TÜRKİYE CUMHURİYETİ
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
ELECTRICAL AND ELECTRONIC ENGINEERING DEPARTMENT

ROBUST END-TO-END SYNTHETIC SPEECH DETECTION WITH DEEP NEURAL NETWORKS AND MASKING

BARIŞ AYDIN

M.Sc.

ADANA 2023

TÜRKİYE CUMHURİYETİ
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
ELECTRICAL AND ELECTRONIC ENGINEERING DEPARTMENT

ROBUST END-TO-END SYNTHETIC SPEECH DETECTION WITH DEEP NEURAL NETWORKS AND MASKING

BARIŞ AYDIN

M.Sc.

THESIS ADVISOR
ASST. PROF. DR. GÖKAY DİŞKEN

ADANA, 2023

DECLARATION OF CONFORMITY

In this thesis study, which was prepared following the thesis writing rules of Adana Alparslan Türkeş Science and Technology University Institute of Graduate School, I declare that I provide all the information, documents, evaluations and results in accordance with scientific ethics and moral codes without resorting to any means or assistance that would be contrary to scientific ethics and traditions. I also declare that I refer to all of the articles I used in this study with appropriate references and accept all moral and legal consequences if a situation is found contrary to my statement regarding my work.

12/08/2023

[Signature]

Barış AYDIN

ÖZET

Sağlam Sonlu Durumlu Yapay Konuşma Algılama: Derin Sinir Ağları ve Maskelerle Güçlendirilmiş Bütünsel Yaklaşım

Barış AYDIN
Yüksek Lisans, Elektrik-ElektronikAnabilim Dalı
Danışman: Dr. Öğr. Üyesi Gökay DİŞKEN
Ağustos 2023, 44 sayfa

Bu tez, gürültülü koşullar altında sentetik konuşma algılamanın i-vectörlerinin sağlamlığını artırmak için bir yöntem önermektedir. İ-vectörler, konuşmacı tanıma sistemlerinde yaygın olarak kullanılan sabit uzunluktaki temsillemelerdir. Ancak, performansları gürültü ile bozulmaktadır. Sistemi korumak için, gürültü maskesi üretmek için evrişimsel sinir ağı (CNN) kullanılması önerilmiştir. Bu maske, gürültü tarafından bozulan konuşma spektrogramındaki güvenilmez bölgeleri bastırır. Maske uygulanan spektrogram daha sağlam i-vectörlerin çıkarılması için kullanılır. Deneyler, eklenmiş gürültü, beyaz gürültü ve araba gürültüsü içeren ASVspoof 2015 veri kümesi kullanılarak yapılmıştır. CNN, her spektrogram çerçevesinde sinyal-gürültü oranını tahmin etmek üzere eğitilir. Bu, i-vectör çıkarılmasından önce uygulanan gürültü maskesini oluşturur. Sonuçlar, önerilen maskeleme yaklaşımının standart i-vectörlerle karşılaştırıldığında eşit hata oranlarını %50'den fazla azalttığını göstermektedir. Ancak, performans, CNN eğitimi sırasında görülmeyen araba gürültüsü üzerinde bozulur. Bu, daha çeşitli eğitim gürültü türlerine ihtiyaç duyulduğunu vurgular.

Sonuç olarak, spektrogram maskesi kullanma tekniği ile derin öğrenme tabanlı bir CNN, gürültülü koşullarda i-vectörlerin sağlamlığını artırabilir. Gürültü maskesi, güvenilmez bölgeleri bastırmaya yardımcı olarak daha iyi sahtecilik karşıtı performans sağlar. Ancak, maske görünmeyen gürültü türlerine iyi genelleşmez. Genel olarak, çalışma, gürültü altında sahtecilik saldırılarına karşı konuşmacı tanıma sistemlerinin güvenliğini artırmak için derin öğrenmeye dayalı maskelerin potansiyelini göstermektedir. Ancak daha fazla araştırma, çeşitli gürültü koşullarını ele alma konusunda gereklidir.

Anahtar Kelimeler: derin öğrenme, evrişimsel sinir ağı, sahte konuşma algılama, konuşmacı tanıma, gürbüz özellikler

ABSTRACT

ROBUST END-TO-END SYNTHETIC SPEECH DETECTION WITH DEEP NEURAL NETWORKS AND MASKING

Barış AYDIN

M.Sc., Department of Electrical and Electronic Engineering

Supervisor: Asst. Prof. Dr. Gökay DİŞKEN

August 2023, 44 pages

This thesis proposes a method to improve the robustness of i-vectors for synthetic speech detection under noisy conditions. I-vectors are fixed-length representations commonly used in speaker recognition systems. However, their performance degrades with noise. In order protect the system, using a convolutional neural network (CNN) to generate a noise mask is proposed. This mask suppresses unreliable regions in the speech spectrogram corrupted by noise. The masked spectrogram is then used to extract more robust i-vectors. Experiments use the ASVspoof 2015 dataset with added babble, white, and car noise. The CNN is trained to estimate the signal-to-noise ratio in each spectrogram frame. This generates the noise mask that is applied before i-vector extraction. Results show the proposed masking approach reduces equal error rates by over 50% compared to standard i-vectors from noisy speech. However, performance degrades on car noise which was not seen during CNN training. This highlights the need for more diverse training noise types.

In conclusion, the proposed spectrogram masking technique using a CNN can increase robustness of i-vectors for synthetic speech detection in noisy conditions. The noise mask helps suppress unreliable regions to provide improved anti-spoofing performance. However, the mask does not generalize well to unseen noise types. Overall, the study shows potential for deep learning-based masking to improve security of speaker recognition systems against spoofing attacks under noise. But more research is needed into handling diverse noise conditions.

Keywords: deep learning, convolutional neural network, spoof detection, speaker recognition, robust features

.

Dedication

To my grandfather

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my esteemed advisor, Asst. Prof. Dr. Gökay DİŞKEN, for his close involvement and unwavering support throughout every stage of my thesis. I am grateful for his profound knowledge and expertise, which have continuously enriched my learning experience by imparting new insights and perspectives.

Another thanks to my cats for their motivational support during the implementation and writing of my thesis.

Lastly, I would like to express my heartfelt gratitude to my father, Mustafa AYDIN, my mother, Fatma AYDIN, and my elder brother, Nuri Burak AYDIN, for their unconditional love and unwavering support throughout my journey toward success, regardless of the circumstances. Their constant encouragement has been invaluable to me, and I am thankful for their unwavering belief in me.

This thesis was partially supported by TÜBİTAK with project number 121E057.

TABLE OF CONTENTS

CERTIFICATION OF APPROVAL Hata! Yer işareti tanımlanmamış.
ÖZET **ii**
ABSTRACT **iii**
Dedication **iv**
TABLE OF CONTENTS **vi**
LIST OF FIGURES **viii**
LIST OF TABLES **ix**
LIST OF ABBREVIATIONS **x**
1. INTRODUCTION **1**
1.1 What is sound? 2
1.2 Speech Recognition 3
1.3 What are the challenges of Speaker Recognition Systems? 6
1.4 Spoofing in Speech 9
2. THEORETICAL FOUNDATIONS AND LITERATURE REVIEW **11**
2.1 Short Time Features 13
2.2 Synthetic Speech Recognition Features 15
2.2.1 MFCC 17
2.2.2 CQCC 20
2.2.3 i-vectors 21
2.3 Speaker Modelling 23
2.3.1 GMM-UBM Model 33
2.4 Artificial Neural Networks 35
2.4.1 Convolutional Neural Networks 38
2.5 Noise 39
2.6 Noise reducing/removing masks 40
3. MATERIALS AND METHODS **42**
3.1 ASVspoof 42
3.2 Noise 44
3.3 CNN Parameters 44
3.4 I-Vector Parameters 44
4. ANALYSIS AND DISCUSSION **47**

5. CONCLUSION **51**

REFERENCES **52**

VİTA Hata! Yer işareti tanımlanmamış.

LIST OF FIGURES

Figure 1.1. General structure of the human voice production mechanism……3
Figure 1.2. Speaker recognition system principle……..…4
Figure 2.1. General speakar recognition system……..….13
Figure 2.2. Mel filter bank……...18
Figure 2.3. Hidden Markov Model……..27
Figure 2.4. An Overview of Statistical Pattern Recognition Techniques for Speaker Verification……...29
Figure 3.1. CNN Architecture……..…46

Figure 4.1. Spectrogram of speech data corrupted with babble 0 dB (upper-left) and car 20 dB (upper-right), and their masks below……..……49

LIST OF TABLES

Table 1.1. Biometric Identification system properties……. 2

Table 3.1. ASVSPOOF 2015 data distribution……...….44

Table 4.1. EERs (%) for the development data……...…….47

Table 4.2. EERs (%) for known attacks of evaluation data (Proposed System/MFCC based i-vector with mask) …….…....48

Table 4.3. (Proposed System/MFCC based i-vector with mask**)** ……..…..49

Table 4.4. Comparison of average EERs (%) for evaluation (K: Known attacks (S1-S5), U: Unknown attacks (S6-S9)……...……..50

LIST OF ABBREVIATIONS

SR	:Speaker Recognition
MFCC	:Mel-frequency cepstral coefficients
CNN	:Convolutional neural network
CQCC	:Constant-Q cepstral coefficient
GMM	:Gaussian Mixture Models
HMM	:Hidden Markov Models
SVM	:Support Vector Machines
RNN	:Recurrent neural networks
DTW	:Dynamic Time Warping
VQ	:Vector Quantization
LBG	:Linde-Buzo-Gray
HMM2s	:Second-order hidden Markov models
JFA	:Joint Factor Analysis
PLDA	:Probabilistic linear discriminant analysis

1. INTRODUCTION

The uniqueness of the speech signal, much like fingerprints, retinas, and palm vein patterns, has garnered significant attentiveness in scientific research. With the progress in computational technology, the study of speech signal has gained prominence. Its applications span across diverse fields, including voice processing, communication, security, intelligence, linguistics, biomedical, control, and psychology. In communication, it includes sound encoding and decoding, while within the domain of security, it involves speaker identification systems. Intelligence applications include functions such as keyword spotting. Linguistics delves into sound production mechanisms, whereas biomedical research centers around speech enhancement applications. Control systems incorporate voice-controlled devices, while in psychology, speech signals are utilized to discern emotional states.

One of the most widespread utilizes in the modern age is the recognition of persons by their voices. Controlling access to protected resources requires personal identity. Classical methodologies such as keys, passwords, or badges, can be vulnerable to theft, loss, or forgery. Nevertheless, numerous unique biometric features that individuals possess cannot be recreated by others. Biometrics makes use of physical qualities such as fingerprints, palm vein geometry as well as human features such as handwriting and voiceprints. While fingerprints and retinal patterns are generally thought to be more accurate for identity verification, voice-based identification has specific advantages for practical application, including the ease of data gathering over the phone.

Biometric recognition systems are used in real-time applications based on criteria such as reliability, ease of use, easy implementation, and low cost. [1] compared well-known and widely used biometric recognition systems according to the specified criteria. These comparisons are shown in Table 1.1. As seen from the table, speech surpasses other methods in many aspects. Furthermore, with the proliferation of mobile communication today, speech has become the most natural source of biometric information that can be utilized [2].

Table 1.1. Biometric Identification system properties [1]

Biometric type	Accuracy	Ease of use	User acceptance	Ease of implementation	Cost
Fingerprint	High	Medium	Low	High	Medium
Hand geometry	Medium	High	Medium	Medium	High
Voice	Medium	High	High	High	Low
Retina	High	Low	Low	Low	Medium
Iris	Medium	Medium	Medium	Medium	High
Signature	Medium	Medium	High	Low	Medium
Face	Low	High	High	Medium	Low

The speech signal contains a lot of information. One of the pieces of information it carries is the message conveyed through perceived and transmitted words at the level of perception. In addition to the transmitted message, the speech signal carries information about the spoken language, the identity of the speaker, their emotional state, and generally about the speaker. During the Audio and Video Based Person Authentication (AVBPA) international conference held in 2003, voice signal emerged as the third prominent biometric feature, following facial and fingerprint recognition [3]. Furthermore, the conference showcased nine distinct multimodal person recognition systems, with six of them incorporating speech recognition.

1.1 What is sound?

An acoustic pressure wave is known as a sound wave, which represents vibrations capable of transmitting from one point to another. To generate sound, a vibrating sound source is necessary within a material medium, and the resulting sound waves require a conducting medium to propagate. In this process, the motion of particles in a material medium is transferred to neighboring particles, facilitating the transmission of vibrations [4].

The production of sound involves the coordinated functioning of various components within the human body, including the larynx, lungs, muscles, skeletal system, and psycho-neurological systems. The speech signal is formed when exhaled air from the lungs passes through the vocal production mechanism. Key components of this mechanism include the vocal folds, palate, teeth, and lips. The journey of sound begins at the larynx and concludes at the lips [5] shaping the speech signal along this path. Essentially, speech formation is influenced by the positions and postures of anatomical structures along the sound pathway,

creating a diverse range of sounds. Initially, as air passes through the larynx, vibrations in the vocal folds produce relatively weak sounds. Subsequently, as the sound travels through the pathway, additional factors come into play, resulting in the sound reaching its basic form and the completion of speech signal production. The length of the sound production pathway is approximately 15 cm in females and around 17 cm in males [6].

The distinction between different speech sounds primarily relies on the events occurring within the sound production mechanism. Accurate mathematical modeling of the human voice production mechanism holds significant importance in speech processing. The figure illustrates the general structure of the human voice production mechanism.

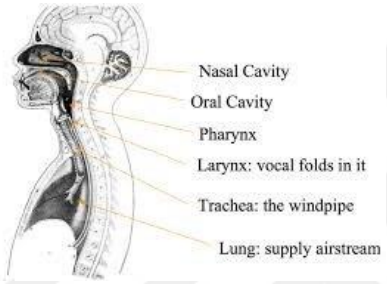

Figure 1.1. general structure of the human voice production mechanism

1.2 Speech Recognition

Speech recognition aims to identify the spoken words through the speech signal. Speaker recognition systems are used in various fields to ensure secure access, ranging from communication services to banking transactions and private records to smart home technologies. These systems rely on automatic speech recognition technology to identify a speaker's identity through their speech signal. They have a wide range of applications, effectively used in areas such as voice calls, telephone shopping, database access services, and security checks. Speaker recognition systems not only enable users to establish secure communication but also facilitate the protection of personal information and streamline authorization processes. Therefore, this widely used technology in our daily lives provides fast and reliable access, meeting the needs of users.

Some of the significant advantages of speaker recognition systems include ease of use, fast processing capability, high level of security, providing personalized authentication, preventing fraud, and reducing error rates. Additionally, the compatibility of these systems with mobile devices and their integration into our daily lives enable users to easily benefit

from them. In the future, it is anticipated that the usage areas of speaker recognition systems will further expand, and they will continue to have a significant impact on various aspects of our lives. The development and widespread adoption of this technology will continue to offer a secure and user-friendly communication and access experience.

Speech recognition can be divided into two parts: speaker verification and speaker identification [7-8]. Both methods utilize reference databases of known speakers and employ similar analysis and decision techniques. Speaker identification involves determining the person to whom an unknown voice sample belongs within a database consisting of voice recordings from specific speakers. Speaker verification, on the other hand, focuses on determining whether an unknown voice sample belongs to the claimed individual. The key distinction between the two methods arises in the decision-making stage. In speaker identification, the number of decisions produced by the system equals to the number of registered speakers in the system. In contrast, speaker verification only has two possible outcomes, acceptance or rejection, regardless of the number of individuals. Consequently, as the number of individuals increases in speaker identification, the recognition rate tends to decrease, while speaker verification achieves a recognition rate that is independent of the number of individuals. [8].

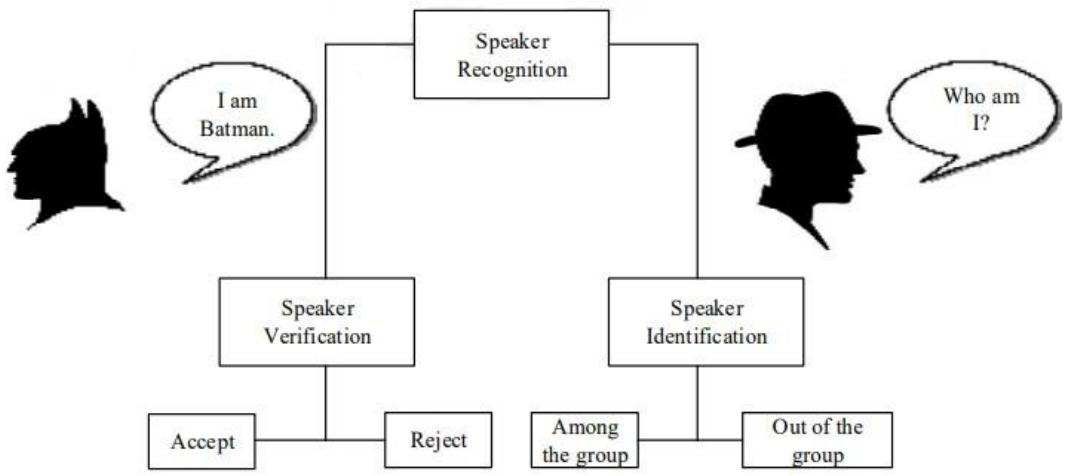

Figure 1.2. Speaker recognition system principle

Text-dependent speech recognition refers to the process of recognizing and transcribing speech that is specifically prompted by a given text or phrase. In this type of speech recognition, the system is trained to recognize and understand speech that is directly related to a specific text or set of texts [9]. The process involves training the system using a large dataset of text-dependent speech samples, where the speech is aligned with the corresponding text. This allows the system to learn the patterns and characteristics of speech that are associated with specific texts. On the other hand, text-independent speech recognition refers to the process of recognizing and transcribing speech without any specific text prompt or constraint. In this type of speech recognition, the system is trained to recognize and understand speech in a more general context, without relying on specific texts or phrases [9]. The process involves training the system using a large dataset of text-independent speech samples, where the speech is not aligned with any specific text. This allows the system to learn the general patterns and characteristics of speech, regardless of the specific content.

Both text-dependent and text-independent speech recognition systems typically involve the use of machine learning algorithms, such as artificial neural networks (ANN), for the classification of speech [9]. These algorithms are trained on large datasets of speech samples, where the speech is represented using various features, such as mel-frequency cepstral coefficients (MFCC) [9]. The training process involves optimizing the parameters of the machine learning model to minimize the error between the predicted transcriptions and the ground truth transcriptions of the speech samples.

In recent years, there have been advancements in speech recognition systems that combine both text-dependent and text-independent approaches. For example, the SpeechT5 framework proposed by [10] utilizes a unified-modal encoder-decoder pre-training approach to improve the capability of speech and textual modeling [10]. This framework pre-trains on a large scale of unlabeled speech and text data and uses a cross-modal vector quantization method to bridge speech and text information [10]. The evaluations of the SpeechT5 framework show its superiority in various spoken language processing tasks, including automatic speech recognition and text-to-speech synthesis [10].

Speaker Recognition (SR) systems consist of two fundamental stages: the front-end and the back-end. The front-end stage involves applying signal processing techniques to the speech signal, with a primary focus on feature extraction. This stage aims to extract pertinent

speaker-specific information while eliminating superfluous data embedded in the raw speech signal. Utilizing these extracted features and employing various machine learning methodologies, speaker models are trained. The back-end stage pertains to the utilization of these machine learning techniques to model and classify speakers, as well as scoring utterances during both the training and testing phases. The trained speaker models are subsequently employed to assess and score unknown utterances in testing scenarios. The origins of SR research can be traced back several decades, with early investigations revealing inter-speaker variations in spectral patterns [11]. Further advancements in SR were explored through subsequent experiments conducted by [12-13].

1.3 What are the challenges of Speaker Recognition Systems?

1. Acoustic Challenges: Speech recognition systems often face difficulties in accurately recognizing speech due to various acoustic challenges. These challenges include background noise, reverberation, and interference from other speakers [14]. Background noise can significantly degrade the quality of the speech signal, making it harder for the system to distinguish the speech from the noise [14]. Reverberation, which is caused by sound reflections in an enclosed space, can distort the speech signal and make it more difficult to recognize [15]. Interference from other speakers, especially in multi-talker scenarios, can further complicate the recognition process [16].

2. Variability Challenges: Variability in speech signals poses another set of challenges for speech recognition systems. Variability can arise from different speakers, accents, speaking rates, and speaking styles [14]. Different speakers may have different vocal characteristics, making it challenging to generalize the recognition system across speakers [17]. Accents and dialects introduce additional variations in pronunciation and phonetic patterns, which can affect the accuracy of recognition [18]. Speaking rates and styles, such as fast or hesitant speech, can also impact the performance of the system [14].

3. Data Limitations: The availability of large and diverse training data is crucial for training accurate speech recognition systems. However, collecting and annotating large-scale speech datasets can be challenging and time-consuming, especially for specific tasks or target populations [19]. For example, collecting data for disordered speech recognition can be

difficult due to the underlying neuro-motor conditions of people with speech disorders and co-occurring physical disabilities [19]. Similarly, collecting data for children's speech recognition can be challenging due to the high fundamental frequency and vocal apparatus changes in children [20].

4. Language and Linguistic Challenges: Speech recognition systems face challenges related to language-specific characteristics and linguistic variations. Different languages have distinct phonetic and phonological properties, which require language-specific models and resources for accurate recognition [18] For example, Arabic speech recognition faces challenges related to the large lexical variety, absence of diacritics, and dialectal variants [18]. Additionally, speech recognition systems need to handle linguistic variations, such as word variations, sentence structures, and vocabulary size, which can vary across different domains and applications [21].

5. Cognitive and Perceptual Challenges: Speech recognition is influenced by cognitive and perceptual factors, which can introduce challenges for accurate recognition. Factors such as attention, fatigue, and cognitive decline can affect the performance of speech recognition systems [22]. Challenging listening conditions, such as degraded or speeded speech, can engage non-auditory systems involved in performance monitoring and attention [22]. Age-related sensory, perceptual, and cognitive declines can increase the relative difficulty of speech recognition tasks [22].

6. Spoofing: Spoofing in speaker recognition refers to the deliberate manipulation or impersonation of a speaker's voice in order to deceive the speaker recognition system. It poses a significant challenge to the accuracy and reliability of speaker verification systems [23]. Spoofing attacks can take various forms, including replay attacks, impersonation, speech synthesis, and voice conversion [23]. These attacks aim to mimic the voice of a genuine speaker or generate artificial speech that can bypass the authentication process [23]. One of the main difficulties in spoofing for speaker recognition is the high level of interspeaker variability, which makes it challenging to distinguish between genuine and spoofed voices [24]. Interspeaker variability refers to the natural differences in speech characteristics among different individuals. These variations can include differences in accent, pronunciation, pitch, and speaking style. The presence of interspeaker variability makes it difficult to develop

robust spoofing detection algorithms that can accurately differentiate between genuine and spoofed voices [24].

To address these challenges and difficulties, researchers have proposed various solutions and methods to improve the performance of speech recognition systems. Some of these solutions include:

1. Noise Reduction and Enhancement Techniques: To mitigate the impact of background noise and reverberation, researchers have developed noise reduction and speech enhancement techniques. These techniques aim to enhance the quality of the speech signal by reducing noise and reverberation, thereby improving the accuracy of speech recognition [14].

2. Acoustic Model Adaptation: Acoustic model adaptation techniques aim to adapt the speech recognition system to specific speakers or acoustic conditions. These techniques involve updating the model parameters based on speaker-specific or environment-specific data, improving the system's ability to handle variability and challenging acoustic conditions [14].

3. Data Augmentation: Data augmentation techniques involve artificially expanding the training dataset by applying various transformations to the existing data. These transformations can include speed perturbation, vocal tract length perturbation, and adding background noise [19]. Data augmentation helps to increase the diversity of the training data and improve the robustness of the speech recognition system [19].

4. Language and Accent Adaptation: Language and accent adaptation techniques aim to adapt the speech recognition system to specific languages or accents. These techniques involve training language-specific or accent-specific models using data from the target language or accent [18]. Language and accent adaptation help to improve the accuracy of recognition for specific linguistic variations.

5. Cognitive and Perceptual Challenges: One approach is to develop speech recognition systems that are robust to challenging listening conditions, such as degraded or speeded speech. This can be achieved through the use of advanced signal processing techniques, such as noise reduction algorithms and time-scale modification methods, which aim to enhance the

quality and intelligibility of the speech signal [25]. Another solution is to incorporate adaptive algorithms that can dynamically adjust the recognition process based on the listener's attention and cognitive state. For example, systems can be designed to monitor the listener's attention level and adapt the recognition process accordingly, by providing additional cues or feedback to enhance comprehension [25]. Furthermore, considering age-related sensory, perceptual, and cognitive declines, personalized approaches can be employed. These approaches take into account individual differences in hearing abilities, cognitive functioning, and attentional resources. By tailoring the speech recognition system to the specific needs and capabilities of the user, performance can be optimized for older adults [26].

1.4 Spoofing in Speech

Speech spoofing refers to the act of manipulating or falsifying speech signals to deceive or trick a listener or a speech recognition system. It involves creating fake or synthetic speech that mimics the voice of a genuine speaker or altering existing speech recordings to produce a desired outcome. Speech spoofing can be used for various purposes, including impersonation, replay attacks, voice conversion, and speech synthesis [27-35].

There are several types of speech spoofing attacks:

1. Impersonation: In this type of attack, the attacker tries to imitate the voice of a specific individual to deceive the listener or a voice recognition system. The attacker may use various techniques, such as voice conversion or speech synthesis, to mimic the target speaker's voice [27-35].

2. Replay Attacks: In a replay attack, the attacker records a genuine speech sample and plays it back to deceive a voice recognition system. The attacker may use pre-recorded speech samples or manipulate existing recordings to mimic the voice of the target speaker [27-35].

3. Voice Conversion: Voice conversion involves modifying the speech characteristics of a source speaker to make it sound like the voice of a target speaker. This technique is often used in impersonation attacks, where the attacker tries to imitate the voice of a specific individual [27-35].

4. Speech Synthesis: Speech synthesis refers to the process of generating artificial speech that sounds like a human voice. Attackers can use speech synthesis techniques to create fake speech samples that mimic the voice of a specific individual or generate entirely new speech [27-35].

To prevent speech spoofing, various countermeasures and detection techniques have been developed. These techniques aim to identify and distinguish between genuine and spoofed speech signals. Some common approaches include:

1. High-dimensional feature analysis: High-dimensional features extracted from speech signals can be used to detect artifacts and anomalies associated with spoofed speech. These features capture subtle differences in speech characteristics that are indicative of spoofing attacks [27].

2. Deep learning-based approaches: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been applied to detect speech spoofing. These models can learn complex patterns and features from speech data, enabling them to identify spoofed speech with high accuracy [32,36-41].

3. Fusion of multiple features: Combining multiple types of features, such as spectral, prosodic, and temporal features, can improve the accuracy of spoofing detection systems. Fusion techniques aim to leverage the complementary information provided by different feature representations [27-40].

4. One-class classification: One-class classification approaches focus on modeling the characteristics of genuine speech and detecting deviations from this model. These methods do [28].

5. Glottal flow analysis: Glottal flow parameters, which capture the characteristics of the vocal folds during speech production, have been used as features for spoofing detection. Analyzing the glottal flow can reveal anomalies and inconsistencies in speech signals [42].

6. Liveness detection: Liveness detection techniques aim to distinguish between live speech and pre-recorded or synthetic speech. These methods often involve analyzing physiological or behavioral cues, such as vocal tract resonances or breathing patterns, to determine the authenticity of the speech signal [33, 34].

It is important to prevent speech spoofing for several reasons. Firstly, speech spoofing can be used for malicious purposes, such as identity theft, fraud, or spreading misinformation. By detecting and preventing speech spoofing, the integrity and trustworthiness of speech-based systems, such as voice assistants, authentication systems, and voice-controlled devices, can be maintained [30,32,36,41,43,44]. Secondly, speech spoofing attacks can undermine the security and reliability of voice-based communication and authentication systems. By developing robust spoof detection systems, the vulnerabilities of these systems can be mitigated [30, 32, 43, 44]. Finally, preventing speech spoofing helps to protect individuals' privacy and personal information. By ensuring that only genuine speakers are granted access to sensitive information or services, the risk of unauthorized access or data breaches can be reduced [30, 32, 41, 43, 44].

In summary, speech spoofing refers to the act of manipulating or falsifying speech signals to deceive or trick listeners or speech recognition systems. It can involve impersonation, replay attacks, voice conversion, or speech synthesis. To prevent speech spoofing, various techniques and countermeasures have been developed, including high-dimensional feature analysis, deep learning-based approaches, fusion of multiple features, one-class classification, glottal flow analysis, and liveness detection. Preventing speech spoofing is important to maintain the integrity and trustworthiness of speech-based systems, enhance security and reliability, and protect individuals' privacy and personal information.

2. THEORETICAL FOUNDATIONS AND LITERATURE REVIEW

Feature extraction is the process of converting a given speech signal into specific vectors that are specific to speech. In this process, non-informative parts about speech are discarded, and the distinctive parameters of the speech signal are emphasized. Mathematically, feature extraction can be defined as the transformation of original high-dimensional vectors to lower-dimensional vectors, $f: R^N \rightarrow R^d$, where $d \ll N$. Feature extraction serves two important purposes. The first is that the number of vectors used for training should be higher than the dimensionality of the original signal in order for the generated speech/speaker models in the training phase to better represent the relevant speech. The required number of vectors

increases exponentially with the dimensionality of the original signal, a phenomenon known as the "curse of dimensionality" in the literature [45-46]. The second purpose is to reduce computation and processing load through feature extraction.

Feature extraction involves discarding irrelevant information in raw speech signals. An analog speech signal is captured by a microphone and digitized using an analog-to-digital converter. After the conversion, the digital speech signal can be processed using a computer, microcontroller, etc. There are various types of features in the literature. The key point is to know the requirements of specific applications.

As mentioned in the introduction, there is information about words, language, speaker, gender, emotions, etc., in the raw speech signal. For example, if word information is sought, emotional features should not be extracted. Therefore, in speaker recognition, the extracted features should discriminate between speakers. The ideal features should have the following characteristics [47-48]:

- The feature should have large variations among speakers and small variations within the same speaker.
- The feature should be easily obtained from the speech signal.
- The feature should be robust against imitations, distortions, and noise.
- The feature should occur frequently and naturally during speech, so that it does not require special training or effort.
- The feature should not be affected by long-term changes such as the speaker's health condition or aging.

The ideal feature does not actually exist. Furthermore, since the human speech production system is a physiological system, deformations caused by health issues or aging affect the produced voice. Therefore, obtaining features close to these ideal characteristics is a challenging and unresolved problem.

Additionally, it is possible to combine different types of features [49,50]. One type of feature should encompass information that the other type does not have. Thus, the features are used complementarily to enhance the system's recognition performance. Features can be categorized into short-term spectral features, prosodic features, high-level features, and others. Various types of features have been proposed in each category. However, short-term

features appear to be the most popular for research purposes and practical applications due to their ease of implementation and good recognition performance. The purpose of short-term features is to capture vocal tract information. Speech identification and speech verification systems consist of the same components (Figure 2.1). Feature extraction is a common process in both the training and testing stages.

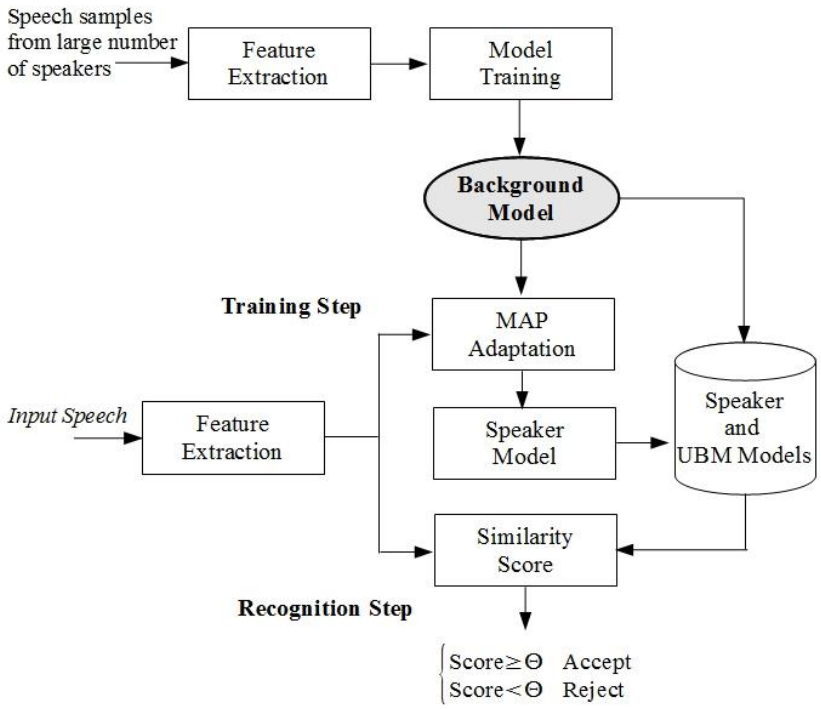

Figure 2.1. General speakar recognition system

2.1 Short Time Features

Short-term features are commonly used in synthetic speech recognition and speech recognition systems. These features capture the spectral and temporal characteristics of the speech signal over short durations, such as phonemes or frames. They provide important information for distinguishing between different speech sounds and can be used for speaker recognition, emotion recognition, and other speech-related tasks.

One widely used short-term feature is the Mel-Frequency Cepstral Coefficients (MFCCs) [51]. MFCCs are derived from the Fourier Transform of the speech signal and represent the spectral envelope of the signal. They are commonly used in speech recognition systems due to their effectiveness in capturing the important spectral features of speech.

Linear Prediction (LP) based coefficients are another type of short-term feature used in speech-related research [51]. LP is used to obtain the spectrum of the signal, and the predictor coefficients are further processed to observe more robust features. Other LP-based features include perceptual linear prediction cepstral coefficients, linear predictive residual, and line spectral [51].

Delta features are another type of short-term feature commonly used in speech recognition systems [52]. Delta features represent the temporal changes in the main feature vectors, such as MFCCs. They are computed by using a few neighboring feature vectors and are concatenated to the main feature vectors. Delta features capture the temporal dynamics of the speech signal and can improve the performance of speech recognition systems.

Glottal source features are segmental analysis features that capture speaker-specific information [52]. These features are more complicated to extract compared to MFCCs or LP-based features. They are based on the glottal excitation of the speech signal and can be used in combination with vocal tract features to achieve better recognition performance.

Spectro-temporal clues, such as energy variations and formant frequency transitions, can also be used as short-term features for synthetic speech recognition [53].These features capture both spectral and temporal information and can provide additional speaker information. In addition to these features, there are other short-term features that have been proposed for synthetic speech recognition and speech recognition systems.

These include wavelet-based features [54], Fourier Transform-based features [55], and features extracted using deep learning techniques [52] These features aim to capture different aspects of the speech signal and improve the performance of speech recognition systems.

Overall, short-term features play a crucial role in synthetic speech recognition and speech recognition systems. They capture important spectral and temporal characteristics of the speech signal and are used for various speech-related tasks. The choice of features depends on the specific application and the desired performance of the system.

2.2 Synthetic Speech Recognition Features

Synthetic speech recognition involves the use of various features to characterize and detect synthetic speech. These features are important for distinguishing between real and synthetic speech, as well as for improving the performance of speech recognition systems.

The history of features used in synthetic speech recognition can be traced back to the early days of speech recognition research. In the past, features such as linear prediction coefficients (LPC) and MFCCs were commonly used for speech recognition [56].These features were effective in capturing the spectral and temporal characteristics of speech signals. However, with the advancement of technology and the availability of large speech corpora, researchers began to explore more advanced features for speech recognition.

The use of MFCCs, cosine phase, and modified group delay features has been explored in the context of synthetic speech detection [51]. These features have shown promising results in distinguishing between real and synthetic speech. The advantages of using these features in synthetic speech recognition include their ability to capture the spectral and temporal characteristics of speech signals. MFCCs, in particular, have been widely used in speech recognition due to their effectiveness in representing speech signals [56]. The use of cosine phase and modified group delay features can provide additional information for detecting synthetic speech [51]. These features can help improve the performance of speech recognition systems by distinguishing between real and synthetic speech.

However, there are also some disadvantages associated with these features. One disadvantage is that they may not be able to capture all the characteristics of synthetic speech. Synthetic speech can have artifacts and inconsistencies that may not be fully captured by these features [57]. Additionally, the performance of these features may vary depending on the specific VC and SS techniques used to generate the synthetic speech [51].

To address the disadvantages of these features, researchers have proposed various novel features for spoof detection. CQCC, extended-CQCC, CQ-est, df-mst, icbc.

Spoof detection is an important task in the field of speaker verification and authentication systems. Various novel features have been proposed to improve the accuracy of spoof

detection algorithms. Some of these features include CQCC, extended-CQCC, CQ-est, df-mst, and icbc.

CQCC (Constant-Q Cepstral Coefficients) is a high-dimensional feature that has been specifically designed for spoof detection tasks in Automatic Speaker Verification (ASV) systems [58]. It has been shown to outperform other features such as Instantaneous Frequency Cosine Coefficients (IFCC), MFCC, and Epoch Features (EF) [58]. CQCC features capture both static and dynamic information from the speech signal, making them effective for distinguishing between natural and spoofed speech [58].

Extended-CQCC is an extension of the CQCC feature that incorporates additional information from the speech signal, such as phase information. This extension improves the discriminative power of the feature and enhances the performance of spoof detection algorithms [27].

CQ-est is another high-dimensional feature that is based on the CQCC representation. It aims to estimate the quality of the CQCC features by considering the artifacts present in the spoofed speech [27]. By incorporating this quality estimation, CQ-est can effectively distinguish between natural and spoofed speech [27].

df-mst (dynamic features and minimum spanning tree) is a feature representation that combines high-dimensional features and dynamic features [27]. Dynamic features capture the temporal variations in the speech signal, which are useful for detecting spoofing attacks [27]. By fusing high-dimensional features and dynamic features, df-mst achieves low equal error rates for various types of spoofing attacks [27].

icbc (Inverse MFCC) is a feature that is proposed as a countermeasure to the limitations of MFCC in spoof detection [59]. MFCC represents the human auditory system and is commonly used in speaker recognition systems, but it may not perform well in the anti-spoofing environment [59]. icbc incorporates feature contents that are absent in MFCC, making it more effective for spoof detection [59].

These novel features have been evaluated in various studies and have shown promising results in improving the accuracy of spoof detection algorithms. They capture different aspects of the speech signal and are designed to be sensitive to the artifacts present in spoofed speech. By incorporating these features into spoof detection algorithms, researchers aim to enhance the security and reliability of speaker verification systems.

2.2.1 MFCC

Mel-Frequency Cepstral Coefficients (MFCC) is a widely used feature extraction technique in speech recognition systems [60]. Mel-frequency cepstral coefficients (MFCCs) are a widely used feature extraction method in speech recognition and analysis. MFCCs were first introduced by Davis and Mermelstein in the 1980s [61]. The MFCC method is based on the observation that the human auditory system is more sensitive to changes in frequency at lower frequencies than at higher frequencies. Therefore, MFCCs aim to mimic the human auditory system's perception of sound by representing the spectral characteristics of speech signals on a mel-frequency scale. It is based on the human auditory system's perception of sound, specifically the Mel scale, which is a perceptual scale of pitches that approximates the human ear's frequency response [60].

The MFCC feature extraction process involves several steps, including framing, windowing, Fourier Transform, Mel filter bank, and Discrete Cosine Transform (DCT) [60]. Firstly the speech signal is preprocessed to enhance its quality and remove any noise or artifacts if it is necesarry. This preprocessing step may include operations such as filtering and normalization [62].

The first step in MFCC feature extraction is framing or frame blocking, where the speech signal is divided into short frames of typically 20-40 milliseconds [60]. This is done to capture the temporal variations in the speech signal. Each frame is then windowed using a window function, such as the Hamming window, to reduce spectral leakage [60]. The windowed frames are then passed through the Fourier Transform to convert them from the time domain to the frequency domain [60].

Next, the Mel filter bank is applied to the power spectrum obtained from the Fourier Transform [60]. The resulting spectrum is then converted to the mel-frequency scale, which is a perceptually linear scale that approximates the human auditory system's response to different frequencies. This conversion is achieved by applying a filterbank that consists of a series of triangular filters spaced evenly on the mel-frequency scale.The Mel filter bank consists of a set of triangular filters that are spaced uniformly on the Mel scale [60]. Each filter in the bank is designed to capture the energy in a specific frequency range. The output of each filter is the weighted sum of the power spectrum values within its frequency range [60].

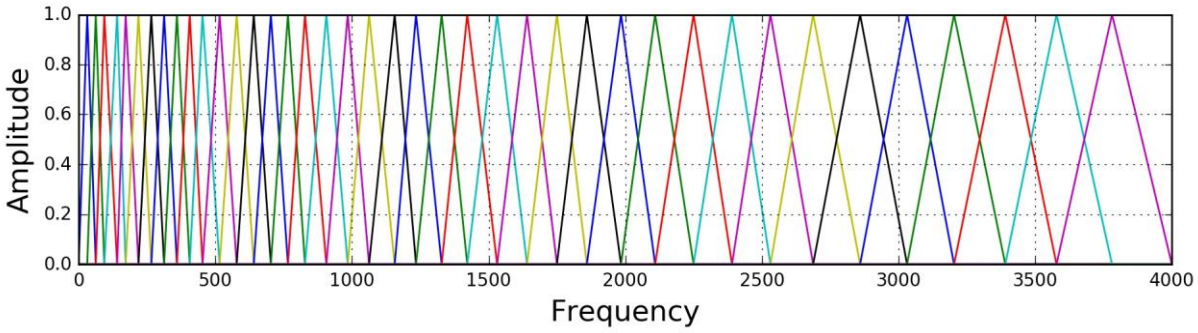

Figure 2.2. Mel filter bank

After applying the Mel filter bank, the logarithm of the filter bank energies is taken to approximate the logarithmic perception of loudness in the human auditory system [60]. This is followed by the Discrete Cosine Transform (DCT), which decorrelates the filter bank energies and produces a compact representation of the spectral features [60]. The resulting coefficients are known as Mel-Frequency Cepstral Coefficients (MFCC) [60]..

The MFCC coefficients capture the spectral envelope of the speech signal and are commonly used as features for speech recognition tasks [60]. They have been shown to be robust to noise and other distortions, making them suitable for synthetic speech recognition [63]. The MFCC feature extraction process can be further enhanced by incorporating techniques such as power normalization, noise suppression, and temporal masking [63].

The mathematical basis of MFCCs lies in signal processing techniques such as the Fourier transform, mel-frequency scaling, and discrete cosine transform. The Fourier transform is used to convert the speech signal from the time domain to the frequency domain, allowing for the analysis of its spectral characteristics. The mel-frequency scaling is applied to approximate the human auditory system's perception of sound by mapping the frequency spectrum onto a mel-frequency scale. This scaling is achieved using a filterbank that consists of triangular filters spaced evenly on the mel-frequency scale. Finally, the discrete cosine transform is applied to the logarithmically compressed filterbank outputs to obtain the MFCCs [61]. These mathematical operations enable the extraction of features that capture the important spectral characteristics of speech signals.

In summary, MFCC is a feature extraction technique that captures the spectral characteristics of speech signals. It involves framing, windowing, Fourier Transform, Mel filter bank, and

DCT. The resulting MFCC coefficients are widely used in speech recognition systems for tasks such as synthetic speech recognition.

The advantages of using MFCCs for speech recognition and analysis are well-documented. MFCCs have been shown to capture important voice characteristics that are crucial for recognition [61]. They are effective in representing speech signals and can capture important information in the voice while producing minimal data loss [61].MFCCs also replicate the auditory perception of sound by the human ear, making them suitable for speech recognition tasks [61].

Additionally, MFCCs are robust to variations in speech conditions, such as background noise and channel distortions [63]. These advantages make MFCCs a popular choice for feature extraction in speech recognition systems.

To detect synthetic speech, various methods can be employed in conjunction with MFCCs. Synthetic speech often exhibits certain characteristics that distinguish it from natural speech. For example, synthetic speech may lack natural prosody, exhibit unnatural spectral characteristics, or have inconsistencies in the phase information . These characteristics can be exploited to develop detection algorithms.

One approach is to train machine learning models, such as Gaussian mixture models (GMMs) or deep neural networks (DNNs), using a combination of real and synthetic speech data. These models can then be used to classify new speech samples as either real or synthetic based on their extracted features, including MFCCs .

Another approach is to analyze specific features or patterns in the MFCCs that are indicative of synthetic speech. For example, the presence of certain artifacts or inconsistencies in the MFCCs may suggest the presence of synthetic speech . By analyzing these features, it is possible to develop algorithms that can accurately detect synthetic speech.

As a result, MFCCs are a widely used feature extraction method in speech recognition and analysis. They were introduced in the 1980s and aim to mimic the human auditory system's perception of sound. The process of extracting MFCCs involves several steps, including

preprocessing, frame division, Fourier transform, mel-frequency scaling, logarithmic compression, and discrete cosine transform. MFCCs have advantages in capturing important voice characteristics, producing minimal data loss, and replicating human auditory perception. To detect synthetic speech, various methods can be employed, including machine learning models and analysis of specific features in the MFCCs. The mathematical basis of MFCCs lies in signal processing techniques such as the Fourier transform, mel-frequency scaling, and discrete cosine transform. These operations enable the extraction of features that capture the spectral characteristics of speech signals.

2.2.2 CQCC

Constant Q Cepstral Coefficients (CQCC) is a feature extraction technique that has been proposed for various applications, including speaker verification and anti-spoofing systems [64]. CQCC is derived from the Constant Q Transform (CQT), which is a modification of the traditional Fourier Transform that uses logarithmically spaced frequency bins [64,65]. The CQT provides a higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies [66]. CQCC features are extracted from the CQT spectrum and are designed to capture the spectral characteristics of the speech signal.

The mathematical explanation of CQCC involves several steps. First, the speech signal is transformed using the CQT, which decomposes the signal into its frequency components [64]. The CQT is computed by convolving the signal with a set of complex-valued filters that are logarithmically spaced in frequency [65]. The resulting CQT spectrum represents the magnitude and phase information of the signal at different frequencies and time frames.

Next, the CQCC features are extracted from the CQT spectrum. This involves taking the logarithm of the magnitude spectrum to approximate the human perception of loudness [66]. The logarithmic transformation is followed by a Discrete Cosine Transform (DCT) to decorrelate the CQT coefficients and obtain a compact representation of the spectral features [64-65]. The resulting coefficients are the CQCC features, which capture the spectral envelope of the speech signal.

One advantage of CQCC is its ability to capture both the magnitude and phase information of the speech signal, which can be useful for distinguishing between genuine and spoofed speech

[64-67]. The phase information can provide additional cues about the temporal characteristics of the signal.

In comparison to Mel-Frequency Cepstral Coefficients (MFCC), CQCC offers some distinct advantages. MFCC is based on the linearly spaced Mel scale and the traditional Fourier Transform, while CQCC uses the logarithmically spaced Constant Q Transform [64]. The logarithmic spacing of the CQT allows for a more perceptually relevant representation of the speech signal, as the human auditory system is known to perceive frequencies logarithmically [64].

Additionally, CQCC has been shown to outperform MFCC in certain applications, such as anti-spoofing systems [64]. However, the choice between CQCC and MFCC depends on the specific application and the characteristics of the speech data.

In terms of preference, CQCC is often preferred in the context of anti-spoofing systems, as it has been shown to achieve superior performance compared to other features [64-68]. However, the choice between CQCC and MFCC ultimately depends on the specific requirements of the application and the characteristics of the speech data.

It is important to consider factors such as the nature of the speech signal, the presence of noise or distortions, and the computational complexity of the feature extraction process.

2.2.3 i-vectors

The i-vector is a popular feature extraction technique used in speech recognition and synthetic speech recognition. It is a representation of the speech signal that captures the speaker-specific information and is widely used in speaker recognition systems.
The concept of i-vectors was first introduced in 2010 by Dehak et al. as an extension of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) framework.

The GMM-UBM framework is a widely used approach for speaker recognition, where a GMM is trained on a large amount of background data to model the speaker-independent characteristics, and the speaker-specific information is captured by adapting the GMM using a small amount of speaker-specific data.

The i-vector approach takes this idea further by representing each speaker as a low-dimensional fixed-length vector, called the i-vector. The i-vector is derived from the GMM-UBM framework by extracting the total variability of the speaker-specific information using factor analysis . This allows for a compact representation of the speaker characteristics, which can be used for speaker recognition tasks.

The process of extracting i-vectors involves several steps. First, a GMM-UBM is trained on a large amount of background data to model the speaker-independent characteristics. Then, for each speaker, a GMM is adapted using a small amount of speaker-specific data. The adapted GMM is used to estimate the sufficient statistics, which capture the speaker-specific information. These sufficient statistics are then used to derive the i-vector using factor analysis.

The i-vector approach has gained popularity in speaker recognition due to its effectiveness in capturing the speaker-specific information and its ability to handle large speaker variability. It has been shown to outperform traditional feature extraction techniques, such as MFCCs, in various speaker recognition tasks . The i-vector approach is also robust to channel and environmental variations, making it suitable for real-world applications.

One advantage of i-vectors is their compact representation, which allows for efficient storage and processing. The fixed-length nature of i-vectors enables easy comparison and matching of speakers. Additionally, i-vectors can be used in combination with other features, such as MFCCs, to further improve the performance of speaker recognition systems .

However, there are also some limitations and challenges associated with i-vectors. One challenge is the need for a large amount of training data to accurately model the speaker-independent characteristics. Another challenge is the sensitivity of i-vectors to the quality and variability of the training data. Variations in recording conditions, channel effects, and background noise can affect the performance of i-vector-based systems .

To overcome these challenges, various techniques have been proposed. One approach is to use data augmentation techniques to increase the amount of training data and improve the robustness of i-vector-based systems . Another approach is to incorporate domain adaptation techniques to handle variations in recording conditions and channel effects [69]. Additionally, deep learning-based methods, such as x-vectors, have been proposed as an alternative to i-vectors, which have shown promising results in speaker recognition tasks [70].

2.3 Speaker Modelling

Speaker modeling refers to the process of creating a mathematical representation or model of a speaker's voice characteristics and patterns. It involves extracting relevant features from the speech signal that are unique to each individual speaker and can be used to distinguish them from others [71]. Speaker modeling is an essential component of speech recognition and synthetic speech recognition systems as it enables the identification and differentiation of speakers based on their unique vocal characteristics [71].

In speech recognition systems, speaker modeling plays a crucial role in various applications such as voice dialing, banking by telephone, telephone shopping, database access services, information services, security control for confidential information areas, voice mail, and remote access to computers [71]. By accurately modeling the characteristics of individual speakers, these systems can provide personalized and secure access to services and information [71]. Speaker modeling is particularly important in forensic applications, where it is used to compare questioned speech with known speech samples to identify potential suspects [72]. It helps in producing a small list of potential suspects and providing evidence in criminal investigations [72].

The primary goal of speaker modelling is to capture the unique characteristics of a speaker's voice that remain consistent across different speech utterances [73]. Various methods are widely used for speaker modelling, including Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs), and neural networks [73-75]. GMMs are commonly used for modelling the statistical distribution of speaker-specific features, while HMMs are used to model the temporal dynamics of speech [75]. SVMs and neural networks are employed for classification and pattern recognition tasks in speaker modelling [74].

Recent trends in speaker modeling focus on improving the robustness and accuracy of speaker recognition systems. One trend is the integration of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to extract high-level representations of speaker-specific features [76]. These deep learning models have shown promising results in capturing complex patterns and improving the discriminative power of speaker models [76]. Another trend is the use of fusion techniques that combine multiple sources of information, such as physiological speech features and behavioral speech features, to enhance the performance of speaker recognition systems [76].

If speaker modeling is not utilized in speech recognition or synthetic speech recognition systems, the systems may struggle to accurately identify and differentiate between different speakers. This can lead to security vulnerabilities, as unauthorized individuals may be able to gain access to restricted services or information. In forensic applications, the absence of speaker modeling can result in reduced accuracy and reliability in identifying potential suspects, potentially leading to wrongful accusations or the failure to identify the true perpetrator [72]. Speaker modeling is therefore crucial for ensuring the integrity, security, and accuracy of speech recognition and synthetic speech recognition systems.

Dynamic Time Warping (DTW) is a matching technique used in speech and speaker recognition [77]. It is a template matching algorithm that aligns sequences that vary in time or speed [78]. DTW is commonly used in speech recognition systems to compare speech samples and identify similarities [79]. It has been shown to greatly improve the accuracy of isolated word speech recognition systems [80]. DTW is often used in combination with other techniques, such as Hidden Markov Models (HMMs), to achieve high accuracy in speaker identification [81]. It is also used in fusion pattern recognition methods to enhance speech recognition performance [82,83]. DTW is particularly useful in scenarios where the time scales of the reference and test patterns are not perfectly aligned [80]. Overall, DTW is a well-known and widely used technique in the field of speech and speaker recognition.

Vector Quantization (VQ) is a technique used in speaker modeling for speech recognition and synthetic speech recognition [84]. It is a lossy data compression method that involves dividing a set of vectors into clusters and representing each cluster by a representative vector called a codeword or centroid [84].

In the context of speaker modeling, VQ is used to create a codebook that represents the speech features of a particular speaker [84,85]. The codebook is generated by training sequences of speech data from the speaker, and it contains a set of codewords that capture the speaker's unique speech characteristics [84,85].

During recognition, the input speech features are compared to the codewords in the codebook to find the best match [85,86]. The goal is to find the codeword that minimizes the distortion between the input features and the codeword [86,87]. This matching process allows for the identification or classification of the speaker [86-88]. VQ-based speaker recognition systems have been shown to achieve high accuracy in speaker identification tasks [85-88].

VQ is often used in combination with other techniques, such as Hidden Markov Models (HMMs), to improve the performance of speaker recognition systems [85-88]. The VQ codebook can be used as the emission probability distribution in an HMM, where each codeword represents a state in the HMM [85-88]. This combination of VQ and HMM allows for more robust and accurate speaker recognition, especially in noisy or multi-speaker environments [85-88].

The design of the codebook is an important aspect of VQ-based speaker modeling [84-89]. Various algorithms, such as the Linde-Buzo-Gray (LBG) algorithm and genetic algorithms, have been used to optimize the codebook design [89-90]. The goal is to create a codebook that efficiently represents the speech features of the speaker while minimizing the distortion between the input features and the codewords [89-90].

In summary, Vector Quantization is a technique used in speaker modeling for speech recognition and synthetic speech recognition. It involves creating a codebook of representative vectors that capture the unique speech characteristics of a speaker. During recognition, the input speech features are compared to the codewords in the codebook to identify or classify the speaker. VQ is often combined with other techniques, such as HMMs, to improve the accuracy and robustness of speaker recognition systems. The design of the codebook is an important aspect of VQ-based speaker modeling, and various algorithms are used to optimize the codebook design.

The hidden Markov model (HMM) is a widely used modeling technique in the fields of speech recognition and speaker recognition [91]. HMMs use Markov chains to model the

changing states of a system over time [91]. In the context of speaker modeling for synthetic speech recognition, HMMs are used to model the speech signal and capture the sequential nature of speech [91]. HMMs are statistically trained and can incorporate dynamic coefficients, such as deltas and double-deltas, which are useful for speech and speaker recognition [51].

HMMs have been applied to improve the recognition performance of speaker identification systems under different conditions, such as the shouted talking condition [91]. In this case, second-order hidden Markov models (HMM2s) have been used to enhance the recognition performance of isolated-word text-dependent speaker identification systems [91]. HMM2s are an extension of HMMs that capture higher-order dependencies in the speech signal [91].

HMMs are also used in the detection of synthetic speech, which can be used to spoof biometric systems based on automatic speaker recognition technology [51]. Synthetic speech signals created using voice conversion and speech synthesis techniques can be detected using appropriate feature extraction techniques [51]. Choosing the right feature extraction technique is crucial for detecting synthetic speech [51]. The flexibility of HMMs is limited by the a-priori choice of the model topology [92]. However, HMMs have been widely used for automatic speech recognition due to their ability to capture the sequential character of the speech signal [92]. HMMs have been used for both isolated and connected speech recognition [92].

In the field of speech-to-text recognition, HMMs are often used in combination with Gaussian Mixture Models (GMMs) [93]. GMM-HMMs, which use GMMs as probability density functions, have been applied to speech recognition since the 1970s [93]. HMMs with GMMs are commonly used for speaker-independent speech recognition [93].

Overall, the hidden Markov model is a powerful tool for speaker modeling in synthetic speech recognition and speech recognition in general. It allows for the modeling of the sequential nature of speech and has been widely used in various applications, including speaker identification, speech synthesis, and speech recognition systems.

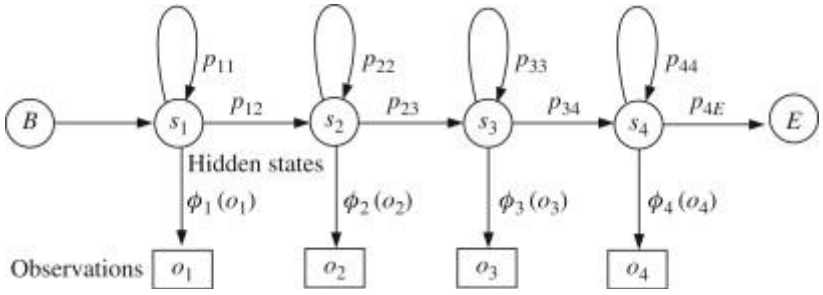

Figure 2.3. Hidden Markov Model

The nearest neighbor algorithm is a classification technique that is commonly used in various fields, including speaker modeling for synthetic speech recognition and speech recognition [94]. The algorithm works by finding the k nearest samples from a given sample based on a distance metric [94]. In the context of speaker modeling, the nearest neighbor algorithm can be used to classify and identify speakers based on their speech characteristics [94].

The algorithm is based on the idea that samples that are close to each other in the feature space are likely to belong to the same class [94]. It is a non-parametric algorithm, meaning that it does not make any assumptions about the underlying distribution of the data [94]. Instead, it relies on the proximity of samples to make predictions [94].

In the field of speech emotion recognition, the nearest neighbor algorithm has been used in combination with other classification techniques, such as support vector machines (SVMs) [94]. The algorithm can be used to classify speech samples into different emotion categories based on their acoustic features [94]. By comparing the acoustic features of a given speech sample with the features of the nearest neighbors in the training set, the algorithm can assign the sample to the most similar emotion category [94].

The performance of the nearest neighbor algorithm can be influenced by various factors, such as the choice of distance metric and the value of k [94]. Different distance metrics, such as Euclidean distance and Manhattan distance, can be used to measure the similarity between samples [94]. The value of k determines the number of nearest neighbors that are considered for classification [94]. Choosing an appropriate value of k is important to balance between overfitting and underfitting the data [94].

In summary, the nearest neighbor algorithm is a versatile classification technique that can be applied to speaker modeling for synthetic speech recognition and speech recognition tasks. It is a non-parametric algorithm that relies on the proximity of samples in the feature space to make predictions. The algorithm has been used in various applications, including speech emotion recognition, and its performance can be influenced by factors such as the choice of distance metric and the value of k.

Support Vector Machine (SVM) is a machine learning algorithm that can be used for speaker modeling in speech recognition and synthetic speech recognition tasks. SVM is a powerful classifier that has been widely used in various pattern recognition applications, including speech recognition [95-99].

In the context of speaker modeling for speech recognition, SVM can be used to classify and identify speakers based on their unique acoustic characteristics. It can learn from a set of labeled training data to create a model that can accurately classify new speech samples [97,100]. SVM has been used in speaker recognition algorithms to improve the accuracy of speaker identification [97,99,100].

In the case of synthetic speech recognition, SVM can be used to distinguish between real and synthetic speech signals. Synthetic speech signals created using voice conversion and speech synthesis techniques can be used to spoof biometric systems based on automatic speaker recognition technology [51]. SVM can be trained on a dataset containing both real and synthetic speech samples to learn the differences between them and classify new speech samples as either real or synthetic [51].

SVM is known for its ability to handle high-dimensional feature spaces and its robustness against overfitting. It works by finding an optimal hyperplane that maximally separates the data points of different classes in the feature space [95-99]. SVM can also handle non-linearly separable data by using kernel functions to map the data into a higher-dimensional space where it becomes linearly separable [96-98].

In summary, SVM is a machine learning algorithm that can be used for speaker modeling in speech recognition and synthetic speech recognition tasks. It can accurately classify and identify speakers based on their unique acoustic characteristics, and it can distinguish between real and synthetic speech signals. SVM is a powerful classifier that has been widely used in various pattern recognition applications, including speech recognition. Its ability to handle high-dimensional feature spaces and its robustness against overfitting make it a suitable choice for speaker modeling in these tasks.

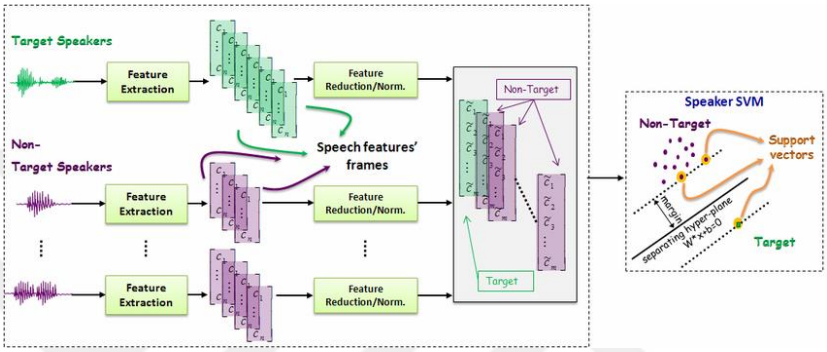

Figure 2.4. An Overview of Statistical Pattern Recognition Techniques for Speaker Verification)

Joint Factor Analysis (JFA) is an approach used for speaker modeling in speech recognition and synthetic speech recognition tasks [101]. JFA aims to model speaker variability and session variability separately [101]. In JFA, speaker variability is modeled by estimating the speaker subspace, which captures the unique characteristics of individual speakers [101]. Session variability, on the other hand, is modeled by estimating the session subspace, which captures the variations caused by different recording conditions [101].

The JFA model has been shown to outperform other speaker modeling techniques, such as eigenchannel modeling, in speaker verification tasks [101]. It has been found that the JFA model can achieve better performance and higher accuracy in speaker verification compared to eigenchannel modeling [101]. The JFA model has also been shown to perform well in speaker verification using a computationally inexpensive decision rule [101].

One advantage of the JFA model is that it can be implemented using the same software as eigenchannel modeling, with the only difference being the enrollment of target speakers [101]. This makes it easier to compare and evaluate the performance of the two approaches

[101]. Additionally, JFA can be combined with other techniques, such as support vector machines (SVM), to further improve speaker verification performance [102].

JFA has also been used in other speech-related tasks, such as overlapped speech recognition and speaker diarization [103,104]. For example, JFA has been integrated into a joint decoding framework for overlapped speech recognition and speaker diarization, where speaker embedding estimation and target-speaker automatic speech recognition (ASR) are performed alternately [104]. This joint decoding framework has shown promising results in recognizing and identifying speakers in overlapped speech [104].

In summary, Joint Factor Analysis (JFA) is an approach used for speaker modeling in speech recognition and synthetic speech recognition tasks. It models speaker variability and session variability separately, capturing the unique characteristics of individual speakers and the variations caused by different recording conditions. JFA has been shown to outperform other modeling techniques, such as eigenchannel modeling, in speaker verification tasks. It can be implemented using the same software as eigenchannel modeling and can be combined with other techniques, such as support vector machines, to further improve performance. JFA has also been applied in other speech-related tasks, such as overlapped speech recognition and speaker diarization, with promising results.

Linear discriminant analysis (LDA) is a technique used in speaker modeling for speech recognition and synthetic speech recognition. LDA is a dimensionality reduction method that aims to find a linear combination of features that maximizes the separation between different classes or speakers [105]. It is commonly used in speaker recognition systems to extract discriminative features from speech signals and improve the accuracy of speaker identification [106,107].

In the context of synthetic speech detection, LDA can be used to distinguish between real and synthetic speech signals. Synthetic speech signals created using voice conversion and speech synthesis techniques can be used to spoof biometric systems based on automatic speaker recognition technology [51]. By applying LDA to the extracted features, it is possible to identify patterns that differentiate between real and synthetic speech, enabling the detection of spoofing attacks [51,108].

LDA is often used in combination with other techniques in speaker recognition systems. For example, gender-independent probabilistic linear discriminant analysis (PLDA) is used to compute the scores of speaker recognition systems [106]. PLDA is a generative model that captures the statistical properties of speaker embeddings and allows for the comparison of speaker similarity [106]. By incorporating LDA into PLDA-based systems, the discriminative power of the speaker embeddings can be enhanced [109].

The effectiveness of LDA in speaker recognition depends on the choice of input features. Traditional dynamic coefficients, such as deltas and double-deltas, are commonly used in LDA-based speaker recognition systems [51]. These coefficients capture the temporal variations in speech signals and can improve the robustness of the system [106]. Other features, such as Mel frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) coefficients, have also been successfully applied in LDA-based speech recognition systems [110].

In summary, linear discriminant analysis (LDA) is a dimensionality reduction technique used in speaker modelling for speech recognition and synthetic speech recognition. It helps extract discriminative features from speech signals and improve the accuracy of speaker identification. LDA can be used to distinguish between real and synthetic speech signals, enabling the detection of spoofing attacks. It is often combined with other techniques, such as probabilistic linear discriminant analysis (PLDA), to enhance the discriminative power of speaker embeddings. The choice of input features, such as dynamic coefficients and MFCC, plays a crucial role in the effectiveness of LDA-based speaker recognition systems.

Within-class covariance normalization (WCCN) is a channel compensation technique used in speaker modelling for speech recognition and synthetic speech recognition [111]. It is applied in the total variability space to remove the nuisance effects caused by channel variations [111]. WCCN aims to normalize the within-speaker variation and enhance the discriminative power of the speaker representations [112].

The process of WCCN involves estimating the within-class covariance matrix, which captures the variations within each speaker class [113]. By normalizing or removing the nuisance subspace containing the session variability, WCCN helps to enlarge the distance between speakers and improve the performance of speaker recognition systems [114]. It is particularly

effective in compensating for channel effects and reducing the impact of channel variations on the recognition system [115].

In the context of synthetic speech detection, WCCN can be used to improve the differentiation between real and synthetic speech signals [112]. By applying WCCN to the extracted features, the within-speaker variations caused by the synthetic speech generation process can be normalized, making it easier to distinguish between real and synthetic speech [112].

WCCN is often used in combination with other techniques, such as linear discriminant analysis (LDA) [111]. The combination of LDA followed by WCCN has been found to yield the best results in speaker verification systems [111]. This combination helps to further enhance the discriminative power of the speaker representations and improve the accuracy of speaker identification [111].

Overall, within-class covariance normalization (WCCN) is a channel compensation technique used in speaker modeling for speech recognition and synthetic speech recognition. It aims to normalize the within-speaker variation and remove the nuisance effects caused by channel variations. WCCN is effective in improving the performance of speaker recognition systems and differentiating between real and synthetic speech signals.

Nuisance attribute projection (NAP) is a method used in the field of speech recognition to eliminate channel interference and solve the problem of channel mismatches [116]. It is also widely used in speaker recognition and face recognition [116]. The goal of NAP is to alleviate the redundant interference caused by multiple channels in speech recognition [117]. It is a technique that eliminates interference information from operating conditions, which can be difficult to alleviate [117]. NAP is effective in compensating for channel effects, which are a central problem in automated speaker recognition systems [118].

It aims to remove subspaces that cause session and channel variability in speaker recognition systems [119]. NAP has been shown to improve performance in speaker verification by compensating for session variability [120]. It can be used to remove session variability due to influences of microphone, environment, etc [121]. NAP is often used in combination with other techniques, such as within-class covariance normalization, to compensate for intersession variability in speaker verification systems [114].

In addition to its application in speech recognition, NAP has also been used in other fields. It has been applied in fault diagnosis of rolling bearings to eliminate interference information and improve feature extraction in variable speed signals [116]. NAP has been used in mispronunciation detection to remove non-pronunciation variations and increase the separation of mispronunciations [119]. It has also been used in language identification to compensate for channel conditions and improve performance [122].

Overall, NAP is a technique used in speech recognition and speaker verification to eliminate channel interference and compensate for session and channel variability. It has been shown to improve performance in these tasks by removing unwanted variability and interference information. NAP can be used in combination with other techniques to further enhance the performance of speaker recognition systems.

2.3.1 GMM-UBM Model

Gaussian Mixture Models (GMMs) are widely used in speaker modeling for speech recognition and synthetic speech recognition. GMMs are a statistical model that represents the probability distribution of a feature vector as a weighted sum of Gaussian components. In the context of speaker modeling, GMMs are used to model the acoustic features of speakers and capture the variations in their speech characteristics [123].

In speech recognition, GMMs are commonly used to model the spectral distribution of speech signals [124]. The GMMs can be trained using a large database of speech recordings, where each speaker is represented by a separate GMM [125]. The GMMs capture the statistical properties of the speech signals for each speaker, allowing for the recognition of the speaker's voice in subsequent speech recognition tasks [126]. GMMs are also used in speaker identification, where the goal is to determine the identity of a speaker given a speech sample. In speaker identification, GMMs are trained to model the speech characteristics of individual speakers [127]. The GMMs capture the unique features of each speaker's voice, allowing for the identification of the speaker based on their speech patterns [123].
In the field of synthetic speech recognition, GMMs are used to model the acoustic features of synthetic speech. GMMs can be trained using a database of synthetic speech recordings, where each synthetic speaker is represented by a separate GMM [128]. The GMMs capture

the statistical properties of the synthetic speech signals, allowing for the recognition and identification of the synthetic speakers [126].

GMMs have several advantages in speaker modeling for speech recognition and synthetic speech recognition. They have a simple structure and their parameters can be easily trained [127]. GMMs are also less susceptible to speaker-dependent variations, making them suitable for modeling different speakers [77]. Additionally, GMMs can effectively capture the statistical properties of speech signals, allowing for accurate recognition and identification of speakers [128].

In summary, GMMs are a widely used model in speaker modeling for speech recognition and synthetic speech recognition. They are used to capture the statistical properties of speech signals and model the unique characteristics of individual speakers. GMMs have proven to be effective in recognizing and identifying speakers in various applications.

In terms of speaker modeling, UBM stands for Universal Background Model. The UBM is a large Gaussian Mixture Model (GMM) that is trained to represent the speaker-independent distribution of features [130]. It serves as a background model against which individual speaker models are compared [131]. The UBM captures the statistical properties of the speech signals in a general sense, without specific speaker information [132].

The UBM is an essential component in speaker verification systems, where the goal is to verify the identity of a speaker based on their speech [133]. In speaker verification, the UBM is used to model the distribution of speech features from a large population of speakers [134]. The UBM is trained using a large database of speech recordings from multiple speakers, and it represents the common characteristics shared by all speakers [131].

To perform speaker verification, the UBM is used to estimate the likelihood of the observed speech features given the background model (Shon et al., 2015). This likelihood is then compared to the likelihoods obtained from speaker-specific models to determine the similarity between the observed speech and the enrolled speaker's model [134]. The UBM provides a reference distribution against which the similarity is measured, allowing for the identification or verification of the speaker [131].

In the context of synthetic speech recognition, the UBM is used to model the acoustic features of synthetic speech [132]. The UBM captures the statistical properties of the synthetic speech

signals, allowing for the recognition and identification of the synthetic speech. By comparing the likelihoods of the observed synthetic speech features with the UBM, the system can determine if the speech is synthetic or natural [133].

The UBM has been widely used in speaker recognition systems and has shown good performance in various applications [131]. However, it is worth noting that the UBM-based systems can be vulnerable to imposture attacks using synthetic speech. Therefore, there is ongoing research to improve the robustness of UBM-based speaker verification systems against synthetic speech imposture [135].

In summary, the UBM is a universal background model used in speaker modeling for speech recognition and synthetic speech recognition. It represents the speaker-independent distribution of speech features and serves as a reference model for comparing individual speaker models. The UBM is used in speaker verification systems to estimate the likelihood of observed speech features and plays a crucial role in distinguishing between different speakers. Additionally, the UBM is used in synthetic speech recognition to model the acoustic features of synthetic speech and detect synthetic speech imposture.

The main difference between GMM-UBM, UBM, and GMM lies in their modeling and adaptation techniques. The UBM is a model that represents the distribution of speech features for a large number of background speakers [136]. It is used as a reference model to compare against the target speaker model in speaker verification tasks [137]. On the other hand, GMM is a probabilistic model that represents the distribution of speech features for a specific speaker [136]. It is used to model the target speaker in the GMM-UBM system [138]. The GMM-UBM system combines the UBM and GMM models to improve the performance of speaker verification systems [139]. It adapts the UBM to capture the specific characteristics of the target speaker, resulting in a more accurate representation of the target speaker's speech features [136].

2.4 Artificial Neural Networks

An artificial neural network (ANN) is a computational model inspired by the structure and function of the human brain. It consists of interconnected nodes, or artificial neurons, that process and transmit information. ANNs are capable of learning from data and making predictions or decisions based on that learning [140].

In the context of speaker modeling, ANNs have been used for various tasks such as voice conversion and speaker recognition. Voice conversion is the process of transforming the speech of one speaker to make it sound like another speaker. ANNs have been used in voice conversion to perform non-linear mapping of spectral features from a source speaker to a target speaker, resulting in transformed speech that has the characteristics of the target speaker. The non-linear mapping ability of ANNs makes them suitable for capturing the complex relationship between vocal tract shapes of different speakers [141].

Speaker recognition, on the other hand, is the task of identifying or verifying the identity of a speaker based on their speech. ANNs have been applied to speaker recognition to capture the individuality of speakers by modeling both static and dynamic features of speech spectra [75]. ANNs have also been used for feature transformation in speaker identification systems, where they transform speaker-specific spectral features from any specific emotion to a neutral state. This allows for robust speaker recognition even in the presence of variations in emotional moods of speakers [142].

In addition to voice conversion and speaker recognition, ANNs have been used in other speaker-related tasks. For example, ANNs have been employed in speaker diarization systems, which aim to segment an audio recording into speaker-specific segments. ANNs have also been used in the analysis of tone production in languages such as Mandarin Chinese, where they can successfully classify tone patterns produced by multiple children [143].

Overall, ANNs have proven to be effective in speaker modeling tasks, providing improved performance compared to other methods such as Gaussian Mixture Models (GMMs) [141, 144]. Their ability to capture non-linear relationships and learn from data makes them a powerful tool for modeling the complex characteristics of speakers' voices.

Speaker recognition for synthetic speech recognition often involves the use of various feature extraction techniques to distinguish between real and synthetic speech. One commonly used feature for speaker recognition is Mel Frequency Cepstral Coefficients (MFCCs). MFCCs are modeled on the filter bank technique and are popular due to their advantages over other features. They capture the spectral characteristics of speech and have been shown to be effective in speaker recognition tasks [145].

In addition to MFCCs, other features such as energy, pitch, zero crossing rate, and delta and double-delta cepstrum coefficients have also been used for speaker recognition. These features provide information about the energy, pitch, and temporal dynamics of the speech signal, which can be useful for distinguishing between speakers [51,145].

One advantage of using these features for speaker recognition is their ability to capture speaker-dependent characteristics that are consistent across different speech utterances. These features can be used to create speaker models that can be used for speaker recognition tasks [145]. Additionally, these features are computationally efficient and can yield low-dimensional representations of the speech signal [146].

However, there are also some disadvantages associated with using these features for speaker recognition. One disadvantage is that these features may not capture all the relevant information for speaker recognition, especially in challenging audio conditions [76].

In such cases, additional techniques, such as DeepTalk-based vocal style encoding, may be used to improve speaker recognition performance [76]. Another disadvantage is that these features may be vulnerable to variations in the speech signal due to factors such as noise and channel effects. This can affect the accuracy and robustness of the speaker recognition system [145].

To address the issue of synthetic speech being used to spoof biometric systems based on automatic speaker recognition technology, various countermeasures have been proposed [51]. These countermeasures aim to detect and distinguish between real and synthetic speech [51]. One approach is to use discrimination methods that analyze the pitch pattern of the speech signal to distinguish between synthetic and natural speech [147]. Another approach is to combine physiological speech feature-based speaker recognition methods with DeepTalk-based vocal style encoding to improve speaker recognition performance [76].

In summary, speaker recognition for synthetic speech recognition often involves the use of various feature extraction techniques, such as MFCCs, energy, pitch, and delta and double-delta cepstrum coefficients. These features capture different aspects of the speech signal and can be used to distinguish between speakers. However, there are advantages and disadvantages associated with using these features, and additional techniques may be

employed to improve speaker recognition performance in challenging audio conditions or when dealing with synthetic speech.

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep learning algorithm that have been widely used in various fields, including biology, medicine, computer science, and petroleum engineering. CNNs are particularly effective in image analysis tasks, such as image classification, object detection, and segmentation [148,149].

The architecture of a CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers perform convolution operations on the input data, which involve applying a set of filters to extract features from the input. These filters are learned during the training process, allowing the network to automatically learn relevant features from the data. The pooling layers downsample the feature maps obtained from the convolutional layers, reducing the spatial dimensions of the data [150].

Finally, the fully connected layers perform classification or regression tasks based on the extracted features [150]. CNNs have shown great success in various applications. For example, in the field of medicine, CNNs have been used to predict treatment response and clinical outcomes in lung cancer patients based on serial medical imaging data. These models have demonstrated significant predictive power for survival, progression, distant metastases, and local-regional recurrence [149]. CNNs have also been applied in the classification of different Parkinsonism patterns of the striatum on positron emission tomography images. The performance of CNNs in these tasks has been evaluated using receiver operating characteristics curve analysis and other statistical methods [148]. In addition to medical applications, CNNs have been used in other fields as well. In petroleum engineering, CNNs have been employed for the recognition of oil and gas reservoir space. These networks have the ability to classify input information based on its hierarchical structure, making them suitable for characterizing learning tasks [150].

In computer science, CNNs have been used for cortical parcellation, a process that involves dividing the cerebral cortex into different regions. The use of spherical harmonics-based convolution in CNNs has improved the accuracy of cortical parcellation on a fine-grained triangle mesh of the cortex [151]. While CNNs have achieved remarkable progress in various

applications, it is important to note that they are not without limitations. Some studies have shown that CNNs can be sensitive to manipulation of the input, which raises concerns about the reliability of their decisions [152]. Therefore, it is crucial to carefully evaluate and interpret the results obtained from CNN models.

In conclusion, CNNs are powerful deep learning algorithms that have been successfully applied in various fields for tasks such as image classification, object detection, and segmentation. They have shown promising results in medical imaging analysis, petroleum engineering, and cortical parcellation. However, it is important to consider the limitations and potential biases of CNNs when interpreting their results. Further research and development are needed to improve the robustness and interpretability of CNN models.

2.5 Noise

Noise in spoof detection refers to unwanted or extraneous signals that can interfere with the detection of spoofing attacks. It can have a significant impact on the performance of spoof detection systems, affecting their accuracy and reliability [153].

There are various types of noise that can affect spoof detection systems. One common type is white noise, which is a random signal with equal intensity at all frequencies. White noise can be caused by various factors, such as electronic components, thermal noise, or environmental factors. It is characterized by a flat power spectral density and can mask or distort the genuine speech signal, making it difficult to detect spoofing attacks [27].

Another type of noise is babble noise, which refers to the background noise or interference in a recording environment. Babble noise can include sounds from other speakers, background music, or environmental noise. It can degrade the quality of the speech signal and make it challenging to distinguish between genuine and spoofed speech [27].

Car noise is another type of noise that can affect spoof detection systems. It refers to the noise generated by the engine, tires, or other components of a vehicle. Car noise can be particularly challenging in scenarios where spoofing attacks are attempted using in-car systems or devices. The noise generated by the vehicle can interfere with the detection of spoofing attacks, making it more difficult to accurately classify the speech signal [153].

The effects of noise on spoof detection systems can be detrimental. Noise can introduce errors or false positives/negatives in the detection process, leading to incorrect classification of genuine and spoofed speech. It can reduce the overall accuracy and reliability of the system, making it more vulnerable to spoofing attacks [153].

To mitigate the effects of noise on spoof detection systems, various solutions have been proposed. One approach is noise reduction or removal techniques. These techniques aim to suppress or eliminate unwanted noise from the speech signal, improving the signal-to-noise ratio and enhancing the performance of the spoof detection system. One common method for noise reduction is the use of noise reduction masks, which estimate the noise characteristics and attenuate the noise components in the speech signal [43].

Another approach is multi-task learning, which involves training the spoof detection system to simultaneously classify different types of noise. By incorporating noise classification into the training process, the system can learn to distinguish between genuine speech, spoofed speech, and different types of noise. This can improve the system's robustness to noise and enhance its ability to detect spoofing attacks in noisy environments [153].

In summary, noise in spoof detection refers to unwanted signals that can interfere with the detection of spoofing attacks. It can include white noise, babble noise, and car noise, among others. Noise can have a detrimental effect on spoof detection systems, reducing their accuracy and reliability. To mitigate the effects of noise, techniques such as noise reduction and multi-task learning can be employed. Noise reduction masks and training the system to classify different types of noise can improve the system's performance in noisy environments [43,153].

2.6 Noise reducing/removing masks

Noise reducing/removing masks in spoof detection systems are techniques or algorithms that aim to suppress or eliminate unwanted noise from the speech signal, improving the signal-to-noise ratio and enhancing the performance of the spoof detection system [43]. These masks are designed to estimate the noise characteristics and attenuate the noise components in the speech signal [43].

There are different types of masks that can be used for noise reduction in spoof detection systems. Three common types of masks are ideal binary masks, ideal ratio masks, and complex ideal ratio masks.

1. Ideal Binary Masks (IBM): An ideal binary mask is a binary mask that indicates which frequency components of the speech signal contain noise and which components contain the desired speech signal. It is defined as a binary function that is equal to 1 for speech components and 0 for noise components. The IBM can be obtained by comparing the magnitude spectrum of the noisy speech signal with the magnitude spectrum of the clean speech signal [43].

2. Ideal Ratio Masks (IRM): An ideal ratio mask is a mask that represents the ratio between the clean speech signal and the noisy speech signal in the frequency domain. It indicates the relative strength of the desired speech signal compared to the noise. The IRM can be obtained by dividing the magnitude spectrum of the clean speech signal by the magnitude spectrum of the noisy speech signal [43].

3. Complex Ideal Ratio Masks (cIRM): A complex ideal ratio mask is an extension of the ideal ratio mask that also considers the phase information of the speech signal. It represents the ratio between the complex spectra of the clean speech signal and the noisy speech signal. The cIRM can be obtained by dividing the complex spectrum of the clean speech signal by the complex spectrum of the noisy speech signal [43].

Deep learning techniques can be used to create masks for noise reduction in spoof detection systems. Deep neural networks can be trained to estimate the noise characteristics and generate masks that can effectively attenuate the noise components in the speech signal. These networks can learn complex patterns and relationships in the data, allowing them to adaptively estimate and remove the noise from the speech signal [43].

Masks work by selectively attenuating or removing the noise components in the speech signal while preserving the desired speech signal. The masks are applied to the frequency domain representation of the speech signal, such as the magnitude spectrum or the complex spectrum. By attenuating the noise components, the masks improve the signal-to-noise ratio and enhance the quality and intelligibility of the speech signal. This, in turn, improves the

performance of the spoof detection system by making it easier to distinguish between genuine and spoofed speech [43].

The performance of masks in spoof detection systems can vary depending on various factors, such as the quality of the masks, the characteristics of the noise, and the specific spoof detection algorithm used. In general, masks can be effective in reducing the impact of noise on spoof detection systems and improving their accuracy and reliability. However, the effectiveness of masks may also depend on the specific noise conditions and the robustness of the spoof detection algorithm to different types of noise [43].

In conclusion, noise reducing/removing masks in spoof detection systems are techniques or algorithms that aim to suppress or eliminate unwanted noise from the speech signal. They can be implemented using different types of masks, such as ideal binary masks, ideal ratio masks, and complex ideal ratio masks. Deep learning techniques can be used to create masks that effectively attenuate noise components. Masks work by selectively attenuating or removing noise while preserving the desired speech signal. The performance of masks in spoof detection systems can vary depending on various factors, but they can generally improve the accuracy and reliability of the system by reducing the impact of noise.

3. MATERIALS AND METHODS

3.1 ASVspoof

The famous spoof speech datasets include the ASVspoof Challenge datasets, such as ASVspoof 2015 [154], ASVspoof 2017 [155], ASVspoof 2019 [156], and ASVspoof 2021 [157]. These datasets contain spoofed speech generated using various techniques, including speech synthesis, voice conversion, and replay attacks. Other datasets mentioned in the references include the VCTK corpus (Lee, 2022), the PartialSpoof database (Zhang et al., 2021), the SAS corpus [154], and the Red Dots and VoicePA datasets (Chadha et al., 2022). These datasets have been used for research and development of spoofing detection systems in the field of automatic speaker verification.

ASVspoof is an initiative that aims to protect automatic speaker verification (ASV) systems from spoofing attacks. It was created to promote the development of countermeasures against

various types of spoofing attacks, such as speech synthesis and voice conversion attacks. The ASVspoof initiative has organized several challenges to evaluate the performance of spoofing countermeasures and provide a platform for researchers to develop and test their algorithms [158].

The ASVspoof 2015 dataset is one of the most popular and widely used datasets for spoof speech detection. It was designed to evaluate the performance of countermeasures against speech synthesis and voice conversion spoofing attacks [159]. The dataset consists of three subsets: a training set, a development set, and an evaluation set. Each subset contains a mix of genuine and spoofed speech [32]. The spoofed speech in the dataset was generated using ten different spoofing algorithms, including voice conversion and speech synthesis algorithms. The attacks in the dataset are labeled as S1, S2, S3, S4, S5, S6, S7, S8, S9, and S10 [159].

The ASVspoof 2015 dataset has been used in various studies to evaluate the performance of different spoof speech detection algorithms [159,32]. It has also been used to measure the impact of countermeasures on the reliability of ASV systems [159]. However, it has been noted that most countermeasures submitted to the ASVspoof 2015 Challenge failed to effectively detect the S10 attack [159]. This highlights the need for further research and development of robust countermeasures against diverse spoofing attack types [43].

In addition to the ASVspoof 2015 dataset, there have been other challenges and datasets in the ASVspoof series. For example, the ASVspoof 2017 challenge focused on replay spoofing attacks [160]. The ASVspoof 2019 challenge included speech synthesis, voice conversion, and replay spoofing attacks [156]. These challenges have provided researchers with opportunities to develop and evaluate countermeasures against different types of spoofing attacks.

Overall, the ASVspoof initiative and its associated datasets have played a significant role in advancing research on spoof speech detection and developing robust countermeasures against spoofing attacks. These datasets have been widely used in the research community to evaluate the performance of different algorithms and to foster the development of generalized countermeasures against spoofing attacks [159,160]. However, there is still a need for further research to improve the detection of diverse spoofing attack types and to enhance the robustness of ASV systems against spoofing attacks [43].

The proposed system was tested using the ASVspoof 2015 dataset [154]. This dataset consists of non-overlapping training, development, and test subsets. The spoof speech attacks are included in this dataset. While there are 10 different attack algorithms, only five of them (S1-S5) are available in the training set. The remaining five (S6-S10) are only available in the test set.

Table 3.1: ASVSPOOF 2015 data distribution

ASVSPOOF 2015	Training Data	Test Data
Real Speech	3750	9404
Spoofed Speech	12625	184000

3.2 Noise

For noise, 'car', 'white', and 'babble' noises from the Noisex-92 database [161] and 'cafe' noise from the QUT-Noise database [161] were used, based on similar studies in the literature [163, 32]). White, babble, and cafe noises were used for mask training. Car noise was only used in the tests to analyze the system's performance under noise types that it had not previously encountered.

3.3 CNN Parameters

The CNN structure used to obtain the mask was trained with a learning rate of 3e-4 and binary cross entropy was chosen as the learning criterion. Only the noisy versions of real speech audios (3750 speech samples) were used as the training data. Each data was corrupted with random noise type selected from white, babble, and cafe noises at a random SNR values between 0 dB and 20 dB. Thus, multi-condition training was performed to prevent the system from focusing on a single noise type and level.

3.4 I-Vector Parameters

The first step in i-vector extraction is obtaining MFCC features. For this, the audio signal is divided into frames of 25 ms length with a frame step of 10 ms. The windowed frames are transformed with a 512-point FFT. The filter bank consists of 32 triangular filters. After

discrete cosine transformation, 32 coefficients are used. In addition, delta and delta-delta features are added to obtain 96-dimensional features per frame.

The CQT is applied with a maximum frequency of Fmax = 8kHz. The number of octaves is 9. The number of bins per octave B is set to 96. Re-sampling is applied with a sampling period of 16 bins in the first octave. Resulting feature vectors are of dimension 19, excluding the C_0 coefficient (C_f 29 coefficients + C_0 for the original system).

The UBM consists of 512 Gaussian components and is trained only on real speakers in the training data. The 600-dimensional T matrix is trained using the entire training data (Hanilçi, 2018). After obtaining i-vectors, whitening, WCCN, and length normalization are applied.

The goal of masking is to improve the robustness against noise by distinguishing less reliable regions of the speech spectrum (more corrupted by noise) and more reliable regions (less affected by noise). Previous studies have shown that applying classic SNR-based masks for spoofed speech detection yields the best results in noisy scenarios [6]. In the proposed system, the CNN structure used in [6] is preferred due to its high performance. The mask creation network is shown in Figure 1. Noisy spectrograms of 31 frames in length, consisting of 15 frames to the right and 15 frames to the left of the central frame are used as inputs to the CNN structure. Therefore, the output of the system (the last linear layer) indicates the signal-to-noise ratio (SNR) for the relevant frame. The sigmoid function is applied to these values to obtain mask values in the range of 0-1. Here, 0 represents completely noise, and 1 represents completely speech information.

The average noise shown in Figure 1 is calculated by averaging the first 10 frames of the corresponding noisy speech data. Typically, it is assumed that the first and last few frames of the speech signal contain only noise information. Therefore, the trained CNN structure has an explicit noise reference, instead of relying only on the spectrogram. During the training phase, the instantaneous SNR target presented to the CNN for each frame is calculated as follows:

$$SNR(\mathrm{t},\mathrm{f}) = 20.\log_{10}\frac{X(\mathrm{t},\mathrm{f})}{N(\mathrm{t},\mathrm{f})} \tag{1}$$

The notation (t, f) represents time-frequency partitions. X and N are the spectrograms of the clean speech and noise, respectively (generated using short-time Fourier transform (STFT)). The values of the target masks to be used in the training phase are obtained using an adjustable sigmoid function given in Equation 2.

$$m_k = \frac{1}{1 - e^{-\alpha(SNR(t,f_k)-\beta)}} \tag{2}$$

Here, α controls the slope of the sigmoid, and β corresponds to the threshold commonly used to define Ideal Binary Masks (IBMs) [18]. Combining these two equations, the Equation 3 is obtained, which calculates the target mask values as $\gamma = 20.\ \alpha / \log(10))$.

$$m_k = \frac{X^{\gamma}}{X^{\gamma} + e^{\alpha\beta}.N^{\gamma}} \tag{3}$$

Figure 3.1, an example spectrogram at the input of the network and the corresponding mask at the output can be seen. The generated mask is multiplied with the noisy spectrogram to obtain a spectrogram with reduced noise. After this stage, a simple speech activity detection system [1] was used to discard frames containing silence or noise.

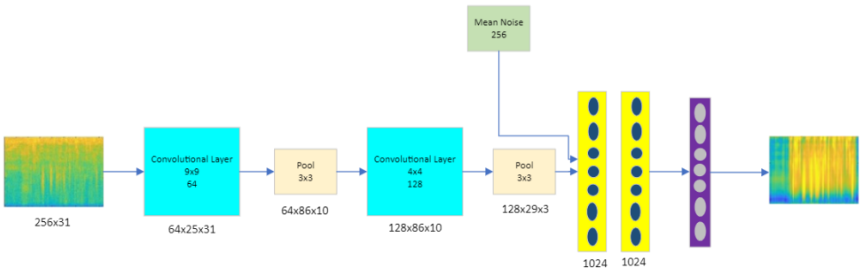

Figure 3.1. CNN architecture

4. ANALYSIS AND DISCUSSION

Table 4.1 shows the performance of the proposed methods on the development data. The EER value represents the average EER values for five different attacks in the development set. For comparison, the results of MFCC-based i-vector without mask [163] are also included in the table. As can be seen, i-vectors enhanced with masks provide significantly better performance than classical i-vectors.

Table 4.1. EERs (%) for the development data

Noise Type	dB	Proposed System with mask				MFCC without mask [163]	MFCC without mask [163]
		MFCC based i-vector	MFCC based i-vector	CQCC based i-vector	CQCC based i-vector		
		COS	PLDA	COS	PLDA	COS	PLDA
White	0 dB	31.05	32.17	40.27	41.56	43.47	43.67
	10dB	25.32	26.60	32.49	33.47	36.35	37.87
	20dB	16.44	18.06	23.57	23.12	26.48	25.32
Babble	0 dB	30.80	31.09	39.67	37.76	45.71	45.82
	10dB	19.44	20.25	28.62	29.66	33.59	33.13
	20dB	10.34	11.49	18.78	19.62	20.94	20.65
Car	0dB	27.54	29.42	32.45	33.82	39.62	38.01
	10dB	24.16	25.38	23.72	24.66	33.67	32.50
	20dB	17.08	18.64	15.54	16.47	24	24.35

As seen in the table, the highest relative improvement in MFCC based i-vector system, with a rate of over 50%, was achieved with the cosine distance classifier at 20 dB level for the babble noise type for encountered noise type. The lowest improvement in MFCC based i-vector system was observed with the PLDA classifier at 20 dB level for car noise type with a 22% improvement rate. The results indicate that masking contributes to robustness compared to the same system without mask. Similar observations can be made for CQCC based system.

Table 4.2 and Table 4.3 show the results for the test data by using MFCC based i-vector system. The noticeable issue here is the low performance for the car noise type. Generally, this type of noise is considered to be the least disruptive [32,163]. However, the masking performance is affected by this noise type because it was not used in the training stage of the mask. As evidence, the examples given on the logarithmic scale in Figure 2. The noisy spectrograms on the top belong to signals corrupted with babble 0 dB and car 20 dB.

Table 4.2. EERs (%) for known attacks of evaluation data (Proposed System/MFCC based i-vector with mask)

Noise Type	dB	S1		S2		S3		S4		S5	
		COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA
White	0	25.85	26.07	36.68	36.72	20.46	20.76	20.15	20.50	28.46	28.21
	10	16.96	18.46	25.96	27.13	6.80	7.38	6.88	7.48	18.05	18.72
	20	5.84	7.14	11.04	12.34	1.48	1.76	1.50	1.83	9.35	10.00
Babble	0	28.97	29.57	38.38	38.51	31.32	31.52	31.47	31.76	31.12	31.55
	10	18.54	20.63	29.96	31.26	16.05	16.97	16.49	17.53	19.59	20.99
	20	10.83	13.55	18.18	20.59	3.55	4.34	3,88	4,71	10,87	12.73
Car	0	28.97	35.95	38.94	42.54	18.48	22.26	19.35	23.24	22.59	26.04
	10	27.58	34.11	39.61	43.87	18.34	21.72	19.30	22.90	19.44	23.58
	20	21.85	26.33	35.40	38.92	9.54	11.48	9.85	11.84	17.44	21.05

Table 4.3. (Proposed System/MFCC based i-vector with mask**)**

Noise Type	dB	S6		S7		S8		S9		S10	
		COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA	COS	PLDA
White	0	33.57	33.70	32.71	33.03	16.76	16.68	29.90	30.10	41.12	41.35
	10	22.55	23.29	23.91	25.28	6.60	6.91	21.92	22.98	35.16	35.40
	20	13.28	14.21	10.05	11.56	2.95	3.39	10.27	12.22	34.29	34.29
Babble	0	32.96	33.08	33.58	34.30	25.86	26	33.33	33.67	40.54	40.45
	10	21.69	22.90	23.24	25.39	16.65	17.53	24.44	26.35	39.84	40.06
	20	14.08	15.79	12.54	15.32	7.61	8.38	14.15	17.14	40.45	40.87
Car	0	26.63	29.95	29.84	34.95	29.42	33.28	40.82	43.21	45.21	47.91
	10	24.73	29.14	28.59	33.58	28.72	32	38.97	42.01	49.82	49.96
	20	22.83	26.83	23.67	27.63	17.53	19.01	29.05	33.33	50	50

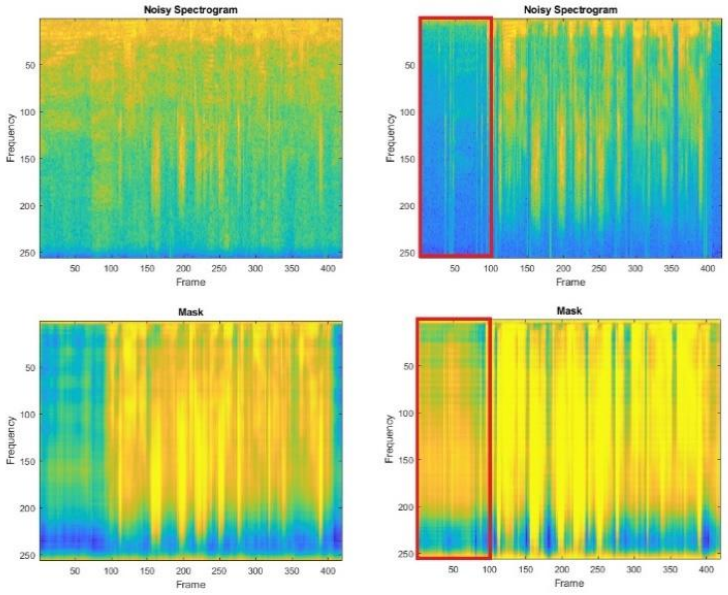

Figure 4.1. Spectrogram of speech data corrupted with babble 0 dB (upper-left) and car 20 dB (upper-right), and their masks below

Table 4.4 shows the performance of the proposed systems and the GMM method for the test data from [163] for comparison. [163] did not analyze the performance of i-vectors since GMM outperformed it. The average EER values are calculated for known attacks (S1-S5), and unknown attacks (S6-S9). Attack type S10 is difficult to detect and has reduced the system performance; therefore, it is not included in the average calculation as in [163].

Table 4.4. Comparison of average EERs (%) for evaluation (K: Known attacks (S1-S5), U: Unknown attacks (S6-S9)

Noise Type	dB	MFCC based i-vector		CQCC based i-vector		MFCC ([163])		RPS ([163])		SCMC ([163])		MGD ([163])	
		K	U	K	U	K	U	K	U	K	U	K	U
		Cos	PLDA	Cos	PLDA								
White	0	26.32	26.45	28.23	28.37	35.07	39.66	44.56	46.64	43.73	42.27	44.42	45.88
	10	14.93	15.83	18.74	19.61	25.45	29.78	42.16	44.98	33.36	32.14	37.42	38.66
	20	5.84	6.61	9.13	10.34	16.43	17.94	38.53	40.62	19.92	15.40	27.25	36.24
Babble	0	32.25	32.58	31.43	31.76	33.54	28.40	40.81	40.66	29.74	25.13	37.59	40.77
	10	20,12	21.47	21.50	23.04	15.59	12.76	21.17	23.71	8.32	5.30	26.30	35.65
	20	9.46	11.18	12.09	14.15	7.48	6.49	6.09	10.62	2.15	1.39	14.20	23.55
Car	0	25.66	30.00	31.67	35.34	17.33	14.69	24.66	25.67	8.59	7.36	30.32	36.63
	10	24.85	29.23	30.25	34.18	7.31	6.03	5.28	9.93	2.16	1.67	15.99	24.44
	20	18.81	21.92	23.27	26.70	3.57	2.83	0.74	3.67	0.79	0.52	9.39	16.12

5. CONCLUSION

This study proposed noise mask based robust i-vector extraction and examined its performance for noisy spoofed speech detection tasks. The results showed that the mask structure is successful in reducing noise effects. This situation reveals that mask structures can be useful in an area where traditional speech enhancement methods have performance-decreasing effects [163].

The results indicate that the proposed approach is more effective for lower dBs. Also. the proposed system delivered the worst results for the unseen noise type (car). This result emphasizes the importance of creating balanced training data. More noise types are necessary to increase the generalization capacity of the network. as noise type may not be known a priori for a practical spoof detection system.

The CNN-based mask. on the other hand. failed against a noise type that was not encountered in the training phase. This situation provides a clue about the necessity of increasing the diversity of noise in the database to prevent memorization.

The i-vector method was chosen due to its high performance for speaker verification. With the proposed method. The same i-vectors can be used for both speaker verification and spoof detection. in a robust manner. However. compared to the state-of-the-art systems in detecting synthetic speech under noise [32,70]. the proposed system was found to be far behind. A reason for this is the low performance of i-vectors in short audio recordings [163]. where the average data length in ASVspoof 2015 dataset is 3.5 seconds. Another reason is the other systems' complexities. For example. The study in [32] designed a system which consists of two different feature types. deep learning models for each feature. and an external classifier in addition to the CNN-based noise mask. Therefore. even though the masking approach performs better than the traditional methods. advanced architectures are necessary for achieving impressive results.

REFERENCES

[1] R. de Luis-García. C. Alberola-López. O. Aghzout. and J. Ruiz-Alzola. “Biometric Identification Systems.” Signal Processing. vol. 83. no. 12. pp. 2539–2557. 2003. doi:10.1016/j.sigpro.2003.08.001

[2] J.-F. Bonastre et al.. “Person authentication by voice: A need for caution.” 8th European Conference on Speech Communication and Technology (Eurospeech 2003). 2003. doi:10.21437/eurospeech.2003-9

[3] Kittler. J.. & Nixon. M. S. (Eds.). (2003). Audio-and Video-Based Biometric Person Authentication: 4th International Conference. AVBPA 2003. Guildford. UK. June 9-11. 2003. Proceedings (Vol. 2688). Springer Science & Business Media.

[4] Uzunçarşılı. M.. 2005. Investigation of Speaker Recognition Algorithms Based on Vector Quantization Techniques. M.Sc.. Ankara University. Institute of Science and Technology. Ankara

[5] Selen. N. (1979). Söyleyiş Sesbilimi. Akustik Sesbilim ve Türkiye Türkçesi. Ankara: Turkish Language Association Publications: 454.

[6] Erzin. E. ve ark.. 2005. Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability. IEEE Multimedia. vol. 7. no. 5. pp. 840-852.

[7] Campbell. J.P.. 1997. Speaker recognition: a tutorial. Proc. IEEE 85. 1437–1462. doi:10.1109/5.628714

[8] Furui. S.. 1997. Recent advances in speaker recognition. Pattern Recognit. Lett. 18. 859–872. doi:10.1016/S0167-8655(97)00073-1

[9] Ramitha. R.. Baburaj. M.. George. S. (2015). Dictionary Learning Based Sparse Coefficients For Speech Recognition In Noisy Environment.. https://doi.org/10.1109/raics.2015.7488405

[10] Junyi. A.. Rui. W.. Long. Z.. Chengyi. W.. Shuo. R.. Yu. W.. … & Furu. W. (2021). Speecht5: Unified-modal Encoder-decoder Pre-training For Spoken Language Processing.. https://doi.org/10.48550/arxiv.2110.07205

[11] Denes. P.. & Mathews. M. V. (1960). Spoken digit recognition using time-frequency pattern matching. The Journal of the Acoustical Society of America. 32(11). 1450-1455.

[12] S. Das and W. Mohn. "A scheme for speech processing in automatic speaker verification." in IEEE Transactions on Audio and Electroacoustics. vol. 19. no. 1. pp. 32-43. March 1971. doi: 10.1109/TAU.1971.1162158.

[13] Li. K. P.. Dammann. J. E.. & Chapman. W. D. (1966). Experimental studies in speaker

verification. using an adaptive system. The Journal of the Acoustical Society of America. 40(5). 966-978.

[14] Cooke. M.. Hershey. J.. Rennie. S. (2010). Monaural Speech Separation and Recognition Challenge. Computer Speech & Language. 1(24). 1-15. https://doi.org/10.1016/j.csl.2009.02.006

[15] Gref. M.. Schmidt. C.. Behnke. S.. Köhler. J. (2019). Two-staged Acoustic Modeling Adaption For Robust Speech Recognition By the Example Of German Oral History Interviews.. https://doi.org/10.1109/icme.2019.00142

[16] Khademian. M.. Homayounpour. M. (2018). Monaural Multi-talker Speech Recognition Using Factorial Speech Processing Models. Speech Communication. (98). 1-16. https://doi.org/10.1016/j.specom.2018.01.007

[17] Yousefi. M.. Hanse. J. (2021). Speaker Conditioning Of Acoustic Models Using Affine Transformation For Multi-speaker Speech Recognition.. https://doi.org/10.48550/arxiv.2111.00320

[18] Alhumsi. M.. Belhassen. S. (2021). The Challenges Of Developing a Living Arabic Phonetic Dictionary For Speech Recognition System: A Literature Review. Adv. J Social Sci.. 1(8). 164-170. https://doi.org/10.21467/ajss.8.1.164-170

[19] Geng. M.. Xie. X.. Liu. S.. Yu. J.. Hu. S.. Liu. X.. … & Meng. H. (2020). Investigation Of Data Augmentation Techniques For Disordered Speech Recognition.. https://doi.org/10.21437/interspeech.2020-1161

[20] Dubagunta. S.. Kabil. S.. Magimai.-Doss. M. (2019). Improving Children Speech Recognition Through Feature Learning From Raw Speech Signal.. https://doi.org/10.1109/icassp.2019.8682826

[21] Zhen. W. (2022). Research On English Vocabulary and Speech Corpus Recognition Based On Deep Learning. Wireless Communications and Mobile Computing. (2022). 1-11. https://doi.org/10.1155/2022/2882964

[22] Eckert. M.. Teubner-Rhodes. S.. Vaden. K. (2016). Is Listening In Noise Worth It? the Neurobiology Of Speech Recognition In Challenging Listening Conditions. Ear and Hearing. 1(37). 101S-110S. https://doi.org/10.1097/aud.0000000000000300

[23] Hanilci. C.. Kinnunen. T.. Sahidullah. M.. Sizov. A. (2015). Classifiers For Synthetic Speech Detection: a Comparison.. https://doi.org/10.21437/interspeech.2015-466

[24] Kenny. P.. Ouellet. P.. Dehak. N.. Gupta. V.. Dumouchel. P. (2008). A Study Of Interspeaker Variability In Speaker Verification. Ieee Transactions on Audio Speech and Language Processing. 5(16). 980-988. https://doi.org/10.1109/tasl.2008.925147

[25] Tun. P. and Wingfield. A. (1999). One Voice Too Many: Adult Age Differences In Language Processing With Different Types Of Distracting Sounds. The Journals of Gerontology Series B. 5(54B). P317-P327. https://doi.org/10.1093/geronb/54b.5.p317

[26] Eckert. M.. Cute. S.. Vaden. K.. Kuchinsky. S.. Dubno. J. (2012). Auditory Cortex Signs Of Age-related Hearing Loss. Journal of the Association for Research in Otolaryngology. 5(13). 703-713. https://doi.org/10.1007/s10162-012-0332-5

[27] Tian. X.. Wu. Z.. Xiao. X.. Chng. E.. Li. H. (2016). Spoofing Detection From a Feature Representation Perspective.. https://doi.org/10.1109/icassp.2016.7472051

[28] Alegre. F.. Amehraye. A.. Evans. N. (2013). A One-class Classification Approach To Generalised Speaker Verification Spoofing Countermeasures Using Local Binary Patterns.. https://doi.org/10.1109/btas.2013.6712706

[29] Wei. L.. Long. Y.. Wei. H.. Li. Y. (2022). New Acoustic Features For Synthetic and Replay Spoofing Attack Detection. Symmetry. 2(14). 274. https://doi.org/10.3390/sym14020274

[30] Wickramasinghe. B.. Irtza. S.. Ambikairajah. E.. Epps. J. (2018). Frequency Domain Linear Prediction Features For Replay Spoofing Attack Detection.. https://doi.org/10.21437/interspeech.2018-1574

[31] Suthokumar. G.. Sriskandaraja. K.. Sethu. V.. Wijenayake. C.. Ambikairajah. E. (2017). Independent Modelling Of High and Low Energy Speech Frames For Spoofing Detection.. https://doi.org/10.21437/interspeech.2017-836

[32] Gomez-Alanis. A.. Peinado. A.. López. J.. Gomez. A. (2018). Performance Evaluation Of Front- and Back-end Techniques For Asv Spoofing Detection Systems Based On Deep Features.. https://doi.org/10.21437/iberspeech.2018-10

[33] Mochizuki. S.. Shiota. S.. Kiya. H. (2018). Voice Liveness Detection Using Phoneme-based Pop-noise Detector For Speaker Verification.. https://doi.org/10.21437/odyssey.2018-33

[34] Shiota. S.. Villavicencio. F.. Yamagishi. J.. Ono. N.. Echizen. I.. Matsui. T. (2016). Voice Liveness Detection For Speaker Verification Based On a Tandem Single/double-channel Pop Noise Detector.. https://doi.org/10.21437/odyssey.2016-37

[35] Korshunov. P. and Marcel. S. (2016). Cross-database Evaluation Of Audio-based Spoofing Detection Systems.. https://doi.org/10.21437/interspeech.2016-1326

[36] Rössler. A.. Cozzolino. D.. Verdoliva. L.. Riess. C.. Thies. J.. Niessner. M. (2019). Faceforensics++: Learning To Detect Manipulated Facial Images.. https://doi.org/10.1109/iccv.2019.00009

[37] Li. Y.. Yang. X.. Sun. P.. Qi. H.. Lyu. S. (2019). Celeb-df: a Large-scale Challenging Dataset For Deepfake Forensics.. https://doi.org/10.48550/arxiv.1909.12962

[38] Kim. J. and Ban. S. (2022). Phase-aware Spoof Speech Detection Based On Res2net With Phase Network.. https://doi.org/10.48550/arxiv.2203.10793

[39] Liu. W.. Sun. M.. Zhang. X.. hamme. H.. Zheng. T. (2022). A Multi-resolution Front-end For End-to-end Speech Anti-spoofing.. https://doi.org/10.21437/odyssey.2022-17

[42] Chadha. A.. Abdullah. A.. Angeline. L. (2021). A Unique Glottal Flow Parameters Based Features For Anti-spoofing Countermeasures In Automatic Speaker Verification. International Journal of Advanced Computer Science and Applications. 8(12). https://doi.org/10.14569/ijacsa.2021.0120894

[43] DİŞKEN. G. (2022). Robust Spoofed Speech Detection With Denoised I-vectors. Gazi University Journal of Science. 1-1. https://doi.org/10.35378/gujs.1062788

[44] Choi. Y.. Jung. Y.. Kim. H. (2021). Neural Mos Prediction For Synthesized Speech Using Multi-task Learning With Spoofing Detection and Spoofing Type Classification.. https://doi.org/10.1109/slt48900.2021.9383533

[45] Jain.A. K.. R. P. W. Duin and Jianchang Mao. "Statistical pattern recognition: a review." in IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 22. no. 1. pp. 4-37. Jan. 2000. doi: 10.1109/34.824819.

[46] Jain. A. and Zongker. D. (1997) Feature Selection: Evaluation. Application. and Small Sample Performance. IEEE Transactions on Machine Intelligence. 19. 153-158.

[47] Kinnunen. T. and Li. H.Z. (2010) An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Communication. 52. 12-40. http://dx.doi.org/10.1016/j.specom.2009.08.009

[48] Wolf. J.J.. 1972. Efficient Acoustic Parameters for Speaker Recognition. J. Acoust. Soc. Am. 51. 2044–2056.doi:10.1121/1.19130657

[49] Li. M.. Kim. J.. Lammert. A.. Ghosh. P.K.. Ramanarayanan. V.. Narayanan. S.. 2016. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. 36. 196–211. doi:10.1016/j.csl.2015.05.00

[50] Venturini. A.. Zao. L.. Coelho. R.. 2014. On speech features fusion. alphaintegration Gaussian modeling and multi-style training for noise robust speaker classification. IEEE/ACM Trans. Audio. Speech. Lang. Process. 22. 1951–1964. doi:10.1109/TASLP.2014.2355821

[51] Sahidullah. M.. Kinnunen. T.. Hanilci. C. (2015). A Comparison Of Features For Synthetic Speech Detection.. https://doi.org/10.21437/interspeech.2015-472

[52] Palaz. D.. Magimai.-Doss. M.. Collobert. R. (2015). Convolutional Neural Networks-based Continuous Speech Recognition Using Raw Speech Signal.. https://doi.org/10.1109/icassp.2015.7178781

[53] Nuesse. T.. Wiercinski. B.. Brand. T.. Holube. I. (2019). Measuring Speech Recognition With a Matrix Test Using Synthetic Speech. Trends in Hearing. (23). 233121651986298. https://doi.org/10.1177/2331216519862982

[54] Biswas. A.. Sahu. P.. Bhowmick. A.. Chandra. M. (2015). Admissible Wavelet Packet Sub-band-based Harmonic Energy Features For Hindi Phoneme Recognition. IET signal process.. 6(9). 511-519. https://doi.org/10.1049/iet-spr.2014.0282

[55] Hibare. R.. Vibhute. A. (2014). Feature Extraction Techniques In Speech Processing: a Survey. IJCA. 5(107). 1-8. https://doi.org/10.5120/18744-9997

[56] Anusuya. M. (2011). Front End Analysis Of Speech Recognition: a Review. Int J Speech Technol. 2(14). 99-145. https://doi.org/10.1007/s10772-010-9088-7

[57] Lotfian. R.. Busso. C. (2019). Lexical Dependent Emotion Detection Using Synthetic Speech Reference. IEEE Access. (7). 22071-22085. https://doi.org/10.1109/access.2019.2898353

[58] Mittal. A. and Dua. M. (2021). Static–dynamic Features and Hybrid Deep Learning Models Based Spoof Detection System For Asv. Complex & Intelligent Systems. 2(8). 1153-1166. https://doi.org/10.1007/s40747-021-00565-w

[59] Chadha. A.. Abdullah. A.. Angeline. L. (2021). A Unique Glottal Flow Parameters Based Features For Anti-spoofing Countermeasures In Automatic Speaker Verification. International Journal of Advanced Computer Science and Applications. 8(12). https://doi.org/10.14569/ijacsa.2021.0120894

[60] Astuti. Y.. Hidayat. R.. & Bejo. A. (2022). A Mel-weighted Spectrogram Feature Extraction for Improved Speaker Recognition System. International Journal of Intelligent Engineering & Systems. 15(6).

[61] Heriyanto. H.. Jayadianti. H.. Juwairiah. J. (2021). The Implementation Of Mfcc Feature Extraction and Selection Of Cepstral Coefficient For Qur'an Recitation In Tpa (Qur'an Learning Center) Nurul Huda Plus Purbayan. cset. 1(1). 453-478. https://doi.org/10.31098/cset.v1i1.417

[62] Benayad. N.. Soumaya. Z. (2022). Features Selection By Genetic Algorithm Optimization With K-nearest Neighbour and Learning Ensemble To Predict Parkinson Disease. IJECE. 2(12). 1982. https://doi.org/10.11591/ijece.v12i2.pp1982-1989

[63] Kim. C.. Stern. R. (2016). Power-normalized Cepstral Coefficients (Pncc) For Robust Speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Process.. 7(24). 1315-1329. https://doi.org/10.1109/taslp.2016.2545928

[64] Todisco. M.. Delgado. H.. Evans. N. (2016). A New Feature For Automatic Speaker Verification Anti-spoofing: Constant Q Cepstral Coefficients.. https://doi.org/10.21437/odyssey.2016-41

[65] Cheng et al.. 2020 A.. Prasanna. S.. Dandapat. S.. Hypernasality Severity Detection Using Constant Q Cepstral Coefficients.. https://doi.org/10.21437/interspeech.2019-2151

[66] Dubey. A.. Prasanna. S.. Dandapat. S. (2019). Hypernasality Severity Detection Using Constant Q Cepstral Coefficients.. https://doi.org/10.21437/interspeech.2019-2151

[67] Mittal. A.. Dua. M. (2021). Static–dynamic Features and Hybrid Deep Learning Models Based Spoof Detection System For Asv. Complex Intell. Syst.. 2(8). 1153-1166. https://doi.org/10.1007/s40747-021-00565-w

[68] Nagarsheth. P.. Khoury. E.. Patil. K.. Garland. M. (2017). Replay Attack Detection Using Dnn For Channel Discrimination.. https://doi.org/10.21437/interspeech.2017-1377

[69] Miley (2017) A.. Prasanna. S.. Dandapat. S. Hypernasality Severity Detection Using Constant Q Cepstral Coefficients.. https://doi.org/10.21437/interspeech.2019-2151

[70] Wang. M.. Liang. H.. Wang. W.. Zhao. S.. Cai. X.. Zhao. Y.. … & Liang. W. (2020). Immune-related Adverse Events Of a Pd-l1 Inhibitor Plus Chemotherapy Versus APd-l1 Inhibitor Alone In First-line Treatment For Advanced Non–small Cell Lung Cancer: A Meta-analysis Of Randomized Control Trials. Cancer. 5(127). 777-786. https://doi.org/10.1002/cncr.33270

[71] J. S.. Y. S.. Nandyala. S. (2014). Automatic Speech Emotion and Speaker Recognition Based On Hybrid Gmm And Ffbnn. IJCSA. 1(4). 35-42. https://doi.org/10.5121/ijcsa.2014.4104

[72] Lei. L.. Kun. S. (2016). Speaker Recognition Using Wavelet Cepstral Coefficient. I-vector. and Cosine Distance Scoring And Its Application For Forensics. Journal of Electrical and Computer Engineering. (2016). 1-11. https://doi.org/10.1155/2016/4908412

[73] Kaberpanthi. N.. Datar. A. (2014). Speaker Independent Speech Recognition Using Mfcc With Cubic-log Compression and Vq Analysis. IJCA. 26(95). 33-37. https://doi.org/10.5120/16962-7081

[74] Hattori. H. (1992). Text-independent Speaker Recognition Using Neural Networks.. https://doi.org/10.1109/icassp.1992.226097

[75] A. A.. Dhonde. S. (2015). Speaker Recognition System Using Gaussian Mixture Model. IJCA. 14(130). 38-40. https://doi.org/10.5120/ijca2015907193

[76] Chowdhury. A.. Ross. A.. David. P. (2021). Deeptalk: Vocal Style Encoding For Speaker Recognition and Speech Synthesis.. https://doi.org/10.1109/icassp39728.2021.9414169

[77] Ouisaadane. A.. Safi. S.. Frikuil. M. (2020). Arabic Digits Speech Recognition and Speaker Identification In Noisy Environment Using A Hybrid Model Of Vq And Gmm. TELKOMNIKA. 4(18). 2193. https://doi.org/10.12928/telkomnika.v18i4.14215

[78] Nath. H.. Baruah. U. (2014). Evaluation Of Lower Bounding Methods Of Dynamic Time Warping (Dtw). IJCA. 20(94). 12-17. https://doi.org/10.5120/16550-6168

[79] Das. T.. Misra. S.. Choudhury. S.. Sah. D.. Baruah. U.. Laskar. R. (2015). Comparison Of Dtw Score and Warping Path For Text Dependent Speaker Verification System.. https://doi.org/10.1109/iccpct.2015.7159305

[80] Myers. C.. Rabiner. L.. Rosenberg. A. (1980). Performance Tradeoffs In Dynamic Time Warping Algorithms For Isolated Word Recognition. IEEE Trans. Acoust.. Speech. Signal Process.. 6(28). 623-635. https://doi.org/10.1109/tassp.1980.1163491

[81] Kao. C.. Chueh. H. (2023). Voice Response Questionnaire System For Speaker Recognition Using Biometric Authentication Interface. Intelligent Automation & Soft Computing. 1(35). 913-924. https://doi.org/10.32604/iasc.2023.024734

[82] Alanazy. A.. Alatawi. M. (2021). Automatic Arabic Speech Recognition- a Comparative Study. IJSRP. 12(11). 444-449. https://doi.org/10.29322/ijsrp.11.12.2021.p12064

[83] Al-Haddad. S.. Samad. S.. Hussain. A.. Ishak. K.. Noor. A. (2009). Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering. American Journal of Applied Sciences. 2(6). 290-295. https://doi.org/10.3844/ajassp.2009.290.295

[84] Gupta. A.. Gupta. H. (2013). Applications Of Mfcc and Vector Quantization In Speaker Recognition.. https://doi.org/10.1109/issp.2013.6526896

[85] Dhonde. S.. Jagade. S. (2016). Comparison Of Vector Quantization and Gaussian Mixture Model Using Effective Mfcc Features For Text-independent Speaker Identification. IJCA. 15(134). 11-13. https://doi.org/10.5120/ijca2016908019

[86] Zhao. Y.. Zhu. L. (2017). Speaker-dependent Isolated-word Speech Recognition System Based On Vector Quantization.. https://doi.org/10.1109/iccnea.2017.103

[87] Shore. J.. Burton. D. (1983). Discrete Utterance Speech Recognition Without Time Alignment. IEEE Trans. Inform. Theory. 4(29). 473-491. https://doi.org/10.1109/tit.1983.1056716

[88] Trivedi. J.. Maitra. A.. Mitra. S. (2005). A Hybrid Approach To Speaker Recognition In Multi-speaker Environment.. 272-275. https://doi.org/10.1007/11590316_38

[89] Ma. D.. Zeng. X. (2012). An Improved Vq Based Algorithm For Recognizing Speaker-independent Isolated Words.. https://doi.org/10.1109/icmlc.2012.6359026

[90] Yan-na. W.. Li-yan. Z. (2013). Codebook Design Method Based On Genetic Algorithm For Chinese Continuous Digital Speech Recognition..
https://doi.org/10.2991/icmt-13.2013.16

[91] Shahin. I. (2005). Improving Speaker Identification Performance Under the Shouted Talking Condition Using The Second-order Hidden Markov Models. EURASIP J. Adv. Signal Process.. 4(2005). https://doi.org/10.1155/asp.2005.482

[92] Bourlard. H.. Wellekens. C. (1990). Links Between Markov Models and Multilayer Perceptrons. IEEE Trans. Pattern Anal. Machine Intell.. 12(12). 1167-1178. https://doi.org/10.1109/34.62605

[93] Zietsman. G.. Zietsman. R. (2022). Modelling Of a Speech-to-text Recognition System For Air Traffic Control And Nato Air Command. Journal of Internet Technology. 7(23). 1527-1539. https://doi.org/10.53106/160792642022122307008

[94] Dujaili. M.. Ebrahimi-Moghadam. A.. Fatlawi. A. (2021). Speech Emotion Recognition Based On Svm and Knn Classifications Fusion. IJECE. 2(11). 1259. https://doi.org/10.11591/ijece.v11i2.pp1259-1264

[95] Sun. L.. Fu. S.. Wang. F. (2019). Decision Tree Svm Model With Fisher Feature Selection For Speech Emotion Recognition. J AUDIO SPEECH MUSIC PROC.. 1(2019). https://doi.org/10.1186/s13636-018-0145-5

[96] Bai. J.. Wang. J.. Zhang. X. (2013). A Parameters Optimization Method Of V-support Vector Machine and Its Application In Speech Recognition. JCP. 1(8). https://doi.org/10.4304/jcp.8.1.113-120

[97] Shen. X.. Zhai. Y.. Wang. Y.. Chen. H. (2014). A Speaker Recognition Algorithm Based On Factor Analysis.. https://doi.org/10.1109/cisp.2014.7003905

[98] Liu. Y.. Yang. H.. Zhou. H. (2009). Speaker Identification Based On Emd.. https://doi.org/10.1109/icnidc.2009.5360889

[99] Faraj. M.. Bigun. J. (2007). Speaker and Digit Recognition By Audio-visual Lip Biometrics.. 1016-1024. https://doi.org/10.1007/978-3-540-74549-5_106

[100] Saod. A.. Sulaiman. S.. Harron. N.. Ahmad. A.. Ramlan. S.. Ramli. D. (2012). Evaluation On Data &Amp;#x2014; Speaker Dependability Approaches For Speech Recognition Tasks.. https://doi.org/10.1109/iccsce.2012.6487151

[101] Kenny. P.. Boulianne. G.. Ouellet. P.. Dumouchel. P. (2007). Joint Factor Analysis Versus Eigenchannels In Speaker Recognition. IEEE Trans. Audio Speech Lang. Process.. 4(15). 1435-1447. https://doi.org/10.1109/tasl.2006.881693

[102] Dehak. N.. Kenny. P.. Glembek. O.. Dumouchel. P.. Burget. L.. Hubeika. V.. … & Castaldo. F. (2009). Support Vector Machines and Joint Factor Analysis For Speaker Verification.. https://doi.org/10.1109/icassp.2009.4960564

[103] Kanda. N.. Gaur. Y.. Wang. X.. Meng. Z.. Chen. Z.. Zhou. T.. … & Yoshioka. T. (2020). Joint Speaker Counting. Speech Recognition. and Speaker Identification For Overlapped Speech Of Any Number Of Speakers.. https://doi.org/10.21437/interspeech.2020-1085

[104] Kanda. N.. Meng. Z.. Lu. L.. Gaur. Y.. Wang. X.. Chen. Z.. … & Yoshioka. T. (2021). Minimum Bayes Risk Training For End-to-end Speaker-attributed Asr.. https://doi.org/10.1109/icassp39728.2021.9415062

[105] Jakovljevic. N.. Miskovic. D.. Janev. M.. Sečujski. M.. Delic. V. (2013). Comparison Of Linear Discriminant Analysis Approaches In Automatic Speech Recognition. ElAEE. 7(19). https://doi.org/10.5755/j01.eee.19.7.5167

[106] Nandwana. M.. Hout. J.. McLaren. M.. Stauffer. A.. Richey. C.. Lawson. A.. … & Graciarena. M. (2018). Robust Speaker Recognition From Distant Speech Under Real Reverberant Environments Using Speaker Embeddings.. https://doi.org/10.21437/interspeech.2018-2221

[107] Madikeri. S.. Himawan. I.. Motlicek. P.. Ferras. M. (2015). Integrating Online I-vector Extractor With Information Bottleneck Based Speaker Diarization System.. https://doi.org/10.21437/interspeech.2015-111

[108] McLaren. M.. Leeuwen. D. (2011). To Weight or Not To Weight: Source-normalised Lda For Speaker Recognition Using I-vectors.. https://doi.org/10.21437/interspeech.2011-144

[109] Cumani. S.. Laface. P. (2016). I–vector Transformation and Scaling For Plda Based Speaker Recognition.. https://doi.org/10.21437/odyssey.2016-6

[110] Jaitly. N.. Hinton. G. (2011). Learning a Better Representation Of Speech Soundwaves Using Restricted Boltzmann Machines.. https://doi.org/10.1109/icassp.2011.5947700

[111] Dehak. N.. Kenny. P.. Dehak. R.. Dumouchel. P.. Ouellet. P. (2011). Front-end Factor Analysis For Speaker Verification. IEEE Trans. Audio Speech Lang. Process.. 4(19). 788-798. https://doi.org/10.1109/tasl.2010.2064307

[112] Hanili. C.. Kinnunen. T.. Sizov. A. (2016). Spoofing Detection Goes Noisy: An Analysis Of Synthetic Speech Detection In the Presence Of Additive Noise. Speech Communication. (85). 83-97. https://doi.org/10.1016/j.specom.2016.10.002

[113] Barkan. O.. Weill. J.. Wolf. L.. Aronowitz. H. (2013). Fast High Dimensional Vector Multiplication Face Recognition.. https://doi.org/10.1109/iccv.2013.246

[114] Lu. L.. Dong. Y.. Zhao. X.. Zhao. J.. Dong. C.. Wang. H. (2008). Analysis Of Subspace Within-class Covariance Normalization For Svm-based Speaker Verification.. https://doi.org/10.21437/interspeech.2008-399

[115] Xing. Y.. Tan. P.. Wang. X. (2019). Speaker Verification Normalization Sequence Kernel Based On Gaussian Mixture Model Super-vector and Bhattacharyya Distance. Journal of Low Frequency Noise. Vibration and Active Control. 1(40). 60-71. https://doi.org/10.1177/1461348419880744

[116] Ma. H.. Li. S.. An. Z. (2019). A Fault Diagnosis Approach For Rolling Bearing Based On Convolutional Neural Network and Nuisance Attribute Projection Under Various Speed Conditions. Applied Sciences. 8(9). 1603. https://doi.org/10.3390/app9081603

[117] Yang. D.. Lv. Y.. Yuan. R.. Li. H.. Zhu. W. (2022). Robust Fault Diagnosis Of Rolling Bearings Via Entropy-weighted Nuisance Attribute Projection and Neural Network Under Various Operating Conditions. Structural Health Monitoring. 6(21). 2890-2909. https://doi.org/10.1177/14759217221077414

[118] Struc. V.. Vesnicer. B.. Pavešić. N. (2010). Removing Illumination Artifacts From Face Images Using the Nuisance Attribute Projection.. https://doi.org/10.1109/icassp.2010.5495203

[119] Huang. S.. Wang. S.. Liang. J.. Xu. B. (2011). Exploring Nuisance Attribute Projection and Score Normalization For Glds-svm Based Automatic Mispronunciation Detection Method.. https://doi.org/10.1109/icassp.2011.5947646

[120] Zhao. X.. Dong. Y.. Zhao. J.. Liu. J.. Wang. H. (2008). Comparison Of Input and Feature Space Nonlinear Kernel Nuisance Attribute Projections For Speaker Verification.. https://doi.org/10.21437/interspeech.2008-400

[121] Yang. H.. Dong. Y.. Zhao. X.. Zhao. J.. Lu. L.. Wang. H. (2007). Cluster Adaptive Training Weights As Features In Svm-based Speaker Verification.. https://doi.org/10.21437/interspeech.2007-163

[122] Richardson. F.. Campbell. W. (2011). Nap For High Level Language Identification.. https://doi.org/10.1109/icassp.2011.5947327

[123] Magsino. E. (2020). Speaker Feature Modeling Utilizing Constrained Maximum Likelihood Linear Regression and Gaussian Mixture Models. IJATCSE. 1(9). 562-566. https://doi.org/10.30534/ijatcse/2020/77912020

[124] Wang. E.. Lee. K.. Ma. B.. Li. H.. Guo. W.. Dai. L. (2011). Factored Covariance Modeling For Text-independent Speaker Verification.. https://doi.org/10.1109/icassp.2011.5947443

[125] Ramoji. S.. & Ganapathy. S. (2018. September). Supervised I-vector Modeling-Theory and Applications. In INTERSPEECH (pp. 1091-1095).

[126] Přibil. J.. Přibil. J.. Matoušek. J. (2014). Gmm Classification Of Text-to-speech Synthesis: Identification Of Original Speaker's Voice.. 365-373. https://doi.org/10.1007/978-3-319-10816-2_44

[127] Matza. A.. Bistritz. Y. (2014). Skew Gaussian Mixture Models For Speaker Recognition. IET signal process.. 8(8). 860-867. https://doi.org/10.1049/iet-spr.2013.0270

[128] Gao. Y.. Han. J.. Lin. L.. Lu. C.. Yang. Q. (2009). Chinese Name Speech Classification Using Fisher Score Based On Continuous Density Hidden Markov Models.. https://doi.org/10.1109/iih-msp.2009.36

[129] Yu. C.. Xie. L.. Hu. W. (2016). Feature Optimization Of Speech Emotion Recognition. JBiSE. 10(09). 37-43. https://doi.org/10.4236/jbise.2016.910b005

[130] Chowdhury. F.. Selouani. S.. O'Shaughnessy. D. (2009). Distributed Automatic Text-independent Speaker Identification Using Gmm-ubm Speaker Models.. https://doi.org/10.1109/ccece.2009.5090157

[131] Bouziane. A.. Kharroubi. J.. Zarghili. A. (2017). Towards An Optimal Speaker Modeling In Speaker Verification Systems Using Personalized Background Models. IJECE. 6(7). 3655. https://doi.org/10.11591/ijece.v7i6.pp3655-3663

[132] Boulkenafet. Z.. Bengherabi. M.. Harizi. F.. Nouali. O.. Mohamed. C. (2013). Forensic Evidence Reporting Using Gmm-ubm. Jfa and I-vector Methods: Application To Algerian Arabic Dialect.. https://doi.org/10.1109/ispa.2013.6703775

[133] Leon. P.. Apsingekar. V.. Pucher. M.. Yamagishi. J. (2010). Revisiting the Security Of Speaker Verification Systems Against Imposture Using Synthetic Speech.. https://doi.org/10.1109/icassp.2010.5495413

[134] Larcher. A.. Bonastre. J.. Mason. J. (2013). Constrained Temporal Structure For Text-dependent Speaker Verification. Digital Signal Processing. 6(23). 1910-1917. https://doi.org/10.1016/j.dsp.2013.07.007

[135] Leon. P.. Pucher. M.. Yamagishi. J.. Hernaez. I.. Saratxaga. I. (2012). Evaluation Of Speaker Verification Security and Detection Of Hmm-based Synthetic Speech. IEEE Trans. Audio Speech Lang. Process.. 8(20). 2280-2290. https://doi.org/10.1109/tasl.2012.2201472

[136] Kinnunen. T.. Sedlak. F.. Bednarik. R. (2010). Towards Task-independent Person Authentication Using Eye Movement Signals.. https://doi.org/10.1145/1743666.1743712

[137] Larcher. A.. Bonastre. J.. Mason. J. (2008). Short Utterance-based Video Aided Speaker Recognition.. https://doi.org/10.1109/mmsp.2008.4665201

[138] Chao. Y.. Tsai. W.. Wang. H. (2008). Discriminative Feedback Adaptation For Gmm-ubm Speaker Verification.. https://doi.org/10.1109/chinsl.2008.ecp.54

[139] Chao. Y.. Tsai. W.. Wang. H. (2009). Improving Gmm–ubm Speaker Verification Using Discriminative Feedback Adaptation. Computer Speech & Language. 3(23). 376-388. https://doi.org/10.1016/j.csl.2009.01.002

[140] Larasati. A.. Dwiastutik. A.. Ramadhanti. D.. Mahardika. A. (2018). The Effect Of Kurtosis On the Accuracy Of Artificial Neural Network Predictive Model. MATEC Web Conf.. (204). 02018. https://doi.org/10.1051/matecconf/201820402018

[141] Desai. S.. Raghavendra. E.. Yegnanarayana. B.. Black. A.. Prahallad. K. (2009). Voice Conversion Using Artificial Neural Networks.. https://doi.org/10.1109/icassp.2009.4960478

[142] Krothapalli. S.. Yadav. J.. Sarkar. S.. Koolagudi. S.. Vuppala. A. (2012). Neural Network Based Feature Transformation For Emotion Independent Speaker Identification. Int J Speech Technol. 3(15). 335-349. https://doi.org/10.1007/s10772-012-9148-2

[143] Xu. L.. Chen. X.. Zhou. N.. Li. Y.. Zhao. X.. Han. D. (2007). Recognition Of Lexical Tone Production Of Children With An Artificial Neural Network. Acta Oto-Laryngologica. 4(127). 365-369. https://doi.org/10.1080/00016480601011477

[144] Xiang. B.. Berger. T. (2003). Efficient Text-independent Speaker Verification With Structural Gaussian Mixture Models and Neural Network. IEEE Trans. Speech Audio Process.. 5(11). 447-456. https://doi.org/10.1109/tsa.2003.815822

[145] Krishna. A. (2019). Speaker Recognition and Gender Identification Using Artificial Neural Network And Support Vector Machine. IJRASET. 7(7). 499-504. https://doi.org/10.22214/ijraset.2019.7078

[146] Anusuya. M. (2011). Front End Analysis Of Speech Recognition: a Review. Int J Speech Technol. 2(14). 99-145. https://doi.org/10.1007/s10772-010-9088-7

[147] Ogihara. A.. Unno. H.. Shiozaki. A. (2005). Discrimination Method Of Synthetic Speech Using Pitch Frequency Against Synthetic Speech Falsification. IEICE Transactions on Fundamentals of Electronics. Communicatio. 1(E88-A). 280-286. https://doi.org/10.1093/ietfec/e88-a.1.280

[148] Choi. B.. Kang. S.. Kim. H.. Kwon. O.. Vu. H.. Youn. S. (2021). Faster Region-based Convolutional Neural Network In the Classification Of Different Parkinsonism Patterns Of The Striatum On Maximum Intensity Projection Images Of [18f]fp-cit Positron Emission Tomography. Diagnostics. 9(11). 1557. https://doi.org/10.3390/diagnostics11091557

[149] Xu. Y.. Hosny. A.. Zeleznik. R.. Parmar. C.. Coroller. T.. Franco. I.. … & Aerts. H. (2019). Deep Learning Predicts Lung Cancer Treatment Response From Serial Medical Imaging. Clinical Cancer Research. 11(25). 3266-3275. https://doi.org/10.1158/1078-0432.ccr-18-2495

[150] Yang. S.. Anjie. J. (2021). Recognition Of Oil and Gas Reservoir Space Based On Deep Learning. E3S Web Conf.. (267). 01038. https://doi.org/10.1051/e3sconf/202126701038

[151] Ha. S.. Lyu. I. (2022). Spharm-net: Spherical Harmonics-based Convolution For Cortical Parcellation. IEEE Trans. Med. Imaging. 10(41). 2739-2751. https://doi.org/10.1109/tmi.2022.3168670

[152] Viering. T. (2019). How To Manipulate Cnns To Make Them Lie: the Gradcam Case.. https://doi.org/10.48550/arxiv.1907.10901

[153] Shim. H.. Jung. J.. Heo. H.. Yoon. S.. Yu. H. (2018). Replay Spoofing Detection System For Automatic Speaker Verification Using Multi-task Learning Of Noise Classes.. https://doi.org/10.1109/taai.2018.00046

[154] Wu et al. (2015): . "SAS: A speaker verification spoofing database containing diverse attacks." in Proceedings of ICASSP. 2015. doi:10.1109/icassp.2015.7178810

[155] Lavrentyeva et al. (2017): . "Audio Replay Attack Detection with Deep Learning Frameworks." in Proceedings of Interspeech. 2017. doi:10.21437/interspeech.2017-360

[156] Alluri & Vuppala (2019): . "IIIT-H Spoofing Countermeasures for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019." in Proceedings of Interspeech. 2019. doi:10.21437/interspeech.2019-1623

[157] Zhao et al. (2022): . "Multi-task Learning-Based Spoofing-Robust Automatic Speaker Verification System." in Circuits Systems and Signal Processing. 2022. doi:10.1007/s00034-022-01974-z

[158] Kinnunen. T.. Sahidullah. M.. Delgado. H.. Todisco. M.. Evans. N.. Yamagishi. J.. & Lee. K. A. (2017). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.

[159] Tan. C.. Hijazi. M.. Mohamad. M.. Nohuddin. P. (2022). Artificial Speech Detection Using Image-based Features and Random Forest Classifier. IJ-AI. 1(11). 161. https://doi.org/10.11591/ijai.v11.i1.pp161-172

[160] Chettri. B.. Benetos. E.. Sturm. B. (2020). Dataset Artefacts In Anti-spoofing Systems: a Case Study On The Asvspoof 2017 Benchmark.. https://doi.org/10.48550/arxiv.2010.07913

[161] Varga. A.. Steeneken. H. J. M. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*. 12(43). 247-251. doi: 10.1016/0167-6393(93)90095-3

[162] Dean. D.. Kanagasundaram. A.. Ghaemmaghami. H.. Hafizur. M.. Sridharan. S. (2015) The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition. *Interspeech 2015*. International Speech and Communication Association. Dresden. doi: 10.21437/Interspeech.2015-685

[163] Hanilci. C.. Kinnunen. T.. Sahidullah. M.. & Sizov. A. (2016). Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. Speech Communication. 85. 83-97.