature extraction with Gammachirp Filterbank

Irino and Patterson [67] and [68] developed the gammachirp filter to better simulate basilar membrane characteristics. Salhi et al. [69] and [70] later applied it to auditory feature extraction in speaker recognition. It provides a more physiologically accurate model of auditory filtering by incorporating a chirp term to account for the frequency-modulation effects in cochlear processing. The frequency response of the gammachirp filter at frequency f is given by:

$$H_{GC}(f) = \frac{a}{(j2\pi(f - f_c) + 2\pi b)^n} \cdot e^{c \cdot \theta(f)} \tag{3.2}$$

where $H_{GC}(f)$ is the frequency response of the gammachirp filter, f_c is the center frequency of the filter, b is the bandwidth parameter, n is the filter order (typically 4), j is the imaginary unit, a is a gain normalization constant, and c is the chirp coefficient (often negative, such as $c = -2$, to create an asymmetric response). The phase-shaping term, $\theta(f)$, is given in Eq. 3.3. The bandwidth normalization coefficient (b_coef = 1.019) is also used to adjust the filter response.

$$\theta(f) = \tan^{-1}\left(\frac{f - f_c}{b}\right) - \tan^{-1}\left(\frac{n}{c}\right) \tag{3.3}$$

The resulting gammachirp spectrogram undergoes a log-power transformation by applying a logarithmic function to enhance perceptual relevance. To obtain GCCCs (gammatone cepstral coefficients), the discrete cosine transform (DCT) is applied to the gammatone-derived features in order to de-correlate them and extract the cepstral coefficients. This process helps in reducing the dimensionality of the feature set while preserving the essential auditory characteristics. The DCT serves to transform the power spectrum of the signal into a more compact and

statistically independent form, making it more suitable for subsequent classification tasks. The resulting cepstral coefficients capture the temporal and spectral properties of the auditory features, as illustrated in Fig. 3.2.

3.5 Acoustic Feature Related to Auditory Perception

The acoustic features related to perception used in this dissertation include timbral features and STM representations, which were applied separately in the experiments. Timbre features play a critical role in music and audio cognition by conveying essential emotional and perceptual information, making them fundamental in many audio processing tasks. In parallel, STM representations offer a two-dimensional representation of sound in terms of spectral and temporal modulation content, capturing how the spectral envelope fluctuates across frequency and how the amplitude envelope varies over time, thereby aligning closely with human auditory perception mechanisms.

3.5.1 Timbral Features

Timbre encompasses a complex array of auditory characteristics that define the identity or quality of a sound. Even when two sounds share the same pitch, loudness, and duration, they can be perceived as distinct due to differences in their timbral attributes [71, 72]. These features include a range of spectral and harmonic components that contribute to the perceived uniqueness of a sound.

Timbre-related attributes fall under the category of psychoacoustic features, each corresponding to a specific sensory experience during auditory perception [73]. Psychoacoustics explores the intricate relationship between sound stimuli and human perception. Importantly, algorithmically generated scores often fail to replicate these subjective characteristics accurately, highlighting the limitations of purely objective models in capturing human auditory experience.

Numerous studies have aimed to model timbre based on psychoacoustic principles and to establish objective measures for its various attributes. A significant effort in this area is the Audio Commons project, which introduced a comprehensive timbre modeling framework. This framework combines low-level descriptors such as spectral centroid, dynamic range, spectral energy ratios, and relates them to eight perceptual timbre dimensions: hardness, depth, brightness, roughness, sharpness, warmth, boominess, and reverberation. Each of these high-level features is quantified on a scale ranging from 0 to 100 [6], and is defined as follows.

Hardness

This quality characterizes the sound generated when a hard object, such as a mallet, strikes a metallic bar. Four primary parameters can be used to predict this characteristic: the mean spectral centroid across time, the time-weighted average spectral centroid, the average spectral level over time, and the slope of the spectral level [74].

The perception of hardness in sound mainly results from the combination of loudness and harshness. Harsh sounds often show irregularities within the 2–5 kHz frequency range, where human hearing is particularly sensitive. As a result, hardness acts as a perceptual indicator of how pleasant or uncomfortable a sound is, reflecting the balance between its loudness and spectral properties in this critical band. In essence, it captures how well the sound's spectral characteristics align with human auditory preferences.

Several studies have investigated the acoustic correlates of perceived hardness. Williams [75] noted that the initial segment of a sound plays a key role in shaping hardness perception. Freed [76] proposed a framework for assessing mallet hardness in percussive sounds, based on four acoustic features: (1) spectral mean level (a form of long-term average spectrum, LTAS), (2) spectral level slope (analogous to cepstral analysis), (3) spectral centroid mean (averaged on the Bark scale), and (4) spectral centroid time-weighted average (TWA).

Solomon's foundational work [77] highlighted the hardness/softness dimension as an important psychological attribute of timbre and suggested a potential link to rhythmic variation, although no formal quantification was provided.

Despite the absence of a standardized model for hardness in existing literature, empirical evidence indicates that perceptual hardness is strongly influenced by the attack portion of a sound and its spectral characteristics. Building on this, a model of hardness was developed using three core metrics: (1) attack time, (2) attack gradient, and (3) spectral centroid during the attack phase. A linear regression model was then employed to estimate perceived hardness based on these features.

Furthermore, Pearce and colleagues [78] identified four suitable predictors for hardness estimation: the mean of the time-varying spectral centroid, the time-weighted average of the spectral centroid, the mean spectral level over time, and the spectral level slope.

Depth

Depth is linked to the perception of sound coming from beneath the surface of its source and is often associated with strong low-frequency components. It is modeled by considering factors such as the spectral centroid, spectral ratio, and the signal's decay time [74].

While the concept of depth in timbre has been discussed in several academic works, no formal model or set of acoustic features has yet been established to represent it. However, insights into this perceptual quality were obtained through an online experiment called Social-EQ, conducted by Pardo and Cartwright [79]. In this study, participants were asked to associate timbral descriptors with corresponding adjustments on a 40-band graphic equalizer.

Among the participants, six individuals selected the descriptor *deep*. Figure 3.3 presents the equalizer settings submitted by each of these participants. The bold black line in the figure represents the average EQ setting across participants, along with 95% confidence intervals.

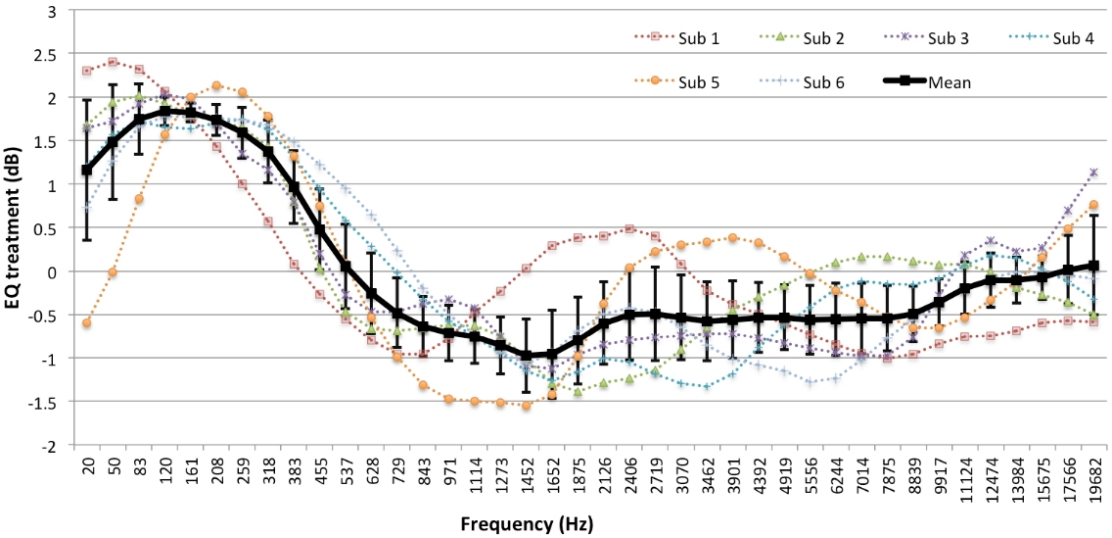


Figure 3.3: Social-EQ graphic equalizer settings for the timbral descriptor *deep*. Original image adopted from [6].

The trend shown in Figure 3.3 clearly indicates that all EQ treatments emphasize boosting the low-frequency components of the signal. The strong consistency across these responses suggests a perceptual association between *depth* and low-frequency energy. Based on this observation, a model for analyzing timbral depth can be proposed, incorporating the following features: (1) the spectral centroid within the lower frequency band, indicating the energy concentration towards low frequencies; (2) the ratio of low-frequency energy relative to the total; and (3) the low-frequency cutoff point of the audio signal, where energy begins to decline.

The lower spectral centroid is computed as follows:

$$\text{Lower spectral centroid} = \frac{\sum_{n(30\,\text{Hz})}^{n(200\,\text{Hz})} f(n) \cdot x(n)}{\sum_{n(30\,\text{Hz})}^{n(200\,\text{Hz})} x(n)}, \tag{3.4}$$

where $n(s)$ denotes the FFT bin number corresponding to the frequency s, $f(n)$ represents the frequency at bin n, and $x(n)$ indicates the magnitude of the spectral bin. The mean lower spectral centroid is subsequently computed by averaging this value across all frames of the signal.

The model additionally computes the lower ratio for each frame, which is defined as the proportion of energy contained within the 30–200 Hz band relative to the total energy spanning from 30 Hz up to the Nyquist frequency:

$$\text{Lower ratio} = \frac{\sum_{n(30\,\text{Hz})}^{n(200\,\text{Hz})} x(n)}{\sum_{n(30\,\text{Hz})}^{n(\text{Nyquist})} x(n)}, \tag{3.5}$$

where $x(n)$ is the magnitude of the nth spectral bin, and $n(\text{Nyquist})$ refers to the bin corresponding to the Nyquist frequency. The average lower ratio is then computed across all frames.

The third metric in the model is the low-frequency limit, which is adapted from the spectral rolloff method implemented in the IRCAM Timbre Toolbox [80]. Here, the low-frequency limit is defined as the frequency below which 5% of the total spectral energy is contained, meaning that 95% of the energy lies above this point. This threshold is determined by locating the frequency bin n that meets the following condition:

$$\frac{\sum_{k=0}^{n} x(k)^2}{\sum_{k=0}^{n(\text{Nyquist})} x(k)^2} \geq 0.05. \tag{3.6}$$

The mean low-frequency limit is then computed over all frames.

Finally, the overall timbral depth is estimated using a linear regression model that incorporates three predictors: (1) the lower spectral centroid, (2) the lower ratio, and (3) the low-frequency limit. The low-frequency limit is derived based on the cumulative spectral energy using $x(n)^2$, following an energy-based thresholding approach.

Brightness

Brightness refers to the perceived clarity or vividness of a sound, which is typically attributed to the presence of higher-frequency components. It is quantified using two primary measures: the ratio of high-frequency energy to the total energy, and the spectral centroid calculated above a defined frequency threshold [81].

As a key aspect of sound quality, brightness has been widely studied. Many researchers have identified the spectral centroid as a dependable indicator of perceived brightness [82,83]. Nonetheless, some studies suggest that the high-frequency energy ratio relative to total spectral energy may provide a more precise prediction. In a recent investigation, Pearce [84] reviewed existing models and introduced a

new approach that combines a modified spectral centroid with the spectral energy ratio. This combined model has since been incorporated into the Audio Commons framework [6], which is discussed in this dissertation.

The computation of brightness involves several signal processing steps. First, the audio signal is segmented into short frames and transformed into the frequency domain using the Fast Fourier Transform (FFT). To enhance the spectral representation, a half-octave smoothing is then applied to the magnitude spectrum of each frame on a sample-by-sample basis. From the smoothed spectrum, two primary features are extracted. The first is the frequency-limited spectral centroid (FLSC), which emphasizes spectral content above 3 kHz:

$$\text{FLSC} = \frac{\sum_{n(3\text{kHz})}^{n(\text{Nyquist})} f(n) \cdot x(n)}{\sum_{n(3\text{kHz})}^{n(\text{Nyquist})} x(n)}, \tag{3.7}$$

where $n(s)$ denotes the FFT bin corresponding to frequency s, $f(n)$ is the center frequency of bin n, and $x(n)$ represents the magnitude of bin n.

The second metric, referred to as the spectral energy ratio, quantifies the proportion of high-frequency energy (above 3 kHz) relative to the total energy:

$$\text{Ratio} = \frac{\sum_{n(3\text{kHz})}^{n(\text{Nyquist})} x(n)}{\sum_{n(20\text{Hz})}^{n(\text{Nyquist})} x(n)}. \tag{3.8}$$

Finally, a linear regression model combines the Ratio and FLSC metrics to predict the perceived brightness B.

$$B = -25.8699 + 64.0127\,(\log_{10}(\text{Ratio})) + 0.44\log_{10}(\text{FLSC}). \tag{3.9}$$

Roughness

Roughness contributes to the perception of a buzzing or harsh sound quality and is primarily associated with amplitude fluctuations in the 16 to 80 Hz range. The perception of roughness is modeled using a sinusoidal approach, which incorporates three variables. Two of these variables pertain to the amplitudes of sinusoidal components and reflect roughness dependence on overall intensity and on the degree of amplitude modulation [85].

The method for calculating roughness in this dissertation follows the procedure described in [6, 86]. The audio signal is segmented into 50ms frames, with each frame processed using a Hanning window and zero-padded to the nearest power of two. A Fast Fourier Transform (FFT) is then applied to each frame, and the resulting magnitude spectra are normalized so that the highest magnitude across

all frames is scaled to 1.0. This normalization ensures consistent spectral comparisons between frames. Next, a peak-picking algorithm identifies the prominent frequency components within each frame's spectrum. Roughness is computed for every pair of spectral peaks in a frame using the following model:

$$R = 0.5X^{0.1}Y^{3.11}Z \tag{3.10}$$

where:

$$X = A_{\min} \cdot A_{\max}, \tag{3.11}$$

$$Y = \frac{2A_{\min}}{A_{\min} + A_{\max}}, \tag{3.12}$$

$$Z = e^{-3.5g(f_{\max} - f_{\min})} - e^{-5.75g(f_{\max} - f_{\min})}, \tag{3.13}$$

$$g = \frac{0.24}{0.0207f_{\min} + 18.96}. \tag{3.14}$$

where, R denotes the roughness between a pair of peaks, $A_{\max}$ and $A_{\min}$ are the peak magnitudes, and $f_{\max}$, $f_{\min}$ represent their corresponding frequencies. The total roughness per frame is obtained by summing roughness values across all peak pairs. Finally, the overall roughness of the audio signal is computed as the average roughness across all frames.

Warmth

Warmth is often linked to the perception of low-frequency sounds, giving the impression of rising warmth or richness. A warmth model was created through iterative selection of low-level timbre features and application of multilinear regression [81].

The perception of warmth in sound has been explored in various studies, each identifying key factors that contribute to this auditory quality. For example, Srensen [87] found that the acoustic characteristics of concert venues, such as their architectural shape and reflective properties, significantly influence how warm orchestral music is perceived. Elements like room size, reverberation time, and the spatial distribution of reflections were shown to enhance a sense of liveness and fullness, particularly in lower frequency ranges closely associated with warmth.

Bromham [88] investigated how audio processing techniques affect the perception of warmth, emphasizing the relevance of the bass ratio (BR) metric, which quantifies energy distribution within the low-frequency spectrum. The study

demonstrated that modifying the BR, for example through equalization, can significantly alter the listener's perception of warmth.

Building on this foundation, Williams [89] developed a method for the perceptual adjustment of warmth by manipulating various timbral characteristics. This approach highlights the complex and multidimensional nature of warmth and shows that it can be actively shaped through precise control of acoustic and signal processing parameters.

Farbood [90] further advanced the understanding of warmth by examining its relationship with perceived musical tension. The study showed that timbral factors such as inharmonicity and roughness influence tension perception, which in turn correlates with the perceived warmth of the sound.

Collectively, these studies underscore the intricate interplay of acoustic, timbral, and perceptual elements in shaping the sensation of warmth. They also stress the importance of considering these factors in practical applications such as music production, sound design, and acoustic engineering to achieve a desired auditory experience.

Sharpness

Sharpness is associated with a piercing or penetrating sensation in sound. In speech, sharpness may be characterized by a short open phase of the vocal glottal cycle. The sharpness model is based on the Klippel sharpness model, which quantifies this sensation [81].

To elaborate, sharpness is a psychoacoustic attribute associated with the perception of acute or piercing auditory sensations, which tend to intensify as the spectral centroid shifts toward higher frequencies. Based on this phenomenon, Zwicker et al. introduced the unit *acum*, defined as the sharpness of a narrowband noise centered at 1,000 Hz with a loudness level of 60 phon [91]. Building on this concept, a mathematical model for sharpness was developed and is expressed as:

In more detail, sharpness is a psychoacoustic quality linked to the perception of high-pitched or piercing sounds, which becomes more pronounced as the spectral centroid moves toward higher frequencies. To characterize this sensation, Zwicker et al. introduced the unit "acum" defined as the sharpness of a narrowband noise centered at 1,000 Hz with a loudness level of 60 phon [91]. Building on this foundation, a mathematical model of sharpness was developed and is expressed as:

$$\text{Sharp} = 0.11 \frac{\int_0^{24\ \text{Bark}} D'(x)\, g(x)\, x\, dx}{\int_0^{24\ \text{Bark}} D'(x)\, dx} \tag{3.15}$$

In this equation, Sharp represents the sharpness value, while $D'(x)$ denotes the

specific loudness density as a function of the critical-band rate x, measured in Bark. Loudness, an essential measure of auditory perception, is expressed in phons. The function $g(x)$ acts as a frequency-dependent weighting factor: it is set to 1.0 for frequencies up to 3,000 Hz and increases steeply to 4.0 for higher frequencies, as established through psychoacoustic studies.

Boominess

Boominess results from low-frequency harmonic resonance, like the sound of a car engine. The boominess model is adapted from the framework developed by Hatano and Hashimoto, and it calculates the weighted average of sound pressure levels across frequency bands important for boominess perception [92], [93].

Extensive studies have examined the phenomenon of surging across a range of fields, including construction machinery [94], automotive systems [95], and interior vehicle acoustics [96]. Surging sensations are commonly linked to low-frequency elements, often originating from engine noise [97]. To better characterize and address this effect, researchers have proposed several objective measures, such as the sound quality index [95] and the weighted sound pressure level [96]. These metrics have proven effective in mitigating the impact of surging, particularly within the domain of in-cabin vehicle noise [96]. Although no standardized model for boominess currently exists in the literature, this dissertation adopts the concept as defined within the Audio Commons framework [6].

Reverberation

Reverberation refers to the lingering presence of sound after the source has stopped, due to reflections in the environment. It represents the acoustic decay of an audio signal and is characterized by several standard metrics, the most common of which is *RT60* an indicator of the time required for the sound level to decay by 60 decibels. While *RT60* is typically measured in controlled environments such as concert halls, accurately estimating it from audio recordings presents significant challenges. To address this, the IEEE organized the Acoustic Characterization of Environments (ACE) Challenge in 2015, aiming to evaluate blind estimation methods for *RT60* and the direct-to-reverberant ratio (DRR) in recorded speech signals [98].

The reverberation algorithm was implemented following the method proposed by Prego *et al.* [99]. The process begins with the computation of the signal's power spectrogram. Although specific parameters such as frame length and window function are not strictly defined in the original method, a Hamming window of length 2048 samples with a 512-sample overlap is selected arbitrarily. The analysis is then confined to the frequency range of 20 Hz to 4 kHz, which is particularly relevant

as it encompasses the majority of speech information and aligns with the standard frequency bands used in building acoustics to estimate *RT60* commonly averaged over the 500 Hz, 1 kHz, and 2 kHz octave bands.

For each identified Signal Frame Decay Region (SFDR), the Schroeder integration is computed using the following equation:

$$\mathrm{c}(k,\ell,n) = 10\log_{10}\left(\frac{\sum_{a=n}^{L} E(k,a)}{\sum_{a=n}^{L} E(k,a)}\right), \tag{3.16}$$

where $\mathrm{c}(k,\ell,n)$ denotes the nth frame within the ℓth SFDR in the kth sub-band, and L is the total number of frames within that SFDR. $E(k,a)$ represents the energy in the kth sub-band at frame a, while n indicates the current frame under analysis.

The Schroeder integral is computed for each SFDR to estimate the corresponding *RT60*. The start of the decay analysis is identified by locating the first frame within the SFDR where the integral consistently begins to decline.

If the most linear segment of the Schroeder integral exhibits a dynamic range of less than 10 dB, the algorithm instead searches for the most linear section with at least a 60 dB dynamic range. If no such segment is found, it proceeds to look for a portion with a minimum range of 40 dB. If this also fails, the threshold is gradually reduced to 20 dB and then to 10 dB until a suitable segment is identified.

Once an appropriate segment has been selected, linear regression is applied to this portion of the SFDR. The SFDR-based *RT60* is then estimated from the regression coefficients, representing the time it would take for the fitted regression line to decay by 60 dB.

The sub-band *RT60* is calculated as the median of all SFDR-based *RT60* estimates within the corresponding sub-band. The overall *RT60* is then approximated as the median of the sub-band *RT60* values. To improve the accuracy of this estimate, an empirical correction is applied by dividing the computed *RT60* by a factor of three.

3.5.2 Spectro-temporal Modulation

In this thesis, STM representations were computed through a multi-step process consisting of gammatone filterbank decomposition, envelope extraction, and two-dimensional Fourier transformation. The calculation processes are depicted in Fig. 3.4.

Initially, a gammatone filterbank is used to decompose the input signal $x(t)$ into a series of frequency bands. In this step, the fammatone filterbank (GTFB) is adopted and implemented using a cascaded IIR filter design. The center frequencies and bandwidths of the gammatone filters are derived based on the speci-

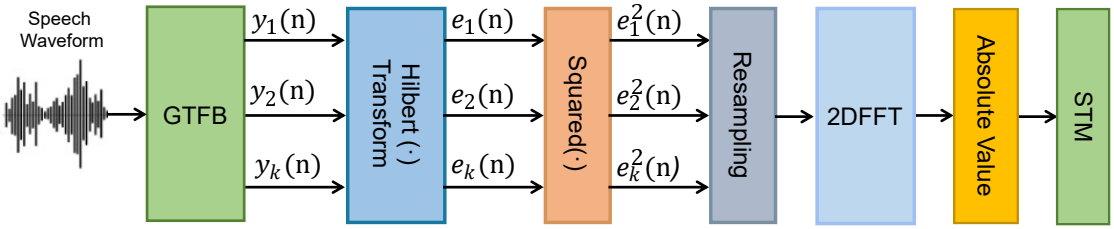


Figure 3.4: Block diagram depicting the steps for STM representation calculation.

fied frequency range and the equivalent rectangular bandwidth (ERB) expression. Therefore, the GTFB is also referred to as an ERB filterbank. In this thesis, the number of channels is set to 64, covering the frequency range from 60 Hz to 7,600 Hz. The impulse response of the k-th gammatone filter can be represented as:

$$g_k(t) = A\, t^{(n-1)} \exp\big(-2\pi b_f\, ERB(f_k)\, t\big)\, \cos(2\pi f_k t), \tag{3.17}$$

with

$$ERB = 24.7\,(4.37\, f_k + 1), \tag{3.18}$$

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

$$y_k(t) = g_k(t) * x(t), \tag{3.19}$$

where $*$ denotes convolution.

Subsequently, the Hilbert transform and squaring operations are implemented to compute the power envelope of each frequency band. The output after low-pass filtering is given by:

$$e_k^2(t) = LPF\big[\,|\,\text{Hilbert}\big(y_k(t)\big)\,|^2\,\big], \tag{3.20}$$

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

$$STM = \big|\,2DFFT\big(e_k^2(t)\big)\big|. \tag{3.21}$$

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

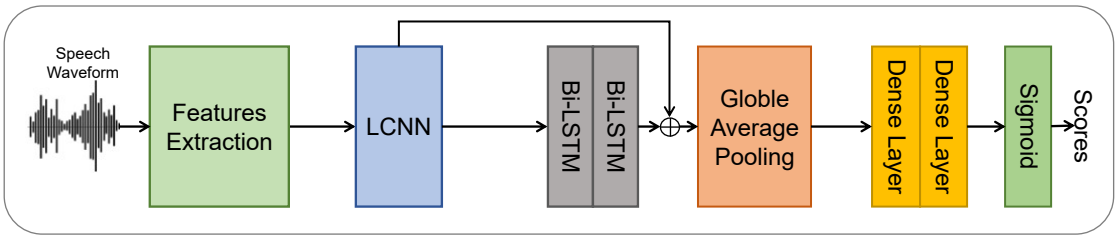


Figure 3.5: LCNN architecture.

3.6 Classification Based on Deep Learning

This study applies various deep learning architectures, LCNN, ResNet18, and VGG16, as back-end classifiers, each offering a distinct method for feature representation and classification.

Recently many studies [56], [62–65] used LCNN for decting the Deepfake speech. LCNN begins with a two-dimensional convolutional layer, followed by a feature reduction mechanism to eliminate redundancies. It then progresses through multiple convolutional layers with normalization and pooling operations, gradually increasing the number of channels while down-sampling. The extracted features are reshaped and processed through bidirectional recurrent layers for temporal modeling. Finally, a pooling operation aggregates the output, followed by fully connected layers for binary classification, as shown in Fig. 3.5. The model is trained using a standard classification-loss function and optimization algorithm.

Another key architecture employed in this study is ResNet-18, a deep convolutional neural network introduced by He et al. as part of the Residual Network (ResNet) family in 2015 [100]. Its core innovation lies in the use of residual connections, short-cut paths that allow gradients to bypass certain layers, effectively mitigating the problem of vanishing gradients in deep networks. ResNet-18 comprises 17 convolutional layers followed by a fully connected layer and is known for its balance between architectural depth and computational efficiency. Although originally developed for image recognition, ResNet-18 has been widely adopted in audio classification tasks [101–105] by transforming one-dimensional audio signals into two-dimensional time-frequency representations such as mel spectrograms, MFCCs, or GTCCs. In the field of spoofed or fake speech detection, ResNet-18 has demonstrated strong performance across multiple studies [106–108], including in ASVspoof challenges, where its ability to learn localized spectro-temporal artifacts has proven effective for distinguishing genuine from manipulated speech.

To implement this approach, a ResNet-18-style architecture is employed as a classifier to distinguish between genuine and human-imitated speech as shown in Fig 3.6. Instead of processing raw audio signals directly, the speech is first transformed into two-dimensional feature representations that encode both spec-

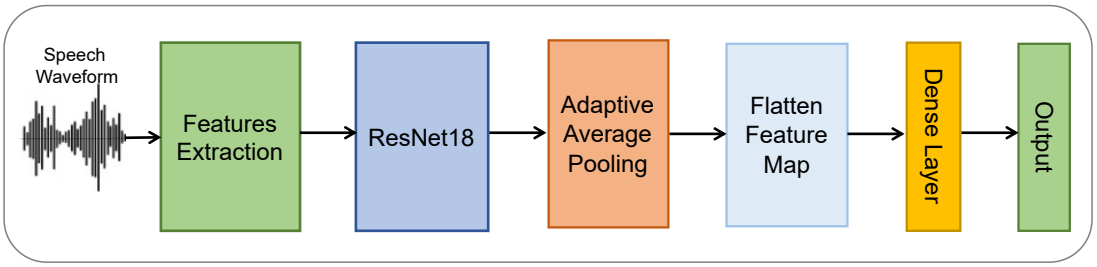


Figure 3.6: ResNet18 architecture.

tral and temporal information. These features are preprocessed and reshaped to match the input format expected by convolutional neural networks. ResNet-18 follows a structured architecture with sequential convolutional blocks, each followed by normalization, activation, and pooling. The network depth increases across residual stages, where each stage comprises two convolutional layers along with a shortcut connection to enable residual learning. As the model progresses, the number of feature channels increases from 64 to 128, 256, and 512, allowing the network to learn increasingly complex and hierarchical representations. The final layers include an adaptive average pooling operation that reduces the spatial resolution to a fixed size, followed by flattening and a fully connected layer to perform binary classification. This design enables efficient gradient flow and enhances the model's capacity to capture subtle spectro-temporal cues for distinguishing between genuine and imitated speech.

Following LCNN and ResNet18, VGG16 is employed in this study as a classifier. It is a deep convolutional neural network architecture developed by Simonyan and Zisserman at the Visual Geometry Group (VGG), University of Oxford, and introduced in 2014 [109]. The hallmark of VGG16 is its simplicity: It consists of 13 convolutional layers using small 3×3 filters and 3 fully connected layers. Although originally designed for image classification, VGG16 has been widely adopted in audio classification tasks by converting audio signals into 2D representations similar to spectrograms (e.g. mel spectrograms, MFCCs), allowing the model to learn patterns in the time-frequency domain [110–112]. In the context of spoofed or fake speech detection, VGG16 and its variants (like VGGish) have been used in multiple studies [113–117] such as ASVspoof challenges and audio deepfake detection articles, demonstrating strong performance in identifying synthetic speech due to their ability to capture subtle artifacts in audio spectrograms.

To adapt VGG16 for this task, the model is applied to classify between genui to distinguish between genuine and human-imitated speech as show in Fig. 3.7. Instead of directly processing raw audio, the speech signals are first transformed into two-dimensional features that represent both spectral and temporal patterns. These features are stored as 2D arrays and padded to a consistent shape, allowing

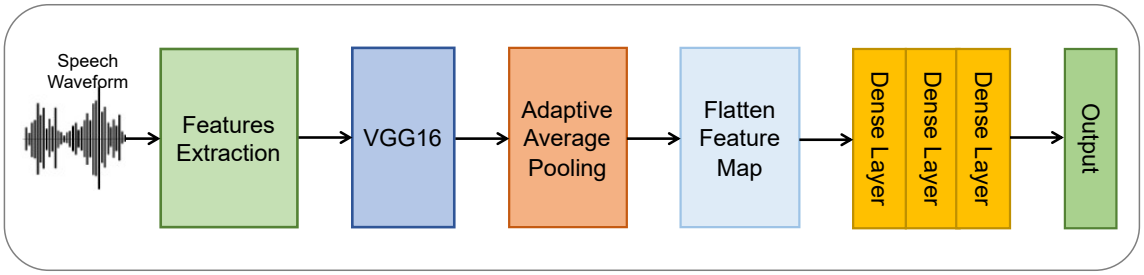


Figure 3.7: VGG16 architecture.

them to be processed similarly to image data. The model architecture follows the original VGG16 structure, which consists of 13 convolutional layers using small 3×3 filters and 3 fully connected layers. These convolutional layers are grouped into five sequential blocks, each followed by batch normalization, ReLU activation, and max pooling. The depth of the network increases progressively, enabling it to learn increasingly complex and hierarchical representations of the input features. Specifically, the five blocks contain 64, 128, 256, and 512 filters, respectively, with multiple convolutional layers per block. An adaptive average pooling layer then reduces the output to a fixed spatial size of 7×7. The final classification stage includes two fully connected layers with 4096 units each, followed by an output layer for binary classification. Each input feature is reshaped to include a channel dimension, ensuring compatibility with the expected input format of the convolutional network. This structured and deep architecture enables the model to effectively learn fine-grained acoustic cues that differentiate genuine speech from human-imitated speech.

Chapter 4

Data Collection and Evaluation Metrics

In the domain of fake speech detection, the quality and design of datasets are fundamental to achieving reliable performance, especially when using acoustic features and deep learning models. This chapter focuses on the process of creating a custom human-imitated speech dataset, specifically developed for this study. It details the data collection procedure, including the design, participants, recording environment, and labeling strategy. Additionally, the chapter introduces the evaluation metrics used to assess the performance of the selected features and deep learning classifiers. These metrics provide a comprehensive understanding of model effectiveness in distinguishing between real and imitated speech.

4.1 Dataset

The creation of the proposed dataset involved the following steps.

4.1.1 Scenario

Several studies have collected data to solve specific imitation-related challenges. Kim et al. [35] gathered vocal imitations to address the problem of sound event retrieval, where imitated sounds help improve search capabilities in audio systems. Ballesteros et al. [7] focused on detecting fake speech, including AI-generated and imitated speech, by collecting synthetic speech datasets for security purposes. Lataifeh et al. [39] worked on identifying authentic and imitated Qur'anic recitations, collecting Arabic audio clips to distinguish religious cantillations. Mehrabi et al. [40] focused on the vocal imitation of percussion sounds, collecting data to study perceptual similarity in musical contexts.

This study aims to solve the problem of distinguishing human-imitated speech from genuine speech in conversational contexts, a critical task in voice authentication and forensic analysis. To address this, a dataset of imitated speech was collected from online sources, where the imitated speech was performed by professional artists across various languages. This dataset supports the development of models capable of detecting imitation, filling a gap in current structured datasets that have not yet explored the challenge of distinguishing human-imitated speech in natural, conversational scenarios.

4.1.2 Data-collection Process

The proposed human-imitated speech dataset was created through a rigorous and carefully controlled process to ensure both audio quality and speaker diversity. The proposed data set includes six languages including English, Turkish, Hindi, Urdu, Dari, and Pashto, and features ten target speakers. All target speakers are well-known politicians and celebrities, making it a diverse and representative collection. The process began with the collection of genuine speech recordings, which were sourced from publicly available platforms, including various online video repositories. These recordings represented natural speech with clear articulation, varying speaking styles, and a broad range of speaker identities and contexts.

For the imitated speech, data was collected from professional mimic artists who possess the ability to accurately replicate speech patterns, intonations, and vocal characteristics of other individuals. These artists were selected for their experience in voice mimicry across different speakers and languages, providing a rich and varied set of imitation samples.

Several challenges were encountered during the data collection process. One major issue was the limited availability of mimicry-specific content, as such performances are relatively niche and not widely archived. Additionally, identifying skilled mimic artists who could contribute high-quality recordings proved difficult and time-consuming. Another significant challenge was the presence of background noise such as audience cheering, clapping, or overlapping conversations, particularly in imitation speech samples sourced from public performances or recordings. To preserve audio clarity and consistency, all recordings were manually reviewed, and noisy or poor-quality samples were excluded from the dataset.

All collected recordings, both genuine and imitated, were processed and uniformly segmented into three-second clips. This segmentation ensured consistency in sample duration, which is essential for training machine learning models on fixed-length inputs.

The final dataset is carefully balanced, containing approximately equal proportions of genuine and imitated speech, making it particularly well-suited for tasks such as fake speech detection, voice imitation analysis, and linguistic feature

extraction. The final dataset not only supports research in deepfake and voice spoofing detection but also serves as a valuable resource for broader studies in speech processing and cross-speaker modeling.

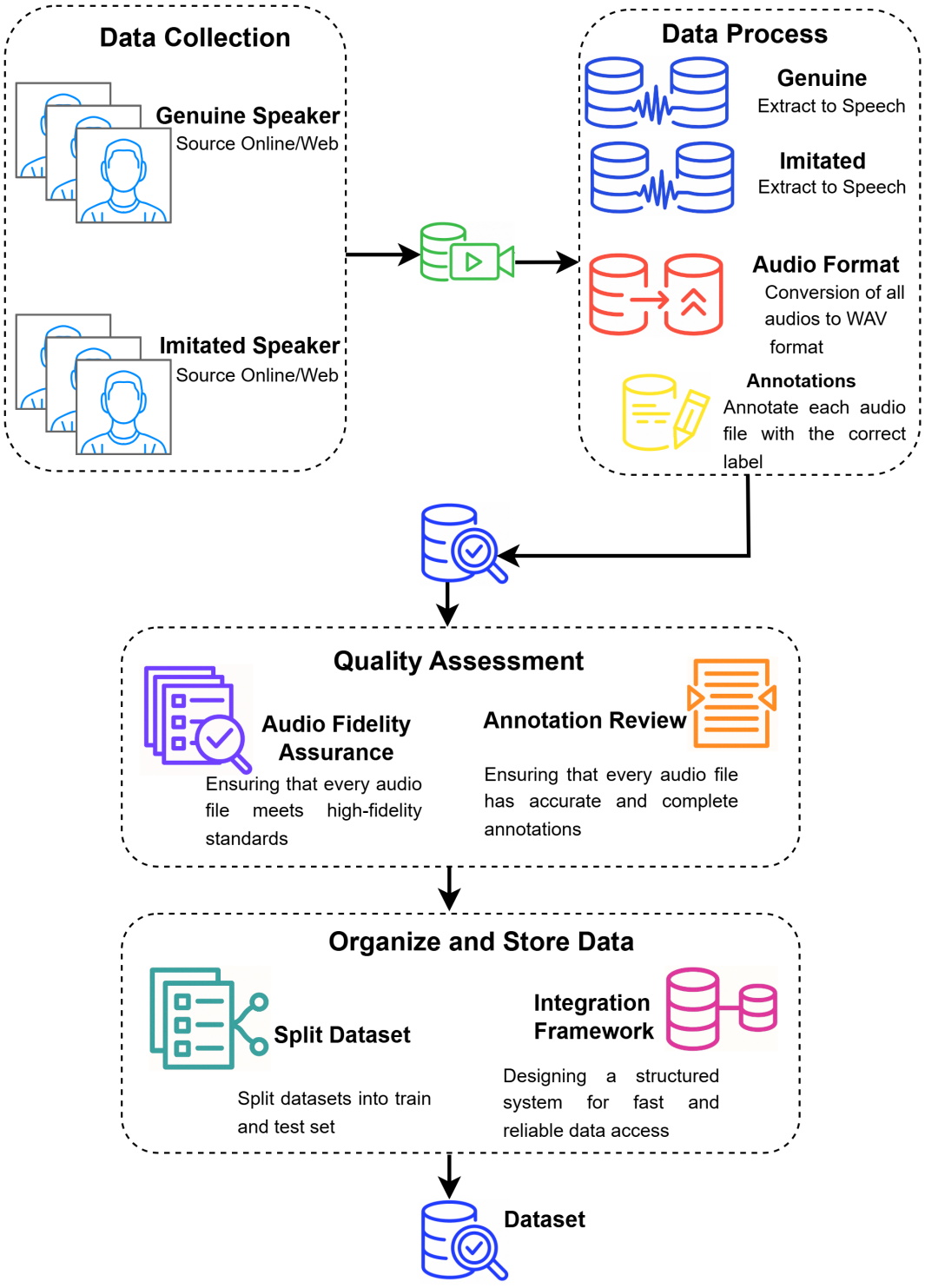


Figure 4.1: Dataset-creation process.

Table 4.1: Number of utterances of genuine and imitated speech.

Splits	**Number of Utterances**		
	Genuine	**Imitated**	**Total**
Train Set	4,944	3,513	8,457
Test Set	5,505	7,117	12,622

4.1.3 Dataset Splits

After completing the data collection phase, the entire dataset was strategically divided into separate training and test sets to enable both effective model training and rigorous evaluation. The complete dataset comprises approximately 17 hours of speech data, which includes 8.70 hours of genuine (real) speech and 8.30 hours of imitation (spoofed) speech. This near-equal distribution of genuine and imitation speech ensures a balanced representation of both classes, which is crucial for training reliable and unbiased models.

To better evaluate the model's generalization capability on unseen data, a deliberate choice was made to allocate a slightly larger portion of the data to the test set. The training set consists of 8,457 utterances, including 4,944 genuine samples and 3,513 imitation samples, while the test set includes 12,622 utterances, made up of 5,505 genuine and 7,117 imitation samples. This division corresponds to approximately 40% of the data assigned for training and 60% reserved for testing.

The rationale behind this data split stems from the objective of conducting a more thorough and detailed evaluation of model performance. By increasing the proportion of the test set, it is ensured that the model's behavior is assessed across a wider range of scenarios and utterance variations, which closely mirrors real-world deployment conditions. At the same time, the training set retains a sufficient number of diverse samples to enable effective learning of discriminative features between genuine and imitation speech.

This design choice also reflects a focus on robustness and generalizability, rather than solely optimizing performance on the training data. This is particularly important in tasks such as fake speech detection, where the model is expected to encounter a wide variety of speech styles, speaking conditions, and deceptive attempts. In this work, fake speech refers specifically to human-imitated speech, where individuals deliberately mimic the voice characteristics of others. Since such imitation can vary widely in quality and style, the ability of the model to generalize well to previously unseen samples is essential. Therefore, allocating a larger portion of data to testing plays a vital role in evaluating the model's performance in real-world scenarios and contributes to the development of more reliable and high-performing detection systems. In addition, a subset of data from

this dataset split was used for a subjective listening experiment, consisting of an equal balance of 50% genuine speech and 50% imitated speech. This subset was carefully selected to evaluate human perception of voice authenticity and to compare it with the model's performance. The creation of the proposed dataset details in Fig. 4.1.

4.2 Evaluation Metrics

This section describes the evaluation metrics applied to assess the performance of the selected features and deep learning classifiers. These metrics offer a comprehensive view of each model's capability to discriminate between genuine and imitated (fake) speech. By incorporating multiple evaluation criteria, the analysis ensures a robust and balanced assessment that considers both overall accuracy and class-specific performance. The computed metrics include accuracy, Equal Error Rate (EER), d-prime, F1-score, and the confusion matrix.

Confusion Matrix

The confusion matrix is a performance evaluation tool commonly used in classification tasks to compare a model's predictions against actual outcomes. It reports counts of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN), offering detailed insights into the model's strengths and weaknesses. In this context, it shows how effectively the model distinguishes between genuine and imitated speech. The confusion matrix is essential for evaluating both overall accuracy and the model's capacity to minimize incorrect classifications by measuring sensitivity (true positive rate) and specificity (true negative rate). Table 4.2 presents the confusion matrix used in this study.

Table 4.2: Confusion matrix.

	Predicted Genuine	Predicted Imitated
Genuine	TP	FN
Imitated	FP	TN

Accuracy

Accuracy refers to the percentage of data points correctly classified by a model, comparing the number of accurate predictions to the total number of predic-

tions [3]. It is a key metric for evaluating the overall performance of a model in classification tasks. Accuracy is calculated using the following equation:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \tag{1}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. However, accuracy can be misleading in imbalanced datasets, where additional metrics like precision, recall, or the F1-score are often necessary to provide a more complete evaluation.

F1-score

The F1-score measures the overall performance of a model by calculating the harmonic mean of precision and recall, offering a balance between the two [3], [118]. It is particularly useful in cases of imbalanced datasets, where accuracy alone may not give a clear picture of performance. The F1-score ranges from 0 to 1, with higher values indicating better performance. It is computed using the following formula.

$$\begin{aligned} \text{F1-score} &= \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \\ &= \frac{TP}{TP + \frac{1}{2}(FP + FN)} \end{aligned} \tag{2}$$

This makes the F1-score an effective metric for evaluating classification models, especially when both false positives and false negatives are significant.

Equal error rate (EER)

The equal error rate (EER) is a critical metric used to evaluate the performance of systems, particularly in biometric recognition tasks. It is the point where the false acceptance rate (FAR) equals the false rejection rate (FRR), providing a balanced measure of the system's ability to avoid both false acceptances and false rejections. Mathematically, it is represented as follows.

$$\text{EER} = \text{FAR}(\text{threshold}) = \text{FRR}(\text{Threshold}) \tag{3}$$

The false acceptance rate (FAR) is the probability that the system incorrectly accepts an unauthorized user. It is calculated as follows.

$$\text{FAR} = \frac{\text{False Acceptances}}{\text{Total Number of Unauthorized Attempts}} \tag{4}$$

The false rejection rate (FRR) is the probability that the system incorrectly rejects an authorized user. It is calculated as follows.

$$\text{FRR} = \frac{\text{False Rejections}}{\text{Total Number of Authorized Attempts}} \tag{5}$$

The EER is often visualized using a receiver operating characteristic (ROC) curve or a detection error tradeoff (DET) curve, where the point of intersection between FAR and FRR represents the EER. A lower EER indicates better system performance, as it minimizes both types of errors.

d-prime

d-prime (d') is a statistical measure used in signal detection theory to assess a system's sensitivity in distinguishing between signal and noise [119], [120]. In the context of genuine and imitated speech classification, d-prime quantifies the system's ability to differentiate between correctly identified genuine speech (hits) and misclassified imitated speech (false alarms). It is calculated as the difference between the Z-scores of the hit rate (HR) and the false alarm rate (FAR), mathematically represented as follows.

$$d' = Z(\text{HR}) - Z(\text{FAR}) \tag{6}$$

The hit rate (HR) is the proportion of genuine speech samples correctly classified as genuine. It is calculated as follows.

$$\text{HR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{7}$$

The false alarm rate (FAR) is the proportion of imitated speech samples incorrectly classified as genuine. It is calculated as follows.

$$\text{FAR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}} \tag{8}$$

d-prime provides a measure of how well the system separates a signal (genuine speech) from noise (imitated speech). A higher d-prime value indicates better sensitivity, meaning the system can more accurately distinguish between genuine and imitated speech.

Chapter 5

Human-Imitated Speech Detection Using Auditory-Based Features and Deep Learning

This chapter presents a study that explores a spoof countermeasure system for distinguishing between genuine and human-imitated speech, with a particular focus on auditory-based feature representations. The evaluation was conducted using the ASVspoof 2019 LA dataset as the synthetic speech benchmark and the entire human-imitated speech dataset developed for this study. This setup was design to highlight the limitations of conventional approaches and the benefits of auditory features.

First, a model was trained on synthetic speech using standard acoustic features, including mel-spectrogram, MFCCs (Mel-Frequency Cepstral Coefficients), and LFCCs (Linear-Frequency Cepstral Coefficients). When tested on synthetic speech, the model performed well; however, its performance dropped significantly when evaluated on human-imitated speech, underscoring the distinct characteristics of human imitation compared to synthetic spoofing.

Next, the same model was trained directly on the proposed human-imitated speech dataset using standard features. This approach improved detection performance on imitated speech, confirming the need for dedicated training data.

Finally, the core contribution of this work involved incorporating auditory-inspired features, specifically gammatone and gammachirp filterbank representations, to better simulate human auditory processing. Models trained and tested on human-imitated speech using these features showed further improvements over conventional spectral representations.

To conduct these experiments, a Light Convolutional Neural Network (LCNN) was employed as the primary architecture, alongside ResNet-18 and VGG-16 models for comparison. This stepwise evaluation demonstrates that auditory-based fea-

tures can enhance the robustness of spoofing countermeasure systems, particularly in detecting challenging human-imitated speech.

5.1 Proposed Method

This section presents the proposed features, deep learning models, and experimental setup used in this study. This study primarily explored gammatone filterbank (GTFB) and gammachirp filterbank (GCFB) features, as illustrated in Fig. 5.1, which closely model human auditory processing for detecting imitated speech. This study follows Slaney's Auditory Toolbox[1] and the study by Nguyen et al. [61]. To benchmark their effectiveness, the mel-spectrogram, Mel-frequency cepstral coefficient (MFCC), and linear-frequency cepstral coefficient (LFCC) were also used as standard features for evaluating a spoofing-countermeasure system [56], [62–65]. Additionally, the gammatone cepstral coefficient (GTCC) and gammachirp cepstral coefficient (GCCC) were introduced for comparative analysis.

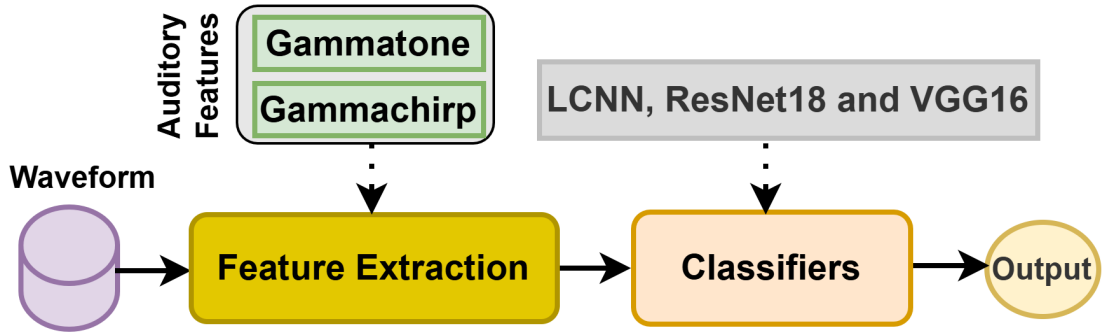


Figure 5.1: Block diagram of proposed method.

5.1.1 Feature Extraction with Standard Features

To benchmark the effectiveness of the auditory features, this study utilized standard features. A mel-spectrogram is derived directly from the raw signal and modeled on the human auditory system. The Mel scale provides a perceptually motivated frequency representation defined as

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.1}$$

[1] https://engineering.purdue.edu/~malcolm/interval/1998-010/

where $M(f)$ represents the Mel frequency corresponding to f, aligning with human pitch perception [8, 121]. MFCCs are extracted by applying the discrete cosine transform (DCT) to the Mel-scale power spectrum:

$$\mathrm{MFCC}_l = \sum_{n=0}^{N-1} S(m) \cos\left(\frac{\pi}{M}\,(m+0.5)\cdot l\right) \tag{5.2}$$

where MFCC_l is the l-th coefficient, $S(m)$ is the magnitude in the m-th Mel-filtered bin, and M is the total number of filters [8]. LFCCs follow a similar process; however, a linearly spaced filter bank instead of the Mel scale is used:

$$\mathrm{LFCC}_l = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi}{M}\,(m+0.5)\cdot l\right) \tag{5.3}$$

5.1.2 Feature Extraction with Gammatone Filterbank

The GTFB was applied using a fast Fourier transform (FFT) size of 512 and a sampling rate of 16,000 Hz to extract perceptually relevant features from speech. The GTFB consists of 64 filters spaced on the equivalent rectangular bandwidth (ERB) scale, covering frequencies from 100 Hz to the Nyquist frequency (maxfreq = $\frac{\mathrm{sr}}{2}$), with a bandwidth-scaling factor (width = 1.0). The short-time Fourier transform (STFT) is computed using a Hann window, window length of 400 samples, and hop size of 100 samples. The key auditory parameters include the ERB quality factor (EarQ = 9.26449), which controls filter spacing, and the minimum bandwidth of 24.7 Hz, ensuring a lower bound on auditory filter widths. The filter order (order = 1) defines how bandwidth scales with frequency, and the gammatone filter order (GTord = 4) specifies the number of cascaded filters shaping the auditory response. The pole radius (r) is determined using an exponential decay function, modeling the natural damping of auditory filters, while filter poles (pole) are computed using complex exponentials to align with human cochlear tuning. The resulting gammatone spectrogram undergoes log-power transformation by applying a logarithmic function to enhance perceptual relevance. Further evaluation was conducted on the GTCCs, where a DCT was applied to decorrelate the features and extract the cepstral coefficients.

5.1.3 Feature Extraction with Gammachirp Filterbank

For the gammachirp filterbank, the same processing setup and auditory parameters as in the gammatone configuration were retained. The key distinction lies in the introduction of a chirp coefficient (chirp_coef = -2.0), which introduces an

asymmetric frequency response by modeling frequency glides observed in the human cochlea. This adjustment enhances the auditory model's ability to simulate the nonlinear and level-dependent characteristics of cochlear filtering. The gammachirp weights are computed using both symmetric (gammatone) and asymmetric components, allowing for improved frequency resolution near the filter center frequencies. The resulting gammachirp spectrogram undergoes a log-power transformation by applying a logarithmic function to enhance perceptual relevance. Further evaluation was conducted on the GTCCs, which were obtained by applying a DCT to decorrelate the features and extract the cepstral coefficients.

Furthermore, standard features including the mel-spectrogram, MFCC, and LFCC were employed as baselines for comparison with auditory-inspired features extracted using gammatone and gammachirp filterbank.

5.1.4 Deep Learning Models

The deep learning models, including light convolutional neural networks (LCNN) [56, 62–65], ResNet18, and VGG16, were used in this study follow different architectural approaches for classification.

LCNN begins with a two-dimensional convolutional layer, followed by a feature-reduction mechanism to eliminate redundancies. It then progresses through multiple convolutional layers with normalization and pooling operations, gradually increasing the number of channels while down-sampling. The extracted features are reshaped and processed through bidirectional recurrent layers for temporal modeling. Finally, a pooling operation aggregates the outputs, followed by fully connected layers for binary classification. The model is trained using a standard classification-loss function and optimization algorithm.

ResNet18 starts with an initial convolutional layer, followed by normalization, activation, and pooling. It comprises four residual blocks, progressively increasing the number of feature channels while reducing spatial dimensions through down-sampling. The final layers include global pooling, flattening, and a fully connected layer for classification.

VGG16 follows a structured architecture with sequential convolutional blocks, each followed by normalization, activation, and pooling. The depth of the network increases with additional convolutional layers, capturing complex hierarchical features. The final layers include global pooling, flattening, and fully connected layers to execute classification.

Experimental Setup In experimental setup, a spoofing countermeasure system was first evaluated using the LCNN model with the ASVspoof 2019 LA dataset and the proposed human-imitated speech dataset. The objective was to analyze the model's when trained on synthetic ASVspoof 2019 LA dataset and tested on

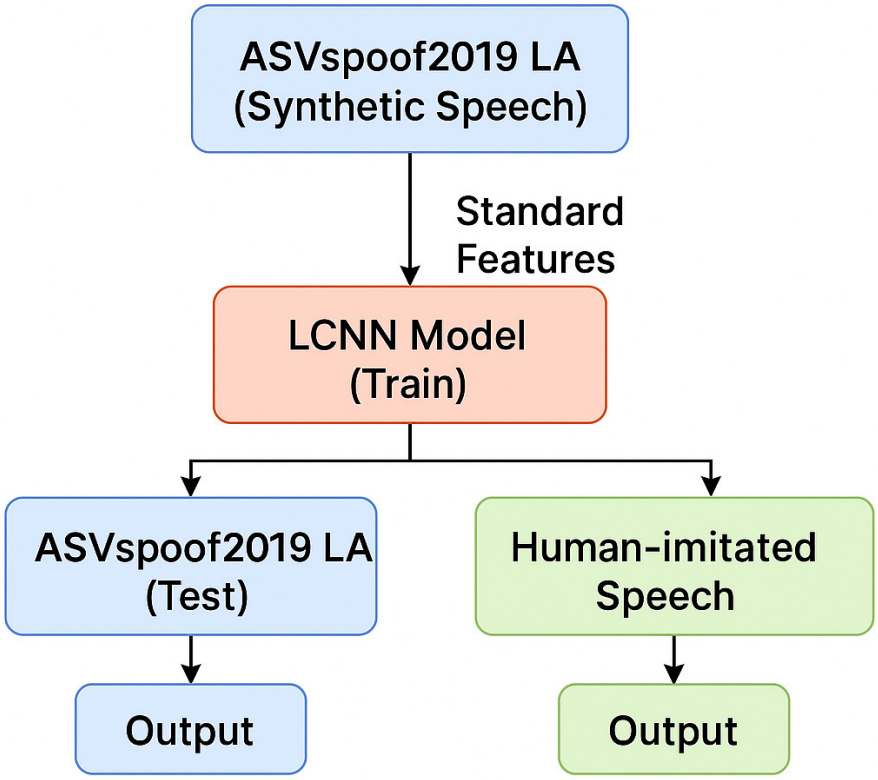


Figure 5.2: Train on ASVspoof2019 LA and test on both ASVspoof2019 LA and proposed human-imitated speech.

both synthetic ASVspoof 2019 LA and proposed human-imitated speech dataset, shown in Fig . 5.2.

Then used the same spoofing-countermeasure system using LCNN model to train and test it exclusively on proposed human-imitated speech dataset. The ResNet18 and VGG16 models were also trained and tested using the same approach.

Building upon the previous study [122], which focused on auditory perception, gammatone and gammachirp features were selected due to their close relation to human auditory processing. Using the same spoofing-countermeasure system, the effectiveness of these features in distinguishing between genuine and spoofed speech was assessed.

These features were then evaluated using various deep learning models, including ResNet18 and VGG16, to determine their impact on spoof-detection perfor-

Table 5.1: Performance of spoof-countermeasure system trained on synthetic speech and tested on synthetic and imitated speech.

Features	Classifier	Dataset	Evaluation Metrics		
			Accuracy (%)	EER (%)	F1 Score (%)
MFCC	LCNN	ASVspoof 2019 LA	93.02	7.20	93.38
		Human-Imitated	61.37	36.57	60.82
Mel-Spec	LCNN	ASVspoof 2019 LA	95.54	6.14	95.63
		Human-Imitated	61.27	35.23	57.91
LFCC	LCNN	ASVspoof 2019 LA	92.57	7.33	93.24
		Human-Imitated	61.32	39.07	61.21

mance.

This study used accuracy, equal error rate (EER), and F1 score as the metrics.

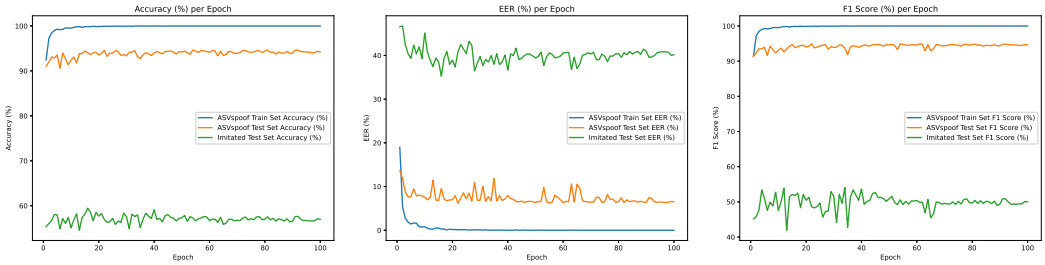
5.2 Evaluations and Results

5.2.1 Train on ASVspoof 2019 LA and Test on ASVspoof 2019 LA and Imitated Speech

Table 5.1 shows that when the spoofing-countermeasure system using LCNN model was trained on ASVspoof 2019 LA dataset and tested on both ASVspoof 2019 LA and proposed human-imitated speech dataset, it showed strong performance on the ASVspoof 2019 LA dataset as shown in Fig.5.3, Fig.5.4 and Fig.5.5 respectively. For instance, when using Mel-spectrogram features with LCNN, the model achieved an accuracy of 95.54%, EER of 6.14%, and F1 score of 95.63%. This robust performance is attributed to the distinct artifacts present in synthetic speech. However, when tested on proposed human-imitated speech dataset, the system's performance deteriorated significantly across all feature sets. With the Mel-spectrogram features, performance decreased to an accuracy of 61.27%, EER of 35.23%, and F1 score of 57.91%. These results highlight the challenge of detecting imitated speech, as it closely resembles genuine human speech and lacks the clear spoofing artifacts found in synthetic speech.

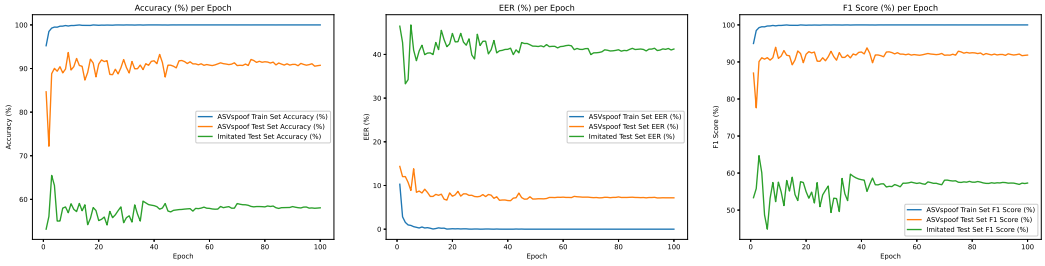
5.2.2 Train on Imitated Speech and Test on Imitated Speech using Stander Features and Deep Learning

Table 5.2 presents the results from when the same spoofing-countermeasure system using LCNN model was trained on the proposed human-imitated speech dataset using stander features such as MFCCs, mel-spectrogram, and LFCCs. LCNN showed significant improvement, achieving 77.40% accuracy, reducing EER to 22.64%, and



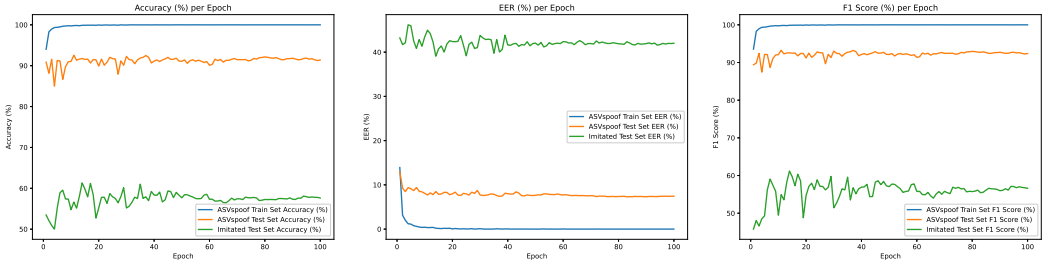
(a) Accuracy (b) EER (c) F1-score

Figure 5.3: Performance comparison when trained on the spoof dataset and evaluated on both spoof and imitated speech using mel-spectrogram with LCNN.



(a) Accuracy (b) EER (c) F1-score

Figure 5.4: Performance comparison when trained on the spoof dataset and evaluated on both spoof and imitated speech using MFCC with LCNN.



(a) Accuracy (b) EER (c) F1-score

Figure 5.5: Performance comparison when trained on the spoof dataset and evaluated on both spoof and imitated speech using LFCC with LCNN.

Table 5.2: Performance of spoof-countermeasure system and standard features for training and testing on imitated speech.

Features	Classifier	Evaluation Metrics		
		Accuracy (%)	EER (%)	F1 Score (%)
MFCC	LCNN	77.40	22.64	77.23
Mel-Spec	LCNN	76.95	23.32	76.92
LFCC	LCNN	76.30	23.79	76.07
MFCC	ResNet18	78.32	21.89	81.85
Mel-Spec	ResNet18	77.60	21.63	78.69
LFCC	ResNet18	78.11	22.17	82.17
MFCC	VGG16	**80.77**	19.87	82.94
Mel-Spec	VGG16	79.84	**19.45**	**83.31**
LFCC	VGG16	79.58	20.27	82.80

attaining a 77.23% F1 score with MFCCs. ResNet18 improves performance across all metrics, achieving an accuracy of 78.32%, an EER of 21.63%, and an F1 score of 82.17%. The VGG16 outperforms both LCNN and ResNet18, achieving an accuracy of 80.77%, an EER of 19.45%, and an F1 score of 83.31%. These results demonstrate the improved ability of these models to detect imitated speech. However, the strong resemblance to genuine speech underscores the need for advanced feature representations to enhance robustness.

5.2.3 Train on Imitated Speech and Test on Imitated Speech using Auditory-Based Features and Deep Learning

Table 5.3 presents the performance evaluation of the GTFB and GCFB, as well as the GTCC and GCCC using LCNN, ResNet18, and VGG16. The results reveal that GCFB with the VGG16 achieved the highest accuracy of 81.07% and lowest EER of 18.60%. The GCCC with ResNet18 performed the best among the cepstral features, with an accuracy of 80.53% and EER of 20.05%, highlighting its strong classification capability. The GTCC consistently exhibited lower performance across all models, with its worst case observed in VGG16 (74.90% accuracy and 25.10% EER). These findings suggest that the GCFB is generally more robust, especially when used with VGG16, while the GCCC can also be effective when paired with ResNet18.

The results highlight the effectiveness of auditory features in human-imitated-speech detection. The superior performance of the GTFB, GCFB, GTCC, and GCCC over traditional spectral features (mel-spectrogram, MFCC, LFCC) suggests that modeling the human auditory system captures more perceptually relevant cues, enhancing classification accuracy and robustness.

Table 5.3: Evaluation of GTFB, GCFB, GTCC, and GCCC using different classifiers.

GTFB & GCFB					GTCC & GCCC				
Features	**Classifier**	**Evaluation Metrics**			**Features**	**Classifier**	**Evaluation Metrics**		
		Accuracy (%)	**EER (%)**	**F1 Score (%)**			**Accuracy (%)**	**EER (%)**	**F1 Score (%)**
GTFB	LCNN	79.56	20.84	**83.86**	GTCC	LCNN	77.74	21.78	81.63
GCFB	LCNN	79.95	20.15	83.50	GCCC	LCNN	79.43	20.84	82.08
GTFB	ResNet18	79.53	20.02	82.64	GTCC	ResNet18	76.14	23.89	79.82
GCFB	ResNet18	79.88	20.51	82.93	GCCC	ResNet18	**80.53**	**20.05**	**84.05**
GTFB	VGG16	78.99	21.07	81.88	GTCC	VGG16	74.90	25.10	80.11
GCFB	VGG16	**81.07**	**18.60**	83.84	GCCC	VGG16	78.90	20.65	81.82

5.3 Summary

The challenge of human-imitated speech, an underexplored spoofing method, was addressed by introducing a novel imitation-based speech dataset. These findings reveal a key limitation in current spoofing-countermeasure systems: model trained solely on synthetic speech fail to effectively generalize to human-imitated attacks. The experimental results indicate that while models perform well on synthetic data, they struggle with imitated speech due to its closer resemblance to genuine speech. Standard features using different deep learning yielded an average accuracy of 78.31%, whereas auditory features, particularly gammatone and gammachirp, significantly improved detection. The GCFB with VGG16 achieved the highest accuracy of 81.07% and lowest EER of 18.60%.

These findings emphasize the need for broader training datasets and the integration of auditory-inspired features to enhance ASV resilience against both AI-generated and human-imitated spoofing attacks. Future research should focus on dataset expansion, the incorporation of auditory-based features, and refinement of deep learning architectures to improve detection capabilities.

Chapter 6

Human-Imitated Speech Detection by Humans and Machines

This chapter presents a study that explores human listening tests and acoustic features related to auditory perception as a framework for distinguishing between genuine and human-imitated speech. Acoustic features related to auditory perception have been extensively studied to understand their characteristics and have been applied in diverse applications to assess sound quality and discriminate sound events. Building on these insights, the present work examines the effectiveness of human listening and perceptually related acoustic features in differentiating genuine speech from human-imitated speech through a systematic listening experiment. Given the lack of suitable publicly available resources, a dedicated dataset of human-imitated speech was developed for this purpose, and a representative sample of this dataset was used in the study. The evaluation followed a three-phase, human-centered approach in which participants were tasked with classifying whether each speech sample was genuine or imitated. One of the most important aspects of this study is the speaker identification task, in which listener familiarity with the target speakers plays a critical role. To ensure that all participants had the same level of familiarity, a three-phase experimental procedure was carefully designed. The first phase involved a training stage in which participants were required to learn and identify all 10 speakers. Only those who achieved over 90% accuracy in identifying the target speakers were allowed to proceed to the final phase, which involved distinguishing between genuine and imitated speech. This requirement was essential to ensure that all listeners had sufficient and uniform familiarity with the target speakers. This design choice also addresses a common limitation in previous studies.

Although other studies [25, 123] claim that their participants were familiar

with the speakers, often because they were native speakers, there is no clear indication of the actual level of familiarity. Whether the participants had high or low familiarity remains uncertain. Without controlling this factor, the similarity with speakers may vary significantly among subjects, leading to unfair or biased evaluation results.

The proposed listening test method directly tackles this issue by ensuring high and consistent familiarity across all subjects, which is necessary for a fair speaker identification task. This realistic test setup represents a significant strength of the current study. Notably, two participants withdrew from the experiment because they were unable to meet the strict training criteria, highlighting the seriousness and rigor of the experimental control. In the human listening test, participants were first exposed exclusively to genuine speech during a training phase to become familiar with the target speakers, before being asked to evaluate both genuine and imitated speech. This setup simulates how listeners typically experience and judge speech in real-world scenarios by relying solely on human auditory perception without prior exposure to imitation attempts.

In addition, a feasibility study was conducted using two sets of features, timbral features, and STM representations for machine-based classification. Both types of features are related to auditory perception and were used to assess their potential to distinguishing genuine and imitated speech. Among the timbral features, attributes such as hardness, depth, brightness, roughness, warmth, sharpness, boominess, and reverb were analyzed. Due to the limited number of available samples, a traditional machine learning approach was adopted rather than deep learning models, which typically require larger datasets to train effectively and generalize reliably. For the machine learning experiments, a Support Vector Machine (SVM) classifier was employed to evaluate the effectiveness of the timbral features. For the STM representations, SVM, k-Nearest Neighbors (KNN), and Extra Trees (ET) classifiers were used. Additionally, general acoustic features such as mel-spectrograms, as well as auditory features like gammatone and gammachirp filterbanks, were tested to benchmark performance in distinguishing between genuine and imitated speech.

To this end, the study builds upon insights derived from the preceding human listening experiments and feasibility analyses. Importantly, whereas the human experiment relied exclusively on exposure to genuine speech, the feasibility study trained a supervised model using labeled examples of both genuine and imitated speech. This created a mismatch between human and machine learning conditions, as the model had access to imitation data that were unavailable to human listeners.

To address this inconsistency and establish a fair comparison, the study introduced one-class classification approaches that more closely mirror the human listening setup while using the same timbral features. Specifically, models were

trained only on genuine speech and evaluated on both genuine and imitated samples, mirroring the conditions faced by human listeners. In this context, one-class SVM, Local Outlier Factor, and Isolation Forest were applied, as they are designed to learn the characteristics of a single class and detect deviations as anomalies. This approach ensures that both human and machine relied exclusively on prior exposure to genuine speech when making classification decisions. Moreover, the public release of the dataset enables other researchers to collaborate, compare, and reproduce results under clearly defined conditions. This contributes to greater transparency and scientific value in the field of speaker imitation detection.

6.1 Experiment

The experiment of the listening test is described in the following subsections.

6.1.1 Stimuli and Apparatus

This experiment took place in a soundproof room, using a PC with two monitors for the participants and experimenter. The setup included the Focusrite Clarett 2Pre audio interface for high-quality, low-latency performance and Sennheiser HD 280 Pro headphones for excellent noise isolation and accurate sound. Both devices are known for their reliability in professional audio settings. The MATLAB R2023b GUI was used to provide a smooth and user-friendly interface for interacting with the participants while playing audio from the PC, as shown in Fig. 6.1. Participants were presented with various audio stimuli to assess their perception, and they received clear instructions on how to navigate the GUI prior to the experiment.

The stimuli for the experiment consisted of speech samples from 10 genuine target speakers. Each stimulus lasted approximately 1.5 minutes, resulting in a total training time of 15 minutes for participants. For the testing phase, participants were presented with 20 speech samples, each lasting 3 seconds. During the final evaluation phase, they were presented with 100 samples (50% genuine, 50% imitated), and each sample also lasted 3 seconds. These stimuli were carefully selected to assess the participants' ability to distinguish between genuine and imitated speech.

6.1.2 Participants

The participants for the experiment came from diverse backgrounds and countries, with a total of 22 individuals initially recruited. Two participants withdrew during the study because they faced difficulty passing the initial test required for the

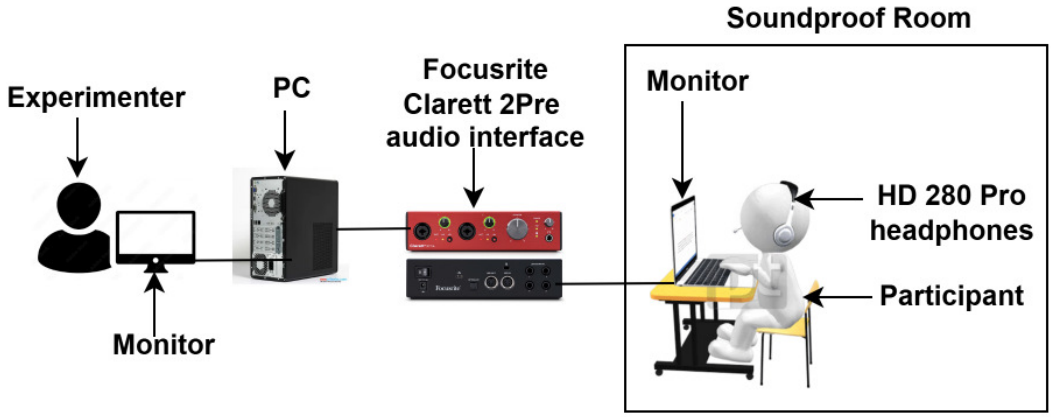


Figure 6.1: Experimental setup.

experiment's final evaluation, resulting in a final evaluation with 20 participants, including 10 males and 10 females, aged between 18 and 35 years. All participants had normal hearing abilities and signed a consent form prior to the beginning of the experiment. This met the requirement for the experiment, ensuring a diverse and representative sample for the analysis.

6.1.3 Condition

The reason for setting the condition was to ensure that suitable participants were selected for this experiment. Participants are asked to evaluate the individuality of the target speakers, which can be challenging since most participants are generally unfamiliar with recognizing speaker-specific characteristics. Therefore, to conduct an effective experiment for speaker identification, all participants had to meet a high level of satisfactory performance. The requirement for the final evaluation was that participants must achieve a score of at least 90% in the test phase, where they identify the target speakers, before proceeding to the final evaluation. This ensured that they were adequately prepared and suitable for the task.

6.1.4 Procedure

The procedure of our experiment is shown in Fig. 6.2. In the first step, a participant was trained using 10 genuine target speakers. Each speaker's speech lasted more than 1.5 minutes, with a total auditory training time of 15 minutes. The participant listened closely to the speech of each target speaker to become familiar with their unique vocal characteristics, which is crucial for this human listening

test. The goal at this stage was to train the participant auditory system to recognize and process each speaker's voice patterns. After the training, the participant was asked if they felt ready to continue or if they needed more time to familiarize themselves with the auditory stimuli. If the participant felt ready, they proceeded to the next step. However, if the participant responded "no" or was unsure, they returned to the training step for further familiarization. Should the participant feel unconfident, they had the option to withdraw from the experiment, as this test was designed to respect individual differences in auditory perception abilities.

In the second step, the participant's ability to distinguish between speech was evaluated by testing with 20 speech samples (stimuli) from the same 10 target speakers. This was a 5-minute auditory testing session, where the participant had to rely entirely on their auditory perception skills to identify the speakers. To ensure a more accurate assessment of auditory processing, the distribution of stimuli was non-uniform, meaning that some speakers provided one stimulus, while others provided two or three. This variability prevented the participants from predicting patterns, requiring them to fully engage their auditory discrimination abilities. They had to achieve a test score of 90% or higher to advance to the final step. If their auditory perception score was below 90%, they were subjected to additional auditory training or could choose to withdraw from the experiment.

In the third step, the final auditory evaluation was conducted to assess the participant's ability to differentiate between genuine and imitated speech from the 10 target speakers. The evaluation consisted of a balanced mix of 50% genuine and 50% imitated speech, and the participant had to rely solely on their auditory processing skills to determine whether the speech was genuine or imitated. This phase included 100 stimuli and lasted for 15 minutes, during which the participant listened to each sample and indicated whether it was genuine (by answering "yes") or imitated (by answering "no"). The participant's final score reflected their ability to perceive subtle differences in vocal patterns and determine the authenticity of the speech.

The experiment schedule was structured as follows. Participants first underwent 15 minutes of training, followed by a 10-minute rest period to absorb the training material. This was followed by a 10-minute testing session, allowing participants to apply what they had learned. After the testing, there was another 10-minute rest period to ensure participants were refreshed before proceeding to the final evaluation. The final evaluation itself lasted for 15 minutes. In total, the experiment was designed to be completed within one hour, assuming participants successfully progressed through all steps on their first attempt. However, the actual time taken by each participant in the experiment is shown in Fig. 6.3.

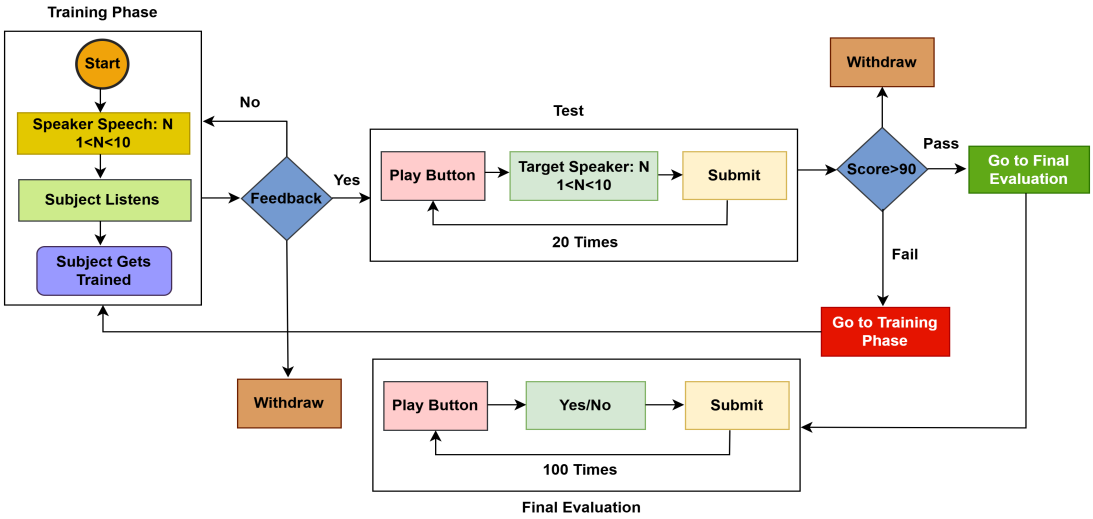


Figure 6.2: Design of subjective tests for discriminating genuine speech and imitated speech.

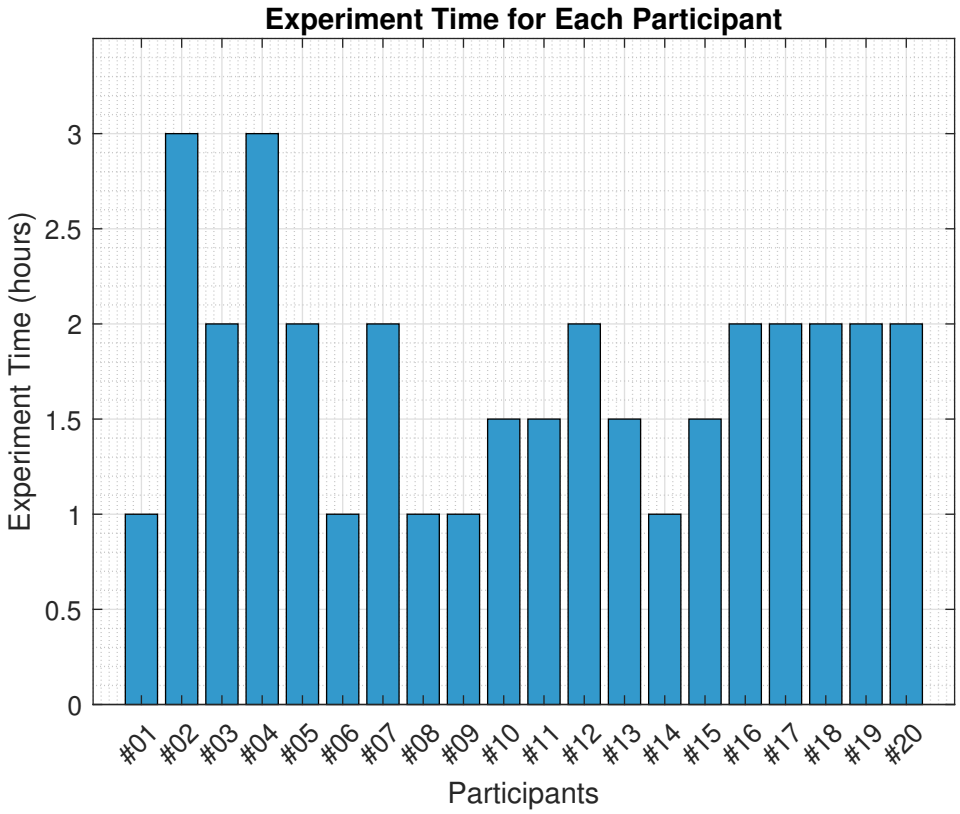


Figure 6.3: Experiment time of each participant.

6.1.5 Results

The human listenig test aimed to measure participants' ability to accurately identify target speakers following their training sessions. The results provided a comprehensive view of their performance. Individual testing was conducted for each participant, and their accuracy is presented in Fig. 6.4. The overall test accuracy for the participants was 95.75%. In addition to individual results, the average accuracy across all participants is shown, along with a representation of the normal distribution of the results. These findings suggest that the participants were well-prepared, exhibiting strong auditory perception abilities in accurately identifying the target speakers.

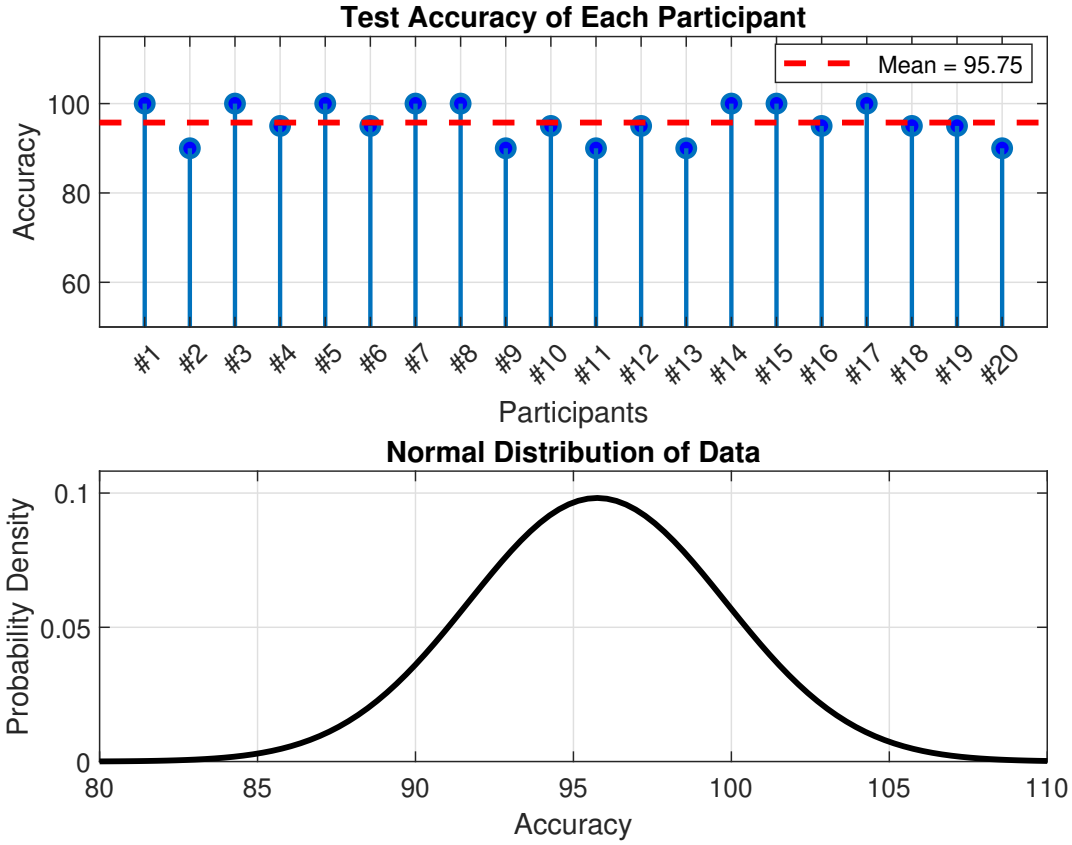


Figure 6.4: Participant test accuracy.

Evaluating human listeners showed that distinguishing between genuine and imitated speech was a challenging task for participants. Figure 6.5 shows the average confusion matrix for all participants, reflecting the overall performance in

distinguishing genuine and imitated speech during the final evaluation. In the confusion matrix, 0 represents imitated speech, and 1 represents genuine speech. From this, it was calculated that human listeners achieved an average accuracy of 70.10% in distinguishing between genuine and human-imitated speech, as illustrated in Fig. 6.6. The differences in individual performance demonstrate the varying degrees of difficulty listeners faced in differentiating between genuine and imitated speech. The average accuracy across all participants is shown, along with a representation of the normal distribution of the results. Figure 6.7 visualizes the d-prime score, which quantifies how well participants distinguish between two classes, like genuine and imitated speech, by measuring the separation between the two distributions. The d-prime score was 1.09. Additionally, the F1-score was 72.70%, and the equal error rate (ERR) was 29.90%, as summarized in Table 6.1. These findings emphasize the effectiveness of human imitation techniques, which produce a naturalness that significantly challenges even trained listeners.

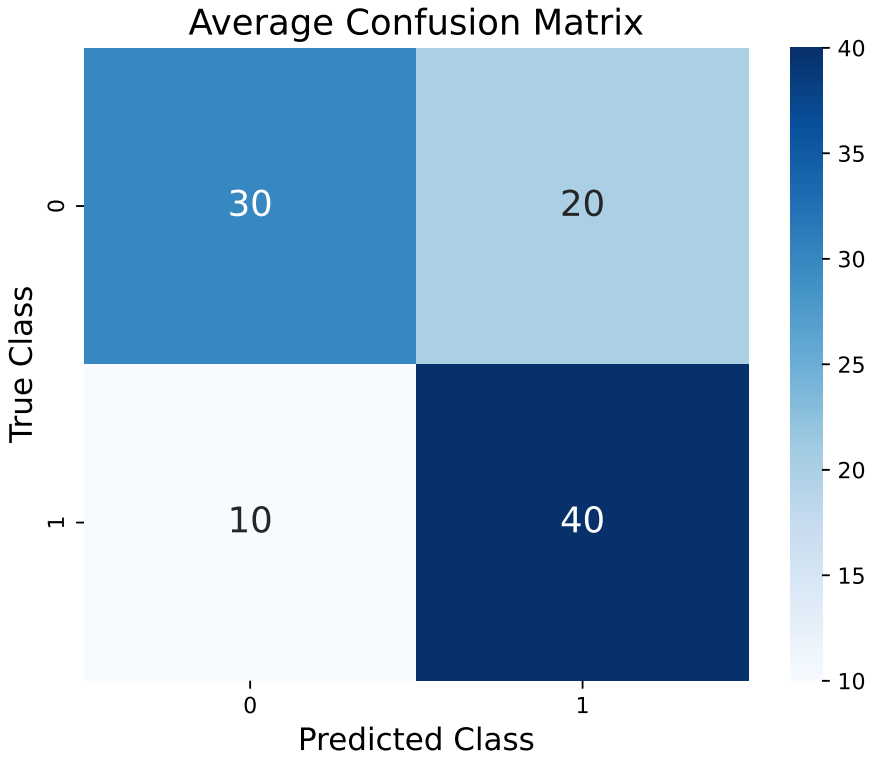


Figure 6.5: Average confusion matrix of all participants.

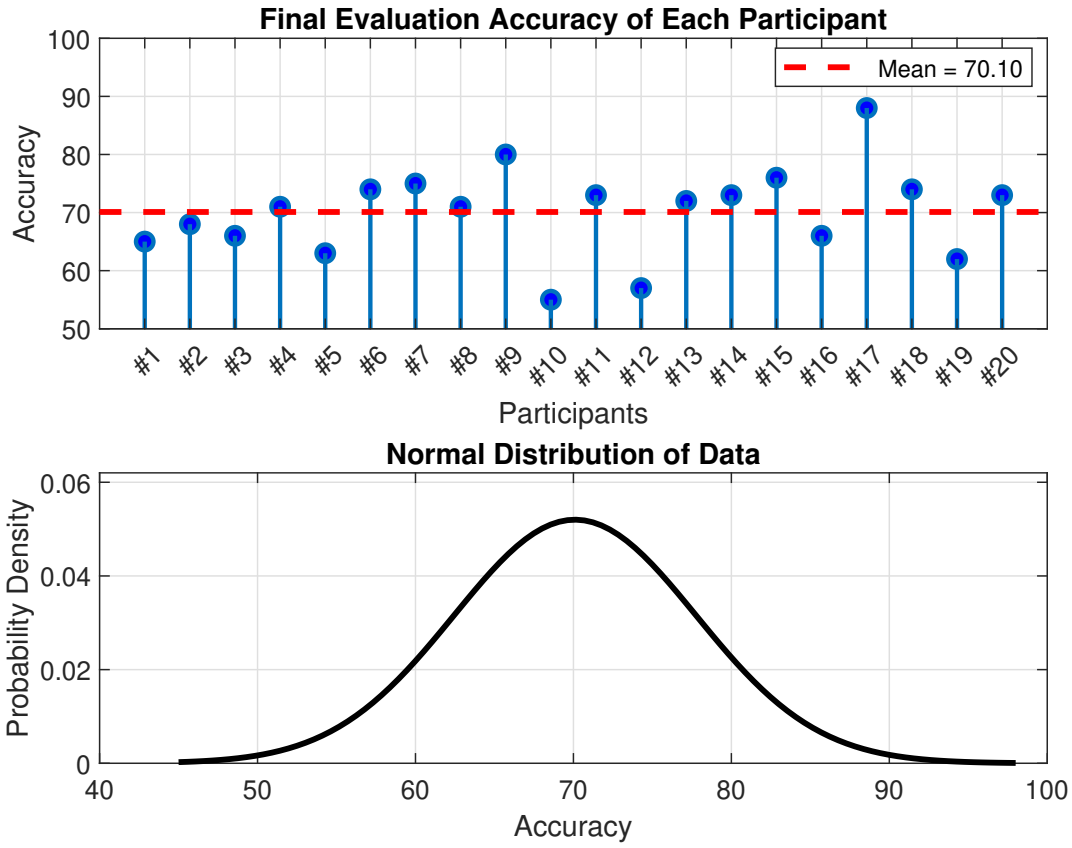


Figure 6.6: Final evaluation accuracy.

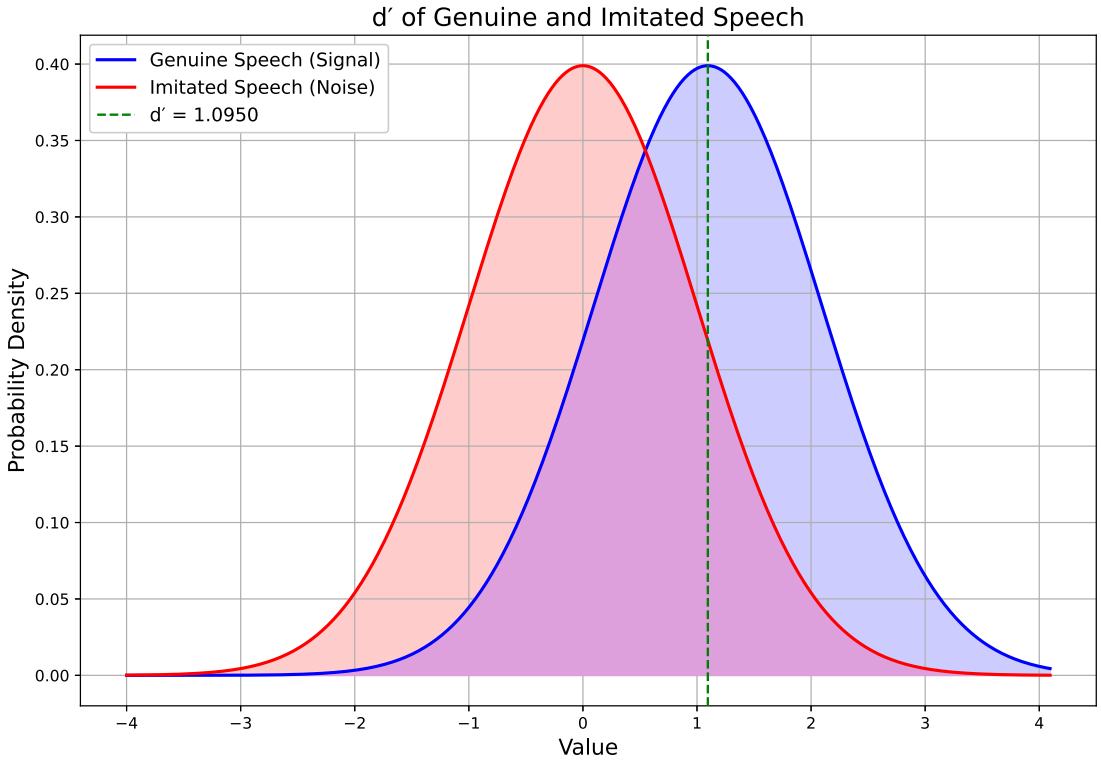


Figure 6.7: d' of genuine and imitated speech.

Table 6.1 summarizes the participant test results for different evaluation metrics.

Table 6.1: Evaluation metrics for genuine and imitated speech classification.

Evaluation Metrics	Accuracy	F1-Score	EER	d'
Overall Score	70.10	72.0	29.90	1.09

6.1.6 Discussion

In this human listening experiment, participants were tested on their ability to distinguish between genuine and imitated speech after undergoing a structured training process. The results indicated that participants were generally successful in identifying target speakers with a high accuracy rate of 95.75% during individual testing, suggesting effective training and strong auditory perception skills. However, distinguishing between genuine and imitated speech in the final evaluation proved more challenging, with an average accuracy of 70.10%. This highlights

the difficulty in perceiving subtle differences between genuine and imitated speech, especially given the naturalness of the imitated speech performed by professional artists. The confusion matrix further reflected these challenges, showing that participants struggled to achieve consistently high performance in this task. The d' score of 1.09 and EER of 29.90% emphasize the complexity of the task, as human listeners face limitations in detecting nuanced differences between genuine and imitated speech. These findings provide valuable insights into the effectiveness of human-imitation techniques, which closely mimic genuine speech and present a significant challenge even for trained listeners.

These findings underscore the importance of the proposed listening test methodology, which ensures that all participants begin with the same high level of familiarity and training. This rigorous and controlled setup helps eliminate variability in prior knowledge and provides a fair, reliable measure of human perceptual capabilities in detecting speech imitation. By controlling for familiarity and using a realistic evaluation scenario, the study offers clearer evidence of how challenging it is to distinguish high-quality imitated speech from genuine speech.

6.2 Feasibility Study for Imitated Speech Detection

This section presents a feasibility study using two sets of feature representations, timbral features, and STM representations for machine-based classification. Both types of features are related to auditory perception and were evaluated for their potential to distinguish between genuine and imitated speech.

6.2.1 Timbral Features

After conducting the human subject test to evaluate how well humans can distinguish between genuine and imitated speech, the study moved to a feasibility study to explore whether a machine-based classification approach could achieve similar results. This study aimed to determine if machines can simulate how human listeners perceive speech by focusing on timbral features, which are known to influence how humans detect speech. Previous research has shown that even when two sounds have the same pitch, loudness, and duration, they can still be perceived differently due to variations in timbral features [124]. Features such as boominess, brightness, depth, reverb, and warmth were selected for this study due to their relevance in human auditory processing. An extraction algorithm from the Audio Commons Project, widely applied in psychoacoustic studies, was used to model these timbral characteristics [125,126]. This algorithm captures key high-level timbral properties by analyzing low-level features such as the spectral

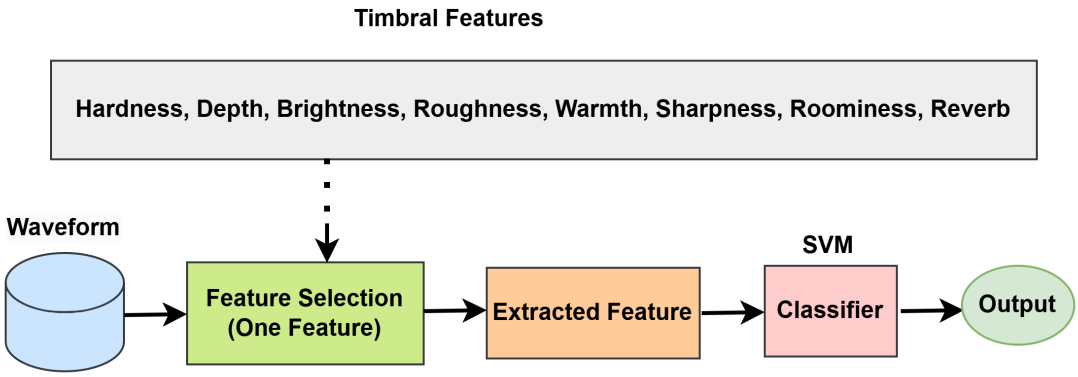


Figure 6.8: Proposed model of feasibility study.

centroid, dynamic range, and spectral energy ratio, which are then rated on a scale from 0 to 100 [81]. Using support vector machine (SVM) classifiers, the aim was to assess the feasibility of using these timbral features to improve machine performance in distinguishing between genuine and imitated speech. Each timbral feature plays a crucial role in shaping auditory perception, and the timbral qualities are categorized into eight distinct attributes, as explained below.

Model Architecture

In this study, a support vector machine (SVM) model was utilized to classify genuine and human-imitated speech. The model processed one timbral feature at a time, such as hardness, depth, brightness, roughness, warmth, sharpness, roominess, or reverb, by selecting the feature, extracting it from the waveform, and then passing it through the classifier , as illustrated in Fig. 6.8. This approach allowed the SVM model to analyze each feature individually before producing an output. SVMs perform well with smaller datasets [127], making them an ideal choice for this experiment. The model was trained on 40 samples, comprising equal numbers of genuine and imitated speech, and evaluated on the same 100 test samples used in the human listening study.

Results

The analysis of various timbral features for distinguishing genuine from imitated speech revealed differing levels of effectiveness. Boominess provided the highest accuracy at 62%, followed by depth at 60%, indicating that these features, which capture resonance and voice texture, are particularly useful for this task. Warmth and reverb showed moderate accuracy levels, at 55% and 54%, respectively, while

roughness achieved 51% accuracy. Brightness and sharpness performed slightly lower, at 48% and 49%, and hardness had the lowest accuracy, reaching only 43%, as presented in Table 6.2. This evaluation was conducted using a linear SVM without scaling or class weighting. In addition, mel-spectrogram feature obtained accuracy of 51%. These results indicate that while features like boominess and depth are highly effective, others such as hardness and brightness are less suited to this classification task.

The analysis of various timbral features for distinguishing genuine from imitated speech revealed differing levels of effectiveness. Boominess provided the highest accuracy at 62%, followed by depth at 60%, indicating that these features, which capture resonance and voice texture, are particularly useful for this task. Warmth and reverb showed moderate accuracy levels, at 55% and 54%, respectively, while roughness achieved 51% accuracy. Brightness and sharpness performed slightly lower, at 48% and 49%, and hardness had the lowest accuracy, reaching only 43%, as presented in Table 6.2. In addition, the mel-spectrogram feature obtained an accuracy of 51%. This evaluation was initially conducted using a linear SVM without scaling or class weighting. When employing a linear SVM with feature scaling and balanced class weights, the performance of boominess further improved, achieving 64% accuracy, an F1-score of 63.94%,, an EER of 36.00%, and a d' of 0.72. The confusion matrix and d-prime for the boominess feature are shown in Fig. 6.9. While the remaining features showed similar performance to the linear SVM without scaling or class weighting.

Table 6.2: Evaluation of timber features using SVM classifier.

Features	Classifier	Accuracy
Hardness	SVM	43
Depth	SVM	60
Brightness	SVM	48
Roughness	SVM	51
Warmth	SVM	55
Sharpness	SVM	49
Boominess	SVM	62
Reverb	SVM	54
All Timbers Features Combined	SVM	58

Table 6.21 summarizes the boominess and depth feature results for different evaluation metrics.

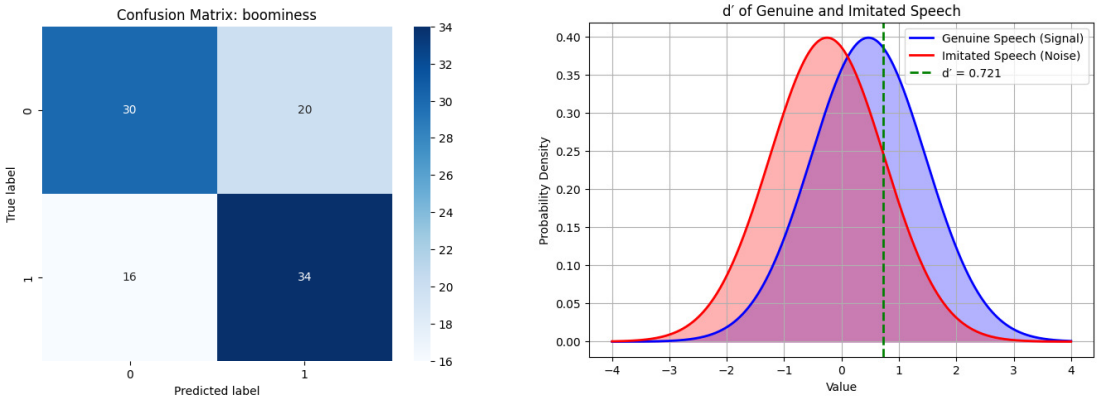


Figure 6.9: Confusion matrix and d' of the boominess feature.

Table 6.3: Evaluation metrics for genuine and imitated speech classification.

Features	Evaluation Metrics			
	Accuracy	F1-Score	EER	d'
Boominess	62.0	62.26	38.0	0.62
Depth	60.0	64.15	40.0	0.51

Discussion

Timbral features like boominess and depth significantly improved machine performance, achieving 64% and 60% accuracy, respectively, demonstrating their importance as perceptual features closely aligned with how humans naturally perceive speech. These features capture critical aspects of voice quality, such as resonance and texture, which are essential in distinguishing genuine from imitated speech. In contrast, mel-spectrogram achieved only 51% accuracy, indicating their limitations in capturing the finer details that timbral features provide. Other features, like roughness and warmth contributed moderately, with accuracies of 51% and 55%, while hardness and brightness underperformed at 43% and 48%. These results emphasize that while timbral features like boominess and depth show great promise in mimicking auditory perception, they alone are not enough to fully replicate human performance. To enhance machine learning models for detecting imitated speech, further refinement and expansion of the auditory feature set are needed, as humans likely rely on a more complex integration of cues. By incorporating a broader range of features, machine models could better approximate how human listeners perceive and judge speech. Although the machine-based approach shows potential, particularly with high-performing timbral features, further refinement

is essential to achieve results comparable to those of human listeners.

6.2.2 Spectro-temporal Modulation

In the feasibility study, STM representations were proposed after timbral features due to the close alignment with auditory perception. The objective was to explore whether machines can simulate and gain insights from human listening tests by leveraging spectro-temporal modulation patterns, which capture characteristic fluctuations of energy over time and frequency. Using machine learning classifiers, the study assessed whether these representations could improve performance in distinguishing between genuine and imitated speech. STM representations provide a detailed description of temporal and spectral dynamics essential for characterizing complex acoustic signals.

Model Architecture

This subsection describes the STM representations extraction process and the classification models used for detecting imitated speechas, shown in Fig. 6.10. STM representations was used to capture dynamic spectral and temporal patterns in speech signals in a way that aligns with human auditory perception.

In this study, STM representations were computed by applying a 64-channel gammatone filterbank (operating in *Normal* mode) that spanned frequencies from 60 Hz to 7,600 Hz. This filterbank approximates the auditory filtering characteristics of the human cochlea. Temporal envelopes were extracted for each channel by computing the squared magnitude of the analytic signal using a Hilbert transform and then resampled to 160 Hz to emphasize modulation patterns relevant to perception and reduce computational complexity. A two-dimensional FFT was applied across time and frequency channels to capture both temporal and spectral modulation energy, highlighting dynamic amplitude fluctuations and spectral changes. For a 3-second recording sampled at 16 kHz, this process produced an STM feature matrix with fixed dimensions of 64×480.

To classify the extracted features, conventional machine learning models were used, including support vector machine (SVM), k-nearest neighbors (KNN), and Extra Trees(ET) classifiers.

Results

The evaluation of the STM representations with different classifiers demonstrated their strong performance in distinguishing genuine from imitated speech. As shown in Table 6.4, the KNN classifier provided the highest accuracy at 68%, followed by the ET classifier at 66% and the SVM at 62%. The F1-scores reflected a similar

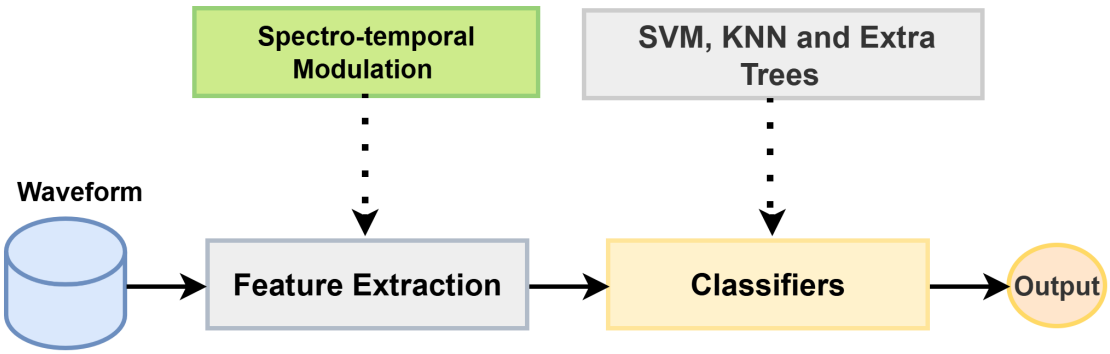


Figure 6.10: Block diagram of the proposed method.

trend, with KNN achieving 68%, indicating a reliable balance between precision and recall across classes.

Overall, these results show that STM representations, which represent how the spectral envelope fluctuates across frequency and how the amplitude envelope varies over time, can effectively capture distinctive patterns in speech that help differentiate between genuine and imitated speech. The consistently high accuracies and F1-scores across models highlight the robustness of this feature representation for this classification task.

Table 6.4: Evaluation metrics for genuine and imitated speech classification.

Features	**Model**	**Evaluation Metrics**			
		Accuracy(%)	**F1-Score(%)**	**EER(%)**	d'
STM	SVM	62.0	60.73	38.0	0.69
STM	KNN	68.0	68.00	32.0	0.93
STM	ET	66.0	65.78	34.0	0.84

Discussion

The evaluation results demonstrate the effectiveness of the proposed STM representations for classifying genuine and imitated speech with three machine learning models: SVM, KNN, and ET.

Overall, the results highlight the effectiveness of STM representations for this challenging detection task. Among the classifiers, KNN achieved the highest accuracy (68%) and F1-score (68%), followed by Extra Trees (66%) and SVM (62%). Notably, the equal error rates remained within a comparable range of 32–38% across different classifiers, highlighting the consistent discriminative effectiveness of the STM representations. The observed results can be attributed to the strong correspondence between STM representations and the processing mechanisms of

the human auditory system. STM captures both spectral and temporal modulation patterns in speech, which are essential cues that listeners use to differentiate genuine from imitated speech. This alignment with perceptually salient characteristics enables the classifiers to achieve performance levels that closely approach the accuracy reported in human listening tests.

In particular, the strong performance of the KNN classifier further suggests that the STM representations space preserves meaningful locality relationships between genuine and imitated speech samples, making simple distance-based methods effective. The overall findings underscore the value of using perceptually motivated acoustic representations, such as STM, as a robust foundation for spoofing detection systems targeting human-imitated speech.

6.3 Discussion

In this human listening experiment, participants first underwent structured training, achieving high individual identification accuracy (95.75%). However, in the final evaluation, distinguishing genuine from imitated speech proved more difficult, with average accuracy dropping to 70.10%. The confusion matrix, $d' = 1.09$, and EER of 29.90% further illustrate the challenge, reflecting the subtlety and naturalness of professional speech imitation. These results highlight the importance of the rigorous training protocol, which ensured consistent familiarity and provided a reliable assessment of listening abilities. Timbral features like boominess and depth notably improved machine classification accuracy using an SVM classifier (64% and 60%), outperforming mel-spectrograms (51%). Other features contributed moderately, while hardness and brightness lagged behind. While these perceptually aligned features capture important voice qualities, they alone were insufficient to match human performance, underscoring the need to expand and refine feature sets to better approximate human auditory judgments. Regarding STM representations, the results demonstrated their strong effectiveness for detecting imitated speech, reflecting STM's alignment with how the human auditory system encodes spectral and temporal cues. Across different classifiers applied to STM representations, accuracy remained consistently high, with KNN achieving the best performance (68% accuracy), followed by ET (66%) and SVM (62%), and all models showing comparable EERs (32–38%). The strong performance of KNN and ET further suggests that STM representations preserve meaningful locality relationships between samples, enabling effective distance-based classification. Overall, these findings highlight the value of acoustical features related to auditory perception as a robust foundation for imitated speech detection.

6.4 Human-imitated Speech Detection by Machine to Mirror Human Listening Evaluation

Concept of Detection Method

This work builds on insights from the earlier human listening and feasibility studies. In the human listening test, participants were exposed exclusively to genuine speech during a training phase to become familiar with the target speakers before evaluating both genuine and imitated speech. This setup simulates how listeners typically perceive and judge speech in real-world scenarios, relying solely on auditory perception without prior exposure to imitated speech.

In contrast, the feasibility study used the same test data and perceptually relevant timbral features but trained a supervised model with labeled examples of both genuine and imitated speech. This introduced a mismatch, as the machine had direct access to imitation data unavailable to human listeners.

To address this inconsistency and enable a fair comparison, the study adopted classification approaches that better reflect the human evaluation setup while using the same timbral features. Specifically, models were trained only on genuine speech and evaluated on both genuine and imitated samples, mirroring the human experiment. This design ensures that both human and machine relied exclusively on exposure to genuine speech when making decisions.

In this context, models that learn the characteristics of a single class and detect deviations as anomalies are particularly valuable. Unlike traditional binary classifiers, these approaches reflect a more realistic scenario where imitation data are not available during training.

Therefore, three such methods were implemented: One-Class Support Vector Machine (SVM), Local Outlier Factor (LOF), and Isolation Forest. For reference, the same models were also evaluated using standard acoustic features, including mel-spectrogram and MFCCs, to assess the relative effectiveness of timbral and conventional features under consistent conditions. The following sections describe each approach in detail along with their implementation and results.

6.4.1 One-Class Support Vector Machine (SVM)

Model Architecture

As the first model in this one-class framework, the One-Class Support Vector Machine (SVM) [128] was employed. The model was trained exclusively on genuine speech, as shown in Fig. 6.11, which also illustrates the model architecture. It utilizes the same timbral features related to auditory perception that were used in

the feasibility study. During evaluation, the model is presented with both genuine and imitated speech, replicating the decision context faced by human listeners.

The SVM constructs a decision boundary that captures the distribution of genuine speech in a transformed feature space. Any test sample that deviates significantly from this learned distribution is classified as an anomaly, indicating a potential imitation. This approach enables the model to detect deviations from the known class without requiring prior exposure to imitated speech. By restricting training to genuine speech only, the SVM offers a margin-based anomaly detection perspective that complements the alternative principles explored in the other models presented in this study.

Results

The results of the SVM model with individual timbral features are presented in Table 6.5. In this setup, the effectiveness of each feature in distinguishing genuine from imitated speech was assessed independently. Among the features tested, roughness achieved the highest classification accuracy of 61%, followed by sharpness with 60% and warmth with 57%. These findings indicate that features related to auditory perception are more effective in capturing the differences between genuine and imitated speech. This underscores the importance of examining each feature individually to assess its reliability for detecting imitation under one-class learning conditions.

Table 6.5: Evaluation of timber features using SVM model.

Features	Model	Accuracy
Hardness	SVM	55
Depth	SVM	56
Brightness	SVM	55
Roughness	SVM	61
Warmth	SVM	57
Sharpness	SVM	60
Boominess	SVM	54
Reverb	SVM	50

6.4.2 Local Outlier Factor (LOF)

Model Architecture

The second model in this framework was implemented using the Local Outlier Factor (LOF) algorithm, proposed by Breunig et al. [129]. LOF identifies outliers based on local density deviations relative to the training distribution. Operating

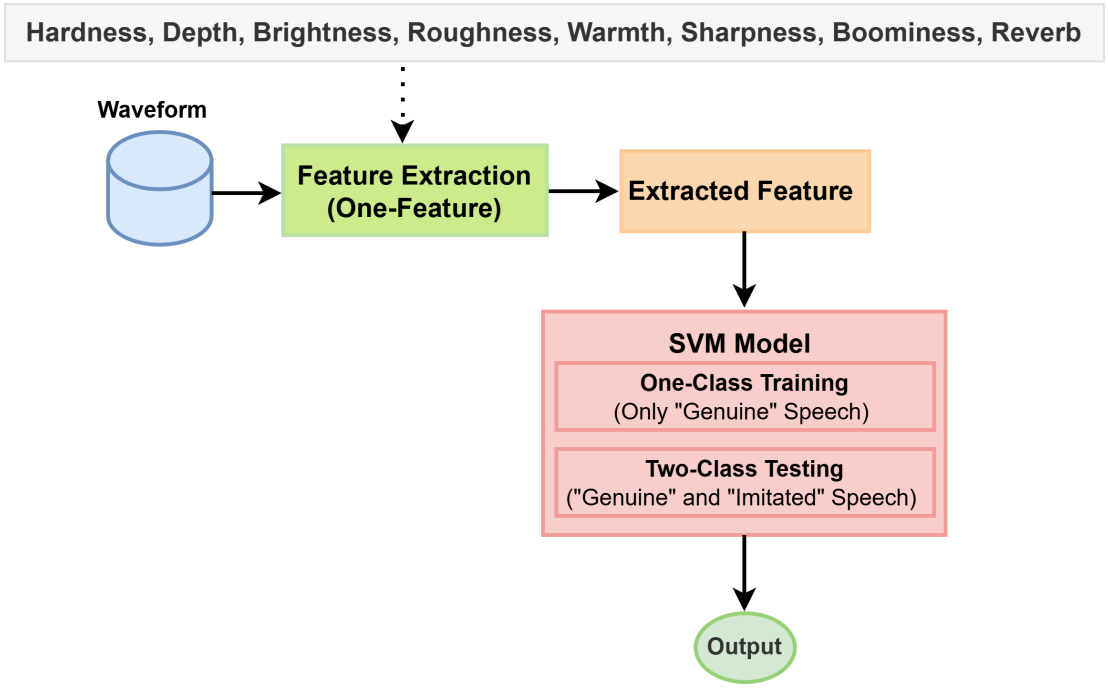


Figure 6.11: Proposed one-class SVM model.

under a one-class learning paradigm, the model was trained exclusively on genuine speech, aiming to detect anomalies that deviate from this known class. The overall model a chitecture framework is illustrated in Fig. 6.12, which mirrors the human training condition, where only genuine speech was encountered before making the final evaluation to distinguish genuine from imitated speech.

Results

The results of the LOF model using individual timbral features are presented in Table 6.6. In this setup, the effectiveness of each feature in distinguishing genuine from imitated speech was assessed independently. Among the features tested, warmth achieved the highest classification accuracy of 63%, followed by boominess at 61% and sharpness at 60%. These findings highlight that certain acoustic features related to auditory perception are more salient and reliable for detecting imitation under one-class constraints. Similar to the SVM-based approach, this method provides an interpretable and perceptually grounded strategy for evaluating speech authenticity when the training data includes only genuine speech.

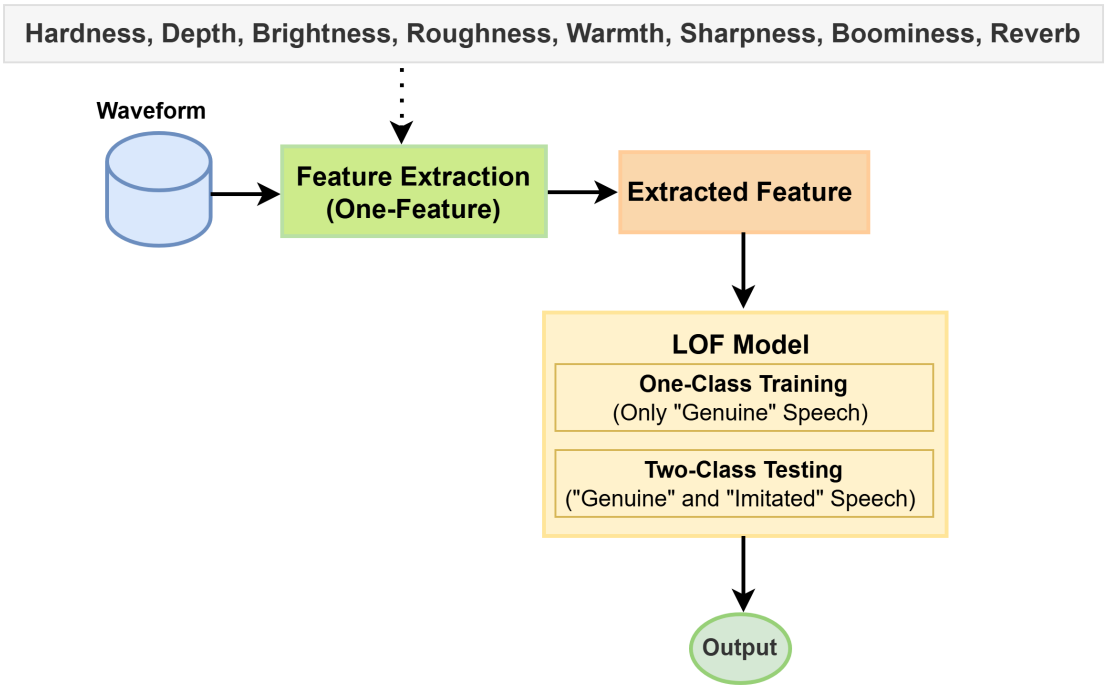


Figure 6.12: Proposed LOF model.

Table 6.6: Evaluation of timber features using LOF model.

Features	Model	Accuracy
Hardness	LOF	60
Depth	LOF	53
Brightness	LOF	57
Roughness	LOF	53
Warmth	LOF	63
Sharpness	LOF	60
Boominess	LOF	61
Reverb	LOF	50

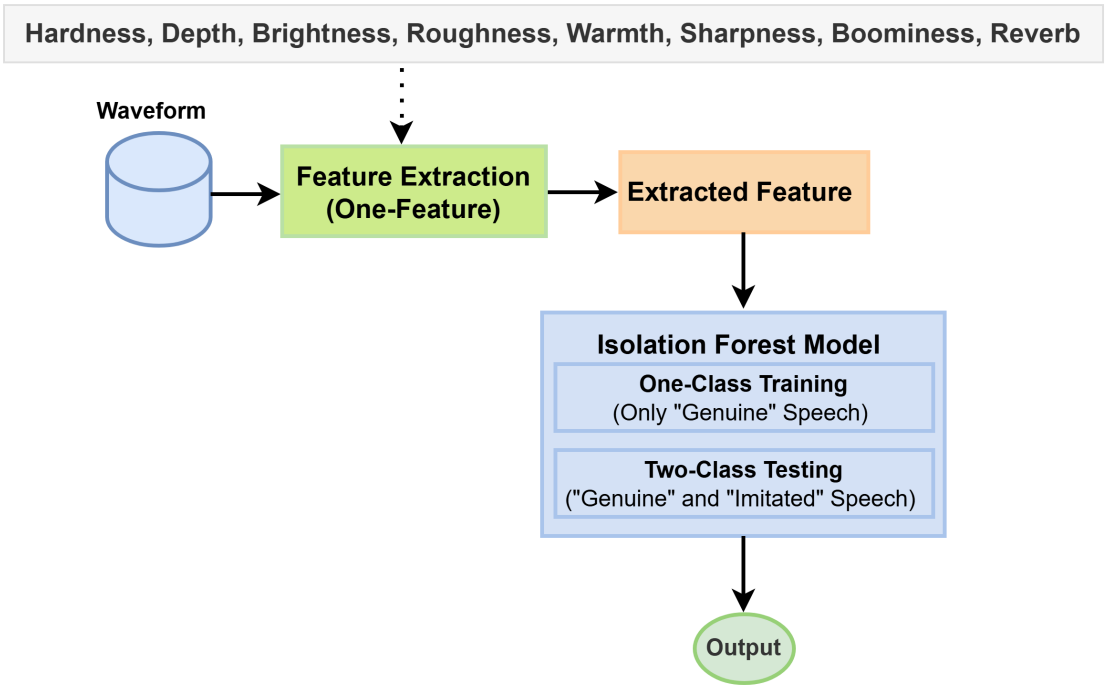


Figure 6.13: Proposed Isolation Forest model.

Table 6.7: Evaluation of timber features using Isolation Forest model.

Features	Model	Accuracy
Hardness	Isolation Forest	59
Depth	Isolation Forest	50
Brightness	Isolation Forest	54
Roughness	Isolation Forest	52
Warmth	Isolation Forest	64
Sharpness	Isolation Forest	59
Boominess	Isolation Forest	60
Reverb	Isolation Forest	50

6.4.3 Isolation Forest Approach

Model Architecture

To complement the LOF-based one-class modeling approach, the Isolation Forest algorithm was also employed for anomaly detection, the model architecture as shown in Fig. 6.13. Unlike density-based methods such as LOF, which rely on local neighborhood comparisons, Isolation Forest adopts a fundamentally differ-

ent strategy by explicitly isolating anomalies. It operates on the principle that anomalies are few and different, and therefore more susceptible to isolation via random partitioning. The algorithm constructs an ensemble of binary trees (isolation trees) and measures how quickly a sample can be separated from the rest of the data shorter average path lengths indicate higher likelihoods of being anomalies. Introduced by Liu et al. [130], this method is computationally efficient and particularly well-suited for high-dimensional data without requiring distance or density calculations.

Results

Using the same setup and timbral features as in the LOF experiments, the performance of Isolation Forest was evaluated to further validate the robustness of perceptually aligned anomaly detection. While using this Isolation Forest-based setup with timbral features, the discriminative power of individual acoustic attributes for imitated speech detection was assessed. As summarized in Table 6.7, the feature warmth yielded the highest classification accuracy at 64%, followed closely by boominess at 60% and sharpness at 59%. These findings indicate that certain acoustic features related to auditory perception, particularly those associated with perceptual richness, offer more robust cues for distinguishing genuine from imitated speech within an unsupervised anomaly detection framework.

6.4.4 Discussion

The classification results across the three models, One-Class SVM, LOF, and Isolation Forest demonstrate the potential of timbral features for distinguishing between genuine and imitated speech. As shown in Tables 6.5, 6.6, and 6.7, the feature warmth consistently achieved the highest accuracy across all models, reaching 57% with One-Class SVM, 63% with LOF, and 64% with Isolation Forest. The other evaluation metrics are also presented in Table 6.8. This finding highlights warmth as an effective acoustic feature related to auditory perception for distinguishing genuine from imitated speech. Similarly, roughness and sharpness showed relatively strong performance across models, indicating their perceptual salience in capturing differences between genuine and imitated speech.

In addition, when comparing the timbral features (Table 6.10) to the standard acoustic representations such as mel-spectrogram and MFCC features, the results further emphasize the superiority of timbral features. For example, the highest accuracy achieved with mel-spectrogram and MFCC features across all models was only 55% and 51%, respectively, which is substantially lower than the 64% obtained using the timbral feature warmth. Moreover, the d-prime values for mel-spectrogram and MFCC were close to zero or negative, indicating poor

Table 6.8: Evaluation metrics for genuine and imitated speech classification using SVM, LOF, and Isolation Forest with timbral features.

SVM Model						LOF Model						Isolation Forest Model					
Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'	Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'	Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'
Hardness	SVM	55.00	48.72	45.00	0.43	Hardness	LOF	60.00	54.04	40.00	1.04	Hardness	Isolation Forest	59.00	52.50	41.00	0.97
Depth	SVM	56.00	50.98	44.00	0.46	Depth	LOF	53.00	52.19	47.00	0.15	Depth	Isolation Forest	50.00	49.68	50.00	0.00
Brightness	SVM	55.00	54.78	45.00	0.25	Brightness	LOF	57.00	53.05	43.00	0.49	Brightness	Isolation Forest	54.00	52.08	46.00	0.23
Roughness	SVM	61.00	56.85	39.00	0.88	Roughness	LOF	53.00	48.68	47.00	0.20	Roughness	Isolation Forest	52.00	47.92	48.00	0.13
Warmth	SVM	57.00	54.13	43.00	0.44	Warmth	LOF	63.00	60.93	37.00	0.82	Warmth	Isolation Forest	64.00	62.50	36.00	0.84
Sharpness	SVM	60.00	57.10	40.00	0.66	Sharpness	LOF	60.00	57.56	40.00	0.63	Sharpness	Isolation Forest	59.00	56.27	41.00	0.58
Boominess	SVM	54.00	52.08	46.00	0.23	Boominess	LOF	61.00	60.52	39.00	0.58	Boominess	Isolation Forest	60.00	59.94	40.00	0.50
Reverb	SVM	50.00	33.33	50.00	0.00	Reverb	LOF	50.00	33.33	50.00	0.00	Reverb	Isolation Forest	50.00	33.33	50.00	0.00

Table 6.9: Evaluation metrics for genuine and imitated speech classification using SVM, LOF, and Isolation Forest with mel-spectrogram and MFCC features.

SVM Model						LOF Model						Isolation Forest Model					
Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'	Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'	Feature	Model	Accuracy (%)	F1 Score (%)	EER (%)	d'
Mel-spectrogram	SVM	48.00	47.92	52.00	-0.10	Mel-spectrogram	LOF	44.00	41.67	56.00	-0.34	Mel-spectrogram	Isolation Forest	55.00	53.96	45.00	0.27
MFCC	SVM	49.00	48.87	51.00	-0.05	MFCC	LOF	48.00	46.26	52.00	-0.11	MFCC	Isolation Forest	51.00	50.16	49.00	0.05

Table 6.10: Performance comparison of human with machine.

Approach	Feature	Accuracy (%)	F1 Score (%)	EER (%)	d'
SVM					
	Mel-spectrogram	48.00	47.92	52.00	-0.10
	MFCC	49.00	48.87	51.00	-0.05
	Roughness	61.00	56.85	39.00	0.88
LOF					
	Mel-spectrogram	44.00	41.67	56.00	-0.34
	MFCC	48.00	46.26	52.00	-0.11
	Warmth	63.00	60.93	37.00	0.82
Isolation Forest					
	Mel-spectrogram	55.00	53.96	45.00	0.27
	MFCC	51.00	50.16	49.00	0.05
	Warmth	64.00	62.50	36.00	0.84
Human Listeners					
	Human Overall Score	71.10	72.00	29.90	1.09

discriminability between genuine and imitated speech. This contrast underscores that timbral features capture perceptual characteristics more effectively than conventional acoustic features, making them better suited for this classification task.

All three models in this analysis were trained using only genuine speech, which mirrors the conditions of the human listening experiment where participants were exposed exclusively to genuine samples before evaluating to classify both genuine and imitated speech. This design ensures a scientifically fair comparison between human and machine judgments. Notably, the best-performing one-class models achieved accuracies (up to 64%) that approach the human benchmark of 70.10%, despite being trained with less information. These findings underscore the value of perceptually motivated timbral features and validate one-class classification as an effective and human-aligned approach for the classification of genuine and imitated speech.

Table 6.11: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 1.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	67.00	Hardness	LOF	67.00	Hardness	Isolation Forest	67.00
Depth	SVM	83.00	Depth	LOF	100.00	Depth	Isolation Forest	100.00
Brightness	SVM	100.00	Brightness	LOF	67.00	Brightness	Isolation Forest	67.00
Roughness	SVM	83.00	Roughness	LOF	67.00	Roughness	Isolation Forest	67.00
Warmth	SVM	83.00	Warmth	LOF	83.00	Warmth	Isolation Forest	83.00
Sharpness	SVM	67.00	Sharpness	LOF	67.00	Sharpness	Isolation Forest	67.00
Boominess	SVM	67.00	Boominess	LOF	83.00	Boominess	Isolation Forest	67.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

6.5 Speaker-Specific Discriminator Modeling

Concept of Detection Method

In order to more closely reflect the mechanism observed in human listeners, where each participant internally forms a speaker-specific discriminator, experiments were also conducted using an individual discriminator setup. In this approach, a separate model was trained for each target speaker using the same timbral features. This allowed for the evaluation of genuine and imitated speech in a speaker-dependent manner. These additional experiments offered complementary insights beyond the generic modeling approach and enabled a closer examination of how classification performance varies from one speaker to another.

6.5.1 Results

The speaker-specific discriminator employed the same three models, SVM, LOF, and Isolation Forest trained and evaluated using the same set of timbral features described earlier. The results for each individual speaker, comparing the performance of all models across different features, are presented in Tables 6.11 to 6.20, and show that performance varied across speakers. Similar to the way each participant internally forms a speaker-specific discriminator during perceptual evaluation, these experiments were conducted using an individual discriminator setup for each speaker.

Table 6.12: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 2.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	50.00	Hardness	LOF	60.00	Hardness	Isolation Forest	70.00
Depth	SVM	50.00	Depth	LOF	50.00	Depth	Isolation Forest	50.00
Brightness	SVM	50.00	Brightness	LOF	60.00	Brightness	Isolation Forest	60.00
Roughness	SVM	70.00	Roughness	LOF	80.00	Roughness	Isolation Forest	80.00
Warmth	SVM	50.00	Warmth	LOF	40.00	Warmth	Isolation Forest	40.00
Sharpness	SVM	70.00	Sharpness	LOF	80.00	Sharpness	Isolation Forest	80.00
Boominess	SVM	20.00	Boominess	LOF	40.00	Boominess	Isolation Forest	40.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.13: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 3.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	50.00	Hardness	LOF	0.00	Hardness	Isolation Forest	0.00
Depth	SVM	50.00	Depth	LOF	50.00	Depth	Isolation Forest	50.00
Brightness	SVM	100.00	Brightness	LOF	50.00	Brightness	Isolation Forest	50.00
Roughness	SVM	100.00	Roughness	LOF	50.00	Roughness	Isolation Forest	50.00
Warmth	SVM	100.00	Warmth	LOF	100.00	Warmth	Isolation Forest	100.00
Sharpness	SVM	0.00	Sharpness	LOF	50.00	Sharpness	Isolation Forest	50.00
Boominess	SVM	50.00	Boominess	LOF	50.00	Boominess	Isolation Forest	50.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.14: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 4.

One-Class SVM			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	0.00	Hardness	LOF	0.00	Hardness	Isolation Forest	0.00
Depth	SVM	50.00	Depth	LOF	0.00	Depth	Isolation Forest	0.00
Brightness	SVM	50.00	Brightness	LOF	50.00	Brightness	Isolation Forest	50.00
Roughness	SVM	50.00	Roughness	LOF	0.00	Roughness	Isolation Forest	0.00
Warmth	SVM	50.00	Warmth	LOF	50.00	Warmth	Isolation Forest	50.00
Sharpness	SVM	50.00	Sharpness	LOF	50.00	Sharpness	Isolation Forest	50.00
Boominess	SVM	50.00	Boominess	LOF	0.00	Boominess	Isolation Forest	0.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.15: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 5.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	0.00	Hardness	LOF	12.50	Hardness	Isolation Forest	12.50
Depth	SVM	50.00	Depth	LOF	87.50	Depth	Isolation Forest	100.00
Brightness	SVM	50.00	Brightness	LOF	25.00	Brightness	Isolation Forest	37.50
Roughness	SVM	37.50	Roughness	LOF	37.50	Roughness	Isolation Forest	25.00
Warmth	SVM	50.00	Warmth	LOF	50.00	Warmth	Isolation Forest	37.50
Sharpness	SVM	25.00	Sharpness	LOF	12.50	Sharpness	Isolation Forest	12.50
Boominess	SVM	62.50	Boominess	LOF	62.50	Boominess	Isolation Forest	62.50
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.16: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 6.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	100.00	Hardness	LOF	100.00	Hardness	Isolation Forest	100.00
Depth	SVM	50.00	Depth	LOF	50.00	Depth	Isolation Forest	50.00
Brightness	SVM	50.00	Brightness	LOF	100.00	Brightness	Isolation Forest	100.00
Roughness	SVM	100.00	Roughness	LOF	100.00	Roughness	Isolation Forest	100.00
Warmth	SVM	50.00	Warmth	LOF	50.00	Warmth	Isolation Forest	50.00
Sharpness	SVM	100.00	Sharpness	LOF	100.00	Sharpness	Isolation Forest	100.00
Boominess	SVM	50.00	Boominess	LOF	50.00	Boominess	Isolation Forest	50.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.17: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 7.

SVM Model			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	64.29	Hardness	LOF	78.57	Hardness	Isolation Forest	85.71
Depth	SVM	42.86	Depth	LOF	42.86	Depth	Isolation Forest	57.14
Brightness	SVM	42.86	Brightness	LOF	92.86	Brightness	Isolation Forest	92.86
Roughness	SVM	28.57	Roughness	LOF	57.14	Roughness	Isolation Forest	71.43
Warmth	SVM	85.71	Warmth	LOF	78.57	Warmth	Isolation Forest	78.57
Sharpness	SVM	92.86	Sharpness	LOF	85.71	Sharpness	Isolation Forest	92.86
Boominess	SVM	64.29	Boominess	LOF	50.00	Boominess	Isolation Forest	50.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.18: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 8.

One-Class SVM			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	61.11	Hardness	LOF	44.44	Hardness	Isolation Forest	38.89
Depth	SVM	66.67	Depth	LOF	72.22	Depth	Isolation Forest	66.67
Brightness	SVM	66.67	Brightness	LOF	66.67	Brightness	Isolation Forest	72.22
Roughness	SVM	44.44	Roughness	LOF	44.44	Roughness	Isolation Forest	44.44
Warmth	SVM	55.56	Warmth	LOF	77.78	Warmth	Isolation Forest	77.78
Sharpness	SVM	55.56	Sharpness	LOF	55.56	Sharpness	Isolation Forest	55.56
Boominess	SVM	61.11	Boominess	LOF	83.33	Boominess	Isolation Forest	77.78
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.19: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 9.

One-Class SVM			LOF Model			Isolation Forest Model		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	33.33	Hardness	LOF	38.89	Hardness	Isolation Forest	36.11
Depth	SVM	38.89	Depth	LOF	36.11	Depth	Isolation Forest	41.67
Brightness	SVM	38.89	Brightness	LOF	36.11	Brightness	Isolation Forest	36.11
Roughness	SVM	41.67	Roughness	LOF	44.44	Roughness	Isolation Forest	44.44
Warmth	SVM	52.78	Warmth	LOF	44.44	Warmth	Isolation Forest	47.22
Sharpness	SVM	33.33	Sharpness	LOF	44.44	Sharpness	Isolation Forest	38.89
Boominess	SVM	33.33	Boominess	LOF	33.33	Boominess	Isolation Forest	44.44
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

Table 6.20: Evaluation of SVM, LOF, and Isolation Forest using timbral features for speaker 10.

One-Class SVM			**LOF Model**			**Isolation Forest Model**		
Feature	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**	**Feature**	**Model**	**Accuracy (%)**
Hardness	SVM	50.00	Hardness	LOF	0.00	Hardness	Isolation Forest	0.00
Depth	SVM	100.00	Depth	LOF	100.00	Depth	Isolation Forest	100.00
Brightness	SVM	50.00	Brightness	LOF	100.00	Brightness	Isolation Forest	0.00
Roughness	SVM	0.00	Roughness	LOF	0.00	Roughness	Isolation Forest	0.00
Warmth	SVM	0.00	Warmth	LOF	0.00	Warmth	Isolation Forest	50.00
Sharpness	SVM	50.00	Sharpness	LOF	50.00	Sharpness	Isolation Forest	50.00
Boominess	SVM	50.00	Boominess	LOF	50.00	Boominess	Isolation Forest	50.00
Reverb	SVM	50.00	Reverb	LOF	50.00	Reverb	Isolation Forest	50.00

6.5.2 Discussion

In order to more closely reflect the mechanism observed in human listeners, where each participant internally forms a speaker-specific discriminator, experiments were also conducted using an individual discriminator setup. In this approach, a separate model was trained for each target speaker using the same timbral features. This allowed for the evaluation of genuine and imitated speech in a speaker-dependent manner, mirroring the way listeners build familiarity with particular speaking characteristics before making judgments. By isolating each speaker as a unique classification problem, this method accounted for inter-speaker variability and the distinct acoustic characteristics that can influence detection performance. These additional experiments offered complementary insights beyond the generic modeling approach and enabled a closer examination of how classification performance varies from one speaker to another, highlighting which features and models were most effective for specific speaking characteristics. The results demonstrated that the discriminative power of timbral features can differ significantly depending on the speaker's speech characteristics, emphasizing the importance of incorporating speaker-dependent strategies in practical detection systems. These findings highlight that the classification of imitated speech is dependent not only on the distinctiveness of the target speaker but also on the proficiency of the imitator in reproducing key acoustic cues. This speaker-dependent variability aligns with the perceptual experiences of human listeners, who naturally adjust their judgments based on familiarity with a speaker's unique vocal traits and the perceived authenticity of the imitation. It underscores the importance of individualized classification frameworks in studying genuine versus imitated speech, as such approaches can capture subtle variations that generic models may overlook. Incorporating speaker-specific discriminators can thus provide more realistic and interpretable insights into imitation detection systems by revealing which speakers are easier or harder to imitate and detect. Furthermore, these additional experiments complemented the generic modeling approach and clearly demonstrated that detection accuracy varies across speakers. They also helped uncover why some speakers are easier to imitate and harder to detect, providing a better understanding of the factors that affect model performance.

6.6 General Discussion

The human listening experiment demonstrated that some participants took significantly longer to finish due to challenges in meeting the required performance threshold of at least 90% accuracy during their training sessions. This extended duration, particularly for participants like 2 and 4, who required 3 hours, was

largely because they struggled to meet the required score, preventing them from finishing within the scheduled one-hour time. Others, such as participants 3, 5, and 7, took approximately 2 hours due to similar difficulties. In contrast, those who completed the experiment within the allotted time had successfully reached the required score and proceeded more efficiently. Despite these variations, distinguishing between genuine and imitated speech remained difficult for many participants. While the overall accuracy in identifying target speakers was high at 95.75%, the specific task of differentiating genuine from imitated speech yielded a lower average accuracy of 70.10%. These results, alongside a d-prime score of 1.09 and EER of 29.90%, underscore the inherent complexity of human listeners when faced with subtle differences between genuine and imitated speech.

Building on these findings, the study explored machine-based classification to simulate human judgments by using acoustic features related to auditory perception, such as timbral features that reflect how listeners perceive speech. Timbral attributes were chosen because they align with the perceptual cues humans use to differentiate between genuine and imitated speech. When applied to SVM classifiers, boominess produced an accuracy of 64%, while depth yielded 60%, demonstrating that these features contribute meaningfully to detecting imitated speech, although their performance remains moderate compared to human listeners.

In addition to timbral features, STM representations were evaluated as an alternative acoustic feature set related to auditory perception, designed to capture dynamic patterns of spectral energy fluctuations over time. STM representations are known to approximate aspects of human auditory processing by modeling how temporal and spectral modulations contribute to perception. In this study, STM representations achieved accuracies of 62%, 66% and 68% with SVM, ET and KNN classifiers, respectively, performing better than timbral features and conventional representations such as mel-spectrograms, GTFB, and GCFB. This result suggests that STM provides a valuable approach for imitated speech detection and shows that richer temporal dynamics can support more accurate differentiation. Notably, among all the evaluated features, STM achieved performance that was closest to human listeners' accuracy in distinguishing genuine from imitated speech.

These results demonstrate that both human listeners and acoustic features related to auditory perception face challenges in classifying genuine and imitated speech. Although machine performance did not fully match human accuracy, the use of timbral and STM representations provided valuable insights into how machines can simulate human perceptual mechanisms. However, humans likely rely on a combination of multiple perceptual cues, rather than isolated features like boominess and depth, to achieve higher accuracy. This broader integration of auditory information may explain humans' superior performance compared with machines in this task.

Table 6.21: Evaluation metrics for genuine and imitated speech classification.

Features	**Evaluation Metrics**			
	Accuracy(%)	**F1-Score(%)**	**EER(%)**	d'
Mel-spectrogram	51.0	49.54	46.0	0.05
GTFB	61.0	60.81	39.0	0.56
GCFB	60.0	59.60	40.0	0.524
Timbral	64.0	63.94	36.0	0.72
STM	68.0	68.0	32.0	0.93

In addition, the classification results using One-Class SVM, LOF, and Isolation Forest demonstrate that timbral features, particularly warmth, effectively distinguish between genuine and imitated speech. The feature warmth achieved the highest accuracy (up to 64%), approaching the human benchmark of 70.10%. Notably, all models were trained solely on genuine speech, mirroring the conditions of the human listening experiment in which participants were first exposed only to genuine samples. This design ensures a fair and scientifically valid comparison between human and machine performance. These findings highlight the strength of acoustic features related to auditory perception and validate one-class classification as a human-aligned approach for detecting imitated speech.

Furthermore, to better reflect human listening evaluation, speaker-specific models were trained using timbral features with a separate discriminator for each target speaker. This approach accounted for inter-speaker variability and revealed that classification performance varies significantly depending on both the distinctiveness of the speaker and the imitator's proficiency. The results emphasize the value of individualized classification frameworks, which can capture subtle variations overlooked by generic models and offer more interpretable insights into imitation detection.

In conclusion, even under challenging conditions, the results demonstrate that both human listeners and acoustic features related to auditory perception such as timbral features and STM representations, play a crucial role in accurately distinguishing between genuine and imitated speech.

6.7 Summary

To summarize, this work introduced approaches to evaluate human and machine performance in detecting human-imitated speech.

In the human listening tests, participants trained exclusively on genuine speech, achieving high speaker identification accuracy (95.75%). However, their final test accuracy in distinguishing genuine from imitated speech dropped to 70.10%, underscoring the difficulty of the task.

Complementing the human listening experiments, machine-based classification experiments were conducted using acoustic features related to auditory perception. Timbral features, such as boominess, achieved accuracies of up to 64%, while STM representations reached 68%, outperforming standard features like GTFB, GCFB, and mel-spectrograms, and approaching human-level performance. To closely mirror the human experiment design in which participants were only exposed to genuine speech one-class classification methods were applied, including One-Class SVM, LOF, and Isolation Forest. Despite being trained without imitated speech, these models demonstrated notable detection capabilities.

Additionally, speaker-specific models were employed to reflect how listeners internally form speaker-specific discriminators. As presented in Tables 6.11 to 6.20, classification accuracy varied considerably across speakers. For instance, Speaker 6 and Speaker 7 (Tables 6.16 and 6.17) consistently yielded higher accuracy, suggesting less skilled imitators, whereas Speaker 4 and Speaker 5 (Tables 6.14 and 6.15) were more difficult to distinguish, indicating higher imitation proficiency.

Overall, this work demonstrates that combining evidence from human listening tests with acoustic features aligned to auditory perception, together with speaker-specific modeling strategies, offers valuable insights and practical foundations for developing systems capable of reliably detecting human-imitated speech in realistic scenarios.

Chapter 7

Conclusion

7.1 Summary

This study shed light on human listening tests and acoustic features grounded in auditory perception for detecting human-imitated speech, an emerging and under-explored category within the fake speech domain. A custom dataset of human-imitated speech was developed to systematically evaluate detection performance from both human and machine perspectives.

Initial experiments demonstrated that models trained exclusively on synthetic speech performed well on synthetic attacks but failed to generalize to imitation, revealing a clear performance gap. Retraining these models directly on the human-imitated dataset improved detection accuracy but remained insufficient for reliable identification. Additional experiments assessed standard acoustic representations and auditory-inspired features, including gammatone and gammachirp filterbanks designed to simulate aspects of human cochlear processing. While these features provided better discrimination than conventional representations, they still did not achieve robust detection performance.

To build on these initial results, this research proposed a two-phase framework. In the first phase, carefully designed listening tests assessed how accurately human listeners could distinguish between genuine and imitated speech, showing that even trained participants struggled when imitation was highly convincing. In the second phase, machine-based experiments applied acoustic features related to auditory perception, including timbral attributes and STM representations, to classify imitated speech. To better match machine-based models to human evaluation conditions, additional experiments were conducted using one-class classification models trained exclusively on genuine speech, as this design reflected the fact that human participants had no prior exposure to imitated speech.

The goal of this research was to study imitated speech detection from both

human and machine perspectives, framed by two key research questions.

- To answer the first question, *How accurately can human listeners distinguish between genuine and imitated speech?*, the findings from the first phase of the study indicate that listeners can distinguish between genuine and human-imitated speech, but only to a limited extent. One of the most important aspects of this study was the speaker identification task, in which listener familiarity with the target speakers played a critical role. To ensure that all participants had the same level of familiarity, a three-phase experimental procedure was carefully designed. The first phase involved a training stage in which participants were required to learn and identify all 10 speakers. Only those who achieved over 90% accuracy in identifying the target speakers were allowed to proceed to the final phase, which involved classifying speech samples as either genuine or imitated. This requirement was essential to ensure that all listeners had sufficient and uniform familiarity with the target speakers.

  In the initial stage, participants achieved high accuracy in identifying target speakers after extensive training, confirming that they became familiar with the voices. However, when the task required classifying speech samples as genuine or imitated, the overall accuracy decreased, even after training. Many participants needed several training sessions and took longer to complete the task, indicating the complexity and difficulty of accurately distinguishing between genuine and imitated speech. Notably, 2 out of the 22 participants were unable to meet the strict training criteria and chose to withdraw due to the challenging nature of the experiment.

  This design also addresses a common limitation in previous studies. Although some authors claim their participants were familiar with the speakers often because they were native speakers, there is often no clear indication of the actual level of familiarity. Whether participants had high or low familiarity remains uncertain. Without controlling this factor, the similarity with speakers may vary significantly among subjects, leading to unfair or biased evaluation results. The proposed listening test method directly addresses this issue by ensuring high and comparable familiarity across all subjects, which is necessary for a fair speaker verification task.

  These results suggest that humans can detect imitated speech, but it remains a complex task requiring a high level of concentration and judgment, especially when the imitated speech closely resembles genuine speech. This realistic and carefully controlled test setup represents a significant strength of the present study.

- To answer the second question, *How effectively can auditory perceptual features be used to detect human-imitated speech by machine/deep learning techniques like human listeners?*, the study evaluated several types of acoustic features related to auditory perception. As part of earlier experiments, auditory-inspired representations derived from gammatone and gammachirp filterbanks were applied to model aspects of human cochlear processing. While these features function as front-end representations inspired by the auditory periphery rather than capturing higher-level perceptual attributes, they performed better than standard acoustic representations but were not sufficient for achieving reliable detection. This limitation underscored the need to explore additional features more closely aligned with human perceptual judgments.

  Building on insights from the human listening tests conducted in the first phase, the study then proposed two sets of acoustic features related to auditory perception. Timbral attributes, including boominess, depth, brightness, and warmth were extracted and assessed in machine learning experiments, while STM representations were also applied within machine learning models to provide a multiscale view of dynamic spectral and temporal patterns in speech. These features demonstrated promising potential, yielding performance trends comparable to those observed in the human evaluation.

Overall, by creating a novel dataset of human-imitated speech and applying both human-centered evaluations and machine-based classification techniques, this study provides valuable insights into the auditory perception of speech and its implications for imitated speech detection. These findings form a solid foundation for future research, particularly in developing robust countermeasure systems that align more closely with human perceptual processes in complex speech-based spoofing scenarios.

7.2 Contribution

This study addresses the challenge of distinguishing between genuine and human-imitated speech by leveraging a human listeners test and acoustic features related to auditory perception. Unlike AI-generated speech, human-imitated speech closely mimics the natural qualities of genuine speech, making it difficult to detect. The proposed methods can be applied to various speech security tasks, such as fake speech detection, speech forensics, and speaker verification spoofing countermeasure challenges.

This research contributes scientifically by distinguishing between genuine and human-imitated speech using human listeners tests and acoustic features related

to auditory perception.

Importantly, this study contributes the first human-imitated speech dataset that is compatible with machine learning analysis. This dataset, created under realistic and diverse acoustic conditions, offers a valuable resource for both subjective experiments and objective evaluation with advanced machine learning models. Moreover, the public release of the dataset enables other researchers to collaborate, compare, and reproduce results under clearly defined conditions. This contributes to greater transparency and scientific value in the field of speaker imitation detection.

One of the key contributions of this study is the demonstration that existing spoofing countermeasure systems fail to reliably detect human-imitated speech, an issue verified through our experimental evaluations. This finding highlights a critical vulnerability in current speech authentication frameworks and underscores the need for new detection strategies tailored to this form of attack.

From a societal perspective, this research contributes significantly to society by protecting voice-based technologies from human-imitated speech attacks, thereby ensuring more secure and trustworthy digital communication in applications such as speaker verification, fake speech detection, and speech forensics.

7.3 Future work

This study has laid the groundwork for advancing the detection of human-imitated speech and its differentiation from genuine speech. However, several key areas of improvement remain to be explored in future research, which will enhance the robustness and applicability of the proposed methods. The future work is organized into the following areas:

One crucial next step is to expand the dataset to ensure broader applicability and robustness of the detection systems. The current dataset, while pioneering, focuses on a limited set of speakers and scenarios. To improve the diversity of the dataset, future work should aim to include a broader range of speakers, incorporating different accents, languages, and age groups. Furthermore, increasing the variability in environmental conditions, such as different background noises, microphone types, and recording environments, will help create a more representative dataset. Additionally, integrating real-world data from various contexts will make the dataset more applicable to practical challenges in speech detection, ultimately improving the generalization capabilities of detection models.

The study has explored timbral features such as boominess, depth, and warmth, which show promise in distinguishing human-imitated speech. However, there is potential for improvement in this area. Future research should focus on investigating additional acoustic features that might reveal more subtle differences between

genuine and imitated speech. Features like spectral variations, non-linear dynamics, and temporal characteristics may provide valuable information for detection. Moreover, refining existing features and combining complementary acoustic features can lead to more accurate and robust detection models, ensuring that the systems can effectively capture the nuanced differences between genuine and imitated speech.

The use of advanced deep learning models could significantly improve the detection of human-imitated speech. In future work, there is a need to explore transformer-based models or attention mechanisms, which have demonstrated superior performance in capturing complex relationships between various features. These models can improve the ability to model temporal and spectral dependencies over long sequences, enhancing the accuracy of detection. Additionally, leveraging transfer learning could allow the models to generalize better across domains, especially when datasets are small or not highly diverse. The integration of multi-task learning can also be considered, enabling models to perform multiple detection tasks simultaneously, further improving their robustness and overall performance.

The development of real-time detection systems is another important avenue for future research. Currently, speech detection models are often designed for offline analysis, but integrating detection systems into real-time voice-based security platforms is crucial for their practical use. Future work should focus on designing efficient algorithms that can operate in real-time, ensuring low-latency detection without compromising accuracy. Additionally, it is important to explore ways to make these systems scalable and efficient to handle large volumes of speech data in real-time, particularly in high-demand applications like digital assistants, automated verification systems, and communication security.

Incorporating multiple modalities into speech detection could significantly enhance the robustness of the systems. Future work could explore multimodal detection systems, combining audio with visual inputs such as lip movement, facial expressions, or even body language. Integrating these visual cues with speech detection could make systems more accurate, as it would allow for the detection of incongruities between the auditory and visual aspects of speech. Additionally, integrating contextual information such as user history or environmental factors could further improve detection accuracy, particularly in dynamic or noisy settings.

Another promising area for future research is the concept of human-AI collaboration in speech detection. While machine learning models have shown significant promise, there are scenarios where human expertise can enhance detection accuracy. By developing systems where machine learning models collaborate with human experts, it may be possible to improve the detection of ambiguous cases. Active learning techniques could be employed, where the detection system requests human feedback for particularly challenging samples, enabling continuous learning

and refinement of the model.

By addressing these key areas, future research can significantly advance the state-of-the-art in speech detection systems, making them more accurate, reliable, and applicable to real-world challenges in speech security, authentication, and privacy protection.

Bibliography

[1] C. L. Brockmann, "A diagram of the anatomy of the human ear," *Retrieved December*, vol. 9, p. 2009, 2009.

[2] R. E. S. Lovett, "Comparisons of unilateral and bilateral cochlear implantation for children: Spatial listening skills and quality of life," Ph.D. dissertation, University of York, 2010.

[3] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106 620–106 649, 2023.

[4] M. Unoki, K. Li, A. Chaiwongyen, Q.-H. Nguyen, and K. Zaman, "Deepfake speech detection: approaches from acoustic features related to auditory perception to deep neural networks," *IEICE Transactions on Information and Systems*, 2024.

[5] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digital Signal Processing*, p. 104956, 2024.

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

[7] 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.

[8] K. Zaman, I. J. Samiul, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Hybrid transformer architectures with diverse audio features for deepfake speech classification," *IEEE Access*, vol. 12, pp. 149 221–149 237, 2024.

[9] T. Kanwal, R. Mahum, A. M. AlSalman, M. Sharaf, and H. Hassan, "Fake speech detection using vggish with attention block," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 35, 2024.

[10] T. Kinnunen, Z. Wu, E. Nicholas Evans, and J. Yamagishi, "Automatic speaker verification spoofing and countermeasures challenge (asvspoof 2015) database," 2018.

[11] T. Kinnunen, M. Sahidullah, E. Héctor Delgado, E. Massimiliano Todisco, E. Nicholas Evans, J. Yamagishi, and K. A. Lee, "The 2nd automatic speaker verification spoofing and countermeasures challenge (asvspoof 2017) database, version 2," 2018.

[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017, 2017, pp. 2–6.

[13] 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.

[14] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

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

[16] 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.

[17] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," *arXiv preprint arXiv:2104.03617*, 2021.

[18] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021.

[19] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[20] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, 2023.

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

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

[23] E. Globerson, T. Elias, N. Kittany, and N. Amir, "Pitch discrimination abilities in classical arab-music listeners," *Applied Acoustics*, vol. 102, pp. 120–124, 2016.

[24] H. Masuda, Y. Hioka, C. J. Hui, J. James, and C. I. Watson, "Performance evaluation of speech masking design among listeners with varying language backgrounds," *Applied Acoustics*, vol. 201, p. 109122, 2022.

[25] E. Zetterholm, "Detection of speaker characteristics using voice imitation," *Speaker classification II: selected projects*, pp. 192–205, 2007.

[26] S. Madureira and M. A. Fontes, "Vocal and visual features in speech imitation."

[27] Y. Deng, K. Chen, H. Li, and J. Zhang, "Matched standard samples method in laboratory listening tests for annoyance perception," *Applied Acoustics*, vol. 224, p. 110103, 2024.

[28] M. Geravanchizadeh and S. Zakeri, "Binaural source separation using auditory attention for salient and non-salient sounds," *Applied Acoustics*, vol. 195, p. 108822, 2022.

[29] Y. Kou, W. Chen, J. Wang, W. Liu, S. Yang, and H. Liu, "Personalized prediction of speech intelligibility for hearing-impaired listeners using a physiological model of the human ear," *Applied Acoustics*, vol. 221, p. 110006, 2024.

[30] R. Shimokura and Y. Soeta, "Estimation of reaction time for birdsongs and effects of background noise and listener's age," *Applied Acoustics*, vol. 194, p. 108785, 2022.

[31] E. De Groote, R. P. Carlyon, J. M. Deeks, and O. Macherey, "Effects of selective stimulation of apical electrodes on temporal pitch perception by cochlear

implant recipients," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 2060–2076, 2024.

[32] J. Vonessen, N. B. Aoki, M. Cohn, and G. Zellou, "Comparing perception of l1 and l2 english by human listeners and machines: Effect of interlocutor adaptations," *The Journal of the Acoustical Society of America*, vol. 155, no. 5, pp. 3060–3070, 2024.

[33] Y. Li, A. Li, J. Tao, F. Li, D. Erickson, and M. Akagi, "Contributions of audio and visual modalities to perception of mandarin chinese emotions in valence-arousal space," *Acoustical Science and Technology*, pp. e24–41, 2024.

[34] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 234–242, 2018.

[35] B. Kim, M. Ghei, B. Pardo, and Z. Duan, "Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology." in *DCASE*, 2018, pp. 148–152.

[36] Y. Zhang and Z. Duan, "Visualization and interpretation of siamese style convolutional neural networks for sound search by vocal imitation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2406–2410.

[37] D. S. Blancas and J. Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[38] Y. Zhang and Z. Duan, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 533–543, 2016.

[39] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: Identification and discrimination of authentic cantillations from imitations," *Neurocomputing*, vol. 418, pp. 162–177, 2020.

[40] A. Mehrabi, S. Dixon, and M. Sandler, "Vocal imitation of percussion sounds: On the perceptual similarity between imitations and imitated sounds," *Plos one*, vol. 14, no. 7, p. e0219955, 2019.

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

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

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

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

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

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

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

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

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

[50] H. Cheng, C. O. Mawalim, K. Li, L. Wang, and M. Unoki, "Analysis of spectro-temporal modulation representation for deep-fake speech detection," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1822–1829.

[51] K. Li, K. Zaman, X. Li, M. Akagi, J. Dang, and M. Unoki, "Machine anomalous sound detection using spectral-temporal modulation representations derived from machine-specific filterbanks," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[52] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI open*, vol. 3, pp. 111–132, 2022.

[53] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," *arXiv preprint arXiv:2303.11607*, 2023.

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

[55] 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.

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

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

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

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

[60] Y. Rodríguez-Ortega, D. M. Ballesteros, and D. Renza, "A machine learning model to detect fake voice," in *International conference on applied informatics*. Springer, 2020, pp. 3–13.

[61] Q.-H. Nguyen and M. Unoki, "Bone-conducted speech enhancement using vector-quantized variational autoencoder and gammachirp filterbank cepstral coefficients," in *Proc. EUSIPCO2022*. IEEE, 2022, pp. 21–25.

[62] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech*, 2021, pp. 4259–4263.

[63] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221819604

[64] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[65] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "Stc antispoofing systems for the asvspoof2019 challenge," in *Interspeech*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:118674374

[66] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit, Cambridge, UK, APU Report 2341, 1988.

[67] T. Irino and R. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, 1997.

[68] T. Irino and M. Unoki, "An analysis/synthesis auditory filterbank based on an iir implementation of the gammachirp," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 6, pp. 397–406, 1999.

[69] K. Salhi, Z. Hajaiej, and N. Ellouze, "A novel approach for auditory spectrum enhancement to improve speech recognition's robustness," in *Proceedings of the 12th IEEE International Multi-Conference on Systems, Signals and Devices*, 2015, pp. 1–5.

[70] A. Abdallah and Z. Hajaiej, "Improved closed set text independent speaker identification system using gammachirp filterbank in noisy environments," in *Proceedings of the 11th International Multi-Conference on Systems, Signals and Devices*, 2014, pp. 1–5.

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

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

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

[74] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Release of timbral characterisation tools for semantically annotating non-musical content," *AudioCommons*, 2020.

[75] D. Williams, *Towards a timbre morpher*. University of Surrey (United Kingdom), 2010.

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

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

[78] A. Pearce, T. Brookes, and R. Mason, "Modelling timbral hardness," *Applied Sciences*, vol. 9, no. 3, p. 466, 2019.

[79] B. Pardo, M. Cartwright, P. Seetharaman, and B. Kim, "Learning to build natural audio production interfaces," *Arts*, vol. 8, p. 110, 08 2019.

[80] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[81] A. Pearce, T. Brookes, and R. Mason, "Evaluation report on the second prototypes of the timbral characterisation tools," *[Online]. Available:http://www.audiocommons.org/materials/.*

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

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

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

[85] P. N. Vassilakis, "Sra: A web-based research tool for spectral and roughness analysis of sound signals," 2007.

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

[87] K. M. Sorensen and M. C. Vigeant, "Study of the perception of warmth in concert halls and correlation with room acoustics metrics," *Journal of the Acoustical Society of America*, vol. 140, no. 4_Supplement, pp. 3176–3176, 2016.

[88] G. Bromham, D. Moffat, M. Barthet, A. Danielsen, and G. Fazekas, "The impact of audio effects processing on the perception of brightness and warmth," in *Proc. Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, 2019, pp. 183–190.

[89] D. Williams and T. Brookes, "Perceptually-motivated audio morphing: Warmth," in *128th Convention, London, UK*, 2010.

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

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

[92] S. Ystad, M. Aramaki, and R. Kronland-Martinet, "Timbre from sound synthesis and high-level control perspectives," *Timbre: Acoustics, Perception, and Cognition*, pp. 361–389, 2019.

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

[94] ——, "Booming index as a measure for evaluating booming sensation," in *Proc. Inter-Noise*, no. 233, 2000, pp. 1–6.

[95] S.-H. Shin, J.-G. Ih, T. Hashimoto, and S. Hatano, "Sound quality evaluation of the booming sensation for passenger cars," *Applied acoustics*, vol. 70, no. 2, pp. 309–320, 2009.

[96] S. Hatano and T. Hashimoto, "On an objective measure of the booming sound factor-modification of the measure by the spectrum pattern and the loudness of the sound," *JSAE Review*, vol. 3, no. 16, pp. 325–326, 1995.

[97] S.-H. Shin and J.-G. Ih, "Prediction of booming sensation and its difference limen for just noticeable change in frequency," *The Journal of the Acoustical Society of America*, vol. 114, no. 4_Supplement, pp. 2351–2351, 2003.

[98] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, vol. 24, no. 10, pp. 1681–1693, 2016.

[99] T. d. M. Prego, A. A. De Lima, S. L. Netto, B. Lee, A. Said, R. W. Schafer, and T. Kalker, "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2811–2816, 2012.

[100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[101] J. Abimbola, D. Kostrzewa, and P. Kasprowski, "Music time signature detection using resnet18," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 30, 2024.

[102] Z. Chen, H. Wang, C.-H. Yeh, and X. Liu, "Classify respiratory abnormality in lung sounds using stft and a fine-tuned resnet18 network," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 233–237.

[103] K. Mittal, K. S. Gill, K. Rajput, and V. Singh, "Enhancing the diagnosis of speech disorders: An in-depth investigation into dysarthria classification using the resnet18 model," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*. IEEE, 2024, pp. 1–5.

[104] A. Elfaki, A. L. Asnawi, A. Z. Jusoh, A. F. Ismail, S. N. Ibrahim, N. F. M. Azmin, and N. N. W. B. N. Hashim, "Using the short-time fourier transform and resnet to diagnose depression from speech data," in *2021 IEEE International Conference on Computing (ICOCO)*. IEEE, 2021, pp. 372–376.

[105] I. Topaloglu, P. D. Barua, A. M. Yildiz, T. Keles, S. Dogan, M. Baygin, H. F. Gul, T. Tuncer, R.-S. Tan, and U. R. Acharya, "Explainable attention resnet18-based model for asthma detection using stethoscope lung sounds," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106887, 2023.

[106] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "Stc antispoofing systems for the asvspoof2021 challenge," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 61–67.

[107] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–5.

[108] M. S. Alam, A. Fathan, and J. Alam, "Audio deepfake detection employing multiple parametric exponential linear units," in *International Conference on Speech and Computer*. Springer, 2023, pp. 307–321.

[109] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[110] A. Ambili and R. C. Roy, "The effect of synthetic voice data augmentation on spoken language identification on indian languages," *IEEE Access*, vol. 11, pp. 102 391–102 407, 2023.

[111] N. Zakaria, F. Mohamed, R. Abdelghani, and K. Sundaraj, "Vgg16, resnet-50, and googlenet deep learning architecture for breathing sound classification: a comparative study," in *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*. IEEE, 2021, pp. 1–6.

[112] A. B. Gumelar, E. M. Yuniarno, W. Anggraeni, I. Sugiarto, V. R. Mahindara, and M. H. Purnomo, "Enhancing detection of pathological voice disorder based on deep vgg-16 cnn," in *2020 3rd International Conference on Biomedical Engineering (IBIOMED)*. IEEE, 2020, pp. 28–33.

[113] H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, and J. Černockỳ, "Detecting spoofing attacks using vgg and sincnet: But-omilia submission to asvspoof 2019 challenge," in *Proc. Interspeech 2019*, 2019, pp. 1073–1077.

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

[115] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Procedia Computer Science*, vol. 219, pp. 211–219, 2023.

[116] R. Vadishetty, "Efficient deep fake detection technique on video and audio dataset using deep learning," in *International Ethical Hacking Conference*. Springer, 2024, pp. 137–155.

[117] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, pp. 1–10.

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

[119] M. Georgeson, "Sensitivity and bias: An introduction to signal detection theory," *Retrieved June*, vol. 15, p. 2005, 2005.

[120] M. J. Hautus, N. A. Macmillan, and C. D. Creelman, *Detection theory: A user's guide*. Routledge, 2021.

[121] B. Vamsi, M. Mahanty, and B. P. Doppala, "An auto encoder-decoder approach to classify the bird sounds using deep learning techniques," *SN Computer Science*, vol. 4, no. 3, p. 289, 2023.

[122] K. Zaman, K. Li, I. J. Samiul, Y. Uezu, S. Kidani, and M. Unoki, "Ability of human auditory perception to distinguish human-imitated speech," *IEEE Access*, vol. 13, pp. 6225–6236, 2025.

[123] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," *Target*, vol. 4000, p. 5000, 2014.

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

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

[126] ——, "Modelling timbral hardness," *Applied Sciences*, vol. 9, no. 3, p. 466, 2019.

[127] H. Goddard and L. Shamir, "Svmnet: Non-parametric image classification based on convolutional ensembles of support vector machines for small training sets," *IEEE Access*, vol. 10, pp. 24 029–24 038, 2022.

[128] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, 2016.

[129] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[130] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

Publications

Main Publications

International Journals

1. K. Zaman, K. Li, I. J. A. M. Samiul, Y. Uezu, S. Kidani, and M. Unoki, "Ability of Human Auditory Perception to Distinguish Human-Imitated Speech," *IEEE Access*, vol. 13, pp. 6225-6236, 2025, https://doi.org/10.1109/ACCESS.2025.3526631.
2. K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, https://doi.org/10.1109/ACCESS.2023.3318015.

Domestic Conference/ Workshop

1. K. Zaman, I. J. Samiul, K. Li, Y. Uezu, S. Kidani, and M. Unoki, "Study on Ability of Human Auditory Perception to Distinguish Human-Imitated Speech," In *the Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report*, Tohoku University, Sendai, Japan, Jan. 2025. https://ken.ieice.org/ken/paper/20250121Hch7/.

Other Publications

International Journals

1. K. Li, K. Zaman, X. Li, M. Akagi, J. Dang, and M. Unoki, "Machine Anomalous Sound Detection Using Spectral-Temporal Modulation Representations Derived From Machine-Specific Filterbanks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2059–2073, 2025, https://doi.org/10.1109/TASLPRO.2025.3570956.

2. K. Zaman, I. J. A. M. Samiul, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Hybrid Transformer Architectures With Diverse Audio Features for Deepfake Speech Classification," *IEEE Access*, vol. 12, pp. 149221-149237, 2024,
https://doi.org/10.1109/ACCESS.2024.3478731.

3. K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and Audio Detection Tasks: An Overview," *Digital Signal Processing*, vol. 158, pp. 149221-149237,
https://doi.org/10.1016/j.dsp.2024.104956.

4. M. Unoki, K. Li, A. Chaiwongyen, Q.-H. Nguyen, and K. Zaman, "Deepfake Speech Detection: Approaches from Acoustic Features Related to Auditory Perception to Deep Neural Networks," *IEICE Transactions on Information and Systems*, 2024,
https://doi.org/10.1587/transinf.2024MUI0001.

International Conference/ Workshop

1. Islam J. A. M. Samiul, Khalid Zaman, Kai Li, Anuwat Chaiwongyen, Shogo Okada, and Masashi Unoki, "Enhancing Model Robustness for Deepfake Singing Voice Detection through Data Augmentation," In *the RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'25)*, Pulau Pinang, Malaysia, February 27–March 2, 2025.
https://ncsp.jp/NCSP25/.