Doctoral Dissertation

Deepfake speech detection using speech-pathological features

Anuwat Chaiwongyen

Supervisor: Masashi UNOKI

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

December, 2024

Abstract

There is a great concern regarding the misuse of deepfake speech technology to synthesize a real person's voice. Therefore, developing speech-security systems capable of detecting deepfake speech remains paramount in safeguarding against such misuse. Although various speech features and methods have been proposed, their potential for distinguishing between genuine and deepfake speech remains unclear. Since speech-pathological features with deep learning are widely used to assess unnaturalness in disordered voices associated with voice-production mechanisms, investigated the potential of speech-pathological features for distinguishing between genuine and deepfake speech.

In this work, two categories of pathological speech features were investigated: perceptual and acoustic features. For perceptual features, eight characteristics were examined: depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation. The acoustic features analyzed included jitter (three types), shimmer (four types), harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio, normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). The proposed method was evaluated on four datasets: Automatic Speaker Verification Spoofing and Countermeasures Challenges (ASVspoof) 2019 and 2021, and Audio Deep Synthesis Detection (ADD) 2022 and 2023.

In the first step, two types of speech-pathological features, perceptual and acoustic, are investigated. The data from the feature extraction for each type of feature were averaged. These averaged features were then fed into a multilayer perceptron neural network for training and evaluating the performance of the model.

After investigation, it was found that acoustic speech-pathological features and perceptual speech-pathological features could effectively detect deepfake speech, except for HNR. To improve the efficiency of the proposed features, the important features from both acoustic and perceptual speech-pathological features were selected. The results indicate that when the important speech-pathological features are combined, the efficiency of the proposed features is improved.

Consequently, aimed to enhance the efficiency of the acoustic speech-pathological features by using segmental frames of analysis. This approach extends the dimension of the features beyond a simple average. The results indicated that using segmental frames of analysis significantly improved the efficiency of the acoustic speech-pathological features.

Therefore, in this work, proposes a method for detecting deepfake speech by using segmental frames of analysis of speech-pathological features. These features include jitter (*local*), jitter ($PPQ3$), jitter ($PPQ5$), shimmer (*local*), shimmer ($APQ3$), shimmer ($APQ5$), shimmer ($APQ11$), GNE, NNE, CHNR. These features are fed into a ResNet-18 for classification, and the results demonstrate that

incorporating these ten features with ResNet-18 significantly improves the efficiency of detecting fake speech.

Moreover, this paper proposes a method of combining two models on the basis of two different dimensions of speech-pathological features to greatly improve the effectiveness of deepfake speech detection, along with mel-spectrogram features, to enhance detection efficiency. The proposed method is evaluated on the ASVspoof 2019, 2021, ADD 2022, and ADD 2023 datasets. It consistently outperforms the baselines in terms of accuracy, recall, F1-score, and F2-score across these datasets. However, the equal error rate for the ADD 2022 test set remains relatively high. Overall, the method demonstrates high performance and effectiveness in deepfake speech detection.

Keywords: Deepfake speech detection, speech-pathological features, acoustical features, perceptual features, and neural network.

Acknowledgment

I would like to express my sincere gratitude to the following individuals and institutions for their invaluable support during my PhD journey.

First and foremost, I am deeply grateful to my doctoral advisor, Professor Unoki Masashi, for his unwavering guidance, insightful suggestions, and constant encouragement throughout my research. His expertise and dedication have been instrumental in shaping my understanding of this field and in the successful completion of this dissertation.

I express my gratitude to Professor Waree Kongprawechnon, who has not only been a good teacher but also everything to me, for her suggestions and guidance, which have been invaluable in my doctoral studies. I feel fortunate to have had the chance to learn under your supervision.

To Dr. Jessada Kanjana and Dr. Suradej Dungpummet, you have not only been my teachers but also supportive brothers to me. Your thoughtful consideration of the challenges I faced in my research and your patient guidance have been invaluable to my academic growth. I am deeply grateful for your contributions.

I am truly grateful to my colleagues and friends who have accompanied me on this academic journey. Together, we engage in academic discussions, envision the future, and support one another. Your companionship has not only enriched this experience with knowledge but also made it enjoyable.

Lastly, my sincere thanks go out to my family: my father and mother, my wife, and my son, Udomchai Chaiwongyen, and Bualat Chaiwongyen, as well as Sumintra Chaiwongyen and Nattachai Chaiwongyen, for their unwavering support and encouragement throughout my doctoral studies. Their belief in me has been a constant source of strength.

Contents

Abstract **i**

Acknowledgment **iii**

List of Figures **viii**

List of Tables **xi**

1 Introduction **1**
1.1 Research Background 1
1.2 Research Problems 4
1.3 Research Motivation 5
1.4 Research Goals 6
1.5 Challenges 7
1.6 Organization of Thesis 8

2 Literature Review **11**
2.1 Introduction 11
2.2 Human Speech Production Mechanism 14
2.2.1 Motor control function 15
2.2.2 Articlulatory motion 16
2.2.3 Sound generation 18
2.3 Voice Disorders Assessment 19
2.3.1 Assessment of Voice Disorders by Clinicians 19
2.3.2 Automatic Voice Disorders Assessment 21
2.4 Perceptual Speech-Pathological Features for Assessing Voice Disorders 24
2.5 Acoustical Speech-Pathological Features for Assessing Voice Disorders 25

3 Contribution of Using Perceptual and Acoustical Speech Pathological Features to Detect Fake Speech **28**
3.1 Concept and Idea of Proposed Features 28

3.2 Philosophy of Utilizing Speech-Pathological Features for Deepfake Speech . . . 31
3.3 Acoustical Speech-Pathological Features . . . 33
3.3.1 Jitter Features . . . 33
3.3.2 Shimmer features . . . 34
3.3.3 Harmonics-to-Noise Ratio (HNR) . . . 35
3.3.4 Cepstral-Harmonics-to-Noise Ratio (CHNR) . . . 36
3.3.5 Normalized noise energy (NNE) . . . 37
3.3.6 Glottal-to-Noise Excitation Ratio (GNE) . . . 38
3.4 Perceptual Speech-Pathological Features . . . 39
3.4.1 Hardness . . . 39
3.4.2 Depth . . . 40
3.4.3 Brightness . . . 42
3.4.4 Roughness . . . 43
3.4.5 Warmth . . . 44
3.4.6 Sharpness . . . 44
3.4.7 Boominess . . . 45
3.4.8 Reverberation . . . 45
3.5 Dataset and Metrics . . . 46
3.5.1 Dataset . . . 46
3.5.2 Metrics . . . 48

4 Deepfake Speech Detection using Acoustical Speech-Pathological Features 51
4.1 Proposed Method using Acoustical Speech-Pathological Features . . 51
4.1.1 Results and discussion in ASVspoof 2019 and 2021 datasets 53
4.1.2 Results and discussion in ADD 2022 and 2023 datasets . . . 56
4.1.3 Ablation study of Acoustical Speech-Pathological Features . 59
4.2 Deepfake Speech Detection using Segmental Frames of Analysis of Acoustical Speech-Pathological Features . . . 64
4.2.1 Proposed Method of using segmental frames of analysis of acoustical speech-pathological features . . . 65
4.2.2 Results and discussion in ASVspoof 2019 and 2021 . . . 69
4.2.3 Results and discussion in ADD 2022 and 2023 datasets . . . 72
4.2.4 Ablation Study of Segmental Frames of Analysis of Speech-Pathological Features . . . 77
4.3 Summary . . . 79

5 Deepfake Speech Detection using Perceptual Speech-Pathological Features 85
5.1 Proposed Method using Perceptual Speech-Pathological Features . . 85

5.2 Results and discussion in ASVspoof 2019 and 2021 datasets 86
5.3 Results and discussion in ADD 2022 and 2023 datasets 90
5.3.1 Results and discussion in ADD 2022 dataset 90
5.3.2 Results and discussion in ADD 2023 90
5.4 Ablation study of Perceptual Speech-Pathological Features 93
5.5 Summary 98

6 Deepfake Speech Detection using Acoustical and Perceptual Speech-Pathological Features 99
6.1 Proposed Method using Acoustical and Perceptual Speech-Pathological Features 99
6.2 Results and discussion in ASVspoof 2019 and 2021 datasets 101
6.3 Results and discussion in ADD 2022 and 2023 datasets 103
6.4 Discussion 103
6.5 Summary 108

7 Conclusion 109
7.1 Summary 109
7.2 Contributions 110
7.3 Remaining Works 111

Bibliography 111

List of Figures

1.1 Organization of this dissertation. 10

2.1 General method for fake speech detection. 11
2.2 Block diagram of human speech production mechanism [1]. 15
2.3 Human vocal apparatus [2]. 17
2.4 Example Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Form. [3]. 20
2.5 Automatic voice disorder detection (AVDD) system [4]. 22

3.1 Concept and idea of using acoustical and perceptual speech pathological features to detect fake speech. 29
3.2 Bar graph comparing four shimmer features of disordered voices, fake speech, and genuine speech. 30
3.3 Philosophy of Utilizing Speech-Pathological Features for Deepfake Speech. 32
3.4 Jitter and shimmer concept illustration. 33
3.5 CHNR calculation process [5]. 36
3.6 NNE calculation process [5]. 37
3.7 GNE calculation process [5]. 38
3.8 Social-EQ graphic equaliser settings representing the timbral descriptor *deep* [6]. 41

4.1 Proposed method using acoustical speech-pathological features with multi-layer perceptron neural networks. 52
4.2 Segmental frames of analysis of acoustical speech-pathological features. 64
4.3 Proposed method, combining of (1) ten pathological-segment features with their first-order and second-order derivatives with ResNet-18 and (2) mel-spectrogram with ResNet-18, through score fusion. . 65

5.1 Proposed method using perceptual speech-pathological features with multi-layer perceptron neural networks. 86

6.1 Proposed method using important acoustic and perceptual speech-pathological features with multi-layer perception neural networks. . 100

List of Tables

3.1 Number of utterances in the ASVspoof 2019 and 2021 datasets [7, 8]. 47
3.2 Number of utterances in the ADD 2022 [9] and 2023 datasets. . . . 47

4.1 Results from applying an average of speech-pathological features with a neural network on the development set of ASVspoo 2019. . . 54
4.2 Results from applying an average of acoustical speech-pathological features with a neural network on the evaluation set of ASVspoof 2019 and 2021. 55
4.3 Results from applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2022. 57
4.4 Results from applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2023. 58
4.5 Ablation study of applying an average of acoustical speech-pathological features with a neural network on the development set of ASV 2019. 60
4.6 Ablation study of applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2022. 61
4.7 Ablation study of applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2023. 62
4.8 Results of using segmental frames of analysis of acoustical speech-pathological features with neural networks on development set of ASVspoof 2019. 66
4.9 Comparison of the proposed method with methods using different features and feature combinations on the ASVspoof 2019 dataset. . 68
4.10 Comparison of results obtained from the proposed method and the baselines on the ASVspoof 2021 dataset. 71
4.11 Results of using segmental frames of analysis of speech-pathological features with neural networks on the adaptation set of ADD 2022. . 73
4.12 Comparison of results obtained from the proposed method and the baselines on the ADD 2022 dataset. 74
4.13 Results of using segmental frames of analysis of speech-pathological features with neural networks on the adaptation set of ADD 2023. . 76

4.14 Comparison of results obtained from the proposed method and the baselines on the ADD 2023 dataset. 78
4.15 Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on ASVspoof 2019 dataset. 80
4.16 Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on on the adaptation set of ADD 2022. . . 81
4.17 Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on on the adaptation set of ADD 2023. . . 82

5.1 Results from applying an average of timbral features with a neural network on the development set of ASVspoof 2019. 87
5.2 Results from applying an average of timbral features with a neural network on the evaluation set of ASVspoof 2019 and 2021. 89
5.3 Results from applying an average of timbral features with a neural network on the adaptation set of ADD 2022. 91
5.4 Results from applying an average of timbral features with a neural network on the adaptation set of ADD 2023. 92
5.5 Ablation study of applying an average of timbral features with a neural network on the development set of ASVSpoof 2019. 94
5.6 Ablation study of applying an average of timbral features with a neural network on the adaptation set of ADD 2022. 96
5.7 Ablation study of applying an average of timbral features with a neural network on the adaptation set of ADD 2023. 97

6.1 Results of using acoustical and perceptual speech-pathological features with multi-layer perceptron neural networks on different datasets. 102
6.2 Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the development set of ASVspoof 2019. 104
6.3 Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the evaluation set of ASVspoof 2021. 105
6.4 Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the adaptation set of ADD 2022. 106

6.5 Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the adaptation set of ADD 2023. 107

List of Symbols/Abbreviations

ADD Audio deepfake detection

AFP Amplitude and frequency perturbation

ANN Artificial neural networks

ASR Automatic speech recognition

ASV Automatic speaker verification

Bi-LSTM Bi-directional long short-term memory

CAPE-V Consensus Auditory-Perceptual Evaluation of Voice

CHNR cepstral harmonics-to-noise ratio

CNN Convolution neural networks

CQCCs Constant Q Cepstral Coefficients

DCT Discrete cosine transform

DFT Discrete Fourier transform

DNN Deep neural network

EER Equal error rate

ERB Equivalent rectangular bandwidth

FFT Fast Fourier transform

HNR Harmonics-to-noise ratio

HILLs Human Log-Likelihoods

GNE Glottal-to-noise excitation ratio

GMMs Gaussian mixture models

KNN K-nearest neighbor

LCNN Light convolution neural network

LFCCs Linear-frequency cepstral coefficients

LPCs Linear prediction coefficients

LSTM Long short-term memory

ML Machine learning

MLP multilayer perceptron neural networks

MFCCs Mel-frequency cepstral coefficients

MSE Mean Squared Error

NNE Normalized noise energy

ResNets Residual neural networks

RF Random Forest

ROC Receiver Operating Characteristic

PD Parkinson disease

STLT Short-term long-term

STFT Short-time Fourier transform

STN Spatial transformer network

SVM Support vector machine

TCN Temporal convolutional network

TDNN Time delay neural network

Chapter 1

Introduction

1.1 Research Background

Recent advances in Artificial Intelligence (AI) are transforming many aspects of life. With its remarkable capabilities, AI brings numerous benefits across diverse fields, spanning from healthcare and finance to transportation and entertainment. Essentially, AI involves computers simulating human intelligence processes, enabling them to undertake tasks that conventionally demand human cognitive abilities, including learning, reasoning, and problem-solving. AI possesses a crucial advantage in boosting efficiency and productivity across various sectors. By automating and optimizing tasks, AI simplifies processes, decreases manual work, and decreases mistakes, resulting in considerable cost reductions and enhanced operational efficiency. In manufacturing, for example, robots powered by AI can execute repetitive and laborious tasks accurately and consistently, thereby augmenting production while maintaining quality control.

Moreover, AI facilitates informed decision-making based on data by analyzing vast amounts of complex information at speeds exceeding human capabilities. By extracting valuable insights and patterns from massive datasets, businesses can gain a competitive edge, improve strategies, and anticipate changes in the market. In the healthcare industry, AI-driven diagnostic platforms can assist healthcare professionals in accurately diagnosing illnesses, recommending personalized treatment plans, and predicting patient outcomes, ultimately saving lives and improving healthcare outcomes. Furthermore, artificial intelligence facilitates the creation of new and advanced products and services designed to meet the changing demands and tastes of consumers. Whether it's virtual assistants and chatbots offering individualized customer assistance or recommendation systems providing customized content and suggestions, AI elevates user satisfaction and encourages interaction with customers across a range of digital medium. Additionally, AI offers the po-

tential to tackle societal issues and foster sustainable development. Utilizing tools such as predictive analytics and intelligent resource management, AI can enhance resource distribution, minimize environmental hazards, and advance energy conservation, thereby fostering a more sustainable future. In the field of speech, AI has revolutionized it in numerous ways, offering a range of advantages and applications. For example, Speech Recognition facilitates precise and effective conversion of spoken language into written text, enabling computers to perform this task. This innovation finds utility in various domains such as voice-operated virtual aides, transcription tools, dictation software, and automated customer support. Second, Natural Language Processing (NLP) - AI-driven NLP algorithms facilitate machines in comprehending and deciphering human language. In speech, NLP assists in endeavors like analyzing sentiment, translating languages, and grasping semantics, fostering more authentic and significant interactions between humans and machines. Third, Speech Synthesis - AI facilitates the creation of natural synthetic speech, which is utilized in various domains such as text-to-speech (TTS) systems. These systems transform written text into spoken language, benefiting visually impaired individuals, language learning applications, and navigation systems in automobiles. Fourth, Healthcare Applications - In this application, AI aids in analyzing speech to diagnose and track medical conditions like speech disorders, cognitive impairments, and neurological diseases. Lastly, Voice Biometrics - AI enables the utilization of voice biometrics for authentication and security objectives. Voiceprints serve to authenticate an individual by their distinct vocal traits, thereby bolstering security measures in realms like access control, banking, and online commerce.

Despite the numerous advantages offered by AI speech, such as its ability to streamline communication processes and enhance accessibility, the misuse of these technologies, known as deepfake speech, poses a significant threat to economies and societies worldwide. For example, criminals take advantage of speech synthesis applications to cheat voice biometric systems, such as automatic speaker verification (ASV) systems. They exploit the advanced capabilities of these applications to mimic the voices of others, circumventing security measures meant to authenticate users based on their unique vocal patterns. Therefore, detecting deepfake speech is crucial for fraud protection and ensuring the reliability of ASV systems.

In this dissertation study, techniques for deepfake speech detection are explored, aimed at improving the efficiency of current methods in detecting deepfake speech.

Detecting deepfake speech has involved using several advanced techniques primarily focusing on two approaches: creating efficient classifiers [10, 11, 12] and exploring acoustic features [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. In the first

approach, various classifiers have been used, including Gaussian mixture models (GMMs) [24], deep neural networks [25], recurrent neural networks (RNNs) [26], convolution neural networks (CNNs) [27], and residual neural networks (ResNets) [28]. The selection of these classifiers might depend on the characteristics and dimensions of the features. For example, features with small dimensions are suitable for traditional machine-learning models, while those with large dimensions are better handled with deep-learning models, such as CNNs, RNNs, and ResNets [29].

The second approach is focused on using speech and acoustic features as front-end features [7]. Numerous features have been used for detecting deepfake speech, including spectrograms, linear-frequency cepstral coefficients (LFCCs) [30], mel-frequency cepstral coefficients [31], constant-Q transform [32], and constant-Q cepstral coefficients [33]. For example, Yi [9] and Wang [34] independently proposed deepfake-detection methods using LFCCs with GMM. These features are represented in phase, power spectrum, and cepstral coefficients. These features, however, were used without thoroughly clarifying their potential for distinguishing between genuine and deepfake speech.

Speech-pathological features, on the other hand, have been introduced to detect the unnatural characteristics of synthesized audio [35, 23]. Speech-pathological features are crucial components closely intertwined with the complex human speech-production mechanisms, representing relevant acoustic, phonatory, and aerodynamic parameters. Speech-language pathologists and otolaryngologists typically use these features to distinguish between normal and disordered voices [36]. In combination with machine-learning algorithms, speech-pathological features are also used in automatic voice assessment and evaluation systems. These systems assist healthcare professionals and medical doctors in classifying, diagnosing, assessing the severity of, and identifying the types of voice disorders [37, 38, 39]. However, the study of pathological features for deepfake speech detection is limited. A method proposed by Kai *et al.* uses only a few features and is used for fake audio detection [23]. Therefore, a comprehensive investigation of the potential of speech-pathological features in distinguishing between genuine and deepfake speech is necessary.

1.2 Research Problems

Nowadays, deepfake speech is a misuse of artificial intelligence (AI) technologies that synthesizes speech through advanced voice conversion (VC) and text-to-speech (TTS) techniques. It generates audio waveforms so realistic that they can fool humans, closely approximating natural human voices. This poses a significant threat to economies and societies. By convincingly replicating voices and producing artificial speech, deepfake technology undermines confidence in audio recordings and verbal exchanges, posing a significant risk of disseminating misinformation and causing widespread confusion. This trend not only threatens the credibility of public discussions but also raises apprehensions regarding privacy violations and the manipulation of prominent individuals.

There are the example of using deepfake to commit fraud. For example, in 2019, criminals have exploited deepfake speech to impersonate a CEO's voice, successfully defrauding over USD $243,000$ [40]. Moreover, in 2020, The bank manager in Hong Kong was contacted by someone resembling a company director with whom he had previously conversed. This individual requested the manager to authorize transfers amounting to $35 million. Trusting their prior interaction, the manager initiated transfers totaling $400,000 before becoming suspicious. Subsequently, it was discovered that the manager had fallen victim to a sophisticated scheme wherein fraudsters utilized deepfake technology to replicate the director's voice [41].

Therefore, the development of robust speech-security systems capable of accurately detecting deepfake speech is crucial. These systems serve as a primary defense, preventing the harmful consequences resulting from the misuse of such technology. In today's digital era, where misinformation spreads rapidly and extensively, there is an urgent need for proactive measures to maintain communication integrity. Consequently, concerted research and innovation efforts are essential to strengthen defenses against this ever-evolving threat. By doing so, trust and authenticity in discourse can be upheld, thereby preserving the fundamental principles of a healthy and well-informed society.

Numerous approaches and methodologies have been proposed to discern between genuine and deepfake speech. However, their effectiveness necessitates significant refinement, especially given the ongoing advancements in speech synthesis technology. With the emergence and evolution of new techniques, it becomes increasingly crucial to integrate linguistic patterns, prosody, and spectral characteristics into detection algorithms. The endeavor to consistently detect deepfakes grows increasingly intricate as creators employ progressively sophisticated techniques to emulate authentic speech patterns and subtleties. Therefore, sustained efforts aimed at enhancing these detection capabilities through interdisciplinary collaboration and technological innovation are imperative to effectively address

the escalating complexity of deepfake production and mitigate its potential societal implications. These motor instructions govern the intricate coordination of various components involved in vocalization, including the lungs, vocal cords, tongue, lips, and jaw. Together, they form what is known as the articulatory phase of speech production. The lungs play a crucial role by facilitating the airflow necessary for sound production. Meanwhile, the vocal cords vibrate to create the initial sound source. Subsequently, the tongue, lips, and jaw work in harmony to shape and modulate this sound into intelligible speech. This intricate process underscores the remarkable coordination required for the human capacity for spoken communication, showcasing the complexity of the vocal mechanism.

1.3 Research Motivation

This research is motivated from the following human speech production mechanism. Because the human speech production mechanism is intriguing, encompassing physiological and neurological elements. It begins with the conceptualization of a linguistic message in the mind, known as the cognitive-linguistic stage. Here, thoughts and ideas are formulated and subsequently translated into linguistic structures. he encoded message is then transformed into a series of motor commands in the motor planning stage, a complex process requiring meticulous coordination. This involves not only the selection and sequencing of specific phonemes, which are the smallest units of sound, but also the orchestration of intricate muscle movements necessary for articulating these sounds into coherent words and sentences. The brain must meticulously decide which muscles to activate, delicately orchestrating their timing, force, and sequence to ensure accurate communication. The final phase in the process of delivering a speech involves the transmission of sound waves containing the spoken message to the listener. As these waves travel through the air, they undergo changes influenced by various factors like environmental conditions, distance, and obstacles. These waves carry not only the literal message but also convey subtle information about the speaker's emotions, culture, and even physical traits. Elements like pitch, rhythm, and tone in speech play a significant role in how the message is understood by the listener, adding complexity to communication. This stage of sound transmission is crucial for conveying meaning and fostering connection beyond mere words, facilitating deeper understanding and resonance between individuals.

From the aforementioned, the complexities inherent in human speech production mechanisms pose a significant obstacle for artificial emulation, as each speaker exhibits a unique constellation of individual characteristics. From the meticulous orchestration of muscles within the vocal tract to the subtle modulation of pitch, tone, and rhythm, the human voice embodies a plethora of nuanced variations

and peculiarities. Moreover, elements such as accent, cadence, and even emotional expression augment the rich mosaic of human communication, rendering each individual's speech production a distinct phenomenon. Consequently, endeavors to replicate this complexity in artificial systems require an in-depth comprehension not only of the physiological processes implicated but also of the intricate interplay between cognition, emotion, and linguistic experience. Thus, the quest to reproduce human speech production in artificial systems constitutes a captivating frontier in the disciplines of artificial intelligence and cognitive science, promising profound revelations about the nature of human communication and the quintessence of individual identity.

Although advances in AI speech are rapidly evolving, with researchers continually developing new techniques and algorithms to improve the realism and expressiveness of synthesized speech, replicating the human speech production mechanism remains constrained and challenging. AI speech struggles to capture the unique characteristics that make each person's voice distinct. Furthermore, they are typically trained on large datasets of recorded speech, which may not encompass the diversity of human speech patterns across different languages, dialects, and cultures.

In this research, speech-pathological features are used to detect deepfake speech, which can also be utilized to identify voice disorders resulting from abnormalities in the human speech production mechanism, and which manifest as unnaturalness. Deepfake speech is characterized by its unnaturalness; hence, the hypothesis suggests that deepfake speech could possibly be the perceived acoustic quality of the disordered voice. Therefore, speech-pathological features can be crucial clues for deepfake speech detection based on human speech production.

1.4 Research Goals

The goal of this research is to propose a scheme of deepfake speech detection based on speech-pathological features. Inspired by the human speech product, speech-pathological features are crucial components closely intertwined with the complex human speech-production mechanisms, representing relevant acoustic, phonatory, and aerodynamic parameters. Speech-language pathologists and otolaryngologists typically use these features to distinguish between normal and disordered voices. This research investigates the possibility of using speech-pathological features that detect voice disorders for distinguishing between deepfake and genuine speech. In this dissertation, the following speech-pathological speech features are investigated: jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). The speech-pathological features mentioned are analyzed to discuss the

advantages and disadvantages of each feature. The significant speech-pathological features are applied with the classifiers, for example, other machine learning models, or deep learning algorithms, to distinguish between genuine and deepfake speech. In this dissertation, there are three research questions as follows:

- To investigate the pathological features used to distinguish between normal and pathological voices whether they can be used to detect deepfake speech.
- To improve the efficiency of using speech-pathological features for detecting deepfake speech.
- To find out the significant speech-pathological feature for detecting deepfake speech among those mentioned above.

1.5 Challenges

There are several challenges in to detecting deepfake speech in currently.

- **Unseen data**

  Unseen data poses significant challenges for deep fake speech detection primarily because it encompasses instances or scenarios that the model hasn't been exposed to during training. Typically, deep fake detection models undergo training on datasets comprising examples of both authentic and fabricated speech. Nonetheless, the sheer diversity of potential modifications in fake speech renders it impractical to train a model on every conceivable variation. Confronted with unseen data, such as novel techniques for generating deep fake speech or alterations that were not encountered during training, the model may encounter difficulty in accurately discerning between genuine and fabricated speech. This limitation arises due to the model's constrained capacity to generalize from the training data to unobserved instances. As a result, the effectiveness of the model in distinguishing between real and fake speech may diminish when confronted with unseen data.

- **Background noise**

  The presence of background noise presents significant challenges in identifying deepfake speech. This noise, which comprises unwanted electrical signals infiltrating a communication system through its medium, disrupts the intended message. Common sources of such interference include human voices, vehicle sounds, rainfall, and wind. Even recorded voices, whether captured indoors or outdoors, are susceptible to real-world disturbances like laughter and rain. Unfortunately, attackers can exploit these natural noises

to manipulate detection systems, underscoring the critical need for robustness in fake voice detection technologies. Regrettably, minimal research has been devoted to this issue within the realm of deepfake detection. Previous endeavors have failed to adequately address the impact of real-world noise using proposed detection methods that are still under development. Addressing this gap could serve as a promising entry point for researchers seeking to develop a robust fake audio detection approach capable of functioning effectively despite encountering noisy data in real-world settings.

1.6 Organization of Thesis

As shown in Fig. 1.1, this thesis consists of seven chapters. Apart from the introduction chapter, the remaining chapters are organized as follows.

Chapter 2 presents a comprehensive literature review related to this study. It begins by introducing the introduction of deepfake speech detection, human speech production mechanism, followed by an introduction to the method of voice disorder assessment. Lastly Chapter 2, the features used for assessing voice disorders include perceptual features and acoustical features are described.

Chapter 3 the contribution of using perceptual and acoustical speech-pathological features to detect fake speech. After that, it describes the philosophy of this dissertation, which involves using speech-pathological features for deepfake speech. Then, the perceptual and acoustical speech-pathological features are explained. Lastly, the dataset and metrics used in this dissertation are presented.

Chapter 4 presents a method for detecting fake speech using acoustical speech-pathological features. It details experiments conducted on ASVspoof 2019, ASVspoof 2021, ADD 2022, and ADD 2023 datasets, analyzing the results and discussions. Additionally, it identifies the key acoustical speech-pathological features crucial for fake speech detection. Furthermore, the chapter introduces a novel approach employing segmental analysis of acoustical speech-pathological features for fake speech detection. This includes the proposed method, results, and discussions based on experiments conducted on ASVspoof 2019 and ASVspoof 2021 datasets, along with additional data from ADD 2022 and ADD 2023. It then identifies the critical segmental frames of analysis within the acoustical speech-pathological features that are most effective for detecting fake speech. Finally, the chapter concludes with a summary.

Chapter 5 demonstrates a method for detecting fake speech using perceptual speech-pathological features and presents the results and discussion of experiments conducted on ASVspoof 2019, ASVspoof 2021, ADD 2022, and ADD 2023. Additionally, it identifies the important of perceptual speech-pathological features in detecting fake speech. Finally, a summary is provided.

Chapter 6 demonstrates a method for detecting fake speech using the important acoustical and perceptual speech-pathological features and presents the results and discussion of experiments conducted on ASVspoof 2019, ASVspoof 2021, ADD 2022, and ADD 2023. Finally, a summary is provided.

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

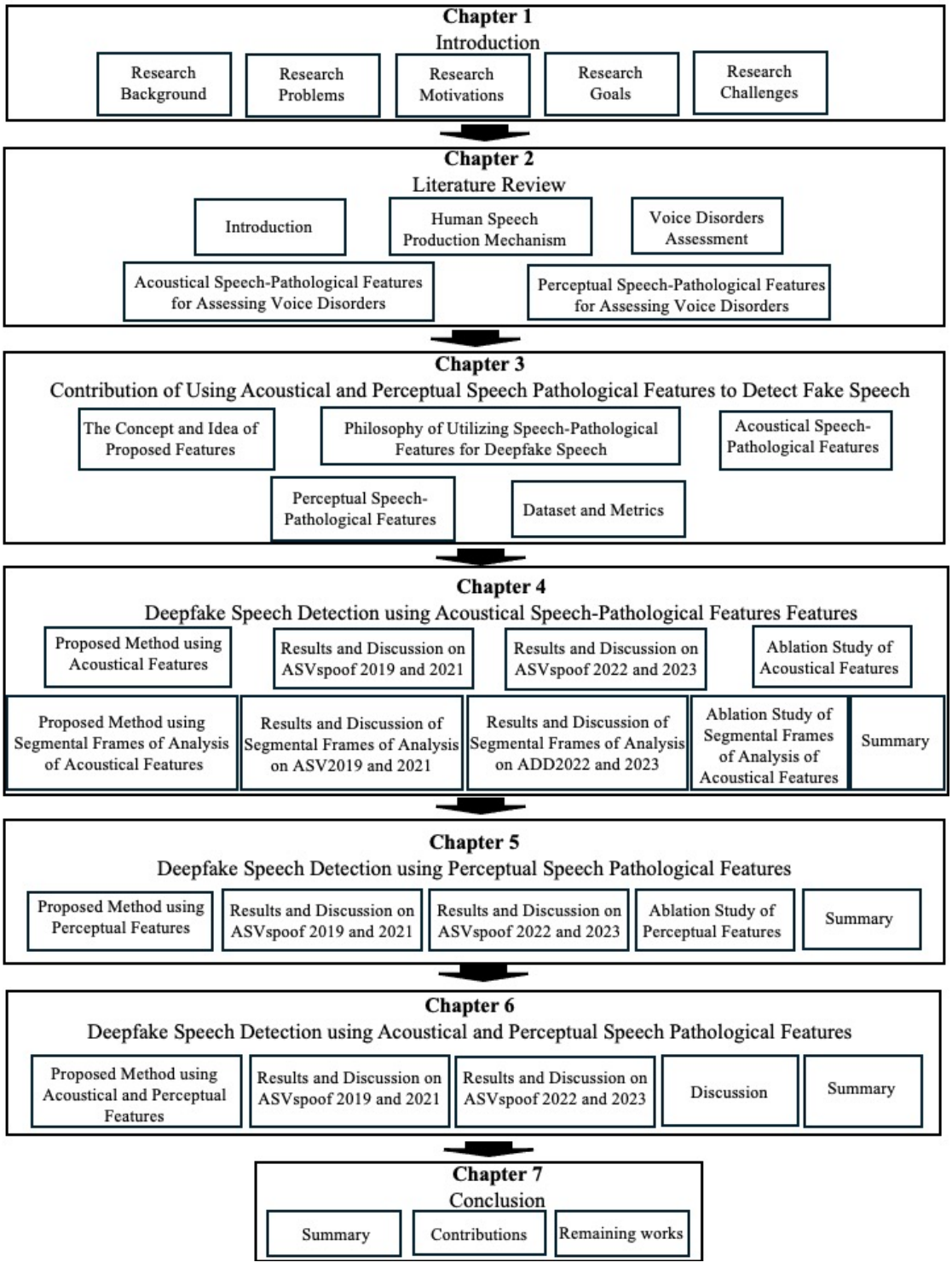

Figure 1.1: Organization of this dissertation.

Chapter 2

Literature Review

2.1 Introduction

Detecting deepfake speech has involved the utilization of several advanced techniques, with primary emphasis on two main parts: the development of efficient classifiers [10, 11, 12] and the exploration of various acoustic features [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. Figure 2.1 shows two main part of the method for fake speech detection.

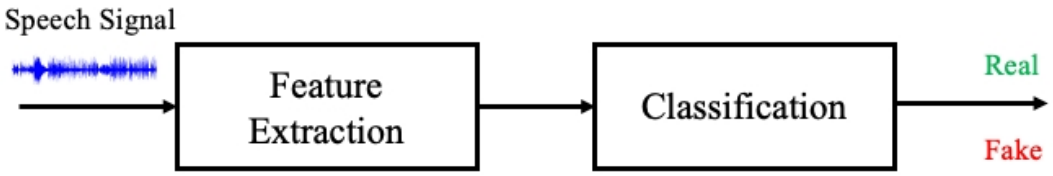

Figure 2.1: General method for fake speech detection.

In the first part, various classifiers have been employed, including GMMs [24], deep neural networks [25], RNNs [26], convolution neural networks (CNNs) [27], and residual neural networks (ResNets) [28]. The selection of these classifiers might depend on the characteristics and dimensions of the features. For example, features with small dimensions are suitable for traditional machine-learning models, while those with large dimensions are better handled with deep-learning models, such as CNNs, RNNs, and ResNets [29].

The second approach is focused on using speech and acoustic features as front-end features [7]. Numerous features have been used for detecting deepfake speech, including spectrograms, linear-frequency cepstral coefficients (LFCCs) [30], mel-

frequency cepstral coefficients [31], constant-Q transform [32], and constant-Q cepstral coefficients [33].

There are various researchers to proposed the method for fake speech detection. For example, Yu et al. [42] proposed a new scoring method called Human Log-Likelihoods (HLLs), which utilizes a Deep Neural Network (DNN) classifier to improve the detection rate. They compared this method with a traditional scoring approach known as Log-Likelihood Ratios (LLRs), which relies on the GMM. DNN-HLLs and GMM-LLRs were evaluated using the ASV Spoof Challenge 2015 dataset, with features extracted automatically. These evaluations confirmed that DNN-HLLs outperformed GMM-LLRs, achieving an Equal Error Rate (EER) of 12.24%. Moreover, Yi [9] and Wang [34] independently proposed deepfake-detection methods using LFCCs with GMM. Borrelli et al. [43] devised a SVM model coupled with Random Forest (RF) to forecast synthetic voice Automatic Speaker Verification (ASV) spoofing challenges using the dataset from the 2019 challenge [7]. Their experiments were based on a novel audio feature termed Short-Term Long-Term (STLT). The findings revealed that the SVM outperformed the RF model by 71% in terms of performance. Liu et al. [20] compared the robustness of SVM with a deep learning method called CNN in detecting fake stereo audio from real ones. The comparison revealed that CNN is more robust than SVM, despite both achieving a high accuracy of 99% in detection. However, SVM encountered issues similar to those faced by the Logistic Regression (LR) model during the feature extraction process.

Based on the works examined up to this point, it is evident that the features in ML models must be extracted manually, requiring intensive preprocessing before training to achieve satisfactory performance. Nonetheless, this process is time-consuming and susceptible to inconsistencies, prompting the research community to explore advanced DL techniques. To address this, Subramani and Rao [44] developed an approach to detect synthetic audio using two CNN models: EfficientCNN and RES-EfficientCNN. In their study, RES-EfficientCNN outperformed EfficientCNN, achieving a higher F1-score of 97.61% compared to 94.14% F1-score for EfficientCNN, as demonstrated on the ASVspoof 2019 dataset. Another adaptation of CNN model, Zhenchun Lei et al. [45], proposed a 1-D CNN and a Siamese CNN for detecting fake audio. In the case of the 1-D CNN, the model's input comprised speech log-probabilities, while the Siamese CNN utilized two trained GMM models. The Siamese CNN consisted of two identical CNNs, akin to the 1-D CNN, but concatenated them using a fully connected layer with a softmax output layer. Both models underwent testing on the ASVspoof 2019 dataset, revealing that the proposed Siamese CNN surpassed the GMM and 1-D CNN by improving the EER when utilizing LFCC features. However, performance slightly declined when employing CQCC features. Additionally, it was noted that the model lacked sufficient

robustness and was tailored to work with specific types of features. One limitation of this study is that it utilized many layers and convolutional networks, resulting in management complexities.

Chintha et al. [46] developed two novel models that depend on a convolutional RNN for audio Deepfake classification. First, the Convolutional Recurrent Neural Network Spoof (CRNN-Spoof) model contains five layers of extracted audio signals that are fed into a bidirectional LSTM network for predicting fake audio. Second, the Wide Inception Residual Network Spoof (WIRE-Net-Spoof) model has a different training process and uses a function named weighted negative log-likelihood. The CRNN-Spoof method obtained higher results than the WIRE-Net-Spoof approach by 4.27% EER in the ASVspoof 2019 dataset. To address this limitation, Shan and Tsai [35] proposed an alignment technique based on classification models: Long Short-Term Memory (LSTM), bidirectional LSTM, and transformer architectures. The technique classifies each audio frame as either matching or nonmatching from a set of 50 recordings. The reported results indicate that bidirectional LSTM outperforms the other models, achieving a 99.7% accuracy and 0.43% EER. However, it is noted that the training process was time-consuming, and the dataset used in the study was small, potentially leading to overfitting.

In regard to transfer learning and unimodal methods, Aravind et al. [47] proposed a new framework based on transfer learning and the ResNet-34 method for detecting faked English-speaking voices. The transfer learning model was pretrained on the CNN network. The ResNet-34 method was employed to address the vanishing gradient problem that commonly arises in deep learning models. The results indicated that the proposed framework achieved the best performance, as measured by EER, with a result of 5.32%. However, it should be noted that training with ResNet-34 can be time-consuming due to its deep architecture. Similarly, Khochare et al. [48] investigated feature-based and image- based approaches for classifying faked audio generated synthetically. New DL models called the Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN) were used in this work. TCN achieved promising outcomes in distinguishing between fake and real audio with 92% accuracy, while STN obtained an accuracy of 80%. Although the TCN works well with sequential data, it does not work with inputs converted to Short-Time Fourier Transform (STFT) MFCC features. A novel audio feature descriptor, introduced by Arif et al. [49] named ELTP-LFCC, combines Local Ternary Pattern (ELTP) and LFCC. This descriptor was integrated into a Deep Bidirectional Long Short-Term Memory (DBiLSTM) network to enhance model robustness for detecting counterfeit audio across various indoor and outdoor settings. The model underwent testing on the ASVspoof 2019 dataset, which encompasses both synthetic and imitation-based fake audio. The results revealed

superior performance on synthetic audio (0.74% EER), while imitation-based samples exhibited lower performance (33.28% EER). A method called ASSERT (Anti-Spoofing with Squeeze-Excitation and Residual neTworks) was introduced by Lai et al. [50], leveraging variations of the Squeeze-Excitation Network (SENet) and ResNet. This approach utilizes log power magnitude spectra (logspec) and CQCC acoustic features for DNN training. Evaluation on the ASVspoof 2019 dataset revealed that ASSERT achieved over a 17% relative improvement in synthetic audio detection. However, during testing, the model exhibited zero EER in a logical access scenario, indicating a significant degree of overfitting.

Based on the literature, the selection of classifiers depends on the characteristics and dimensions of the features. For instance, features with small dimensions are suitable for traditional machine learning models, whereas those with large dimensions are better handled by deep learning models such as CNNs, RNNs, and ResNets. Additionally, these deep learning models are commonly used in combination for detecting fake speech.

In the field of features extraction, these features are represented in phase, power spectrum, and cepstral coefficients, and other acoustic characteristics. These features, however, were used without thoroughly clarifying their potential for distinguishing between genuine and deepfake speech.

In this study, an in-depth investigation was conducted into the potential features associated with the mechanism of speech production. The aim was to enhance the detection of deepfake speech, a growing concern in today's digital communication landscape. By understanding these features, researchers hope to develop more robust and reliable methods for identifying and mitigating the impact of deepfake speech.

2.2 Human Speech Production Mechanism

Speech, as the innate mode of human communication, stands as the fundamental and widely employed means of conveying thoughts and ideas. To the average person, speech may simply constitute the audible vibrations emitted from the mouth and interpreted through the ears. However, its generation involves intricate processes. Understanding the mechanisms behind human speech production and perception holds paramount significance, essential for advancing technologies such as hearing aids, cochlear implants, speech recognition, enhancement, simulation, and modeling. The mechanism of speech production comprises three primary functions, depicted in Fig. 2.2 through a block diagram.

Motor control pertains to the cognitive process orchestrated by the human brain, initiating the formation of speech and subsequently transmitting control signals via sensory nerves to the speech production organs. Upon receipt of these

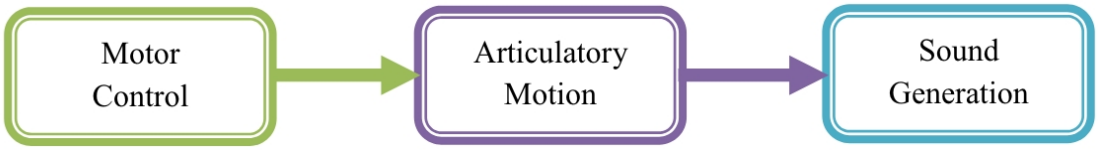

Figure 2.2: Block diagram of human speech production mechanism [1].

signals from the motor control unit, the speech production organs adapt and assume the appropriate configurations to articulate the intended words or sounds. Next, the concept of articulatory motion will be further elucidated in subsequent paragraphs. The third element of the human speech production mechanism is speech generation, involving the expulsion of air from the mouth and nasal cavity, producing acoustic waves released into the surrounding space. Concerning speech perception, the acoustic waves generated by the mouth and nasal cavity reach the human ear and are interpreted through sensory nerves linking the ear to the brain. This paper exclusively concentrates on the speech production mechanism, excluding discourse on speech perception.

2.2.1 Motor control function

The process of controlling speech production entails the coordination of numerous muscles and structures responsible for generating speech sounds. It is a multifaceted process that harmonizes neural signals from the brain with muscular actions to produce distinct sounds and speech patterns. The process of motor control unfolds as follows: speech production commences with the activation of specific brain regions, predominantly within the left hemisphere for the majority of right-handed individuals (including Broca's area and the motor cortex). These regions take charge of planning and initiating the movements necessary for speech. The brain devises a sequence of movements required to articulate the intended speech sounds, meticulously selecting the appropriate muscles and coordinating their actions. Once the motor plan is set, signals are dispatched from the brain to the muscles involved in speech production, traversing through the motor pathways in the nervous system until they reach the muscles of the vocal tract. Upon receiving signals from the nervous system, these muscles contract in synchronization to execute the desired movements. Distinct muscles are engaged depending on the specific speech sounds being articulated. For instance, the lips, tongue, jaw, and vocal cords each contribute significantly to shaping the vocal tract to produce varied sounds. Throughout the speech production process, the brain continually receives feedback from diverse sources, including auditory feedback (such as hearing one's own voice), proprioceptive feedback (sensations from muscles and joints),

and visual feedback (observing one's own articulatory movements). This feedback is utilized to monitor and adjust speech production in real-time, ensuring precision and clarity. Moreover, speech production involves fine-tuning of movements based on contextual factors like speaking rate, emphasis, and coarticulation (the interplay between adjacent sounds). This meticulous adjustment enables smooth and natural speech production.

2.2.2 Articlulatory motion

Several organs play a crucial role in the production of speech and sound in human beings. These organs possess flexibility and can change their shape and size in response to motor control signals from the brain, depending on the type of speech or sound needed. The lungs supply the necessary airflow to create sound in the form of acoustic waves. This airflow travels through the vocal tract, which connects the lungs to the throat, as well as through the vocal cords, glottis, epiglottis, and other mouth organs before being emitted through the mouth and nasal cavities as acoustic waves. Figure 2.3 illustrates the various organs through which airflow passes during the process of generating speech and sound.

The process of motor control unfolds in the following manner: during speech, air expelled from the lungs ascends through the trachea and enters the larynx. Within the larynx, the air encounters vocal cords, which are paired lip-like tissues responsible for determining speech pitch. These vocal cords, characterized by their pearly white appearance, are attached at one end to the arytenoid cartilages at the rear and to the thyroid cartilage at the front. Typically, males possess larger larynxes, resulting in lower-pitched voices, whereas females tend to have smaller larynxes, leading to higher-pitched voices. The length of vocal folds varies between genders, ranging from 17 to 25 mm in males and 12.5 to 17.5 mm in females. Voiced speech is produced by the vibration of the vocal folds, while unvoiced speech is generated by their momentary restriction. In addition to basic speech, vocal cords adjust their opening and closing in various patterns, known as phonemes, enabling airflow through the upper part of the vocal tract. This tract, resembling a tube, extends from the glottis to two openings: the oral and nasal cavities. Its non-uniform cross-section spans approximately 17 cm in males, branching out at the soft palate halfway through the tract and opening up at the nostrils as a secondary branch, measuring approximately 13 cm. As air exits the vocal cords, it enters the pharyngeal, oral, and nasal cavities, where resonance occurs. These cavities amplify certain frequencies and dampen others, thereby shaping the sound according to the intended words. Additionally, various organs within the mouth including the soft palate, teeth, tongue, lips, and jaw, adjust their positions to regulate airflow through the mouth and nose, thus modulating the shape and volume of the sound. Due to differences in the size and shape

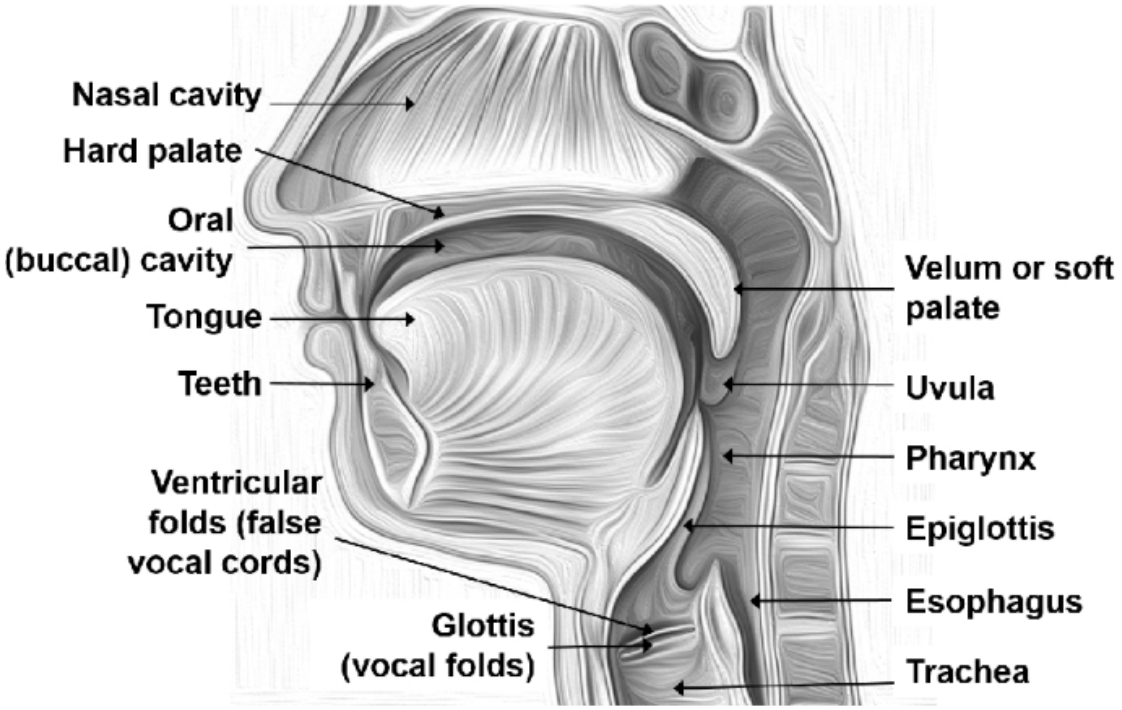

Figure 2.3: Human vocal apparatus [2].

of these speech production organs, each individual's speech possesses a unique quality. Remarkably, despite the complexity of these articulatory movements, the system reacts rapidly to adjust to changing speech parameters. Furthermore, the epiglottis and false vocal cords below the pharynx play a crucial role in preventing food from entering the larynx and isolating the esophagus from the vocal tract acoustically [1].

2.2.3 Sound generation

In the sound generator unit, often referred to in linguistic terms as phonemes, plays a crucial role in forming spoken words. Phonemes represent the most basic sounds within a language, capable of altering meaning. The phonemes of a language encompass contextual nuances, emotions, and the unique traits of the speaker, all of which contribute to their pronunciation, though such intricacies are not necessary in written text. These phonemes are primarily constructed based on the articulatory gestures of the vocal tract.

The phonetics of any language typically consist of two main types of phonemes: vowels and consonants. Vowels are always voiced sounds, while consonants can be either voiced or voiceless. Voiced sounds occur when the vocal cords vibrate regularly as air passes through them, with a fundamental frequency of about 110 Hz for men, 200 Hz for women, and 300 Hz for children. Apart from the fundamental frequency, the movements of the speech production organs generate resonance frequencies specific to each phoneme. These resonance frequencies, denoted as N number, including $F1$, $F2$, ..., Fn, are termed as Formant Frequencies. For adult males, the typical range of formant frequencies is as follows: $F1$ = 180 to 800 Hz, $F2$ = 600 to 2500 Hz, $F3$ = 1200 to 3500 Hz, and $F4$ = 2300 to 4000 Hz. Conversely, unvoiced sounds exhibit a completely random nature. During the generation of voiceless sounds, the vocal cords can be either fully open, fully closed, or partially open. Vowel phonemes are generated by the frequent vibration of the vocal cords. These phonemes are categorized into three types based on the position of the tongue in the oral cavity: *Front*, including sounds like /IY/, /IH/, /EY/, and /EH/; *Mid*, such as /AA/ and /ER/; and *Back*, which comprises sounds like /AE/, /AO/, /UH/, /OW/, and /AH/. Consonants can be categorized as either voiced or unvoiced, and they are further divided into Nasal, Stop or Plosive, Fricative, and Affricate sounds. Nasal sounds, such as /M/, /N/, and /NG/, occur when the mouth cavity is closed, and the air passes through the nasal cavity via the open velum. Plosive sounds, like /P/, /B/, /T/, /K/, /D/, and /G/, occur when pressure builds up behind the vocal cords and is suddenly released upon momentary closure. Among these, /B/, /D/, and /G/ are voiced, while /P/, /T/, and /K/ are unvoiced. Fricative sounds, such as /HH/, /F/, /V/, /TH/, /DH/, /S/, /Z/, /SH/, and /ZH/, are produced when the mouth cavity is not

fully blocked, allowing for a quasi-periodic flow of air due to vocal cord vibrations. Affricate sounds, like /CH/ and /JH/, result from a combination of plosive and fricative actions [1].

2.3 Voice Disorders Assessment

A voice disorder is a condition that affects the ability to produce a clear, normal voice. Voice disorders can impact a person's ability to communicate effectively in their daily life. Common causes of voice disorders include overuse of the vocal cords, such as from excessive yelling, singing, or speaking loudly for long periods, growths or abnormalities on the vocal cords, and neurological issues that affect the muscles involved in speech production. When the vocal cords are unable to vibrate properly, it results in alterations in voice quality such as hoarseness, breathiness, strain, or challenges in maintaining control over vocal volume and pitch. These voice changes can negatively impact a person's ability to communicate clearly and be understood by others during speech production.

Research has delved into voice disorders and speech production. Deary *et al.* [51] discovered a correlation between self-reported voice issues and personality traits as well as psychological distress, indicating a subjective interpretation of voice quality. While, Niimi *et al.* [52] examined the influence of neuromuscular diseases on voice and speech disorders, emphasizing the crucial role of the neuromuscular system in speech production. Next, Chen *et al.* [53] enhanced speech production models through an analysis of disordered speech in diverse groups, identifying distinctive characteristics. The research of Dietrich *et al.* [54] delved deeper into the neural control of phonation under stress, uncovering how limbic-motor interactions affect speech production. Together, these studies highlight the intricate nature of voice disorders and their connection to speech production.

2.3.1 Assessment of Voice Disorders by Clinicians

The evaluation of disordered voice generally requires a thorough, multidisciplinary strategy, integrating both standardized and non-standardized assessment. The auditory-perceptual evaluation method is widely utilized in clinical settings and is often regarded as the benchmark for assessing voice quality [55].

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) is extensively employed as a standardized instrument, offering structured guidelines for the perceptual assessment of atypical vocal characteristics [3].

The example of consensus CAPE-V form as shown in Fig. 2.4. This assessment serves as the foundation of the evaluation process, during which a clinician assesses

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:_____ **Date:**_____

The following parameters of voice quality will be rated upon completion of the following tasks:
1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant
MO =Moderately Deviant
SE = Severely Deviant

			SCORE
Overall Severity	_____ MI MO SE	C I	/100
Roughness	_____ MI MO SE	C I	/100
Breathiness	_____ MI MO SE	C I	/100
Strain	_____ MI MO SE	C I	/100
Pitch	(Indicate the nature of the abnormality): _____ _____ MI MO SE	C I	/100
Loudness	(Indicate the nature of the abnormality): _____ _____ MI MO SE	C I	/100
_____	_____ MI MO SE	C I	/100
_____	_____ MI MO SE	C I	/100

COMMENTS ABOUT RESONANCE: NORMAL OTHER (Provide description):_____
_____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

Clinician:_____

Figure 2.4: Example Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Form. [3].

various aspects of voice quality, including overall severity, roughness, breathiness, strain, pitch, and other characteristics.

Another crucial aspect of voice assessment involves the utilization of stroboscopy, a diagnostic technique enabling thorough examination of the oscillating vocal folds. This visual inspection yields valuable insights into vocal fold performance, aiding in the detection of root causes behind voice disorders. The evaluation of voice disorders is most effectively carried out through a team effort, with collaboration between otolaryngologists (specialists in ear, nose, and throat conditions) and speech-language pathologists. Through the utilization of an extensive array of assessment tools and the collaboration of a multidisciplinary team, clinicians can gain a comprehensive insight into the patient's voice disorder, thereby facilitating the development of more efficient treatment and management approaches.

2.3.2 Automatic Voice Disorders Assessment

Instead of solely relying on clinicians' expertise to assess voice disorders, the utilization of machine learning (ML) algorithms holds promise in aiding clinicians in analyzing and evaluating the effectiveness of treatments for voice disorders. One important approach entails utilizing classifier model algorithms to detect and assess voice abnormalities through analyzing acoustic speech signals.

Researches have investigated various machine learning methods, including support vector machines, neural networks, and decision trees, to categorize voice recordings as either disordered or healthy. These algorithms usually gather various speech characteristics such as pitch, formants, cepstral coefficients, and measures of voice quality. These collected features are subsequently utilized as inputs for the classification models [56, 4]. They introduce an automatic voice disorder detection (AVDD) system, which can be helpful for both patients and laryngologists. This work explains the process of developing the AVDD system using machine learning, as shown in Fig. 2.5. The essential procedures entail several key stages: initially, designated voice recordings stored as audio files undergo manual labeling by experts to distinguish between healthy and pathological voices. Following this, the audio data within each file are segmented into brief frames, with each frame being processed to extract pertinent features. Subsequently, the gathered set of features extracted from all frames serves as input for machine learning algorithms. The dataset is then divided into training and testing sets, with observations randomly chosen from both normal and pathological voice categories. The training set is used to develop the ML model, while the testing set is employed to evaluate its performance. Throughout the assessment phase, the classification accuracy is computed, serving as a metric to evaluate the efficacy of various AVDD systems.

The two key points of the AVDD systems are the feature extraction and ML

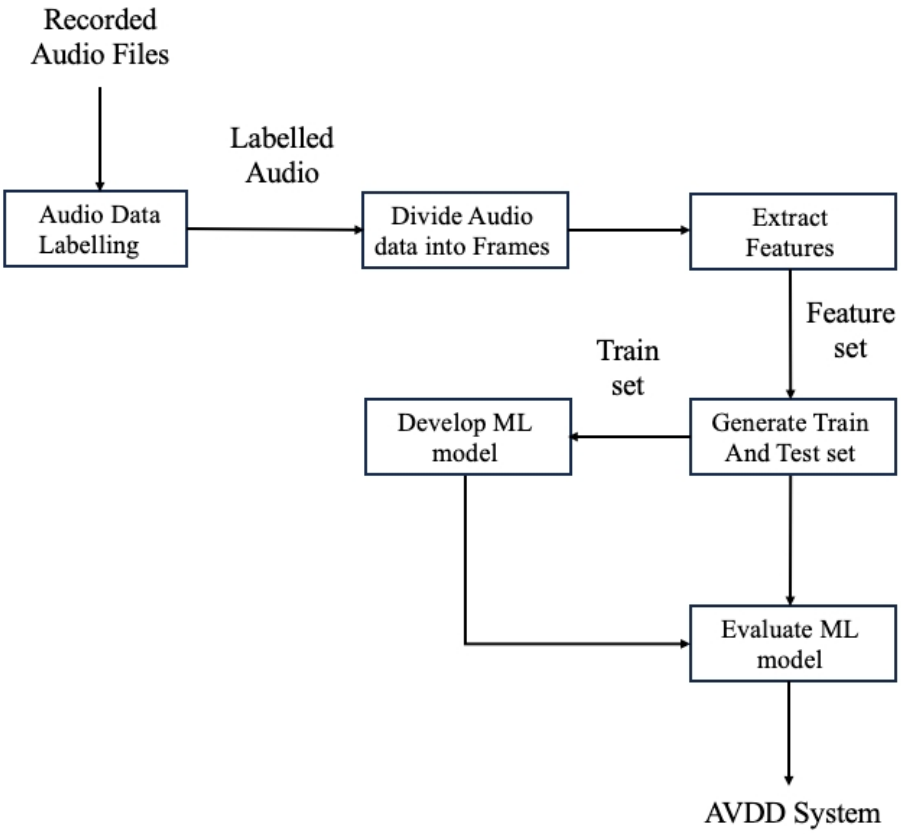

Figure 2.5: Automatic voice disorder detection (AVDD) system [4].

models. Several research studies are developing methods to evaluate voice disorders. Behroozmand and Almasganj [57] have investigated the role of energy and entropy features extracted using wavelet packet decomposition for the speech signal with unilateral vocal fold paralysis. The extracted features are optimized using a genetic algorithm and classified using Support vector machine (SVM) with a linear kernel. An accuracy of 100% is achieved with entropy features and 93.62% with energy features. Cairns *et al.* [58] have proposed a noninvasive method to detect hypernasality in speech based on a nonlinear operator. They classified normal and hypernasal voices based on probability distribution functions. The maximum classification accuracy was 94.7%. Gomez *et al.* [59]. has conducted experiments to detect pathological voices in the larynx (such as polyps, nodules, cysts, sulcus, edemas, carcinoma, etc.) using two neural network classification approaches, namely Multilayer perceptron (MLP) and learning vector quantization with Mel-frequency cepstral coefficient (MFCC) features. They have reported that the learning vector quantization method outperformed MLP, achieving a classification accuracy of 96%. Ali [60] *et al.*. have proposed a system for voice disorder detection using Gaussian Mixture Model (GMM), where they detect voice disorders by determining the source signal from speech through linear prediction (LP) analysis. The spectrum, computed using the features obtained from LP analysis, provides the distribution of energy in both normal and pathological voices, which will be used to differentiate them. The system is tested with both sustained vowels and running speech and achieved accuracies of 99.94% and 99.75%, respectively. Hadjitodorov and Mitev [61] developed a methodology for pathology detection using acoustic parameters such as jitter, shimmer, several harmonics-to-noise ratios (HNRs), along with new parameters for estimating turbulent noise in voice signals (Turbulent Noise Index) and characterizing "breathy" voice. The classification accuracy achieved was up to 96.1%.

The above-mentioned information pertaining to the AVDD systems reveals that almost all proposed methods demonstrate exceptional performance when it comes to evaluating voice disorders, boasting an accuracy rate exceeding 93%. This indicates that the AVDD system holds significant promise for aiding both patients and clinicians in the comprehensive assessment of voice disorders. With its high level of accuracy, the system can provide invaluable insights into the diagnosis and treatment of various vocal impairments, thereby facilitating more effective and targeted interventions for individuals experiencing such conditions. Consequently, the utilization of the AVDD system stands to greatly benefit not only patients seeking reliable diagnostic measures but also healthcare professionals striving for enhanced precision and efficacy in their clinical evaluations.

2.4 Perceptual Speech-Pathological Features for Assessing Voice Disorders

The assessment of voice disorders is based on auditory perception, which evaluates the quality of voice by examining perceptual characteristics such as pitch, loudness, resonance, and overall timbre. This evaluation encompasses various factors that can affect a person's voice quality, including vocal fold anatomy, respiratory support, and vocal technique. In numerous instances, alterations in voice quality, such as hoarseness, breathiness, roughness, or strained vocal quality, may result from voice disorders. Nonetheless, not all variations in voice quality signify a voice disorder. Temporary shifts in voice quality may arise from factors such as acute illness, fatigue, or vocal strain, whereas voice disorders usually entail persistent or recurrent abnormalities in voice production. Evaluating voice quality is crucial in the diagnosis and treatment of voice disorders, and it involves the utilization of diverse tools and methods by speech-language pathologists and otolaryngologists (ENT doctors), such as perceptual assessment, acoustic analysis, and laryngeal imaging. Therefore, perceptual features can be crucial clues for voice order detection based on auditory perception.

There are research related to detecting voice disorders that evaluates voice quality by examining perceptual characteristic. Sasou [62] introduces a technique to evaluate pathological voice quality automatically using the grade-roughness-breathiness-asthenia-strain (GRBAS) categorization, which is a widely accepted standard in the field. It achieves an impressive average F-measure of 87.25% for tasks related to identifying speakers, demonstrating its efficacy in accurately discerning subtle nuances in vocal characteristics. The second work reported by Hidaka *et al.* [63] explores automatic estimation of voice quality using an Recurrent Neural Network (RNN) with non-parametric features extracted from the amplitude and phase spectrograms. This automated method has the potential to significantly enhance reproducibility in laryngological practice, offering a standardized approach to voice quality assessment. Non-parametric features extracted from amplitude and phase spectrograms, especially when transformed into the mel scale, demonstrated enhanced efficacy in evaluating hoarseness. Specifically, temporal phase variation along the mel scale proved effective for assessing Grade, Rough, Breathy, and Strained qualities, while log mel amplitude emerged as a reliable indicator for Asthenic characteristics. Then, Jouaiti *et al.*. [64] conducted a detailed analysis of speech data collected over time from individuals with speech impairments, aiming to gain insights into cognitive changes. They manually labeled the speech data and matched these labels to acoustic features using openSMILE, identifying critical features that contribute to perceptual ratings related to phonation, breathiness, roughness, asthenia, and strain. The experiment resulted

in successful assessments of phonation aspects, with an F1-score of 55%; breathiness, 71%; roughness, 60% ; asthenia, 65%; and strain, 0.74%. Finally, Krue et al. [65] proposed that using musical features extracted from voice recordings improved the accuracy of discriminating Parkinson's disease (PD) patients from healthy individuals. This work applied musical features consisting of Dynamics, Rhythm, Timbre, Pitch, and Tonality. They applied these features to K-Nearest Neighbors (KNN) and SVM, and the results indicated that they outperformed existing studies in terms of accuracy.

From the above-mentioned works, the feature related to evaluating the voice quality of abnormal voices found that those features can potentially detect voice disorders. Timbre features, such as roughness and brightness attributes, are utilized in these studies. Consequently, it can be inferred that timbre features possess the capability to identify voice disorders effectively.

2.5 Acoustical Speech-Pathological Features for Assessing Voice Disorders

Pathological features used for identifying voice disorders denote unusual attributes observed in an individual's voice, suggesting a possible underlying medical issue or impairment. These attributes encompass alterations in pitch, volume, and resonance, alongside occurrences such as voice breaks, hoarseness, or breathiness. Several speech pathology features and classifiers are utilized for detecting voice disorders. These methods aim to automate the assessment of voice disorders, assisting clinicians in improving their expertise to enhance the performance of detecting voice disorders.

Various research studies are developing methods to assess voice disorders. Zhang and Jiang [66] examined the acoustic properties of sustained and running vowels in both healthy individuals and patients with laryngeal disorders. They utilized perturbation techniques, such as jitter and shimmer, along with signal-to-noise ratio analysis, and nonlinear dynamic methods, such as correlation dimension and second-order entropy, to analyze these vowels. The results indicated a significant statistical distinction between the voices of individuals with laryngeal pathologies and those with normal vocal function. Watts *et al.* [67]. assessed various aspects of vocal performance in a professional singer with vocal fold edema both before and after medication. They found that following medication, there was a notable rise in the fundamental frequency ($F0$), along with significant reductions in jitter, shimmer, long-term frequency, and amplitude variability. Shama and Cholayya [68] conducted a study aimed at detecting various laryngeal pathologies such as adductor paralysis, cysts, leukoplakia, vocal fold polyps, degenerative

polyps, vocal fold edema, and vocal nodules. They utilized the HNR measure and critical-band energy spectrum as features for this purpose. HNRs were estimated across four frequency bands and served as one set of features. The normalized energies were obtained by filtering voiced speech signals through 21 critical band-pass filters, mimicking human auditory neurons, and constituted another set of features. The set of HNR features achieved an accuracy of 94.28%, while the critical energy spectrum features achieved 92.38% accuracy when employed with a KNN classifier. The results indicate that these features can complement the perceptual evaluation of speech in detecting suspected laryngeal pathologies. Importantly, this method requires a shorter length of speech data for analysis and is computationally less expensive compared to extracting fundamental frequency and noise measures. Parsa and Jamieson [69] examined various glottal noise metrics, including signal-to-noise ratio, HNR, normalized noise energy, frequency domain HNR, pitch amplitude, and spectral flatness ratio, to distinguish between healthy and pathological voices. They categorized these metrics into two groups and compared them based on (1) the probability distribution, (2) the ranking, and (3) the Receiver Operating Characteristic (ROC) of each metric. The highest achieved classification accuracy was 96.50%. Hadjito-Dorov *et al.* [70] presented a technique focused on creating prototype distribution maps (PDM) to model the probability density functions of input vectors from both normal and pathological speakers. This method utilizes characteristics like pitch period, pitch pulse shape, HNR, and low-to-high energy ratio. The results showed a classification accuracy of 95.10%. Teixeira *et al.* [71]. examined the role of various acoustic features, including jitter, shimmer, and HNR, in evaluating voice disorders through artificial neural networks (ANN). Their analysis encompassed samples from both male and female voices, culminating in a remarkable achievement: they achieved a perfect accuracy rate of 100% for female voices and an impressive 90% accuracy for male voices. Gomez *et al.* [72] introduced a method for automatically detecting voice disorders utilizing the glottal-to-noise excitation ratio (GNE) and cepstral harmonics-to-noise ratio (CHNR). To classify the data, this study employed ordinal regression and Gaussian regression. The most effective automatic detector, trained using the Saarbrücken voice disorders database, achieves an Area Under The Curve (AUC) of 88%. WU *et al.* [73] investigated the application of glottal flow waveform in conjunction with a random forest classifier for detecting voice pathology. The study yielded high accuracies in detecting voice disorders by utilizing a combined feature set derived from the glottal source signal. The results showcased enhanced performance through the utilization of glottal flow waveform, surpassing existing methods. The accuracy rates for voice pathology detection were 88.52% for the Saarbrücken Voice Database and 100.00% for the Massachusetts Eye and Ear Infirmary Database in this study.

The aforementioned details pertain to the utilization of acoustical speech-pathological features for automated voice disorder detection. The findings indicate that a majority of the proposed methods demonstrate notably high performance in assessing voice disorders. These methods achieve accuracy rates surpassing 88% and even reaching up to 100% on certain datasets. Such results underscore the promising potential of integrating speech-pathological features with classifiers. This integration holds considerable promise for assisting both patients and clinicians in conducting comprehensive assessments of voice disorders, thereby enhancing diagnostic accuracy and facilitating tailored treatment plans.

Chapter 3

Contribution of Using Perceptual and Acoustical Speech Pathological Features to Detect Fake Speech

3.1 Concept and Idea of Proposed Features

The concept and idea of utilizing speech-pathological features represent hypotheses regarding the perceived acoustic quality of a disordered voice. The hypothesis of this research is that deepfake speech could potentially simulate the perceived acoustic quality of a disordered voice. Moreover, voice disorders are indeed abnormalities in the speech production mechanism. They occur when there are issues with the structures involved in producing voice, such as the vocal cords, larynx, and lungs. These disorders can affect the quality, pitch, volume, and other characteristics of a person's voice. The abnormalities of voice disorders represent unnaturalness. Deepfake speech is synthesized from a machine generator, and it is also unnatural speech. Therefore, the speech pathological features can be clues to detect deepfake speech, as depicted in Fig. 3.1. There are several features used to detect voice disorders. In this research, we focus on two concepts: acoustic and perceptual features.

The idea of using acoustical speech-pathological features is that these features are widely used to distinguish a healthy voice from disordered voices, thereby providing invaluable insights into the diagnosis, treatment, and management of various speech disorders and conditions. Acoustical speech-pathological features are crucial components closely intertwined with the complex human speech-production mechanisms, representing relevant acoustic, phonatory, and aerodynamic param-

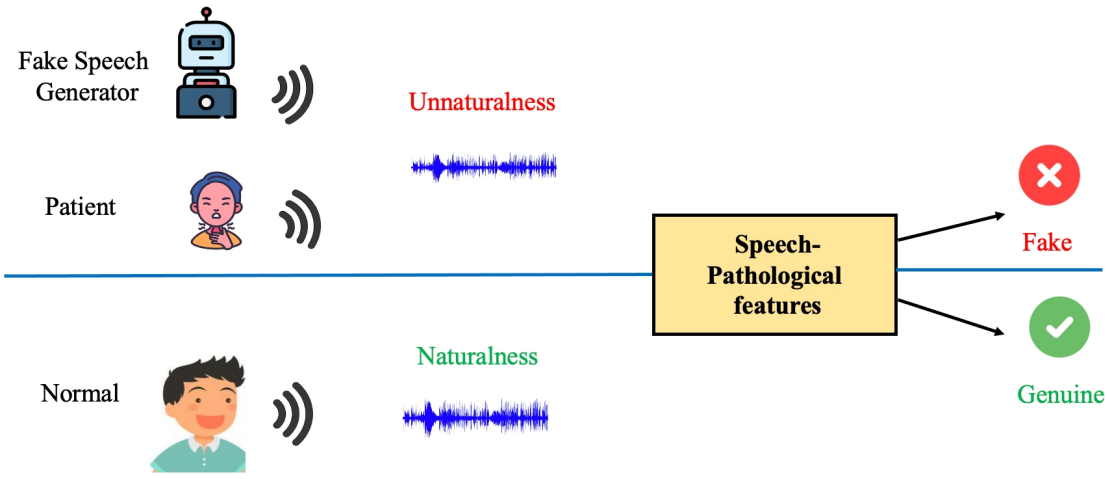

Figure 3.1: Concept and idea of using acoustical and perceptual speech pathological features to detect fake speech.

eters that not only illuminate the intricacies of communication disorders but also offer invaluable insights into the diagnosis, treatment, and rehabilitation of individuals facing speech and language challenges. Both disordered voice and deepfake speech represent unnaturalness. Therefore, acoustical speech-pathological features can be crucial clues for deepfake speech detection.

Perceptual features are based on auditory perception, which evaluates voice quality by examining characteristics such as pitch, loudness, resonance, and overall timbre. Research has been conducted using perceptual features to evaluate the voice quality of disordered voices [62, 67, 68]. The results indicate that perceptual features could potentially effectively detect disordered voices. The concept of using auditory perceptual to detect deepfake speech is that deepfake speech could possibly be the perceived acoustic quality of the disordered voice. Because voice disorder and deepfake speech exhibit unnaturalness, so perceptual features possible be to detect deep fake speech.

The relationship of speech-pathological features derived from disordered voice (Hyperkinetic dysarthria) [74], deepfake, and genuine speeches is investigated. Figure 3.2 shows an example of shimmer features: shimmer (*local*), shimmer ($APQ3$), and shimmer ($APQ5$). These features exhibit notable distinctions. The feature values of disordered voice and deepfake speech are close to each other, whereas the feature values of genuine speech are different. Therefore, these speech-pathological features, particularly the shimmer features, might be crucial indicators for detecting deepfake speech so that we investigate the potential of the 11 speech-pathological features in more detail in Section Acoustical Speech-Pathological Fea-

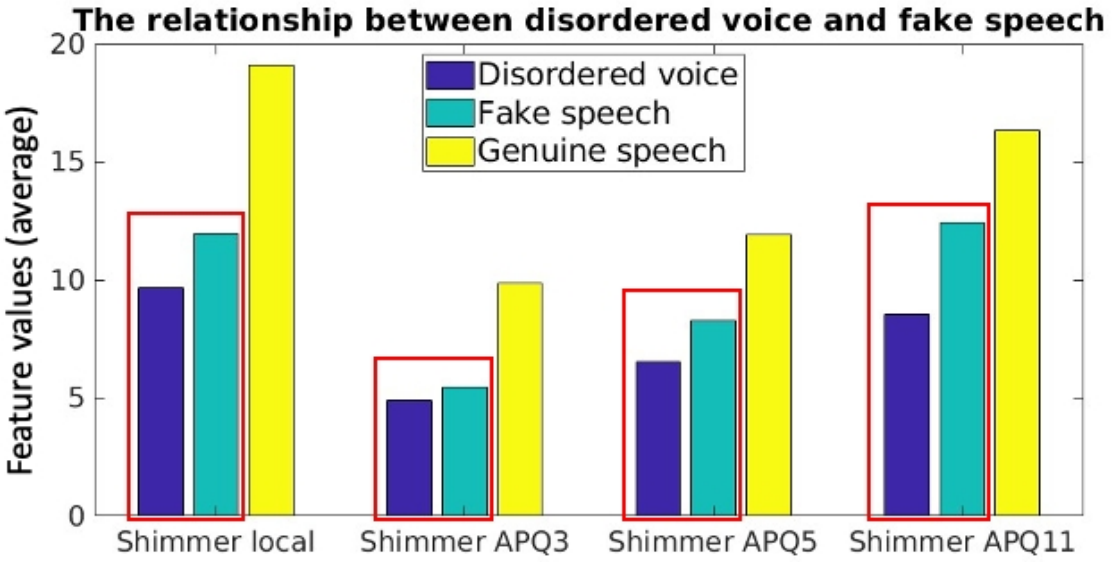

Figure 3.2: Bar graph comparing four shimmer features of disordered voices, fake speech, and genuine speech.

tures.

The perceptual features investigated in this dissertation comprise eight attributes including Hardness, Depth, Brightness, Roughness, Warmth, Sharpness, Boominess, and Reverberation are examined using statistical techniques. The predominant features are subsequently employed to differentiate between authentic and deepfake speech. These significant features are then integrated into MLP for the accurate identification of genuine and deepfake speech.

In this dissertation, six acoustical speech-pathological features are investigated: jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE). These speech-pathological features are analyzed using statistical methods. The insignificant features are removed, and the dominant features are then utilized to distinguish between genuine and deepfake speech. The identified significant features are applied with MLP to effectively differentiate between genuine and deepfake speech.

3.2 Philosophy of Utilizing Speech-Pathological Features for Deepfake Speech

The aim of this dissertation is to propose a framework for detecting deepfake speech based on speech-pathological features , which are typically utilized to differentiate between normal and pathological voices [62] and diagnose various diseases such as Parkinson's disease [75], neck and head cancers [76], and organic pathologies [77].

This study seeks to explore the speech-pathological traits utilized in distinguishing between normal and pathological voices [68, 73], with a focus on their potential application in identifying deepfake speech. Therefore, the primary objective of this research is to identify and analyze the critical indicators of pathological features necessary for effectively detecting deepfake speech. Through this exploration, significant features will be carefully selected and employed to differentiate between fake and genuine speech. As a result, this research aims to propose a comprehensive framework rooted in pathological speech processing for the precise detection of deepfake speech.

The hypothesis suggests that utilizing speech-pathological features for speech detection may indicate a potential correlation between deepfake speech and the perceived acoustic quality of a disordered voice. It proposes that deepfake speech detection via speech-pathological features could be connected to the perceived acoustic qualities of a disordered voice. Additionally, both deepfake speech and disordered voice exhibit unnatural characteristics [35, 23] and have been utilized to identify the artificial attributes of synthetic audio. Furthermore, the lack of naturalness in synthesized speech often stems from limitations in capturing and reproducing diverse prosodic elements, which include the non-linguistic aspects of voice. Addressing this challenge is crucial for enhancing the quality of synthesized speech.

While the natural human speech production mechanisms are complex and difficult to replicate artificially, the tiny variations in the speech production mechanism are unique to individual speakers. The first step is motor control function, which discusses how the brain controls the muscles involved in speech production. It includes the neural pathways and processes that govern the coordination and timing of muscle movements necessary for speech. The second step is articulatory motion; this step focuses on the physical movements of the articulators, such as the tongue, lips, and jaw, involved in shaping sounds during speech production. This section delves into how these movements are precisely coordinated to produce specific phonemes and sequences of phonemes. The last step is sound generation, which examines how the movements of the articulators result in the production of speech sounds. It encompasses the airflow, vocal cord vibrations, and resonance in the vocal tract that contribute to the creation of different speech sounds, including

vowels and consonants [1].

Although advances in speech synthesis technologies have made it possible to create increasingly realistic speech, it remains constrained and challenging to replicate. Therefore, speech-pathological feature can be crucial clues for deepfake speech detection based on speech production described in the above. The philosophy of this dissertation is shown in Fig. 3.3.

The novelty of this research lies in the application of speech-pathological features for the detection of deepfake speech. This approach mirrors the methods employed by medical professionals when diagnosing speech disorders in patients. By utilizing these pathological features, a novel and effective strategy is created for identifying deepfake speech, thereby enhancing the security and authenticity of digital communication.

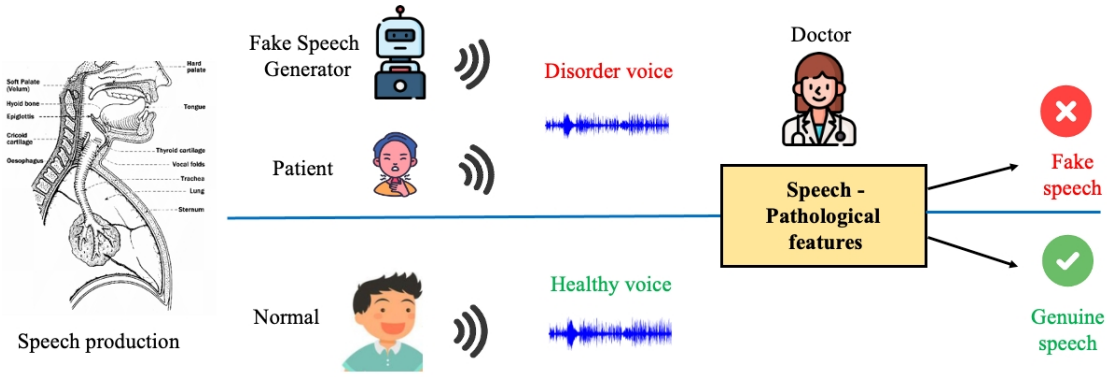

Figure 3.3: Philosophy of Utilizing Speech-Pathological Features for Deepfake Speech.

In this dissertation, two kinds of speech-pathological features were investigated. The first involves acoustic features, including jitter, shimmer, HNR, CHNR, NNE, and GNE. The second category comprises perceptual features, including Hardness, Depth, Brightness, Roughness, Warmth, Sharpness, Boominess, and Reverberation.

3.3 Acoustical Speech-Pathological Features

Acoustical speech-pathological features are typically used to distinguish between normal and pathological voices [62]. These features are crucial for diagnosing various medical conditions, including Parkinson's disease [75], neck and head cancers [76], organic pathologies [77], infections, autoimmune disorders, and genetic diseases. Pathologists analyze tissue samples obtained through biopsies, surgeries, or autopsies to identify these features and provide insights into the underlying disease processes.

This section describes the derivation of acoustical speech-pathological features, which have the potential to distinguish between genuine and deepfake speech. In this dissertation, six acoustical speech-pathological features are discussed: jitter, shimmer, harmonics-to-noise ratio (HNR), cepstral-harmonics-to-noise ratio (CHNR), normalized noise energy (NNE), and glottal-to-noise excitation ratio (GNE).

3.3.1 Jitter Features

Jitter measures the period variation from cycle to cycle of a speech signal [78, 71], as shown in Fig. 3.4. Since Jitter can be defined by several methods, this work focused on three definitions as follows.

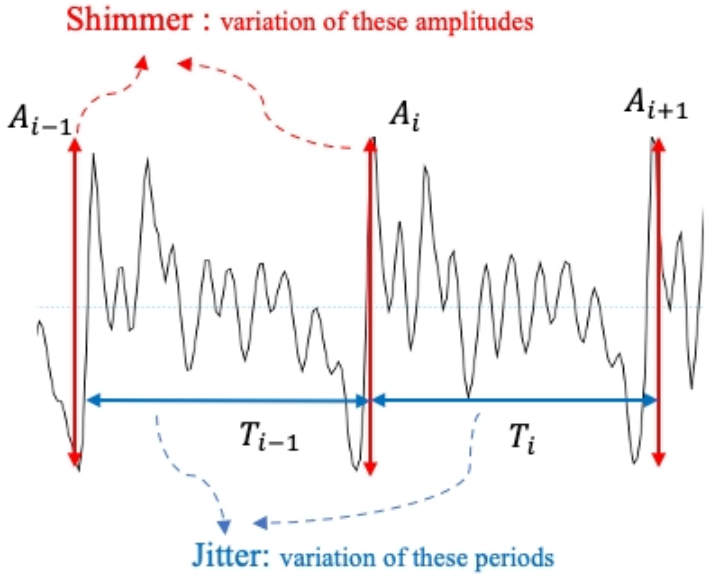

Figure 3.4: Jitter and shimmer concept illustration.

Jitter (*local*)

Jitter (*local*) is the percentage of the average absolute difference between consecutive periods divided by the average period, that is:

$$\text{Jitter } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100, \tag{3.1}$$

where T_i is the period lengths of the extracted fundamental frequency (F_0), and N is the number of F_0 periods [71].

Jitter ($PPQ3$)

Jitter ($PPQ3$), also known as jitter rap, is the percentage of the average absolute difference between a period and the average of that period with its two neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (PPQ3) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - (\frac{1}{3}\sum_{i=i-1}^{i+1} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \tag{3.2}$$

Jitter ($PPQ5$)

Jitter ($PPQ5$) is the percentage of the average absolute difference between a period, and the average of that period with its four neighbors, divided by the average period. It is defined as:

$$\text{Jitter } (PPQ5) = \frac{\frac{100}{N-1}\sum_{i=2}^{N-2}|T_i - (\frac{1}{5}\sum_{i=i-2}^{i+2} T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100. \tag{3.3}$$

3.3.2 Shimmer features

Shimmer measures the amplitude variation, resulting from irregular vocal fold vibrations, as shown in Fig. 3.4. Research in [79] demonstrated that the shimmer has significant differences in speaking styles. This feature can be used to assess the vocal quality and potentially indicate a voice disorder [71]. Since there are various ways to identify shimmer characteristics, focused on two types of shimmer features as follows.

Shimmer (*local*)

Shimmer (*local*) refers to the percentages of the average of absolute differences between the source signal amplitude related in each index (A_i) and its next neighbor (A_{i+1}), divided by the average of the signal amplitudes. It is defined as:

$$\text{Shimmer } (local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \tag{3.4}$$

where N is the number of F_0 periods, and A_i denotes the signal amplitude at index i.

Shimmer (x-point amplitude perturbation quotients, $APQx$)

Shimmer x-point amplitude perturbation quotients, Shimmer ($APQx$), are defined similarly to Shimmer (*local*). However, shimmer considers the absolute difference between the amplitude of each index (A_i) and the average of the x-point closest neighbors around A_i. It is defined as:

$$\text{Shimmer } (APQ\text{x}) = \frac{\frac{1}{N-m+1}\sum_{i=m}^{N-m}|A_i - (\frac{1}{x}\sum_{n=i-m}^{i+m} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \tag{3.5}$$

where $m = \frac{x-1}{2}$. In this study, three x-point shimmer features were investigated: $APQ3$, $APQ5$, and $APQ11$.

3.3.3 Harmonics-to-Noise Ratio (HNR)

HNR is a metric that quantifies the balance between the harmonic and noisy elements present in speech. Calculating the noise component (ι_{En}) involves computing the energy of the residual signal obtained by subtracting the average waveform from each cycle. The harmonic energy (γ_{En}) is derived from the energy of an average waveform created from a frame pitch that is synchronized with approximately ten consecutive glottal cycles. Therefore, this feature relies on an earlier estimation of F_0 [5]. The HNR is defined as:

$$\text{HNR} = 20\log\frac{\gamma_{En}}{\iota_{En}}. \tag{3.6}$$

It is computed for each frame of analysis. The final HNR is calculated averaging the values obtained for each frame.

3.3.4 Cepstral-Harmonics-to-Noise Ratio (CHNR)

CHNR, as called cepstral-HNR (CHNR) is employed to compute the HNR by quantifying the disparity in energy levels between the overall spectrum and the energy attributed to noise. In this context, noise energy represents the portion of energy that cannot be attributed to the original signal's spectrum [5]. The CHNR computation involves several steps for each analysis frame: (1) calculating the cepstrum of the signal; (2) identifying the harmonic components of the signal as periodic rahmonics (which correspond to harmonics in the cepstrum); (3) removing these components through a liftering operation to extract the equivalent noise energy; and (4) determining CHNR by comparing the previously calculated noise energy with the total cepstral energy. The CHNR calculation procedure is shown in Fig. 3.5.

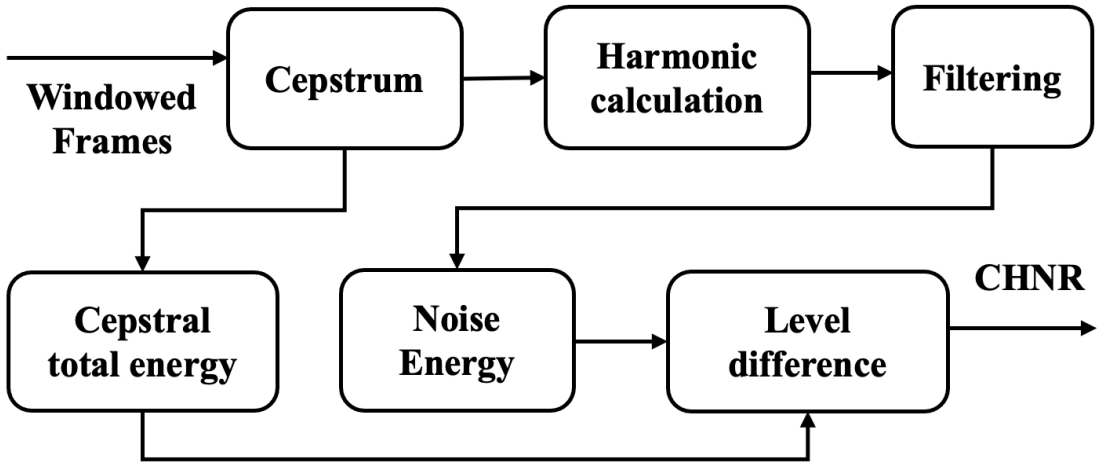

Figure 3.5: CHNR calculation process [5].

3.3.5 Normalized noise energy (NNE)

NNE measures the extended additive noise. The NNE is determined by comparing the energy of the noise to the overall energy of the signal within each analyzed frame [5]. To summarize the computation of NNE, the process involves: (1) calculating the F_0 of the signal and its log-spectrum; (2) directly computing the noise energy in the valleys from the spectrum, while estimating the noise energy in the harmonic peaks by interpolating the minima of neighboring valleys; and (3) ultimately determining NNE as the difference in level between the spectral total energy and the noise energy. The NNE calculation procedure is presented in Fig. 3.6.

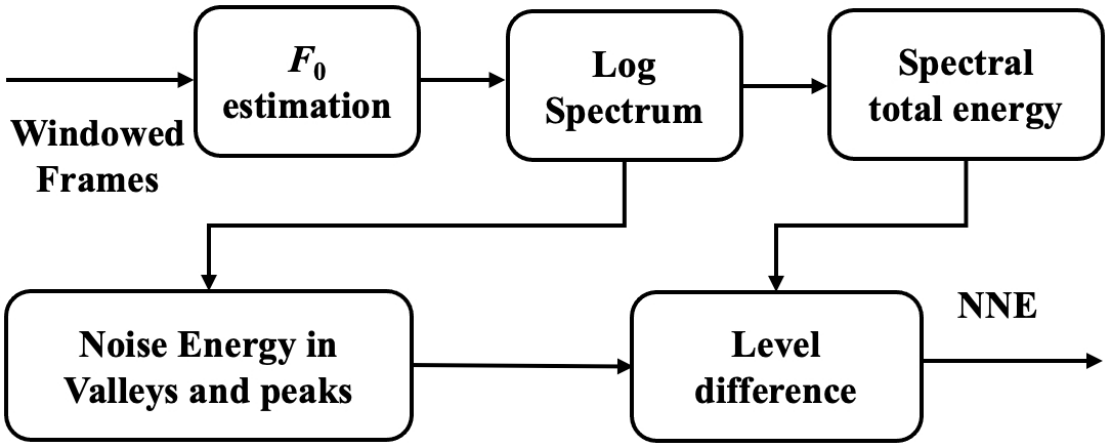

Figure 3.6: NNE calculation process [5].

3.3.6 Glottal-to-Noise Excitation Ratio (GNE)

GNE characterizes turbulent noise in speech, disregarding modulation effects [80]. It assumes that glottal pulses generate simultaneous and synchronous excitation across multiple frequency channels, as evidenced by the correlation observed in the Hilbert envelopes of these distinct frequency bands [5]. Computing NNE can be summarized as follows: (1) downsampling the input speech to 10 kHz; (2) applying an inverse filtering operation to obtain an estimate of the speech source; (3) calculating Hilbert envelopes for various frequency bands with fixed bandwidths and different center frequencies; (4) computing the cross-correlation for each pair of envelopes where the difference in their center frequencies is equal to or greater than half the bandwidth; (5) determining the GNE value as the maximum absolute value among all correlation functions. The calculation of the GNE is shown in Fig. 3.7.

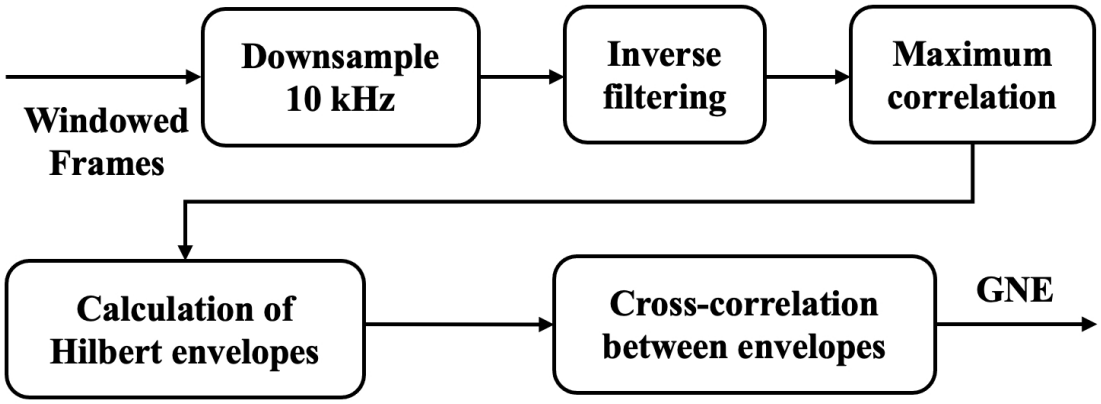

Figure 3.7: GNE calculation process [5].

3.4 Perceptual Speech-Pathological Features

The perceptual feature used in this dissertation is timbre features, because it is of significant importance in music and audio cognition, conveying essential emotional and perceptual data crucial in different domains of audio processing.

It represents a complex range of auditory qualities that together determine the essence or nature of a sound. Generally, timbre encompasses a variety of spectral and harmonic elements that impart unique traits to a sound [81]. The definition of timbre states that even if two sounds have the same pitch, loudness, and duration, they can still be perceived as different because they have different timbre features [82].

The timbral qualities are part of psychoacoustic features, with each relating to a distinct sensation experienced when listening to music [83]. Psychoacoustics, dedicated to understanding the complex interplay between sound and psychology in human perception, elucidates this relationship. It's worth noting that algorithm-generated scores fail to faithfully reproduce these attributes due to the inherent subjectivity in human sensory experiences.

Several studies have explored the modeling of timbre based on psychoacoustics, and the creation of objective metrics for each timbral characteristic. An influential example is the framework of the Audio Commons project, which formulated a timbral model. This model expresses eight high-level timbre features consisting of hardness, depth, brightness, roughness, sharpness, warmth, boominess, and reverberation, by integrating low-level timbre features, such as the spectral centroid, dynamic range, and spectral energy ratios, and rating them from 0 to 100 [6]. These high-level timbre features are described as follows.

3.4.1 Hardness

The perception of sound hardness primarily results from a combination of two factors: loudness and harshness. When a sound is labeled as harsh, it indicates an irregularity in its audio core. In this context, the audio core pertains to the frequency spectrum ranging from 2kHz to 5kHz, which corresponds to the frequency range where human hearing is most sensitive. Hence, hardness serves as a measure to assess the pleasantness or discomfort of a sound by examining the coherence between its loudness and frequency range, as perceived by the human ear. Essentially, it quantifies how effectively the balance between these two aspects conforms to human auditory preferences.

Several studies have identified factors contributing to the perception of hardness. For instance, Williams [84] proposed that the initial segment of a sound influences how hardness is perceived. Furthermore, Freed [85] introduced a framework for understanding the perception of mallet hardness in individual percussive

sounds, considering four acoustic parameters: (1) spectral mean level (a type of long-term average spectrum, LTAS); (2) spectral level slope (similar to cepstrum analysis); (3) spectral centroid mean (average spectral centroid over time, measured on the bark scale); and (4) spectral centroid TWA (time-weighted mean of the spectral centroid).

Solomon's research [86] also recognized the perceptual dimension of hardness/softness as a significant psychological aspect of timbre. This study proposed that this characteristic may be linked to rhythmic distinctions among stimuli, although it did not present any specific methods for quantifying this relationship.

Although no explicit model of hardness exists in the literature, there is an indication that the attack portion of a sonic event determines the apparent hardness, along with the spectral content of the attack. Therefore, a model of hardness was developed which employs three metrics: (1) attack time; (2) attack gradient; and (3) spectral centroid of attack. A linear regression model was then used to estimate the apparent hardness from these parameters.

There are four appropriate parameters for hardness prediction: the mean of the time-varying spectral centroid, the time weighted-average of the time-varying spectral centroid, the mean across time of the time-varying spectral level, and the slope of the spectral level [87].

3.4.2 Depth

While numerous academic papers discuss the concept of depth, none have presented a model or proposed any acoustic features associated with it. Nonetheless, in an online experiment named Social-EQ conducted by Pardo and Cartwright Pardo [88], participants were requested to provide a timbral descriptor along with a suitable adjustment on a 40-band graphic equalizer to illustrate that descriptor.

Six participants opted to use the term *deep*. Figure 3.8 displays the 40-band equalization treatment submitted by each participant. The average equalization across all participants, along with 95% confidence intervals, is depicted by the bold black line.

The trend depicted in Fig. 3.8 illustrates that all subjects' EQ treatments prioritize enhancing the low frequency components of the signal. Given the considerable similarity among these EQ treatments, it is probable that emphasizing low frequency content correlates with timbral depth. Hence, it is proposed that an appropriate model for depth analysis should include: 1) assessing the spectral centroid of lower frequencies (indicating energy concentration towards the low-end); 2) evaluating the ratio of low frequency energy; 3) considering the low-frequency limit of the audio excerpt (referring to the point at which low frequencies begin to roll off).

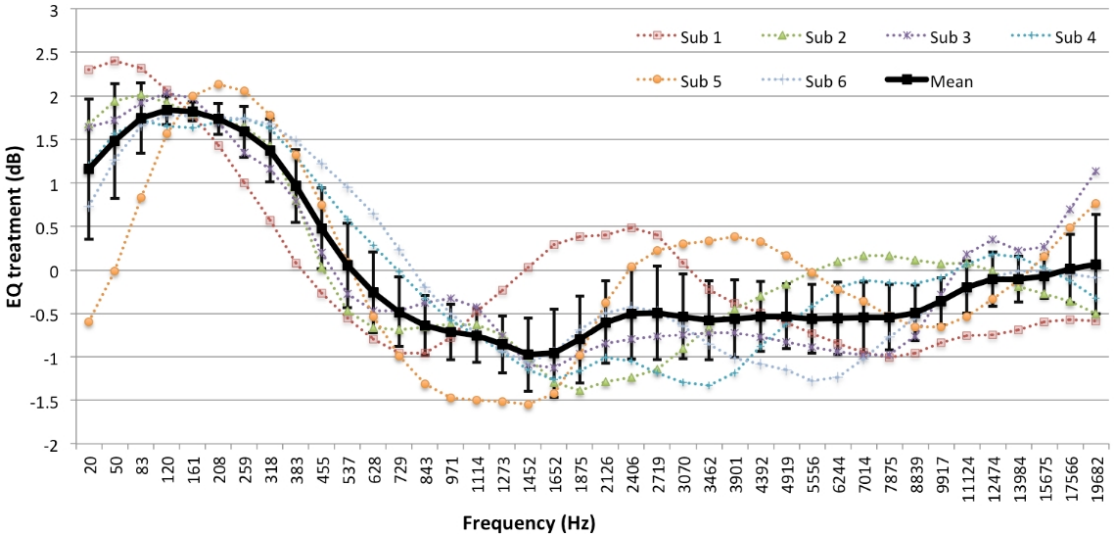

Figure 3.8: Social-EQ graphic equaliser settings representing the timbral descriptor *deep* [6].

$$\text{Lower spectral centroid} = \frac{\sum_{n(30Hz)}^{n(200Hz)} f(n)\ x(n)}{\sum_{n(30Hz)}^{n(200Hz)} x(n)}, \tag{3.7}$$

where $n(s)$ is the bin number relating to frequency s, is the frequency of the nth bin, and is the magnitude of the nth bin. The mean lower spectral centroid is then calculated across all frames.

The model also computes the lower ratio for each frame, which is the ratio of energy between 30Hz and 200Hz in comparison to the total energy between 30Hz and the Nyquist frequency:

$$\text{Lower ratio} = \frac{\sum_{n(30Hz)}^{n(200Hz)} f(n)\ x(n)}{\sum_{n(30Hz)}^{n(200Hz)} x(n)}, \tag{3.8}$$

where n(Nyquist) represents the frequency bin corresponding to the Nyquist frequency, the mean lower ratio is subsequently computed across all frames.

The final metric computed by the model pertains to the low-frequency limit. This calculation mirrors the methodology employed for the 'spectral rolloff' metric in the IRCAM timbre toolbox [89]. In this context, the low-frequency limit is delineated as the frequency wherein 95% of the spectral energy is situated above. The mean low-frequency limit is then calculated across all frames.

$$\mathrm{x(n)}^2 = \frac{\sum_{n(30Hz)}^{n(200)} x(n)}{\sum_{n(30Hz)}^{n(Nyquist)} x(n)}, \tag{3.9}$$

The depth model was calculated using a linear regression based on lower spectral centroid, lower ratio, and low-frequency limit ($\mathrm{x(n)}^2$).

3.4.3 Brightness

Brightness, a characteristic of sound quality, has received considerable attention in research. Several studies have indicated that the spectral centroid is a metric that aligns with perceived brightness [90, 91]. Nevertheless, certain studies propose that the proportion of high frequencies compared to the total energy may serve as a more accurate predictor. In a recent study, Pearce examined existing models and formulated a novel brightness model that integrates both a variant of spectral centroid and spectral energy ratio [92]. This model is used in Audio Commons [6], which is introduced in the dissertation.

The calculation of brightness is as follows : First, the audio signal is segmented into small frames and transformed into the frequency domain through FFT. Next, a sample-by-sample half-octave smoothing method is applied to refine the magnitude frequency response of each audio frame. This refined response is then utilized to calculate two metrics, one of which is the frequency-limited spectral centroid ($FLSC$), designed to focus on frequencies exceeding 3 kHz and ratio.

$$\text{Frequency-limited spectral centroid} = \frac{\sum_{n(3kHz)}^{n(Nyquist)} f(n)\ x(n)}{\sum_{n(3kHz)}^{n(Nyquist)} x(n)}, \tag{3.10}$$

$$\text{Ratio} = \frac{\sum_{n(3kHz)}^{n(Nyquist)} x(n)}{\sum_{n(20Hz)}^{n(Nyquist)} x(n)}, \tag{3.11}$$

In this context, $n(s)$ signifies the bin number corresponding to a frequency s, $f(n)$ represents the frequency linked with the nth bin, 'x(n)' denotes the magnitude of nth bin, $x(n)$ and $n(Nyquist)$ stands for the bin number corresponding to the Nyquist frequency. At last, a linear regression analysis is utilized to obtain findings regarding brightness.

$$B = 25.8699 + 64.0127((\log_{10}(Ratio) + 0.44\log_{10}(FLSC)). \tag{3.12}$$

3.4.4 Roughness

Roughness gives the sense of buzzing and produces a harsh feeling. Rough sound is perceived on the basis of the amplitude fluctuations of voice modulated between 16 and 80 Hz. The roughness model is constructed by considering three variables from a sinusoidal model approach study. The first two variables, which relate to the amplitude of two sinusoidal components, are the dependence of the roughness on intensity and the dependence of roughness on the amplitude-fluctuation degree [93].

The calculation of roughness in this dissertation is introduced in [94, 6]. Initially, the audio signal is divided into frames, each lasting 50 milliseconds. These frames undergo windowing using a Hanning window and are then zero-padded to the closest power of two after each frame. Following this, each frame undergoes Fast Fourier Transform (FFT), and the magnitudes of the frequency components in all frames are standardized. This standardization guarantees that the highest magnitude across all frequencies and frames is set to 1.0, enabling uniform comparisons and evaluations. After applying the FFT and normalization procedures, a peak-picking algorithm is utilized on individual frames to detect peaks within the frequency spectrum. Roughness is then computed for each pair of detected peaks within a frame.

$$R = 0.5X^{0.1}Y^{3.11}Z \tag{3.13}$$

with:

$$X = A_{min} * A_{max}, \tag{3.14}$$

$$Y = \frac{2A_{min}}{A_{min} + A_{max}}, \tag{3.15}$$

$$Z = e^{(-3.5g(f_{max}-f_{min}))} - e^{(-5.75g(f_{max}-f_{min}))}, \tag{3.16}$$

where R denotes to roughness, A_{max} and A_{min} indicate the maximum and minimum magnitudes of the peak pairs, while f_{max} and f_{min} correspond to the maximum and minimum frequencies of the two peaks, respectively. The total roughness of a frame is determined by adding up all the pairs of roughness values within it. The total roughness of an audio file is subsequently computed as the average of roughness values across all frames.

3.4.5 Warmth

The perception of warmth has been examined in several studies, which have identified various factors that contribute to it. For example, Sorensen [95] discovered that the acoustic characteristics of concert venues, including their shape and the way sound bounces, greatly influence how warm orchestral music sounds to listeners. Certain aspects of the room, such as its size, how long sound echoes, and where reflections occur, were demonstrated to enhance the feeling of liveness and fullness in the lower frequencies of the music, which are strongly linked to warmth perception.

Bromham [96] delved into the impact of various audio effect processing methods on warmth. The research emphasized the significance of the bass ratio (BR) metric, which gauges the energy distribution in the low-frequency spectrum. Alterations to the BR, achieved through methods such as equalization, were shown to markedly affect the perceived warmth of the audio.

Expanding upon this idea, Williams [97] created a method enabling the precise adjustment of warmth in sound through the manipulation of diverse timbral characteristics. This system illustrates the intricate, multifaceted aspect of warmth perception and the capacity to mold it through meticulous examination of various acoustic and signal processing element

Farbood [98] expanded comprehension of warmth by exploring its correlation with the perception of tension in music. Distinct timbral attributes such as inharmonicity and roughness were demonstrated to impact the perception of tension, a factor closely associated with the overall warmth of the sound.

These studies highlight the nuanced and multifaceted nature of warmth perception in sound, involving the interplay of acoustic, timbral, and perceptual factors. The research underscores the importance of considering these various elements when aiming to create a desired sense of warmth in audio applications, such as music production and sound design.

3.4.6 Sharpness

Sharpness is a measure linked to the perception of acute or piercing sensations. It intensifies when the central frequency shifts to a higher range. Zwicker et al. introduced the notion of "acum" based on this observation. An acum is defined as a unit of narrowband noise centered at 1,000 Hz with a loudness level of 60 phon [99]. Following this, a model for sharpness was formulated, calculated as follows.

$$\text{Sharp} = 0.11 \frac{\int_0^{24Bark} D'(x) g(x) x dx}{\int_0^{24Bark} D' dx} \tag{3.17}$$

In the equation, Sharp represents sharpness, while $D^{'}(t)$ signifies the density of loudness within the critical-band rate t. Loudness, an inherent characteristic of auditory perception, is measured in phons. The function $g(t)$ represents the weighting factor for Sharp at the given rate. Through psychoacoustic studies, this weighting factor is established as 1.0 for frequencies up to 3,000 Hz, with a sharp increase to 4.0 for frequencies above this range.

3.4.7 Boominess

Research has extensively delved into the surging phenomenon across various domains, including construction machinery [100], automobiles [101], and interior vehicle noise [102]. The surging sensation typically arises from low-frequency elements, notably those associated with the engine [103]. In response, studies have proposed objective metrics, such as a sound quality index [101] and a weighted sound pressure level [102], to assess and quantify the surging sensation. These metrics have proven effective in mitigating the surging sensation, particularly concerning interior vehicle noise [102]. Although there is no explicit model of boominess documented in the literature, this dissertation utilized it within the Audio Commons framework [6].

3.4.8 Reverberation

Reverberation is a prolonged sound that continuously lasts even after the source has already stopped. Reverberation may be caused by the reflection of multiple sounds from the environment. The term reverberation refers to the acoustic fading of an audio signal and encompasses various standard measures. Among these, the most prevalent is RT60, which gauges the duration of decay. Although RT60 is often assessed in concert halls and similar settings, estimating it from recordings poses challenges. In 2015, the IEEE conducted the ACE challenge to attempt blind estimation of RT60 and the direct-to-reverberant ratio in recorded speech signals [104].

The reverb algorithm was executed according to the method outlined by Prego *et al.* [105]. Initially, a power spectrogram of the signal is computed. Without specifying frame length or window function, a Hamming window of 2048 length is arbitrarily selected, with a frame overlap of 512 samples. Subsequently, analysis is performed within the frequency range of 20 Hz to 4 kHz. This range is significant as it encompasses the majority of information in speech signals and aligns with the frequency spectrum typically utilized for measuring and specifying RT60 in building acoustics (averaged across 500 Hz, 1 kHz, and 2 kHz octave bands). Each identified SFDR entails the calculation of Schroeder integration using the provided equation:

$$\mathrm{c(k,l,n)} = 10 \log_{10} \frac{\sum_{a=n}^{L} E(k,a)}{\sum_{a=n}^{L} E(k,a)}, \tag{3.18}$$

c(k, l, n) represents the nth frame within the ℓ-th SFDR in the kth sub-band, where L denotes the total number of frames within the SFDR. $E(k, a)$ signifies the energy in the kth sub-band at frame a, with n indicating the current frame under analysis. The Schroeder integral is examined for each Signal-to-Noise and Distortion Ratio (SFDR) to estimate the SFDR RT60. Initially, the beginning of the analysis period is determined as the first frame within the SFDR where the Schroeder integral starts.

If the most linear portion of the Schroeder integral exhibits a dynamic range of less than 10dB, then select the most linear portion that demonstrates a dynamic range of at least 60dB. In cases where a 60dB dynamic range cannot be found for any segment, the algorithm seeks out the most linear segment with a minimum dynamic range of 40dB. If no suitable segment is identified under this criterion, the search threshold is progressively lowered to 20dB and then 10dB.

A linear regression analysis is performed on the chosen segment of the SFDR. Subsequently, the SFDR RT60 is determined from the regression coefficients, representing the duration for the linear regression line to decay by 60dB.

Following this, the sub-band RT60 is computed as the median value among all SFDRs within the respective sub-band. The overall RT60 is then approximated as the median value derived from all sub-band RT60s. To improve the accuracy of the estimated RT60, an arbitrary adjustment is applied by dividing the calculated RT60 by 3.

3.5 Dataset and Metrics

3.5.1 Dataset

The datasets from the ASVspoof 2019 [7] and ASVspoof 2021 [106] challenges were used to evaluate the performance of the proposed method.

The ASVspoof is a series of bi-annual, competitive challenges where the goal is to develop countermeasures capable of distinguishing between genuine and spoofed or deepfake speech since 2015. The ASVspoof 2019 is the first edition focusing on countermeasures for logical access related to spoofing attacks on speech synthesized by using text-to-speech and voice conversion techniques. The dataset is divided into three subsets: training set, development set, and evaluation set. The datasets from the ASVspoof 2019 [7] and ASVspoof 2021 [106] were utilized to assess the effectiveness of the proposed method.

Table 3.1: Number of utterances in the ASVspoof 2019 and 2021 datasets [7, 8].

Dataset		Number of utterances		
		Genuine	spoofed	Total
ASVspoof 2019	Training	2,580	22,800	25,380
	Development	2,548	22,296	24,844
	Evaluation	7,355	64,578	71,933
ASVspoof 2021	Evaluation	18,452	163,114	181,566

Similarly, the ASVspoof 2021 challenge extends the 2019 challenge. The evaluation set aims to assess the robustness of channel variation of the detection. The statistical information of both ASVspoof 2019 and ASVspoof 2021 datasets is shown in Table 3.1. The training set was used to train the models, while the development and evaluation sets were used for evaluation.

Table 3.2: Number of utterances in the ADD 2022 [9] and 2023 datasets.

Dataset		Number of utterances		
		Genuine	spoofed	Total
ADD 2022	Training	3,012	24,072	27,084
	Development	2,307	21,295	23,602
	Adaptation	300	700	1,000
	Test	-	-	109,199
ADD 2023	Training	3,012	24,072	27,084
	Development	2,307	26,017	28,324
	Test	-	-	118,477

The datasets from the ADD 2022 [9] and ADD 2023 [107] challenges were chosen to evaluate the effectiveness of the proposed method. These challenges are aimed at influencing the future trajectory of detecting deep synthetic and manipulated audio in multimedia. In ADD 2022, all tracks utilize identical training and development datasets, with each track having an individual adaptation dataset provided for fine-tuning and evaluation. On the other hand, ADD 2023 only includes the training and development datasets. Test datasets for both ADD 2022 and ADD 2023 are accessible online, containing unseen audio samples generated from various speech synthesis systems. This dissertation utilizes data from the low-quality FAD (LF) track in ADD 2022 and the audio fake game detection (FG-D) track in ADD 2023. The distinction between these datasets lies in the competition system settings; notably, the FG-D track in the ADD 2023 challenge incorporates two rounds of

testing. Given that the second round is more challenging, this paper focuses solely on the results from the second round. Statistical information regarding these datasets is provided in Table 3.2.

3.5.2 Metrics

In this dissertation, various metrics were employed to evaluate the efficacy of deepfake speech, encompassing accuracy, balanced accuracy, precision, recall, F1-score, F2-score, and the EER. These metrics serve as vital tools in quantifying the performance of deepfake detection systems.

Accuracy

Accuracy refers to the extent to which predictions made by a model are correct compared to the total number of predictions it makes. It is a crucial metric in evaluating the performance of a model. Mathematically, accuracy can be defined as the ratio of correct predictions to the total number of predictions. In essence, accuracy provides insight into how well a model is able to make accurate predictions across its entire dataset. It serves as a fundamental measure in assessing the reliability and effectiveness of a model's predictions. In mathematical terms, accuracy is represented as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.19}$$

where TP is the number of true positive, TN is the number of true negative, FP is the number of false positive, and FN is the number of false negative.

Balanced Accuracy

Balanced accuracy is a metric commonly used to evaluate the performance of binary classifiers, especially when dealing with imbalanced datasets. It takes into account both the true positive rate (sensitivity) and the true negative rate (specificity). Specifically, it is defined as the average recall obtained on each class.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{3.20}$$

Precision

Precision is a crucial metric in evaluating the performance of a model, as it quantifies the accuracy of positive predictions. Specifically, precision measures the ratio of true positive predictions to all positive predictions made by the model. In essence, it highlights the model's capability to correctly identify relevant instances while minimizing false positives. A higher precision value indicates a lower rate of false positives, signifying the model's effectiveness in discerning true positives from the overall positive predictions. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.21}$$

Recall (Sensitivity)

Recall measures the ratio of correctly identified positive predictions to the total number of actual positive instances in the dataset. Essentially, it gauges how effectively the model captures all relevant positive cases. This metric is crucial in assessing the model's sensitivity to identifying positives, making it valuable in scenarios where comprehensively capturing all positives is vital, such as medical diagnoses or anomaly detection. A high recall score signifies that the model is adept at minimizing false negatives, ensuring that few positive cases slip through undetected. Conversely, a low recall indicates that the model is missing a considerable number of positive instances, which could lead to critical oversights or incomplete analyses. Thus, achieving a balance between precision and recall is essential for optimizing model performance and ensuring reliable results. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3.22}$$

F1-score

The F1-score, regarded as the harmonic mean of precision and recall, offers a nuanced evaluation by striking a balance between these two metrics. This becomes particularly beneficial in scenarios marked by disparate class distributions or where the implications of false positives and false negatives diverge in significance. Essentially, it serves as a robust measure, encapsulating both the ability to correctly identify relevant instances (precision) and the ability to capture all relevant instances (recall) within its calculation. It is defined as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.23}$$

This balanced assessment proves invaluable in various fields, including but not limited to machine learning, where an accurate understanding of model performance is paramount for effective decision-making.

F2-score

The F2-score operates much like the F1-score but places greater emphasis on recall. This emphasis is particularly beneficial in scenarios where recall holds greater significance than precision, as in medical diagnosis or anomaly detection. In such cases, capturing as many true positives as possible is paramount, even if it means tolerating some false positives. The F2-score is appropriate for this scenario since reducing the false negative rate is more important than reducing the false positive rate. It is defined as:

$$\text{F2-score} = \frac{(1 + 2^2) \times \text{Precision} \times \text{Recall}}{(2^2 \times \text{Precision}) + \text{Recall}}. \tag{3.24}$$

Equal Error Rate (EER)

Equal Error Rate (EER) serves as a standard performance measure frequently employed in biometric technologies like facial recognition, fingerprint identification, and speaker verification. It signifies the threshold at which the False Acceptance Rate (FAR) matches the False Rejection Rate (FRR). In biometric systems, the objective is to reduce both FAR and FRR, though often there exists a compromise between these two rates. It is defined as:

$$\text{EER} = \min_{T} \{\text{FAR}(T) = \text{FRR}(T)\} \tag{3.25}$$

where $\text{FAR}(T)$ is the False Acceptance Rate at threshold T, which is the proportion of impostor attempts incorrectly accepted by the system. FRR (T) is the False Rejection Rate at threshold T, which is the proportion of genuine attempts incorrectly rejected by the system. T is the decision threshold of the biometric system.

Chapter 4

Deepfake Speech Detection using Acoustical Speech-Pathological Features

4.1 Proposed Method using Acoustical Speech-Pathological Features

This chapter conducts preliminary studies to investigate the potential of acoustical speech-pathological features for distinguishing between genuine and deepfake speech. The fundamental effectiveness of each pathological feature is first analyzed based solely on their average values. These acoustical speech-pathological features are then incorporated into a basic classifier.

Jitter and shimmer are first derived using the instantaneous robust algorithm for pitch tracking (IRAPT) [108], while the HNR, CHNR, NNE, and GNE are extracted using the AVCA-ByO toolbox [5]. The speech signals are set to 4 s, with a sample rate of 16 k. Note that to ensure all signals are 4 s long, signals shorter than 4 s are repeated from the beginning, whereas signals longer than 4 s are truncated.

A classifier is implemented using a neural network. This classifier is structured with ten nodes in the input layer, ten nodes in the hidden layer, and another single node in the output layer. The hidden layer utilizes the ReLU function for activation, while the output layer employs the sigmoid function as its activation function. The classifier's training setup includes a maximum of 100 epochs, a learning rate set at 0.0001, and a batch size of 128. Binary cross-entropy serves as the loss function, and the optimization is carried out using the Adam optimizer.

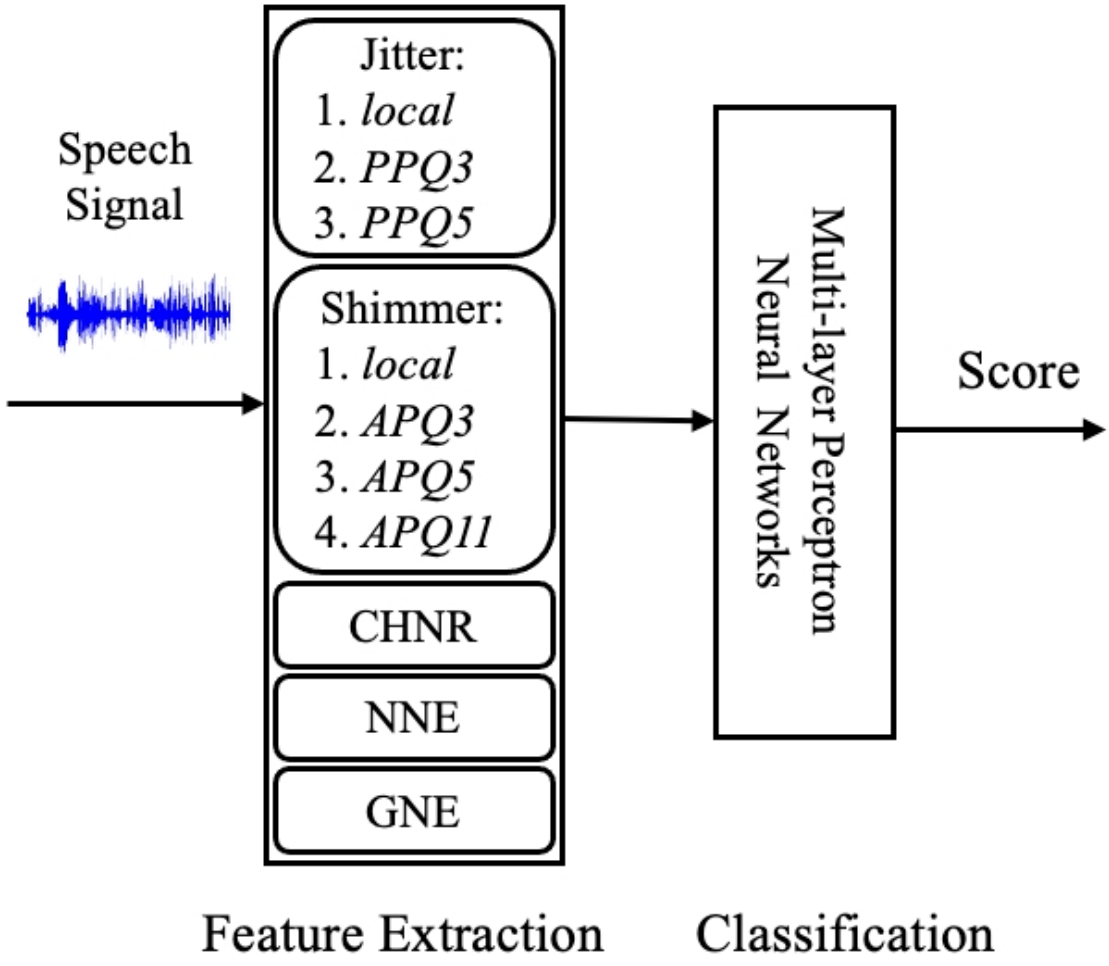

Figure 4.1: Proposed method using acoustical speech-pathological features with multi-layer perceptron neural networks.

4.1.1 Results and discussion in ASVspoof 2019 and 2021 datasets

As depicted in Table 4.1 provides a clear demonstration of the effectiveness of acoustical speech-pathological features when they are utilized in conjunction with a neural network model on the development set of the ASVspoof 2019 dataset. Through careful observation and analysis, it has been noted that two specific features, namely shimmer ($APQ11$) and shimmer ($APQ3$), exhibit particularly strong performance in terms of recall. They achieve impressive rates of 95.97% and 94.03%, respectively, which underscores their superior ability to detect fake speech when compared to other features. Moreover, the NNE achieves dominant performance in terms of balanced accuracy and precision, reaching 62.84% and 92.95%, respectively, compared to other features.

Interestingly, this study found that a combination of 10 features (excluding HNR) yields better results than when all 11 features are combined. This somewhat counterintuitive finding suggests that the inclusion of HNR may not contribute positively to the performance of the model. When these 10 features are combined (excluding HNR), the results indicate an accuracy of 89.94%, a balanced accuracy of 61.82%, a precision of 92.04%, a recall of 97.20%, an F1-score of 94.55%, and an F2-score of 96.12%.

These findings not only provide valuable insights into the role of individual features in detecting deepfake speech but also suggest that acoustical speech-pathological features hold significant promise for effectively tackling this challenge. This could potentially pave the way for more robust and reliable deepfake detection systems in the future.

As shown in Table 4.2, The performance of integrating the 10 acoustical speech-pathological features (excluding HNR) with the neural network on the evaluation sets of both the ASVspoof 2019 and 2021 datasets has been observed.The particular focus on the ASV2021 evaluation set stems from the fact that its training and development sets are identical to those of ASVspoof 2019. This consistency allows for a more accurate and reliable comparison of the results.

The results from this integration are quite promising. They indicate that the combination of these 10 features has the potential to effectively detect deepfake speech. In fact, this combination achieved a notable recall rate of 97.86%, which is a significant accomplishment in the field of deepfake detection. This high recall rate suggests the model, with the selected features, is highly sensitive and capable of identifying a large proportion of actual deepfake instances.

These findings reinforce the importance of feature selection in the design of effective deepfake detection systems. They also highlight the potential of acoustical speech-pathological features in enhancing the performance of such systems. As the model is further refined and other potential features are explored, there is hope

Table 4.1: Results from applying an average of speech-pathological features with a neural network on the development set of ASVspoo 2019.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	68.19	54.56	90.93	71.70	80.18	74.74
Jitter ($PPQ3$)	61.15	60.28	92.94	61.31	73.93	65.68
Jitter ($PPQ5$)	66.24	55.53	91.24	69.01	78.57	72.54
Shimmer (*local*)	75.33	51.90	90.17	81.38	85.56	83.00
Shimmer ($APQ3$)	85.64	53.17	90.37	94.03	92.16	93.27
Shimmer ($APQ5$)	58.16	50.25	89.82	60.20	78.08	65.45
Shimmer ($APQ11$)	86.81	52.00	90.13	95.97	92.87	94.60
CHNR	84.21	54.61	90.67	91.84	91.25	91.61
NNE	73.51	**62.84**	**92.95**	76.26	83.78	79.11
GNE	85.56	59.75	91.73	92.21	91.97	92.11
HNR	21.11	55.92	99.74	12.13	21.63	14.71
Combining 10 features (except HNR)	**89.94**	61.82	92.04	**97.20**	**94.55**	**96.12**

Table 4.2: Results from applying an average of acoustical speech-pathological features with a neural network on the evaluation set of ASVspoof 2019 and 2021.

10 Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
ASVspoof 2019	80.59	52.49	90.27	87.85	89.01	88.32
ASVspoof 2021	90.23	60.36	91.83	97.86	94.74	96.68

for further improvement in the accuracy and reliability of deepfake detection.

4.1.2 Results and discussion in ADD 2022 and 2023 datasets

Results and discussion in ADD 2022 datasets

Table 4.3 provides a detailed depiction of the performance of Speech-pathological features when they are employed in conjunction with a neural network model on the adaptation set of the ADD 2022 dataset. Three specific features stand out: shimmer ($APQ3$) with its strong recall rate of 99.42%, GNE for its balanced accuracy and precision, and again, shimmer ($APQ11$) for its dominant F1-score

These impressive rates indicate their superior capability in identifying fake speech, setting them apart from other features in the model.

Interestingly, this study found that a combination of 10 features (excluding HNR) yields better results than when all 11 features are utilized. This finding suggests that the inclusion of HNR may not contribute positively to the performance of the model. When these 10 features are integrated (excluding HNR), the results show an accuracy of 71.10%, a balanced accuracy of 52.50%, a precision of 71.07%, a recall of 99.00%, an F1-score of 82.75%, and an F2-score of 91.78%.

These findings are quite revealing. They show that acoustical speech-pathological features hold significant potential for effectively detecting deepfake speech. This could potentially lead to the development of more robust and reliable deepfake detection systems in the future. As the model is further refined and other potential features are explored, there is hope for further improvement in the accuracy and reliability of deepfake detection.

Results and discussion in ADD 2023 datasets

Table 4.4 presents a detailed analysis of the performance of acoustical speech-pathological features when they are utilized in conjunction with a neural network on the adaptation set of the ADD 2023 dataset. One specific features, the jitter ($PPQ3$) stand out due to their robust recall rates of 97.27%. This high recall rate is indicative of their exceptional ability to identify fake speech, thereby outperforming other features in the same category. Another feature to consider is NNE, as it demonstrates dominant performance with a balance accuracy of 78.24% and a precision of 97.64%.

An interesting observation made during the experiment was that the combination of 10 selected features, excluding HNR, yielded superior results compared to the utilization of all 11 features. This suggests that the exclusion of HNR from the feature set could potentially enhance the performance of the model.

Table 4.3: Results from applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2022.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	69.20	51.33	70.59	96.00	81.36	89.56
Jitter ($PPQ3$)	68.10	51.24	70.55	93.43	80.39	87.74
Jitter ($PPQ5$)	69.50	51.55	70.68	96.43	81.57	89.88
Shimmer (*local*)	68.10	50.07	70.32	95.14	80.67	88.78
Shimmer ($APQ3$)	69.90	50.21	70.09	**99.42**	82.20	91.75
Shimmer ($APQ5$)	69.30	50.26	70.11	97.86	81.69	90.68
Shimmer ($APQ11$)	68.10	43.83	69.50	97.00	**89.89**	89.98
CHNR	66.80	51.24	70.58	90.14	79.17	85.41
NNE	68.50	50.83	70.37	95.00	80.85	88.79
GNE	64.80	**56.00**	**73.39**	78.00	75.62	77.03
HNR	42.40	45.42	65.27	37.85	47.92	41.32
Combining 10 features (except HNR)	**71.10**	52.50	71.07	99.00	82.75	**91.78**
Combining all 11 features	64.20	55.29	72.98	77.57	75.21	76.61

Table 4.4: Results from applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2023.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	89.23	59.24	93.33	95.07	94.19	94.72
Jitter ($PPQ3$)	90.48	55.61	92.73	97.27	94.94	96.32
Jitter ($PPQ5$)	89.34	58.31	93.18	95.38	94.27	94.31
Shimmer (*local*)	87.15	47.43	91.45	94.88	93.13	94.17
Shimmer ($APQ3$)	88.83	53.33	92.38	95.75	94.03	95.05
Shimmer ($APQ5$)	88.49	53.05	92.33	95.59	93.84	94.76
Shimmer ($APQ11$)	87.21	50.93	91.50	94.91	93.17	94.20
CHNR	78.71	76.52	97.20	78.50	86.86	81.64
NNE	77.77	**78.24**	97.64	77.68	86.52	80.99
GNE	85.10	50.77	91.98	91.79	91.89	91.83
HNR	37.54	62.62	**98.02**	32.67	49.00	37.69
Combining 10 features (except HNR)	**95.58**	77.61	96.22	**99.08**	**97.62**	**98.49**
Combining all 11 features	93.37	69.95	95.01	97.93	96.44	97.33

When these 10 selected features are integrated with a neural network, the results are quite impressive. The model achieved 95.58% accuracy, 77.61% balanced accuracy, 96.22% precision, 99.08% recall, 97.62% F1-score, and 98.49% F2-score. These metrics provide a comprehensive evaluation of the model's performance, demonstrating its effectiveness and reliability.

These findings underscore the significant potential of acoustical speech-pathological features in the field of deepfake speech detection. It suggests that these features, when used appropriately, can contribute significantly to the development of more effective and reliable deepfake detection systems. This opens up new avenues for further research and development in this field. This study serves as a stepping stone towards the development of more sophisticated and accurate deepfake detection models. It highlights the importance of feature selection in improving the performance of such models and paves the way for future research in this area.

4.1.3 Ablation study of Acoustical Speech-Pathological Features

Table 4.5 provides a comprehensive illustration of an ablation study conducted on the development set of ASVspoof 2019. This study meticulously analyzes the performance of each acoustical speech-pathological feature when integrated with a neural network. In this unique approach, one acoustical speech-pathological feature was systematically removed at a time. This method was employed to assess the importance and potential contribution of each individual feature towards the detection of deepfake speech.

This study's findings reveal that shimmer ($APQ3$) and GNE play a pivotal role in spoofed speech detection. Removing these features from the set resulted in the most significant performance decline across all metrics, including accuracy, balanced accuracy, precision, F1-score, and F2-score on the development set.

This observation underscores the critical importance of shimmer ($APQ3$) and GNE as features in the detection of deepfake speech. It suggests that the inclusion of shimmer ($APQ3$) in the feature set can significantly enhance the effectiveness of deepfake detection systems. This insight could be instrumental in guiding future research and development efforts in the field of deepfake speech detection. It highlights the need for a deeper understanding of the role and impact of individual features in the performance of deepfake detection models.

Table 4.6 depicts the results of an ablation study conducted on the adaptation set of ADD 2022. This study meticulously examines the performance of each acoustical speech-pathological feature when it is integrated with a neural network. In this insightful analysis, one acoustical speech-pathological feature was system-

Table 4.5: Ablation study of applying an average of acoustical speech-pathological features with a neural network on the development set of ASV 2019.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	69.99	71.34	95.76	69.64	80.63	73.69
Jitter ($PPQ3$)	86.81	62.95	92.38	92.97	92.67	92.85
Jitter ($PPQ5$)	82.85	60.36	91.95	88.66	90.27	89.93
Shimmer (*local*)	79.27	61.57	92.36	83.84	97.90	85.42
Shimmer ($APQ3$)	**57.78**	64.66	94.84	**55.00**	**70.42**	**61.00**
Shimmer ($APQ5$)	85.94	53.84	90.50	94.23	92.33	93.46
Shimmer ($APQ11$)	87.88	62.41	92.23	94.44	93.32	94.00
CHNR	89.30	61.44	92.01	95.23	93.59	94.57
NNE	90.65	67.61	93.22	96.60	94.88	95.90
GNE	78.11	**51.25**	**90.02**	85.04	87.46	85.99

Table 4.6: Ablation study of applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2022.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	70.70	53.55	71.58	96.43	82.17	90.17
Jitter ($PPQ3$)	70.20	52.33	71.03	97.00	82.00	90.39
Jitter ($PPQ5$)	70.60	52.05	70.88	98.43	82.42	91.33
Shimmer (*local*)	70.00	52.80	71.05	96.43	81.82	90.00
Shimmer ($APQ3$)	70.40	54.71	72.10	94.14	81.66	88.72
Shimmer ($APQ5$)	70.30	52.98	71.32	96.29	81.95	89.99
Shimmer ($APQ11$)	**69.40**	**51.00**	**70.43**	97.00	81.61	90.20
CHNR	70.50	52.36	71.03	97.71	82.26	90.88
NNE	70.10	51.14	70.49	99.29	82.44	91.79
GNE	69.70	53.79	71.74	**93.57**	**81.22**	**88.20**

Table 4.7: Ablation study of applying an average of acoustical speech-pathological features with a neural network on the adaptation set of ADD 2023.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	93.04	71.46	95.28	**97.24**	96.25	**96.84**
Jitter ($PPQ3$)	95.11	79.74	96.62	98.11	97.36	97.81
Jitter ($PPQ5$)	94.46	75.85	95.97	98.08	97.02	97.65
Shimmer (*local*)	93.86	74.42	95.76	97.64	96.69	97.26
Shimmer ($APQ3$)	93.59	70.72	95.23	98.05	96.57	97.45
Shimmer ($APQ5$)	94.09	73.53	95.60	98.09	96.82	97.58
Shimmer ($APQ11$)	93.87	66.68	94.44	99.17	96.75	98.18
CHNR	92.61	64.12	94.06	98.15	96.06	97.30
NNE	93.69	69.88	94.98	98.32	96.62	97.64
GNE	**92.40**	**62.34**	**93.77**	98.25	**95.96**	97.32

atically eliminated at a time. This method was employed to evaluate the significance and potential contribution of each individual feature towards the detection of deepfake speech.

This study's results are particularly insightful. They reveal that shimmer ($APQ11$) and GNE emerge as crucial features for the detection process. The removal of GNE from the feature set provides the strongest evidence for this, as it results in the most significant performance drop across all metrics on the adaptation set.

This observation underscores the critical importance of shimmer ($APQ11$) and GNE in the detection of deepfake speech. It suggests that the inclusion of shimmer ($APQ11$) and GNE in the feature set can significantly enhance the effectiveness of deepfake detection systems. This insight could be instrumental in guiding future research and development efforts in the field of deepfake speech detection. It highlights the need for a deeper understanding of the role and impact of individual features in the performance of deepfake detection models.

Table 4.7 provides a comprehensive illustration of an ablation study conducted on the development set of ADD 2023. This study meticulously investigates the performance of each individual speech-pathological feature when it is integrated with a neural network. In this insightful analysis, each feature was systematically removed one at a time. This method was employed to evaluate the significance and potential contribution of each individual feature towards the detection of deepfake speech.

The results of this study are quite revealing. They underscore the jitter *local* and GNE as the crucial features in the detection process. This is evidenced by the fact that the exclusion of jitter *local* and GNE from the features set results in the significant drop in performance across all metrics on the development set.

This observation underscores the critical importance of jitter *local* and GNE in the detection of deepfake speech. It suggests that the inclusion of jitter *local* and GNE in the features set can significantly enhance the effectiveness of deepfake detection systems. This insight could be instrumental in guiding future research and development efforts in the field of deepfake speech detection. It highlights the need for a deeper understanding of the role and impact of individual features in the performance of deepfake detection models. This could potentially lead to the development of more sophisticated and accurate deepfake detection systems in the future.

4.2 Deepfake Speech Detection using Segmental Frames of Analysis of Acoustical Speech-Pathological Features

Although the average of acoustical speech-pathological features has the potential to distinguish between genuine and deepfake speech, it might be not adequate. For instance, if the disparity between genuine and fake speech lies in consistency, with approximately 70%–80% consistency, while the remaining portions of the speech exhibit significant fluctuations, the average between genuine and fake speech becomes inconsequential. Therefore, instead of deriving the average of acoustical speech-pathological features from the speech signal as the conventional method does, a segmented frames of analysis technique is proposed for deriving acoustical speech-pathological features.

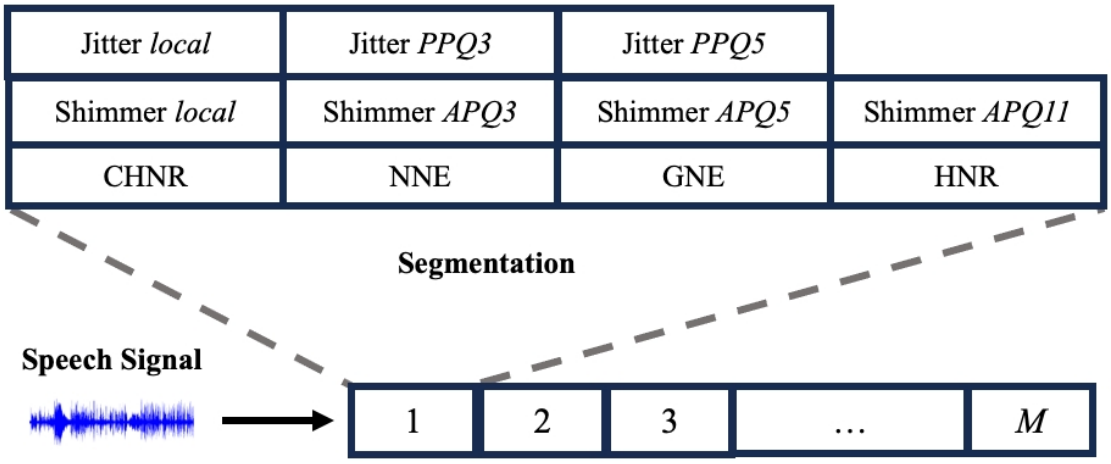

Figure 4.2: Segmental frames of analysis of acoustical speech-pathological features.

The process of deriving speech-pathological features using segmented frames of analysis is illustrated in Fig. 4.2. The process involves receiving a speech signal and segmenting it into frames. The acoustical speech-pathological feature is then extracted from each frame. This derivation process starts from the first frame to the M-th frame. Thus, the number of acoustical speech-pathological features depends on the number of frames.

The effectiveness of applying segmented frames of analysis for acoustical speech-pathological features was evaluated. Each feature is derived frame by frame, with a window frame of 50 ms and an overlap of 25 ms. Thus, for a 4-s signal with a sampling rate of 16 k, each acoustical speech-pathological segmental feature has a dimension of 159, i.e., the 4-s signal consists of 159 frames. These features are

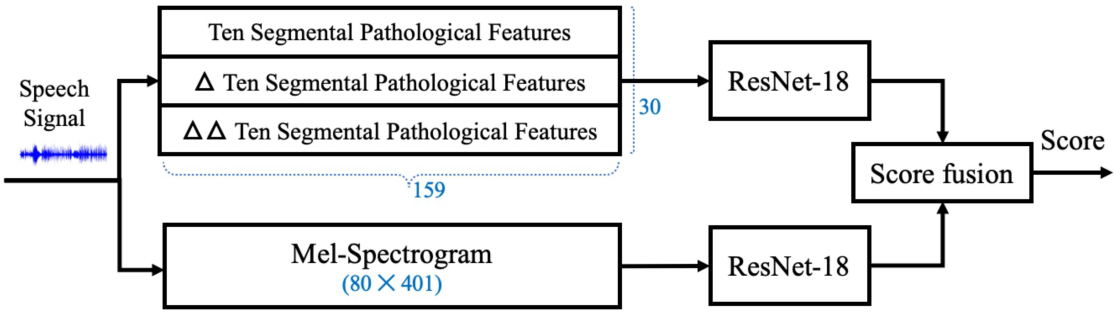

Figure 4.3: Proposed method, combining of (1) ten pathological-segment features with their first-order and second-order derivatives with ResNet-18 and (2) mel-spectrogram with ResNet-18, through score fusion.

inputted into a neural network similar to the previous study. The classifier model consists of three layers: an input layer with 159 nodes corresponding to the new dimension of the feature, hidden layer with 159 nodes, and single node for the output. The hidden layer is activated with the ReLU function, whereas the output layer is activated with the sigmoid function. The classifier's training settings included up to 100 epochs, learning rate set to 0.0001, and batch size of 128. Binary cross-entropy served as the loss function, and the Adam optimizer was used.

Table 4.8 lists the results of applying segmental frames of analysis with the acoustical speech-pathological features. The results indicate that extending the dimensions of ten speech-pathological features, excluding HNR, through segmental frames of analysis significantly improves performance compared with using the average method (as shown in Table 4.1) as follows: accuracy from 74.60 to 87.79%, recall from 79.00 to 95.70%, F1-score from 84.00 to 93.60%, and F2-score 81.20 to 95.00%.

4.2.1 Proposed Method of using segmental frames of analysis of acoustical speech-pathological features

Although the ten segmental speech-pathological features are effective for distinguishing between genuine and deepfake speech, there is still room for improvement. Therefore, this method combines two models to enhance the effectiveness of deepfake speech detection: 1) PF+Δ+ΔΔ with ResNet-18, and 2) mel-spectrogram with ResNet-18. These two models are integrated using score fusion.

The proposed method is illustrated in Fig. 4.3. The method involves using PF+Δ+ΔΔ with the ResNet-18 model as the primary model, while the mel-spectrogram with the ResNet-18 model is the secondary model. If the prediction score from the primary model exceeds a predetermined threshold of 0.5, it is con-

Table 4.8: Results of using segmental frames of analysis of acoustical speech-pathological features with neural networks on development set of ASVspoof 2019.

Speech-pathological feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Jitter (*local*)	85.95	64.57	92.78	91.46	91.11	91.72
Jitter ($PPQ3$)	85.44	64.13	92.70	90.94	91.81	91.27
Jitter ($PPQ5$)	87.44	59.37	91.60	94.68	93.12	94.05
Shimmer (*local*)	89.84	60.61	91.79	97.39	94.50	96.21
Shimmer ($APQ3$)	90.30	63.21	92.31	97.27	94.73	96.25
Shimmer ($APQ5$)	89.61	60.74	91.83	97.07	94.38	95.97
Shimmer ($APQ11$)	88.24	54.27	90.54	97.03	93.68	95.66
CHNR	**90.93**	**64.69**	92.60	97.70	**95.08**	**96.63**
NNE	88.67	61.57	92.06	95.67	93.80	94.91
GNE	88.62	54.49	89.98	**98.47**	93.95	96.61
HNR	21.16	55.95	**99.74**	12.78	21.70	14.77
Average of 10 features (except HNR)	87.79	60.07	92.00	95.70	93.60	95.00

sidered the final decision. However, if the score is below the threshold, the final score is determined by averaging the outputs from both the primary and secondary models. Segmental speech-pathological features for this study were derived using the following methods.

For jitter and shimmer, the IRAPT algorithm [108] are used. For the HNR, CHNR, NNE, and GNE, the AVCA-ByO toolbox [5] are used. The segmental speech-pathological features were derived from a speech signal of 4 s with a sampling rate of 16 k. The length of window frames was set to 50 ms with an overlap of 25 ms. Consequently, the total frames of a speech signal were 159. The segmental features of ten of the pathological features were concatenated. Hence, the dimension of these features was 10×159. The design of the features is illustrated in Fig. 4.3.

The mel-spectrogram and LFCC were derived using the *torchaudiolibrary* [109]. The dimensions of the mel-spectrogram are 80×401, while the dimensions of the LFCC are 60×265. The LFCC and its first-order and second-order derivatives were baseline features in ASVspoof 2019 [7].

A ResNet [110], used as a classifier, is an effective deep neural network architecture that addresses the vanishing gradient problem, wherein the gradients during backpropagation become excessively small. Numerous studies have leveraged ResNet in audio and speech-signal processing [111, 112, 113], including detecting synthetic speech [114, 115, 116]. The learning of the residual function of the residual block, which incorporates an intermediate input into the output of a sequence of convolutional blocks, is defined as:

$$\mathbf{out} = \mathcal{F}(\mathbf{in}) + \mathbf{in}, \tag{4.1}$$

where **in** and **out** denote the input and output from the previous layer, respectively, and $\mathcal{F}(\mathbf{in})$ is a component of a CNN comprising several convolutional blocks. Residual blocks are available across multiple layers, ranging from 10 to over 100 layers, with each layer containing a distinct number of residual blocks. However, excessive features were not deemed necessary for this study. The decision was made to use 18 residual layers (ResNet-18 model) for classification.

The ResNet-18 models were utilized as a classifier. The training process consisted of 100 epochs, a learning rate of 0.0001, and a batch size of 32. The Adam optimizer was employed. The binary cross-entropy between the predictions and the targets was used as the loss function. The output score was computed using the output of the "fake" node at the last fully connected layer before the softmax operation.

Table 4.9: Comparison of the proposed method with methods using different features and feature combinations on the ASVspoof 2019 dataset.

Method	Development set (%)							Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60×265)	**96.86**	85.02	96.69	99.91	**98.27**	99.25	4.57	90.10	**89.94**	**98.27**	90.15	94.23	91.73	10.06
2. Mel-spectrogram (80×401)	96.79	84.41	96.56	**99.99**	98.24	**99.28**	**3.30**	94.36	86.71	97.33	96.36	96.84	98.36	**8.44**
3. PF (10×159)	96.67	**85.81**	96.60	99.47	98.17	98.95	8.89	91.36	72.99	94.33	96.14	95.23	96.77	11.33
4. Δ (10×159)	94.47	75.94	94.83	99.26	96.91	98.34	15.10	92.69	72.64	94.17	97.91	96.01	97.14	13.39
5. $\Delta\Delta$ (10×159)	93.63	72.28	94.08	99.14	96.54	98.09	15.37	91.65	70.17	93.67	97.26	95.43	96.52	12.69
6. PF+Δ (20×159)	95.81	82.25	96.15	99.31	97.70	98.60	6.78	92.77	72.63	94.16	98.01	96.05	97.21	12.44
7. PF+$\Delta\Delta$ (20×159)	96.59	85.03	**96.72**	99.57	98.13	98.99	6.55	93.59	74.65	94.55	98.64	96.65	97.80	10.34
8. Δ+$\Delta\Delta$ (20×159)	93.63	70.80	93.76	99.53	96.56	98.32	9.80	92.48	71.91	94.02	97.82	95.89	97.04	12.86
9. PF+Δ+$\Delta\Delta$ (30×159)	96.72	85.05	96.71	99.74	98.20	99.17	8.72	93.96	73.86	94.36	99.19	96.71	98.51	10.22
10. **Proposed method**	96.75	83.82	96.43	**99.99**	98.17	99.25	8.62	**95.06**	77.70	97.30	**99.46**	**97.30**	**98.59**	10.19

4.2.2 Results and discussion in ASVspoof 2019 and 2021

Table 4.9 presents the experimental results obtained using the ASVspoof 2019 dataset. When comparing the LFCC and mel-spectrogram on the development set, these two features were comparable. However, on the evaluation set, the mel-spectrogram was better than the LFCC in terms of accuracy, recall, F1-score, F2-score, and EER, while the LFCC was slightly better only in terms of balanced accuracy and precision. The reason is that the LFCC correctly detects genuine speech better than the mel-spectrogram but correctly detects deepfake speech less effectively than the mel-spectrogram on an imbalanced dataset. Therefore, the mel-spectrogram showed significantly better results than LFCC. When comparing the efficiency of the mel-spectrogram in both development and evaluation sets, the results were similar, with high accuracy, balanced accuracy, F1-score, and F2-score and low EER.

The third to fifth rows display the results of PF, Δ, and $\Delta\Delta$ with dimensions of 10×159. In the comparison between Δ and $\Delta\Delta$, the results indicate that Δ outperforms $\Delta\Delta$ in almost all metrics on both the development and evaluation sets, except for EER. However, it's important to note that the difference in EER between Δ and $\Delta\Delta$ is less significant. Nonetheless, the method using PF outperforms both methods with Δ and $\Delta\Delta$. Thus, PF is considered to be the most contributing feature among them in terms of performance.

The results of the combinations of the segmental speech-pathological features: PF+Δ, PF+$\Delta\Delta$, and Δ+$\Delta\Delta$, each with a dimension of 20×159 are presented in the sixth to the eighth rows. The row no. 9 presents the results from the combination of PF+Δ+$\Delta\Delta$, which has a dimension of 30×159.

When comparing the combination of PF+$\Delta\Delta$ with PF+Δ+$\Delta\Delta$, the results indicate that PF+Δ+$\Delta\Delta$ was better in terms of accuracy, recall, F1-score, and F2-score on both datasets. The differences in the rest of the metrics are not significant; this is because PF+Δ+$\Delta\Delta$ has more dimensions than Δ+$\Delta\Delta$. Among the ten segmental speech-pathological features listed from the third to the ninth rows, PF+Δ+$\Delta\Delta$ was the most effective at detecting fake speech. The results of the PF+Δ+$\Delta\Delta$ on the development and evaluation sets are quite similar. Its efficiency was high in terms of accuracy, recall, F1-score, and F2-score. The rest of the metrics were also similar, except for the balanced accuracy, which differed significantly.

In comparison with the LFCC and mel-spectrogram, the findings indicate that PF+Δ+$\Delta\Delta$ performed better than using the LFCC in terms of accuracy, recall, F1-score, and F2-score. Conversely, PF+Δ+$\Delta\Delta$ marginally underperformed relative to the mel-spectrogram. However, these differences are not statistically significant, as they are less than 1%, except the EER.

The results on the ASVspoof 2019 evaluation set indicate that the proposed

method is comparable in efficiently detecting deepfake speech to the mel-spectrogram in terms of accuracy, recall, F1-score, and F2-score. However, its balanced accuracy exhibits a degree of decline.

These results highlight two interesting aspects: (1) the dimensionality of the features and (2) classification of speech as genuine or synthetic. Since the dimensions of $PF{+}\Delta{+}\Delta\Delta$ are relatively small, i.e., 30×159, compared with those of the LFCC and mel-spectrogram, i.e., 60×265 and 80×401, respectively. However, the efficiency of them was comparable. Thus, it might be possible to enhance the performance of the proposed method by extending its resolution, such as reducing the length of window frames. These results also indicate that the mel-spectrogram was more effective for correctly detecting genuine speech, whereas $PF{+}\Delta{+}\Delta\Delta$ was more effective for correctly detecting fake speech.

Table 4.10 lists the experimental results on ASVspoof 2021. In the comparison between LFCC and mel-spectrogram, the results indicate that the mel-spectrogram provided better results than LFCC in terms of accuracy, recall, F1-score, and F2-score. $PF{+}\Delta{+}\Delta\Delta$ slightly outperformed the mel-spectrogram in terms of accuracy, balanced accuracy, precision, F1-score, and particularly the EER. However, $PF{+}\Delta{+}\Delta\Delta$ exhibited only a slight decrease compared with the mel-spectrogram in terms of recall and F2-score.

When $PF{+}\Delta{+}\Delta\Delta$ is combined with the mel-spectrogram and ResNet-18, i.e., the proposed method, the results indicate that the performance of the proposed method surpasses that of the individual components in terms of recall, F1-score, and F2-score. However, balanced accuracy and precision showed a decrease. The reason for this is that both $PF{+}\Delta{+}\Delta\Delta$ and the mel-spectrogram exhibited similar characteristics, resulting in high performance in correctly detecting fake speech but lower performance in correctly detecting genuine speech. Although the proposed method, which combines these two models, did not improve in terms of all metrics, it showed high recall rates. The advantages of high recall are crucial for preventing unauthorized access and impersonation. In tasks involving sensitive scenarios in which unauthorized access carries a significant cost, prioritizing high recall is crucial for deepfake speech detection.

As evident from the third row of Table 4.10, the accuracy, balanced accuracy, and precision exhibited a slight decrease compared with the results obtained on the ASVspoof 2019 dataset. The effectiveness of $PF{+}\Delta{+}\Delta\Delta$ has limitations in detecting synthetic audio in environments involving communication over telephony and Voice over Internet Protocol (VoIP) networks, particularly due to various coding and transmission effects [8].This scenario will be further investigated.

Table 4.10: Comparison of results obtained from the proposed method and the baselines on the ASVspoof 2021 dataset.

Method	Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60×265)	85.22	**83.44**	**97.58**	85.68	91.24	87.82	16.55
2. Mel-spectrogram (80×401)	92.50	66.37	92.96	99.16	95.96	97.86	20.92
3. PF+Δ+$\Delta\Delta$ (30×159)	**92.60**	67.61	93.12	99.09	96.01	97.84	**15.97**
4. **Proposed method**	91.87	59.97	91.69	**99.96**	**96.65**	**98.18**	**15.97**

4.2.3 Results and discussion in ADD 2022 and 2023 datasets

Results and discussion in ADD 2022 dataset

Table 4.11 provides a detailed performance analysis of segmental frames in the context of acoustical speech-pathological features. These features were employed in conjunction with a neural network on the adaptation set of the ADD 2022 dataset. Among the various features analyzed, the GNE feature stands out with its impressive metrics. It boasts a recall rate of 99.14%, an F1-score of 83.72%, an F2-score of 92.33%, and an ERR of 30.00%.

These remarkable results underscore the superior capability of the GNE feature in detecting fake speech, outperforming other features in the same category. As a result, the analysis of segmental frames of acoustical speech-pathological features demonstrates significant potential for effectively detecting deepfake speech, a growing concern in today's digital age.

On the other hand, the HNR feature underperformed in comparison. Its lower performance metrics led us to conclude it may not be the best choice for the proposed study. Therefore, the use of HNR is not recommended for this particular research endeavor.

Tables 4.12 show the comparision the proposed ten segmental speech-pathological features (PF), the first order derivative of PF (Δ), the second order derivative of PF ($\Delta\Delta$), and the combinations of them. and their combinations. LFCC is the baseline feature in this study. The mel-spectrogram is a famous feature that is also used in voice disorder detection.

The results of the experiment presented in Table 4.12 were obtained using the ASVspoof 2019 dataset. When comparing the LFCC and mel-spectrogram on the adaptation set, these two features were comparable. However, on the test set, the mel-spectrogram performed better than the LFCC in terms of EER. The reason is that the mel-spectrogram has fewer false negatives in anonymized noisy speech than the LFCC.

The third to fifth rows display the results of PF, Δ, and $\Delta\Delta$ with dimensions of 10×159. In the comparison between PF, Δ and $\Delta\Delta$, the results indicate that PF outperforms Δ and $\Delta\Delta$ in almost all metrics on both the adaptation and test sets. Thus, PF is considered to be the most contributing feature among them in terms of performance.

The results of the combinations of the segmental speech-pathological features: PF+Δ, PF+$\Delta\Delta$, and Δ+$\Delta\Delta$, each with a dimension of 20×159 are presented in the sixth to the eighth rows. The row no. 9 presents the results from the combination of PF+Δ+$\Delta\Delta$, which has a dimension of 30×159.

When comparing the combination of PF+$\Delta\Delta$ with PF+Δ+$\Delta\Delta$, the results indicate that PF+Δ+$\Delta\Delta$ was better in all metrics on both datasets. While

Table 4.11: Results of using segmental frames of analysis of speech-pathological features with neural networks on the adaptation set of ADD 2022.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)	EER(%)
Jitter (*local*)	69.50	53.14	71.44	93.29	80.92	87.91	45.53
Jitter ($PPQ3$)	69.90	52.02	70.89	96.71	81.81	90.46	47.97
Jitter ($PPQ5$)	69.20	52.67	71.21	94.00	81.03	88.35	46.00
Shimmer (*local*)	70.30	53.34	71.50	95.71	81.86	89.64	40.00
Shimmer ($APQ3$)	69.30	52.64	71.20	94.29	81.13	88.54	43.33
Shimmer ($APQ5$)	71.10	54.40	71.94	98.14	83.02	91.48	32.67
Shimmer ($APQ11$)	72.40	**55.62**	72.51	97.57	83.19	91.26	35.67
CHNR	71.30	53.21	71.40	98.43	82.76	91.50	39.85
NNE	69.20	53.33	71.53	93.00	80.87	87.74	45.00
GNE	**73.00**	55.57	72.44	**99.14**	**83.72**	**92.33**	**30.00**
HNR	52.80	55.52	**75.11**	48.71	59.09	52.39	45.00

Table 4.12: Comparison of results obtained from the proposed method and the baselines on the ADD 2022 dataset.

Method	Adaptation set (%)							Test set (%)
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER	EER
1. LFCC (60×265)	92.40	91.05	94.70	94.42	94.56	94.48	8.33	38.57
2. Mel-spectrogram (80×401)	91.40	85.67	89.06	**100**	94.21	97.60	4.33	**33.55**
3. PF (10×159)	87.60	80.38	85.91	98.42	91.74	95.64	11.00	48.72
4. Δ (10×159)	76.30	60.70	74.81	99.71	85.49	93.49	16.33	48.75
5. $\Delta\Delta$ (10×159)	84.70	77.17	84.32	96.00	89.78	93.41	21.00	48.93
6. PF+Δ (20×159)	90.04	84.48	88.42	99.28	93.54	96.90	7.00	48.49
7. PF+$\Delta\Delta$ (20×159)	87.40	79.76	85.54	98.86	91.66	95.84	11.00	48.70
8. Δ+$\Delta\Delta$ (20×159)	77.10	62.31	75.63	99.86	85.85	93.44	16.00	48.72
9. PF+Δ+$\Delta\Delta$ (30×159)	93.80	89.67	91.86	**100**	95.76	98.20	3.90	47.38
10. Proposed method	**97.60**	**96.00**	**96.69**	**100**	**98.31**	**99.31**	**1.00**	35.54

PF+Δ was better than PF+$\Delta\Delta$ and Δ+$\Delta\Delta$ in all metrics on adaptation set and EER in test set.

In comparison with the LFCC and mel-spectrogram, the findings indicate that PF+Δ+$\Delta\Delta$ performed better than using the LFCC and mel-spectrogram in terms of accuracy, recall, F1-score, and F2-score on adaptation set, however on test PF+Δ+$\Delta\Delta$ performed less than the mel-spectrogram and LFCC. This is because pathological speech gradually degrades in noisy environments due to anonymous audio noise.

These results highlight the dimensionality of the features. The dimensions of PF+Δ+$\Delta\Delta$ are relatively small, i.e., 30×159, compared with those of the LFCC and mel-spectrogram, which are 60×265 and 80×401, respectively. However, the efficiency of them was comparable in adaptation set. The EER for the test set, which uses pathological speech features, is worst in various anonymous audio noise. This may be because pathological speech is used to detect voice disorders. However, the ADD dataset is an audio file with several background noises. Therefore, this issue will be studied further.

Results and discussion in ADD 2023 dataset

Table 4.13 presents a comprehensive evaluation of the effectiveness of segmental frames analysis of speech-pathological features. This analysis was conducted using a neural network on the adaptation set of the ADD 2023 dataset. The results are quite revealing. The shimmer (*local*) feature, in particular, demonstrates a robust performance with an accuracy of 95.70%, a balanced accuracy of 81.14%, and an F1-score of 97.68%. These impressive metrics underscore the superior capability of the shimmer (*local*) in detecting fake speech, outperforming other features in the same category. These findings suggest that the segmental frames analysis of speech-pathological features holds significant potential for effectively detecting deepfake speech. The shimmer (*local*) emerges as a particularly effective feature in this regard. However, the HNR did not perform as well in comparison. Its effectiveness in detecting deepfake speech was found to be inferior. Consequently, HNR was not recommended for the proposed study. This decision was based on the comparative analysis of the performance of different features in detecting deepfake speech.

Table 4.14 presents the experimental results obtained from the adaptation set of the ADD 2023 dataset. The focus of this analysis was a comparison between two features: LFCC and mel-spectrogram. The results of this comparative study indicate that the mel-spectrogram slightly outperformed LFCC across all metrics. These metrics included accuracy, balanced accuracy, precision, recall, F1-score, F2-score, and EER. Notably, the mel-spectrogram achieved a high rate of over 99.90% across these metrics. The performance underscores the effectiveness of the

Table 4.13: Results of using segmental frames of analysis of speech-pathological features with neural networks on the adaptation set of ADD 2023.

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)	EER(%)
Jitter (*local*)	93.71	70.24	95.04	98.28	96.63	97.61	14.82
Jitter ($PPQ3$)	92.98	71.07	95.22	97.19	96.19	96.79	16.12
Jitter ($PPQ5$)	93.13	67.70	94.64	98.08	96.32	97.37	16.21
Shimmer (*local*)	**95.70**	**81.14**	96.84	98.53	**97.68**	98.19	9.83
Shimmer ($APQ3$)	94.98	78.46	96.41	98.19	97.29	97.83	10.27
Shimmer ($APQ5$)	95.07	79.90	96.66	98.03	97.34	97.75	**8.63**
Shimmer ($APQ11$)	95.53	79.77	96.61	**98.60**	97.59	**98.20**	9.62
CHNR	93.04	70.32	95.08	97.46	96.26	96.98	18.86
NNE	93.09	74.87	95.87	96.63	96.25	96.48	18.47
GNE	94.15	74.22	95.07	98.03	96.86	97.56	11.96
HNR	37.65	63.59	**98.55**	32.60	48.99	37.63	41.81

mel-spectrogram in the analysis. This suggests that the mel-spectrogram might be a more reliable feature for the study compared to LFCC.

When compare two speech-pathological features: PF and PF+Δ+$\Delta\Delta$. The results indicate PF+Δ+$\Delta\Delta$ outperformed PF across all metrics. These metrics included accuracy, balanced accuracy, precision, recall, F1-score, F2-score, and EER. Notably, the PF+Δ+$\Delta\Delta$ achieved a high rate of over 99.90% across these metrics like mel-spectrogram.

In comparison with the LFCC and mel-spectrogram, the findings indicate PF+Δ+$\Delta\Delta$ that performed slightly lower. However, these differences are not statistically significant, as they are less than 1%. These results highlight two interesting aspects: (1) the dimensionality of the features and (2) the classification of speech as genuine or synthetic. The dimensions PF+Δ+$\Delta\Delta$ of are relatively small, i.e., 30×159, compared with those of the LFCC and mel-spectrogram, i.e., 60×265 and 80×401, respectively. Despite this, their efficiencies were comparable. Thus, it might be possible to enhance the performance of the proposed method by extending its resolution, such as reducing the length of window frames.

When PF+Δ+$\Delta\Delta$ is combined with the mel-spectrogram and ResNet-18, i.e., the proposed method, the results indicate that the performance of the proposed method surpasses that of the individual components in all metrics: the acuracy, balanced accuracy, presision, F1-score, F2-score, and EER are 99.99%, 99.92$, 99.96%, 100%, 99.98%, 99.99%, and 0.00%, respectively.

The reason for the higher efficiency of all features in this dataset might be twofold. First, the training set and the adaptation set share similar characteristics. Second, this dataset has less background noise, specifically white noise and reverberation. In contrast, ADD 2022 contains various background noises such as background music, car engines, and people chatting. Additionally, while the training set for ADD 2022 consists of clean speech without noise, the adaptation set includes significant background noise. Consequently, the efficiency of all features on ADD 2023 surpasses that of ADD 2022.

4.2.4 Ablation Study of Segmental Frames of Analysis of Speech-Pathological Features

The ablation study of the proposed features on the ASVspoof 2019 dataset, as shown in Table 4.15. ResNet-18 was a classifier, and the datasets were the development and evaluation sets. In this study, one speech-pathological feature was removed at a time to assess the importance and potential of each feature for deepfake speech detection. The results of the baselines, which use all speech-pathological features, are presented in the last row. The findings indicate that the CHNR is the most important feature since its removal leads to the lowest performance in terms

Table 4.14: Comparison of results obtained from the proposed method and the baselines on the ADD 2023 dataset.

Method	Adaptation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60×265)	99.92	99.84	**99.98**	99.94	99.96	99.95	0.02
2. Mel-spectrogram (80×401)	**99.99**	99.91	**99.98**	**100**	**99.99**	**100**	**0.00**
3. PF (10×159)	99.80	99.47	99.91	99.87	99.89	99.88	0.47
4. PF+Δ+$\Delta\Delta$ (30×159)	99.87	99.69	99.90	99.92	99.91	99.95	0.03
5. **Proposed method**	**99.99**	**99.92**	99.96	**100**	99.98	99.99	**0.00**

of accuracy, recall, F1-score, F2-score, and EER on the development set. These trends were also observed in the accuracy and F1-score on the evaluation set.

Table 4.16 presents an ablation study conducted on the adaptation set of the ADD 2022 dataset. This study examines the performance of each feature derived from the segmental frames analysis of speech-pathological characteristics when integrated with a ResNet-18 model. In this analysis, one feature from the segmental frames analysis of speech-pathological characteristics is removed at a time. This method allows for the evaluation of the significance and potential of each individual feature in the context of detecting deepfake speech. The results of this study are quite revealing. They show that the shimmer ($APQ11$) feature emerges as the most crucial element in this context. When this feature is removed from the analysis, the accuracy, recall, F1-score, and F2-score on the adaptation set all drop to their lowest levels. This finding underscores the importance of the shimmer ($APQ11$) feature in the detection of deepfake speech. It suggests that this feature plays a pivotal role in the performance of the ResNet-18 model when applied to the task of deepfake speech detection.

Table 4.17 presents the results of an ablation study that was conducted on the adaptation set of the ADD 2023 dataset. This study investigates the performance of each feature derived from the segmental frames analysis of speech-pathological characteristics when they are incorporated with a ResNet-18. In this analysis, a systematic approach is taken where one feature from the segmental frames analysis of speech-pathological characteristics is removed at a time. This method allows for a thorough assessment of the significance and potential of each individual feature in the context of detecting deepfake speech. The findings from this study are quite revealing. They highlight the GNE feature as the most important element in this context. When the GNE feature is excluded from the analysis, the accuracy, balanced-accuracy, recall, F1-score, and F2-score on the adaptation set all drop to their lowest levels. This finding underscores the importance of the GNE feature in the detection of deepfake speech.

4.3 Summary

This chapter explores the use of acoustical speech-pathological features for deepfake speech detection. Two methods are proposed: the first utilizes the average value of these features with a Multi-Layer Perceptron (MLP) neural network, and the second leverages segmental frames of analysis with a ResNet-18 architecture.

For the first method, the potential of acoustical speech-pathological features for distinguishing between genuine and deepfake speech is investigated. The six features are investigated consis of 3 jitter, 4 shimmer, HNR, CHNR, NNE, and GNE. After feature extraction, the average value of these features is calculated.

Table 4.15: Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on ASVspoof 2019 dataset.

Excluded Feature (9×159)	Development set (%)							Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
Jitter (*local*)	96.09	82.16	96.10	99.67	97.86	98.95	7.57	93.66	78.47	95.42	97.61	96.50	97.16	**15.15**
Jitter (PPQ3)	96.49	84.49	96.61	99.59	98.07	98.97	6.83	92.94	**70.72**	**93.74**	98.73	96.17	97.69	10.45
Jitter (PPQ5)	95.99	82.42	96.17	99.49	97.80	98.81	7.18	92.81	83.15	96.61	**95.32**	95.96	**95.57**	11.64
Shimmer (*local*)	95.71	**80.68**	**95.80**	99.59	97.66	98.81	7.03	93.27	71.22	93.83	99.01	96.35	97.92	10.97
Shimmer (APQ3)	95.82	81.45	95.97	99.52	97.71	98.79	7.50	93.75	84.01	96.73	96.29	96.51	96.37	10.18
Shimmer (APQ5)	96.55	86.57	97.09	99.13	98.10	98.71	8.16	94.46	81.17	95.98	97.92	96.94	97.53	9.35
Shimmer (APQ11)	95.82	88.78	95.81	99.70	97.72	98.90	6.43	93.06	71.47	93.89	98.68	96.23	97.68	12.37
CHNR	**95.06**	84.88	96.81	**97.70**	**97.26**	**97.53**	**8.43**	**92.35**	79.25	95.71	95.76	**95.74**	95.75	10.88
NNE	95.93	82.42	96.18	99.42	97.78	98.76	8.35	93.45	76.74	95.04	97.80	96.40	97.24	11.54
GNE	96.10	82.00	96.06	99.74	97.87	98.99	6.12	94.28	81.36	96.05	97.64	96.84	97.31	9.80
All Features (10×159)	96.67	85.81	96.60	99.47	98.17	98.95	8.89	91.36	72.99	94.33	96.14	95.23	96.77	11.33

Table 4.16: Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on on the adaptation set of ADD 2022.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)	EER(%)
Jitter (*local*)	80.30	71.17	80.93	94.00	86.98	91.06	**23.00**
Jitter ($PPQ3$)	81.60	72.29	81.39	95.57	87.91	92.35	18.34
Jitter ($PPQ5$)	**80.00**	72.45	81.78	93.57	87.28	90.95	19.00
Shimmer (*local*)	80.80	71.05	80.68	95.43	87.43	92.06	**23.00**
Shimmer ($APQ3$)	81.30	72.17	81.40	95.00	87.67	91.93	20.67
Shimmer ($APQ5$)	80.50	72.17	81.68	93.00	86.79	90.49	22.00
Shimmer ($APQ11$)	**80.00**	72.57	82.22	**91.14**	**86.45**	**89.20**	21.33
CHNR	80.80	70.67	**80.38**	96.00	87.50	92.41	22.33
NNE	93.69	**69.88**	94.98	98.32	96.62	97.64	15.86
GNE	83.60	77.62	85.26	92.57	88.77	91.01	19.86

Table 4.17: Ablation study of segmental frames of analysis of speech-pathological features with ResNet-18 on on the adaptation set of ADD 2023.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)	EER(%)
Jitter (*local*)	98.91	99.31	99.80	98.83	99.40	99.06	0.35
Jitter ($PPQ3$)	97.28	98.34	**99.64**	97.07	98.50	97.64	1.00
Jitter ($PPQ5$)	96.13	97.83	99.98	95.79	97.84	96.60	0.78
Shimmer (*local*)	97.81	98.75	99.98	97.63	98.79	98.09	0.47
Shimmer ($APQ3$)	96.17	97.9	99.99	95.84	97.87	96.64	**1.83**
Shimmer ($APQ5$)	94.97	97.24	99.99	94.53	97.18	95.57	0.56
Shimmer ($APQ11$)	98.29	99.01	99.98	98.15	99.06	98.51	0.98
CHNR	97.98	98.68	99.95	97.84	98.89	98.26	0.99
NNE	97.87	98.76	99.98	97.70	98.82	98.14	0.05
GNE	**89.78**	**94.41**	99.99	**88.89**	**94.11**	**90.90**	0.56

These features are then fed into an MLP neural network for training and testing efficiency. The proposed method is evaluated using four datasets: ASVspoof 2019, ASVspoof 2021, ADD 2022, and ADD 2023 datasets.

Upon investigation, it was discovered that HNR is not useful for detecting deepfakes in all datasets. Consequently, in this chapter 10 speech-pathological features are combined, which include: jitter (*local*), jitter ($PPQ3$), jitter ($PPQ5$), shimmer (*local*), shimmer ($APQ3$), shimmer ($APQ5$), shimmer ($APQ11$), GNE, NNE, CHNR.

The results from ASVspoof 2019 indicate that the accuracy, recall, F1-score, and F2-score are 89.94%, 97.20%, 94.55, and 96.12%, respectively. For ASVspoof 2021, the accuracy, recall, F1-score, and F2-score are 84.56%, 92.19%, 91.43%, and 91.88%, respectively. The experimental outcomes from ADD 2022 demonstrate that the performance metrics accuracy, recall, F1-score, and F2-score stand at 70.10%, 99.00%, 82.75%, and 91.78% respectively. This indicates a robust model performance. While, the experiments conducted on ADD 2023 reveal even more impressive results. The accuracy, recall, F1-score, and F2-score are 95.58%, 99.08%, 97.62%, and 98.49% respectively. The above results suggest that the proposed method could effectively detect deepfake speech.This is achieved by utilizing a mere ten features in combination with a simple neural network.

In the search for the most significant attribute among acoustical speech-pathological features, it was discovered that shimmer ($APQ3$) holds the utmost importance in ASVspoof2019. Similarly, GNE is of paramount importance in ADD2022 and ADD2023 datasets.

For the second method, the goal is to improve the efficiency of speech-pathological features in detecting deepfake speech. Previous methods used an average of each feature, which limited their effectiveness. In this work, a new method is developed that extends the dimension of speech-pathological features by using segmental frames of analysis. This approach enhances the performance of speech-pathological features, making them more effective in detecting deepfake speech.

After applying the segmental frames of analysis to all speech-pathological features, it was found that HNR is not useful for detecting deepfakes in all datasets. The results from ASVspoof 2019 indicate that segmental frames of analysis significantly improves performance compared with using the average method in accuracy from 74.60% to 87.79%, recall from 79.00% to 95.70%, F1-score from 84.00% to 93.60%, and F2-score 81.20% to 95.00%, respectively, on development set.

For ASVspoof 2021, the performance of segmental frames of analysis improves performance compared with using the average method in accuracy, from 92.23% to 92.60%, recall from 97.86% to 99.09%, F1-score from 94.74% to 96.65%, and F2-score 96.68% to 97.84%, respectively, on development set.

The experimental outcomes from ADD 2022 demonstrate that segmental frames

of analysis improves performance compared with using the average method in accuracy from 70.10% to 87.60%, recall from 99.00% to 99.28%, F1-score from 82.75% to 93.54%, and F2-score 91.78% to 91.90%,respectively on adaptation set.

While, the experiments conducted on ADD 2023 reveal segmental frames of analysis improves performance compared with using the average method in accuracy from 95.58% to 99.80%, recall from 99.08% to 99.87%, F1-score from 97.62% to 99.89%, and F2-score 98.49% to 99.88%,respectively on adaptation set. The above results suggest that the proposed method could effectively detect deepfake speech.

Chapter 5

Deepfake Speech Detection using Perceptual Speech-Pathological Features

5.1 Proposed Method using Perceptual Speech-Pathological Features

This chapter proposes a method for detecting deepfake speech by using simple neural networks with a combination of the eight timbre features base on auditory perception as illustrated in Fig. 5.1. For, timbral feature extraction processes must be applied to the input speech signal. The result from the timbral feature consists of eight values. Each value is the average of each attribute, which includes: depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation.

For timbre feature extraction, Python code from the Audio Common timbre model [1] is implemented. After the eight timbre features were extracted, they were saved and imported to the Python program for training and testing.

The speech signals are set to 4 s, with a sample rate of 16 k. Note that to ensure all signals are 4 s long, signals shorter than 4 s are repeated from the beginning, whereas signals longer than 4 s are truncated.

A multilayer perceptron neural network (MLP) is then used as a classifier. The structure of the classifier comprises one node in the input layer, eight nodes in the hidden layer, and eight nodes in the output layer. The hidden layer is activated with the ReLU function, and the sigmoid function is the activation function in the output layer. The training configurations of the classifier consisted of a maximum of 100 epochs, learning rate of 0.0001, and batch size of 128. The loss function

[1] https://github.com/AudioCommons/ac-audio-extractor

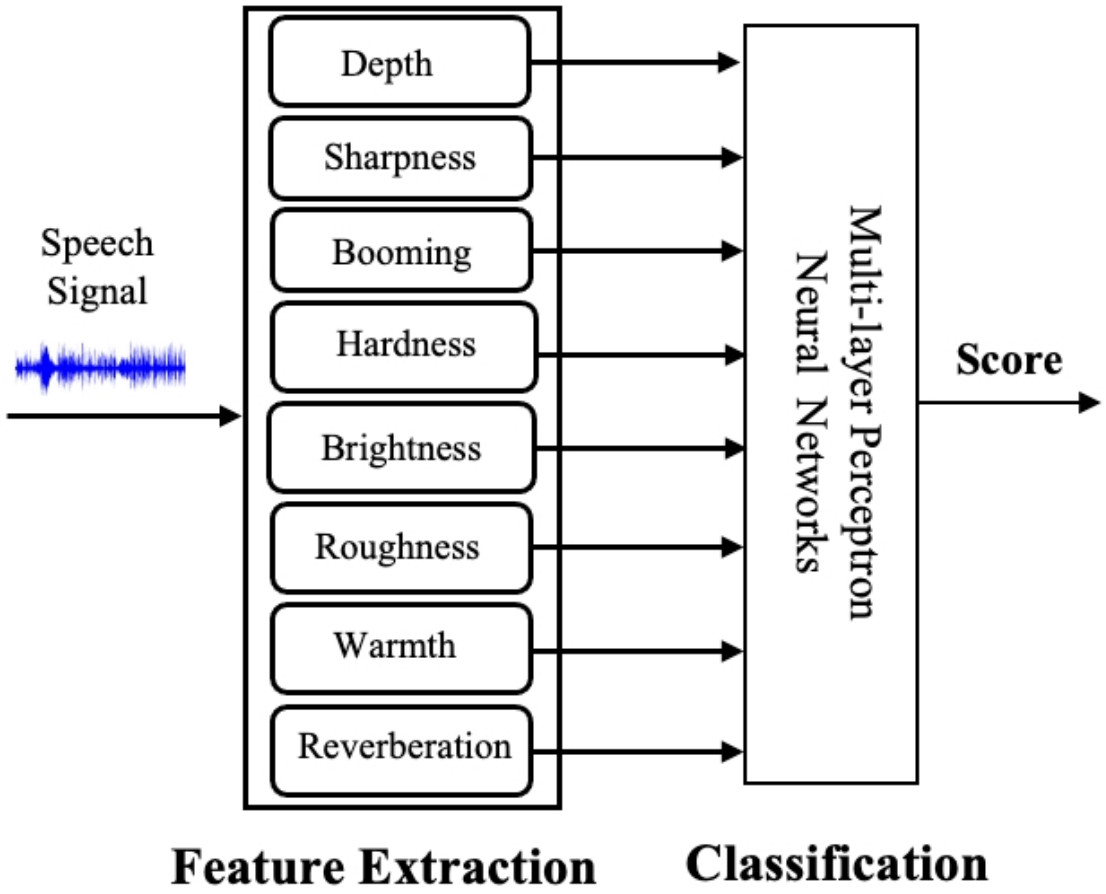

Figure 5.1: Proposed method using perceptual speech-pathological features with multi-layer perceptron neural networks.

was binary cross-entropy, and the Adam optimizer was used.

5.2 Results and discussion in ASVspoof 2019 and 2021 datasets

The effectiveness of the perceptual timbral features is shown in Table 5.1 when used with the neural network on the development set of the ASVspoof 2019 dataset. Among the timbral features, hardness exhibit particularly high performance in recall, achieving 92.74%. This superior performance can be attributed to the ability of hardness to correctly identify fake speech more effectively than other timbral features. Moreover, hardness suggests strong performance compared to other features with a balanced accuracy of 61.94% .

Table 5.1: Results from applying an average of timbral features with a neural network on the development set of ASVspoof 2019.

Timbral features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	73.10	**61.94**	75.84	**92.74**	83.44	78.71
Depth	80.00	51.90	75.92	90.21	82.45	78.21
Brightness	78.91	51.93	85.87	90.15	87.96	86.69
Roughness	66.72	50.17	70.99	89.97	79.29	74.09
Warmth	72.61	61.15	75.57	92.53	83.20	78.45
Sharpness	63.41	50.58	66.65	89.98	76.58	70.30
Boominess	66.49	58.99	68.42	92.23	78.56	72.15
Reverberation	77.38	58.82	82.17	91.76	86.70	83.92
Combining all features	**82.03**	61.57	**87.65**	92.24	**89.89**	**88.53**

Another feature is brightness, which exhibits a particularly high precision performance, achieving 85.87%.

When all timbral features are combined and used with the neural network, the results are impressive, demonstrating 82.03% accuracy, 61.57% balanced accuracy, 87.65% precision, 92.24% recall, 89.89% F1-score, and 88.53% F2-score. These metrics highlight the robustness and reliability of the timbral features in detecting deepfake speech.

These findings suggest that timbral features, when integrated with neural networks, have significant potential in the realm of deepfake speech detection. The high recall rates, in particular, indicate a strong capability to identify false positives, thereby enhancing the overall effectiveness of the detection system.

Table 5.2 show the effectiveness of the overall timbral features when used with the neural network on the evaluation sets of the ASVspoof 2019 and 2021 datasets. The reason for presenting the performance on the evaluation set of ASV2021 is that the training and development sets of ASVspoof 2021 consist of the same data as those in ASVspoof 2019. Consequently, the performance of the proposed method on the evaluation set of ASVspoof 2021 is the focus to provide a clear and distinct assessment.

These results indicate that timbral features have the potential to effectively detect deepfake speech, demonstrated by a high recall of 92.19%. This high recall rate underscores the method's ability to correctly identify a large proportion of fake speech instances, minimizing false negatives. This performance is crucial for practical applications where accurate and reliable detection of deepfake speech is essential.

Furthermore, the integration of timbral features with a neural network not only enhances recall but also contributes to the overall robustness of the detection system. By leveraging the unique characteristics of timbral features, the neural network can more accurately distinguish between genuine and fake speech, thereby improving the system's reliability and effectiveness in detecting deepfake speech.

Table 5.2: Results from applying an average of timbral features with a neural network on the evaluation set of ASVspoof 2019 and 2021.

All timbral features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
ASVspoof 2019	77.91	48.10	89.26	85.67	87.43	86.37
ASVspoof 2021	84.56	56.49	90.67	92.19	91.43	91.88

5.3 Results and discussion in ADD 2022 and 2023 datasets

5.3.1 Results and discussion in ADD 2022 dataset

Table 5.3 demonstrates the effectiveness of utilizing timbral features with the neural network on the adaptation set of the ADD 2022 dataset. Among the timbral features, roughness shows particularly high performance in recall, achieving 99.57%. Sharpness exhibits stronger performance than other features in terms of accuracy, balanced accuracy, and precision, achieving 72.00%, 55.81%, and 72.63%, respectively.

When all timbral features are combined with the neural network, the results are as follows: 75.10% accuracy, 62.50% balanced accuracy, 76.07% precision, 94.00% recall, 84.09% F1-score, and 89.76% F2-score. These results indicate that timbral features have significant potential for effectively detecting deepfake speech.

The high recall rates of roughness and depth suggest that these features are particularly adept at identifying instances of fake speech, reducing the likelihood of false negatives. This is crucial for applications that require high sensitivity in detecting deepfake speech to ensure the integrity and authenticity of audio content.

Additionally, the combined use of all timbral features with the neural network enhances the overall performance of the detection system. The metrics demonstrate a well-rounded capability, with strong precision and F-scores, indicating a balanced approach to both identifying fake speech and minimizing false positives. This comprehensive performance highlights the robustness and reliability of the proposed method in practical, where accurate deepfake detection is essential.

5.3.2 Results and discussion in ADD 2023

Table 5.4 highlights the effectiveness of integrating timbral features with the neural network on the development set of the ADD 2023 dataset. The sharpness feature exhibits strong accuracy and recall rates of 91.57% and 98.60%, respectively, compared to other timbral features, indicating their high capability in effectively identifying fake speech. Depth exhibits strong performance, achieving a balanced accuracy of 61.23% and a precision of 93.56%, which are higher than those of other features.

Upon combining all timbral features with the neural network, the results are as follows: the accuracy reaches 93.37%, the balanced accuracy stands at 67.54%, precision reaches 94.59%, recall achieves 98.40%, and both the F1-score and F2-score attain 96.46% and 97.62%, respectively.

These results demonstrate that the integration of timbral features with the

Table 5.3: Results from applying an average of timbral features with a neural network on the adaptation set of ADD 2022.

Timbral features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	70.10	52.35	70.00	96.71	81.91	90.19
Depth	71.80	53.95	71.72	98.57	83.03	91.70
Brightness	70.00	53.33	71.51	95.00	81.59	89.14
Roughness	70.00	50.29	70.12	**99.57**	82.29	**91.85**
Warmth	63.60	51.04	70.54	82.43	76.07	79.74
Sharpness	72.00	55.81	72.63	96.28	82.80	90.39
Boominess	66.50	47.88	69.07	94.42	79.78	87.97
Reverberation	70.70	54.40	72.00	95.14	81.96	89.39
Combining all features	**75.10**	**62.50**	**76.07**	94.00	**84.09**	89.76

Table 5.4: Results from applying an average of timbral features with a neural network on the adaptation set of ADD 2023.

Timbral features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	88.71	54.65	92.59	95.34	93.94	94.78
Depth	89.59	61.23	93.56	95.11	94.37	94.81
Brightness	88.03	52.62	92.27	94.93	93.58	94.38
Roughness	88.82	48.50	91.62	96.67	94.07	95.61
Warmth	79.20	51.61	92.14	87.57	88.19	85.99
Sharpness	91.57	55.44	92.68	**98.60**	95.56	97.36
Boominess	89.66	54.08	92.49	96.58	94.49	95.73
Reverberation	85.06	49.62	91.79	91.96	91.88	91.92
Combining all features	**93.37**	**67.54**	**94.59**	98.40	**96.46**	**97.62**

neural network significantly enhances the system's performance in detecting deepfake speech. The high recall rates of sharpness and roughness underscore their importance in identifying fake speech, reducing the likelihood of false negatives, which is critical for ensuring the integrity and reliability of audio content.

Furthermore, the combined use of all timbral features with the neural network results in a well-rounded detection system. The high precision rate of 94.59% indicates a strong ability to minimize false positives, while the robust F1 and F2 scores reflect a balanced and effective approach to deepfake speech detection.

The reason for the higher efficiency of all features in this dataset might be twofold. First, the training set and the adaptation set share similar characteristics. Second, this dataset has less background noise, specifically white and babble noises and, reverberation. In contrast, ADD 2022 contains various background noises such as background music, car engines, and people chatting. Additionally, while the training set for ADD 2022 consists of clean speech without noise, the adaptation set includes significant background noise. Consequently, the efficiency of all features on ADD 2023 surpasses that of ADD 2022. Moreover, from the observations, both the training and development sets exhibited a high signal-to-noise ratio (SNR). In contrast, the adaptation and test sets demonstrated a low SNR, characterized by a high level of various types of real-world background noise. Therefore, this research randomly applied data augmentation techniques, including the introduction of reverberation, babble, and music noise, during the extraction of the speech-pathological feature to enhance the diversity of the training set.

5.4 Ablation study of Perceptual Speech-Pathological Features

Table 5.5 depicts an ablation study on the development set of ASVspoof 2019, analyzing the performance of each timbral feature when utilized with a neural network. In this study, one timbral feature was removed at a time to assess the importance and potential of each feature for deepfake speech detection. The findings indicate that sharpness and brightness are the important features, as their removal led to the lowest performance across all metrics on the development set.

This ablation study highlights the critical role that the sharpness and brightness features play in the detection process. When sharpness and brightness are excluded, the neural network's ability to accurately identify and classify fake speech diminishes significantly. This reduction in performance metrics underscores sharpness's essential contribution to the overall effectiveness of the detection system. Moreover, this analysis provides valuable insights into the individual impact of each timbral feature, guiding future improvements and optimizations in deepfake

Table 5.5: Ablation study of applying an average of timbral features with a neural network on the development set of ASVSpoof 2019.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	75.83	55.73	81.01	91.07	85.74	82.83
Depth	78.43	66.90	81.41	93.73	87.14	83.60
Brightness	79.16	**47.35**	87.37	**89.13**	88.27	87.73
Roughness	71.22	63.28	73.27	93.21	82.04	76.51
Warmth	69.86	56.26	73.37	91.34	81.37	76.37
Sharpness	**68.33**	50.39	**72.96**	89.84	**80.52**	**75.80**
Boominess	76.53	63.28	79.94	92.91	85.94	82.39
Reverberation	84.51	53.62	92.49	90.47	91.47	92.08

speech detection methodologies. By understanding the relative importance of each feature, researchers and developers can prioritize the most influential ones, ensuring the development of more robust and reliable detection systems. Overall, the results from Table 5.5 emphasize the necessity of including the sharpness and brightness features for achieving high detection performance, reinforcing its value in the ongoing efforts to combat deepfake speech.

Table 5.6 illustrates an ablation study on the adaptation set of ADD 2022, evaluating the performance of individual timbral features when integrated with a neural network. Throughout this investigation, one timbral feature was eliminated at a time to assess the significance and efficacy of each feature for detecting deepfake speech. The results show that brightness emerges as the important feature, as its removal results in the lowest recall and F1-score on the adaptation set. This ablation study underscores the critical role that the brightness feature plays in the detection of deepfake speech. Furthermore, the analysis provides valuable insights into the individual impact of each timbral feature, allowing researchers to prioritize features that contribute most significantly to the detection process. By identifying and focusing on the most influential features, such as brightness, it becomes possible to enhance the overall effectiveness of deepfake detection systems.

As depicted in Table 5.7, an ablation study on the adaptation set of ADD 2023 provides a comprehensive evaluation of the performance of individual timbral features when they are incorporated into a neural network model. This study is not a cursory examination, but a thorough investigation where each timbral feature is systematically removed in a step-by-step process. The objective of this methodical approach is to gauge the significance and effectiveness of each feature in the complex task of detecting deepfake speech, a growing concern in the realm of digital communication. The results of this study are illuminating, particularly in the case of the reverberation feature. The findings underscore the critical role that reverberation plays in the detection process. When this feature is absent from the model, a notable decrease in the performance metrics is observed. Specifically, the accuracy, balanced-accuracy, and precision on the adaptation set all reach their lowest levels without the reverberation feature. This compelling evidence highlights the indispensable role of reverberation in the successful detection of deepfake speech, reinforcing its importance in the design of the neural network model.

Table 5.6: Ablation study of applying an average of timbral features with a neural network on the adaptation set of ADD 2022.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	73.00	57.28	73.32	96.57	83.35	90.81
Depth	**70.06**	55.09	72.36	93.86	**81.72**	88.59
Brightness	72.60	60.14	75.00	**91.29**	82.34	**87.49**
Roughness	72.90	57.98	73.07	95.29	83.11	90.01
Warmth	71.90	55.93	72.68	95.86	82.69	90.12
Sharpness	73.80	61.46	75.64	92.29	83.14	88.40
Boominess	70.70	**54.69**	**72.14**	94.71	81.90	89.14
Reverberation	72.60	56.05	72.71	97.43	83.27	91.23

Table 5.7: Ablation study of applying an average of timbral features with a neural network on the adaptation set of ADD 2023.

Excluded Feature	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Hardness	93.20	65.64	94.25	98.60	96.38	97.70
Depth	92.72	63.73	93.99	98.36	96.13	97.45
Brightness	92.85	63.97	94.02	98.47	96.20	97.55
Roughness	93.28	70.51	95.10	**97.72**	96.39	**97.18**
Warmth	92.24	63.40	93.94	98.01	95.93	**97.18**
Sharpness	92.91	62.68	93.81	98.80	96.24	97.76
Boominess	93.11	66.23	94.39	98.36	96.32	97.52
Reverberation	**92.15**	**60.75**	**93.51**	98.27	**95.83**	97.28

5.5 Summary

In this chapter, the potential of timbral features for distinguishing between genuine and deepfake speech is investigated. The eight attributes—depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation—are considered. After feature extraction, the average value of these features is calculated and fed into an MLP neural network for training and testing efficiency. The proposed method is evaluated using four datasets: ASVSpoof 2019, ASVSpoof 2021, ADD 2022, and ADD 2023. The ASVSpoof 2019 and 2021 datasets consist of speech samples without background noise. The model is trained using the training set and tested with the development set from ASVSpoof. However, for ASVSpoof 2021, evaluation is performed using the evaluation set since this dataset only has an evaluation set. The training and development sets remain the same as those used for ASVSpoof 2019. The ADD 2022 and 2023 datasets contain audio samples with various types of background noise, particularly in ADD 2022. These two ADD datasets are trained with the training set and tested with the adaptation set.

The results from ASVspoof 2019 indicate that the accuracy, recall, F1-score, and F2-score are 82.03%, 92.24%, 89.89%, and 88.53%, respectively. For ASVspoof 2021, the accuracy, recall, F1-score, and F2-score are 84.56%, 92.19%, 91.43%, and 91.88%, respectively. The experimental outcomes from ADD 2022 demonstrate that the performance metrics accuracy, recall, F1-score, and F2-score - stand at 70.10%, 94.00%, 84.09%, and 89.76% respectively. This indicates a robust model performance. While, the experiments conducted on ADD 2023 reveal even more impressive results. The accuracy, recall, F1-score, and F2-score have significantly improved to 93.37%, 98.40%, 96.46%, and 97.62% respectively.

The above results suggest that the proposed method could effectively detect deepfake speech.This is achieved by utilizing a mere eight features in combination with a simple neural network.

In the search for significant attributes among timbral features, hardness, sharpness, and brightness have been shown to be important in ASVspoof 2019. Similarly, sharpness, brightness, and roughness are of paramount importance in ADD 2022. Finally, depth, sharpness, and reverberation take precedence in ADD 2023.

Chapter 6

Deepfake Speech Detection using Acoustical and Perceptual Speech-Pathological Features

6.1 Proposed Method using Acoustical and Perceptual Speech-Pathological Features

This chapter selects the important acoustic and perceptual speech-pathological features from Chapters 5 and 6. After investigation, it is found that six significant acoustical speech-pathological features are utilized: jitter (*local*), shimmer ($APQ3$), shimmer ($APQ11$), GNE, NNE, and CHNR. Additionally, six important perceptual features are incorporated: sharpness, hardness, brightness, roughness, depth, and reverberation, as shown in Fig. 6.1. These features are combined and fed to an MLP for detecting deepfake speech. This approach aims to assess the effectiveness of the proposed method in identifying deepfake speech based on both acoustic and perceptual features.

A neural network-based classifier was implemented for this task. The network architecture consists of an input layer with 12 nodes, a single hidden layer with 12 nodes, and an output layer with a single node. The hidden layer utilizes the ReLU activation function and The output layer employs the sigmoid function. The training process was configured with a maximum of 100 epochs. A learning rate of 0.0001 was used to control the magnitude of weight updates. A batch size of 128 samples was chosen to balance computational efficiency and gradient estimation accuracy. Binary cross-entropy was chosen as the loss function to measure the discrepancy between the predicted and actual labels. The Adam optimizer was employed for efficient optimization of the network weights during training.

The results presented in Table 6.1 demonstrate the performance of the proposed

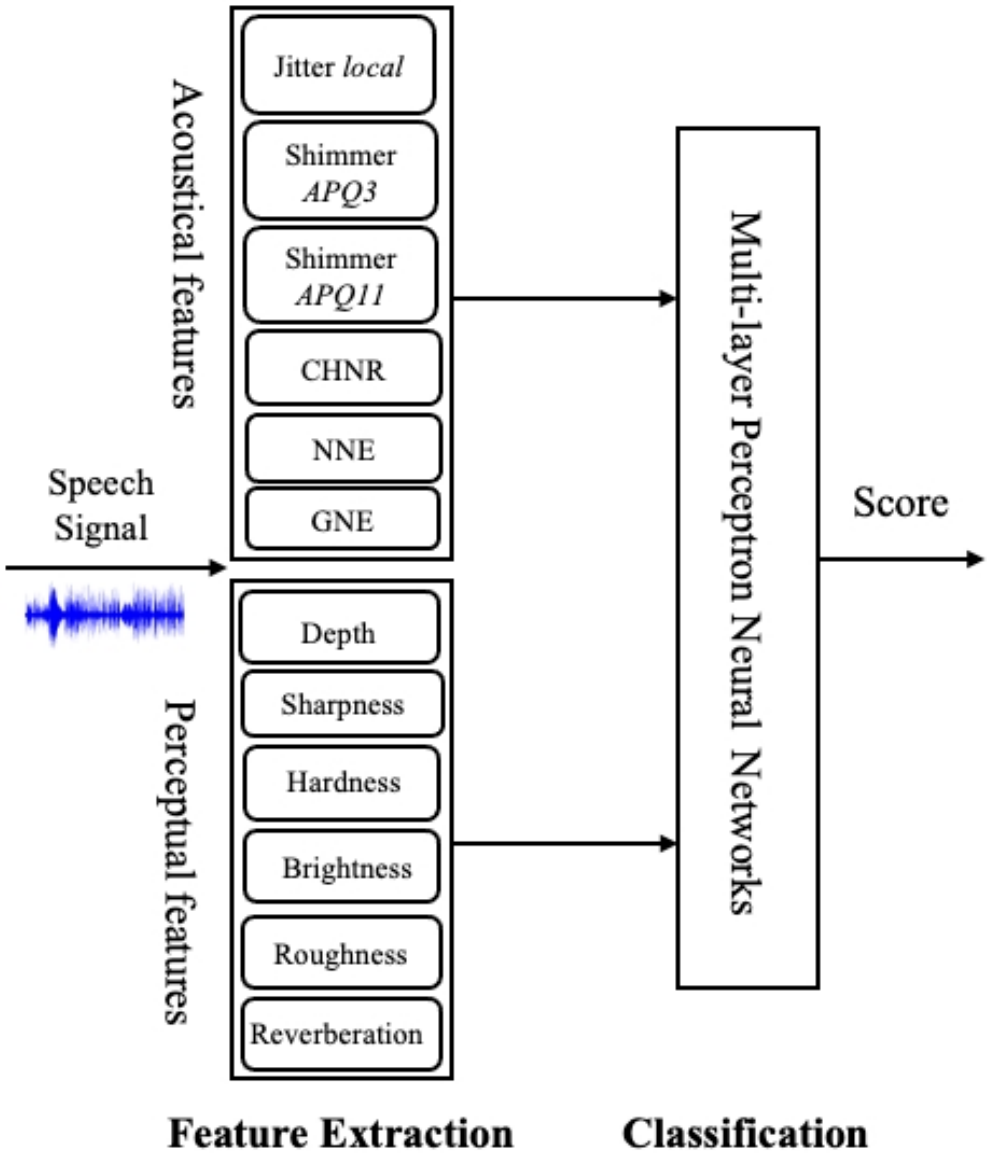

Figure 6.1: Proposed method using important acoustic and perceptual speech-pathological features with multi-layer perception neural networks.

method on four datasets: ASVspoof 2019, ASVspoof 2021, ADD 2022, and ADD 2023. The analysis reveals that the proposed method achieves high performance on ASVspoof 2019, ASVspoof 2021, and ADD 2022. In these datasets, accuracy surpasses 90% and recall exceeds 98%, indicating strong detection capabilities. However, the performance on the ADD 2023 dataset is considerably lower. Here, accuracy drops to around 70%, and other metrics also fall below those observed in the other datasets. The difference in performance between ADD 2022 and ADD 2023 datasets might be attributed to the presence of background noise. ADD 2022 exhibits a higher prevalence of diverse background noises, including engine sounds, music, and people talking. Conversely, ADD 2023, while not entirely free of background noise, features a significantly lower rate of such distractions. In contrast, the ASVspoof 2019 and ASVspoof 2021 datasets contain clean speech recordings devoid of any background noise.

6.2 Results and discussion in ASVspoof 2019 and 2021 datasets

The results in table 6.2 shows the comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the development set of ASVspoof 2019. The experimental results showed that when comparing the performance of acoustic features and perceptual features, acoustic features performed better in terms of accuracy, balanced accuracy, recall, F1-score, and F2-score. Furthermore, the experiments demonstrated that combining the important acoustic and perceptual features significantly improved performance metrics such as accuracy, recall, F1-score, and F2-score. This improvement was observed compared to using only acoustic features or perceptual features on their own.

Table 6.3 presents a comparison between combining acoustical and perceptual features and using each feature set individually. This analysis is based on the performance of a neural network trained on the ASVspoof 2019 dataset. The experimental results demonstrated that when comparing the performance of acoustic features and perceptual features, acoustic features outperformed perceptual features across all metrics. Additionally, the findings indicated that combining the important acoustic and perceptual features led to significantly better performance in terms of accuracy, recall, F1-score, and F2-score compared to using perceptual features alone. However, when comparing the combined features to acoustic features alone, it was observed that the combination performs better in terms of recall and F2-score, while the other metrics showed a slight decline, which was not statistically significant. This suggests that the proposed method enhances the de-

Table 6.1: Results of using acoustical and perceptual speech-pathological features with multi-layer perceptron neural networks on different datasets.

Datasets	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
ASVspoof 2019	90.61	58.30	91.32	98.94	94.97	97.3
ASVspoof 2021	90.07	57.23	91.05	98.61	94.68	97.00
ADD 2022	70.10	57.12	73.51	89.57	80.75	85.82
ADD 2023	95.62	84.20	97.18	98.04	97.61	97.87

tection of fake speech more effectively, particularly in a highly imbalanced dataset where fake speech is more prevalent than genuine speech.

6.3 Results and discussion in ADD 2022 and 2023 datasets

Table 6.4 compares the results of combining the important acoustical and perceptual features with the individual results of each feature type, using a neural network on the ADD 2022 adaptation set. The experimental results demonstrated that when comparing the performance of acoustic features and perceptual features, perceptual features outperformed acoustic features in terms of accuracy, balanced accuracy, precision, and F1-score. This indicates that perceptual features perform well across various levels of background noise. When combining the important acoustic and perceptual features, the proposed method showed better accuracy, balanced accuracy and precision compared to using acoustic features alone, although the other metrics were lower. Furthermore, when comparing the combination of features to individual perceptual features, the combination was found to be less effective. This suggests that the proposed method's performance decreases with varying levels of background noise.

Table 6.5 presents a comparison between the results obtained by combining acoustical and perceptual features and the results from each feature type individually, utilizing a neural network on the ADD 2023 development set.

The experimental results revealed that acoustic features consistently outperformed perceptual features across all evaluation metrics. Furthermore, combining these informative acoustic features with key perceptual features yielded significant improvements in performance on nearly all metrics compared to using perceptual features alone. Notably, the proposed combination of features surpassed the performance of acoustic features in terms of accuracy, balanced accuracy, precision, and F1-score. This suggests that the combined feature set offers a more robust approach for discriminating between genuine and fake speech, particularly in the presence of low level of background noise.

6.4 Discussion

This chapter investigates the efficacy of combining important acoustic and perceptual features within a MLP neural network for deepfake speech detection. The proposed method achieved superior performance on datasets with clean or low background noise (ASVspoof 2019, ASVspoof 2021, and ADD 2023). This suggests

Table 6.2: Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the development set of ASVspoof 2019.

Features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Acoustical Features	89.94	**61.82**	92.04	97.20	94.55	96.12
Perceptual Features	82.03	61.57	**92.24**	87.65	89.89	88.53
Proposed features	**90.61**	58.30	91.32	**98.94**	**94.97**	**97.31**

Table 6.3: Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the evaluation set of ASVspoof 2021.

Features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Acoustical Features	**90.23**	**60.36**	**91.83**	97.86	**94.74**	96.68
Perceptual Features	84.56	56.49	90.67	92.19	91.43	91.88
Proposed features	89.12	52.14	89.73	**99.17**	94.22	**97.13**

Table 6.4: Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the adaptation set of ADD 2022.

Features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Acoustical Features	71.10	52.50	71.07	**99.00**	82.75	**91.78**
Perceptual Features	**75.10**	**62.50**	**76.07**	94.00	**84.09**	89.76
Proposed features	71.40	55.67	72.60	95.00	82.30	89.48

Table 6.5: Comparison of the results from the combination of acoustical and perceptual features with the individual results of each acoustical and perceptual feature using a neural network on the adaptation set of ADD 2023.

Features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
Acoustical Features	95.58	77.61	96.22	**99.08**	**97.62**	**98.49**
Perceptual Features	93.37	67.54	94.59	98.40	96.46	97.62
Proposed features	**95.64**	**84.34**	**97.21**	98.04	**97.62**	97.87

that both acoustic and perceptual features can effectively capture characteristics of genuine speech when the environment is relatively noise-free.

However, the performance drop observed in the ADD 2022 dataset highlights the limitations of the current approach when dealing with significant background noise. The high prevalence of various noises, such as engine car, pepole chating, music noise, white, pink, and babble noise, likely disrupts the informative content within the speech signal, making it more challenging for both acoustic and perceptual features to accurately distinguish genuine from spoofed speech. This warrants further investigation into feature extraction techniques specifically designed to be robust against background noise.

Exploring feature normalization or noise reduction techniques for audio data pre-processing before feature extraction presents a promising avenue for future research. Additionally, investigating alternative feature sets or feature fusion methods specifically designed for noisy environments could be fruitful. Addressing these challenges will enhance the overall robustness and generalizability of the proposed method for real-world deepfake speech detection scenarios.

6.5 Summary

This chapter proposes of combining important acoustical and perceptual speech-pathological features for deepfake speech detection using a MLP neural network. Six acoustical speech-pathological features are utilized, including jitter (*local*), shimmer ($APQ3$), shimmer ($APQ11$), GNE, NNE, and CHNR. Additionally, six perceptual features are incorporated, encompassing sharpness, hardness, depth, brightness, roughness, and reverberation.

The proposed method was evaluated using four different datasets. The results demonstrated that when the important acoustic and perceptual features were combined, the performance improved across almost all datasets compared to when these features were used individually. This suggests that the integration of acoustic and perceptual features enhances the method's ability to accurately analyze and interpret the data. Therefore, the utilization of speech-pathological features can be employed for the detection of deepfake speech.

Chapter 7

Conclusion

7.1 Summary

Deepfake speech refers to a synthesized human voice generated using advanced voice conversion and text-to-speech techniques. It finds applications in various domains, such as audio books, customer services, and virtual assistants. However, the misuse of deepfake speech poses a significant threat to economies and societies. Therefore, detecting deepfake speech is crucial for fraud protection and ensuring the reliability of automatic speaker verification (ASV) systems.

Detecting deepfake speech has involved using several advanced techniques primarily focusing on two approaches: creating efficient classifiers and exploring acoustic features. In this dissertation, aim to investigate speech-pathological features to detect deepfake speech, which may also aid in identifying voice disorders caused by abnormalities in the human speech production mechanism. These disorders often manifest as unnaturalness. Since deepfake speech is also characterized by unnaturalness, the hypothesis suggests that it might mimic the perceived acoustic quality of a disordered voice. Therefore, speech-pathological features can be crucial clues for deepfake speech detection. Inspired by the human speech production mechanism, this process is complex and difficult to replicate artificially. Moreover, the tiny variations in speech production mechanism are unique to speaker individuality. Although advances in deepfake speech have made it possible to create increasingly realistic speech, it is still constrained and challenging to replicate.

The goal of this research is to propose a method for detecting deepfake based on speech-pathological features with three research questions.

- To answer the first question of whether speech pathological features is used to detect disorder voices can detect deepfake speech. In this study, two types of speech-pathological features were investigated: acoustic and perceptual features. The acoustic speech-pathological features include three types of jitter, four types of shimmer, CHNR, NNE, and GNE. The perceptual speech-pathological features consist of depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation. When these two sets of speech-pathological features were used with an MLP neural network, the results indicated that acoustic features and perceptual features could effectively detect deepfake speech.

- To address the second question of whether speech-pathological features can be enhanced to detect deepfake speech, extending the range of acoustic speech-pathological features proves more efficient compared to others. By utilizing segmental frames of analysis techniques in acoustic speech-pathological features without HNR, the detection of deepfake speech can be significantly improved.

- To answer the last research question: which features are important in detecting deepfake speech. The results of the experiment show that shimmer ($APQ11$), GNE, NNE, CHNR, sharpness, hardness, and brightness are important for detecting deepfake speech without background noise, while jitter *local*, shimmer ($APQ3$), shimmer ($APQ11$), GNE, brightness, sharpness, depth, roughness, and reverberation are important for detecting deepfake speech in noisy environments.

7.2 Contributions

This study focuses on deepfake speech detection by utilizing speech-pathological features, which are traditionally used to detect voice disorders, drawing inspiration from human speech production mechanisms. By integrating these features, the study aims to enhance the accuracy and robustness of identifying deepfake speech. Therefore, this research contributes significantly to society by improving the security and reliability of digital speech communication systems. For example, by analyzing voiceprints for authenticity, these systems can enhance security measures in various applications, such as voice banking, access control, and forensic investigations. This leads to more reliable biometric authentication and safeguards

against fraudulent activities. Furthermore, this research sheds light on the fundamental mechanisms of human speech production, advancing knowledge beyond the realm of security and technology. Its findings hold significant implications for diverse fields, including speech emotion recognition, the biological basis of language processing, linguistic theories of speech production and perception, the psychology of auditory attention, and the cognitive processes underlying communication.

7.3 Remaining Works

To improve the efficiency of using speech-pathological features in detecting deepfakes, future work will focus on the following areas.

- Analyzing the feature extraction of speech-pathological features for detecting deepfake speech in more detail, consider the following aspects. First, investigate the speech length—determining the optimal duration for detecting deepfake speech. Second, explore the sampling rate of the speech signal, aiming to identify a suitable rate that significantly contributes to deepfake detection. Lastly, examine the frame length, seeking an appropriate frame size for effective deepfake speech detection.

- Investigating the effectiveness of additional acoustic speech-pathological features in detecting deepfake speech. These features include the largest Lyapunov exponent (LLE), rate of points above linear average (RALA), and correlation dimension (D2). These features have been shown to be efficient in detecting voice disorders, suggesting their potential applicability in uncovering the subtle manipulations characteristic of deepfakes. Therefore, hypothesize that these features might perform well in differentiating between genuine and deepfake speech.

- Applying segmental frames of analysis to perceptual speech pathological features, similar to how they are used with acoustic speech-pathological features, can be beneficial. These perceptual features, including depth, sharpness, booming, hardness, brightness, roughness, warmth, and reverberation, can then be investigated for their importance in detecting deepfake speech.

Bibliography

[1] H. C. Mahendru, "Quick review of human speech production mechanism," *International Journal of Engineering Research and Development*, vol. 9, no. 10, pp. 48–54, 2014.

[2] R. Singh and R. Singh, "Production and perception of voice," *Profiling Humans from their Voice*, pp. 27–83, 2019.

[3] A. S.-L.-H. Association *et al.*, "Consensus auditory-perceptual evaluation of voice (cape-v)," *Rockville: ASHA Special Interest Division*, vol. 3, pp. 1–3, 2002.

[4] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947–e11, 2019.

[5] J. Gómez-García, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part iii: Review of acoustic modelling strategies," *Biomedical Signal Processing and Control*, vol. 66, p. 102049, 2021.

[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] 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.

[8] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, *et al.*, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[9] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, X. Yan, L. Xu, Z. Wen, and H. Li, "Add 2022: the first audio deep synthesis detection challenge," in *Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220, 2022.

[10] W. Ge, M. Panariello, J. Patino, M. Todisco, and N. Evans, "Partially-connected differentiable architecture search for deepfake and spoofing detection," *arXiv preprint arXiv:2104.03123*, 2021.

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

[12] Z. Wang, S. Cui, X. Kang, W. Sun, and Z. Li, "Densely connected convolutional network for audio spoofing detection," in *Proc. 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1352–1360, IEEE, 2020.

[13] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[14] Z. Zhang, X. Yi, and X. Zhao, "Fake speech detection using residual network with transformer encoder," in *Proc. Proceedings of the 2021 ACM workshop on information hiding and multimedia security*, pp. 13–22, 2021.

[15] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, p. 102622, 2020.

[16] 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.

[17] K. Li, Y. Wang, M. Le Nguyen, M. Akagi, and M. Unoki, "Analysis of amplitude and frequency perturbation in the voice for fake audio detection," in *Proc. 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 929–936, IEEE, 2022.

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

[19] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *arXiv preprint arXiv:1906.09890*, 2019.

[20] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9226–9230, IEEE, 2022.

[21] Z. Lv, S. Zhang, K. Tang, and P. Hu, "Fake audio detection based on unsupervised pretraining models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9231–9235, 2022.

[22] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9241–9245, 2022.

[23] K. Li, X. Lu, M. Akagi, and M. Unoki, "Contributions of jitter and shimmer in the voice for fake audio detection," *IEEE Access*, 2023.

[24] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, "Efficient volume exploration using the gaussian mixture model," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 11, pp. 1560–1573, 2011.

[25] S. Duraibi, W. Alhamdani, and F. T. Sheldon, "Replay spoof attack detection using deep neural networks for classification," in *Proc. 2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 170–174, IEEE, 2020.

[26] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[27] J. Wu, "Introduction to convolutional neural networks," *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495, 2017.

[28] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: a survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[29] 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.

[30] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, pp. 2087–2091, 2015.

[31] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *Proc. 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 1–3, IEEE, 2019.

[32] J. Yang and R. K. Das, "Low frequency frame-wise normalization over constant-Q transform for playback speech detection," *Digital Signal Processing*, vol. 89, pp. 30–39, 2019.

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

[34] 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, pp. 101–114, 2020.

[35] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: disadvantages and limitations," *Int J Res Eng Technol (UAE)*, vol. 7, pp. 234–239, 2018.

[36] V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? an introduction to speech production and description for speaker classification," *Speaker Classification I: Fundamentals, Features, and Methods*, pp. 1–20, 2007.

[37] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[38] Z. Xie, C. Gadepalli, J. Farideh, B. M. Cheetham, and J. J. Homer, "Machine learning applied to grbas voice quality assessment," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 6, pp. 329–338, 2018.

[39] T. Kojima, S. Fujimura, K. Hasebe, Y. Okanoue, O. Shuya, R. Yuki, K. Shoji, R. Hori, Y. Kishimoto, and K. Omori, "Objective assessment of

pathological voice using artificial intelligence based on the grbas scale," *Journal of Voice*, 2021.

[40] K. Hartmann and K. Giles, "The next generation of cyber-enabled information warfare," in *Proc. 2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300, pp. 233–250, IEEE, 2020.

[41] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, "Warning: humans cannot reliably detect speech deepfakes," *Plos one*, vol. 18, no. 8, p. e0285333, 2023.

[42] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 10, pp. 4633–4644, 2017.

[43] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, pp. 1–14, 2021.

[44] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proc. Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5859–5866, 2020.

[45] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using gaussian probability feature for spoofing speech detection.," in *Proc. Interspeech*, pp. 1116–1120, 2020.

[46] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[47] P. Aravind, U. Nechiyil, N. Paramparambath, *et al.*, "Audio spoofing verification using deep convolutional neural networks by transfer learning," *arXiv preprint arXiv:2008.03464*, 2020.

[48] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, pp. 1–12, 2021.

[49] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162857–162868, 2021.

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

[51] I. Deary, J. Wilson, P. Carding, and K. Mackenzie, "The dysphonic voice heard by me, you and it: differential associations with personality and psychological distress," *Clinical Otolaryngology & Allied Sciences*, vol. 28, no. 4, pp. 374–378, 2003.

[52] S. Niimi, "Voice and speech disorders due to neuro-muscular diseases," *Nihon Kikan Shokudoka Gakkai Kaiho*, vol. 42, no. 5, pp. 394–399, 1991.

[53] M. Y. Chen, K. N. Stevens, H.-K. J. Kuo, and H. Chen, "Contributions of the study of disordered speech to speech production models," *Journal of Phonetics*, vol. 28, no. 3, pp. 303–312, 2000.

[54] M. Dietrich, R. D. Andreatta, Y. Jiang, and J. C. Stemple, "Limbic and cortical control of phonation for speech in response to a public speech preparation stressor," *Brain Imaging and Behavior*, vol. 14, pp. 1696–1713, 2020.

[55] J. Oates, "Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[56] Y. Liu, *Automatic Assessment of Disordered Voice Based on Diverse Speech Tasks*. The Chinese University of Hong Kong (Hong Kong), 2019.

[57] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474–485, 2007.

[58] D. A. Cairns, J. H. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE transactions on biomedical engineering*, vol. 43, no. 1, p. 35, 1996.

[59] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[60] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in smart cities," *Ieee Access*, vol. 5, pp. 3900–3908, 2017.

[61] S. Hadjitodorov and P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Medical engineering & physics*, vol. 24, no. 6, pp. 419–429, 2002.

[62] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *Proc. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1243–1247, IEEE, 2017.

[63] S. Hidaka, Y. Lee, K. Wakamiya, T. Nakagawa, and T. Kaburagi, "Automatic estimation of pathological voice quality based on recurrent neural network using amplitude and phase spectrogram.," in *Proc. INTERSPEECH*, pp. 3880–3884, 2020.

[64] M. Jouaiti, P. Kirby, and R. Vaidyanathan, "Matching acoustic and perceptual measures of phonation assessment in disordered speech-a case study," in *Proc. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, pp. 4508–4512, 2023.

[65] İ. Kurt, S. Ulukaya, and O. Erdem, "Musical feature based classification of parkinson's disease using dysphonic speech," in *Proc. 2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pp. 1–4, IEEE, 2018.

[66] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *Journal of Voice*, vol. 22, no. 1, pp. 1–9, 2008.

[67] C. R. Watts, R. Clark, and S. Early, "Acoustic measures of phonatory improvement secondary to treatment by oral corticosteroids in a professional singer: a case report," *Journal of Voice*, vol. 15, no. 1, pp. 115–121, 2001.

[68] K. Shama, A. Krishna, and N. U. Cholayya, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2006.

[69] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research*, vol. 43, no. 2, pp. 469–485, 2000.

[70] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 1, pp. 68–73, 2000.

[71] "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[72] J. A. Gómez García, *Contributions to the design of automatic voice quality analysis systems using speech technologies*. PhD thesis, Telecomunicacion, 2018.

[73] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang, and Z. Tao, "Investigation and evaluation of glottal flow waveform for voice pathology detection," *IEEE Access*, vol. 9, pp. 30–44, 2020.

[74] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[75] D. Meghraoui, B. Boudraa, T. Merazi-Meksen, and P. G. Vilda, "A novel preprocessing technique in pathologic voice detection: Application to parkinson's disease phonation," *Biomedical Signal Processing and Control*, vol. 68, p. 102604, 2021.

[76] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66749–66776, 2020.

[77] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.

[78] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition.," in *Proc. 8th Annual Conference of the International Speech Communication Association*, pp. 778–81, 2007.

[79] J. J. Jiang, D. B. Wexler, I. R. Titze, and S. D. Gray, "Fundamental frequency and amplitude perturbation in reconstructed canine vocal folds," *Annals of Otology, Rhinology & Laryngology*, vol. 103, no. 2, pp. 145–148, 1994.

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

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

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

[83] A. Pooransingh and D. Dhoray, "Similarity analysis of modern genre music based on billboard hits," *IEEE Access*, vol. 9, pp. 144916–144926, 2021.

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

[85] 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.

[86] 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.

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

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

[89] 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.

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

[91] 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.

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

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

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

[95] 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.

[96] 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*, pp. 183–190, 2019.

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

[98] 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.

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

[100] S. Hatano and T. Hashimoto, "Booming index as a measure for evaluating booming sensation," in *Proc. Inter-Noise*, no. 233, pp. 1–6, 2000.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

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

[110] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[111] M. Yousefi and J. H. Hansen, "Speaker conditioning of acoustic models using affine transformation for multi-speaker speech recognition," in *Proc. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 283–288, 2021.

[112] S. M.B and S. Vijayan, "Parkinson's disease prognosis using the resnet-50 model from speech features," in *Proc. 2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD)*, pp. 282–286, 2022.

[113] A. Kumar, S. S. Mahmoud, Y. Wang, S. Faisal, and Q. Fang, "A comparison of time-frequency distributions for deep learning-based speech assessment of aphasic patients," in *Proc. 2022 15th International Conference on Human System Interaction (HSI)*, pp. 1–5, 2022.

[114] R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou, and X. Li, "Audio deepfake detection system with neural stitching for add 2022," in *Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9226–9230, 2022.

[115] M. Hafizur Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, and A. Lawson, "Detecting synthetic speech manipulation in real audio recordings," in *Proc. 2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2022.

[116] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "Resnet and model fusion for automatic spoofing detection.," in *Proc. Interspeech*, pp. 102–106, 2017.

Publications

International Journal

1. Anuwat Chaiwongyen, Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, Masashi Unoki, "Potential of Speech-Pathological Features for Deepfake Speech Detection", *IEEE Access*, vol. 12, pp. 121958-121970, 2024, *DOI:10.1109/ACCESS.2024.3447582*.

2. 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. Institute of Electronics, Information and Communications Engineers (IEICE). 2024, *DOI:10.1587/transinf.2024mui0001*.

Domestic Journal (Research note)

1. Anuwat Chaiwongyen, Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, Masashi Unoki, "Replay Attack Detection in Automatic Speaker Verification Using Gammatone Cepstral Coefficients and ResNet-Based Model", J. Signal Processing, vol. 26, no. 6, pp. 171-175, Nov. 2022.

International Conference

1. Anuwat Chaiwongyen, Pinkeaw, K., Kongprawechnon, W., Karnjana, J., and Unoki, M. (2021, December). "Replay Attack Detection in Automatic Speaker Verification Based on ResNeWt18 with Linear Frequency Cepstral Coefficients". In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing* (iSAI-NLP) (pp. 1-5).

2. Anuwat Chaiwongyen, Songsriboonsit, N., Duangpummet, S., Karnjana, J., Kongprawechnon, W., and Unoki, M. (2022, November). "Contribution of timbre and shimmer features to deepfake speech detection". In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC) (pp. 97-103).

3. Anuwat Chaiwongyen, Duangpummet, S., Karnjana, J., Kongprawechnon, W., and Unoki, M. (2023, November). "Deepfake-speech Detection with Pathological Features and Neural Networks". In *2023 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC).

4. Kasorn Galajit, Thunpisit Kosolsriwiwat, Candy Olivia Mawalim, Pakinee Aimmanee, Waree Kongprawechnon, Win Pa Pa, Anuwat Chaiwongyen, Teeradaj Racharak, Hayati Yassin, Jessada Karnjana, Surasak Boonkla and Masashi Unoki, "ThaiSpoof: A Database for Spoof Detection in Thai Language", In *Proceedings of the 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing*(iSAI-NLP 2023), Bangkok, Thailand, November 27-29, 2023.