ANTI-SPOOFING USING RESNET50 WITH LINEAR DISCRIMINANT ANALYSIS FOR AUTOMATIC SPEAKER VERIFICATION

BY

PEEMAPOT UPARAKOOL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (ENGINEERING TECHNOLOGY)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2025

THAMMASAT UNIVERSITY
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THESIS

BY

PEEMAPOT UPARAKOOL

ENTITLED

ANTI-SPOOFING USING RESNET50 WITH LINEAR DISCRIMINANT ANALYSIS FOR AUTOMATIC SPEAKER VERIFICATION

was approved as partial fulfillment of the requirements for
the degree of Master of Engineering (Engineering Technology)

on June 17, 2025

Chairperson _____
(Kanokvate Tungpimolrut, Ph.D.)

Member and Advisor _____
(Associate Professor Waree Kongprawechnon, Ph.D.)

Member and Co-Advisor _____
(Jessada Karnjana, Ph.D.)

Member _____
(Assistant Professor Seksan Laitrakun, Ph.D.)

Director _____
(Associate Professor Kriengsak Panuwatwanich, Ph.D.)

Thesis Title	ANTI-SPOOFING USING RESNET50 WITH LINEAR DISCRIMINANT ANALYSIS FOR AUTOMATIC SPEAKER VERIFICATION
Author	Peemapot Uparakool
Degree	Master of Engineering (Engineering Technology)
Faculty/University	Sirindhorn International Institute of Technology/ Thammasat University
Thesis Advisor	Associate Professor Waree Kongprawechnon, Ph.D.
Thesis Co-Advisor	Jessada Karnjana, Ph.D.
Academic Years	2025

ABSTRACT

Deep-learning-based models have shown significant potential in speech spoof detection, which is crucial to ensuring the authenticity of speech signals. This work aims to expand the knowledge about deep learning-based spoof detection by integrating ResNet50 with linear discriminant analysis (LDA) to reduce the dimensionality. Using the logical access (LA) subset from the ASVspoof 2019 dataset, we generated mel-spectrogram and gammatone spectrogram representations of the speech signals. ResNet50 was used to extract deep features from these spectrograms, and subsequently LDA was applied to reduce feature dimensionality and improve classification accuracy. Our method significantly outperformed the baseline ResNet50 model by reducing the equal error rate (EER) by 43.55% and increasing balanced accuracy by 48.59% for duplicated mel-spectrogram tensor, 8.95% and 15.52% for differentiated mel-spectrogram tensor, and 44.14% and 44.77% for differentiated gammatone spectrogram tensor, respectively. These results demonstrate the effectiveness of combining ResNet50 with gammatone spectrograms and LDA, providing a more robust solution for audio spoof detection.

To further investigate our approach, we extended the evaluation by applying traditional classifiers such as Random Forest (RF), k-Nearest Neighbors (KNN), and

Naïve Bayes (NB)—on the deep features extracted by ResNet50 and reduced by LDA or PCA. Among all combinations, the LDA-reduced features paired with Naïve Bayes classifier achieved the best result, reaching 88.18% balanced accuracy and 2.80% EER. These findings confirm that our proposed framework not only improves spoof detection performance under a threshold-based scheme but is also compatible with various machine learning classifiers, making it a flexible and effective solution for audio spoof detection tasks.

Keywords: Anti-spoofing, Automatic speaker verification, Linear discriminant analysis, Principal component analysis (PCA), ResNet50.

ACKNOWLEDGEMENTS

This work is supported by Graduate Scholarship Program for Excellent Thai Students (ETS), Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU). The ASEAN IVO project Spoof Detection for Automatic Speaker Verification, was involved in the production of the contents of this presentation and was financially supported by NICT.

Peemapot Uparakool

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Abbreviations	Terms
ASV	Automatic Speaker Verification
Bal	Balance Accuracy
CQCC	Constant Q Cepstral Coefficient
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DECRO	DEepfake CROss-lingual
DF	Deepfake
DNN	Deep Neural Network
EER	Equal Error Rate
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FRR	False Rejection Rate
GMM	Gaussian Mixture Model
GT	Gammatone
KNN	K-Nearest Neighbors
LA	Logical Access
LDA	Linear Discriminant Analysis
LFCC	Linear Frequency Cepstral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
NB	Naive Bayes
PA	Physical Access
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
ResNet	Residual Networks
RNN	Recurrent Neural Network
RF	Random Forest
RFCC	Rectangular Filter Cepstral Coefficient
SENet	Squeeze-and-Excitation Network
SVM	Support Vector Machine

TTS	Text-to-speech
t-DCF	Tandem Detection Cost Function
VC	Voice Conversion
VCC	Voice Conversion Challenge

CHAPTER 1
INTRODUCTION

1.1 Background

Biometric identification is a method of verifying an individual using physiological or behavioral characteristics called biometric information, such as fingerprints, facial features, iris patterns, and voice. These characteristics are difficult to replicate and unique for each individual, making biometric systems more secure and convenient for users, which are widely used in various applications, including mobile devices, security systems, and access control environments.

In this study, we focus specifically on voice-based biometric authentication, also known as automatic speaker verification (ASV). ASV systems have been used in security applications to verify an individual's identity based on their vocal characteristics (Wu, Evans, et al., 2015). The system typically receives both the user's speech signal input and a corresponding registration ID, and then determines whether to accept or reject the user by comparing the input with the registered voiceprint, as shown in Fig. 1.1.

However, despite the growth of technology, human voices can still be easily spoofed by machines or through various types of attacks with the purpose of mimicking or impersonating legitimate users in order to bypass the security mechanisms of ASV systems (Sahidullah et al., 2019). Spoof attacks can take multiple forms, including replay attacks, voice conversion, and voice synthesis (Li et al., 2024). In these cases, an attacker can send a manipulated speech signal together with a legitimate user's registration ID to the system, deceiving it to accept those spoofed inputs, as shown in Figure 1.2. These attacks represent one of the most critical challenges in the field of ASV and are the primary focus of this research.

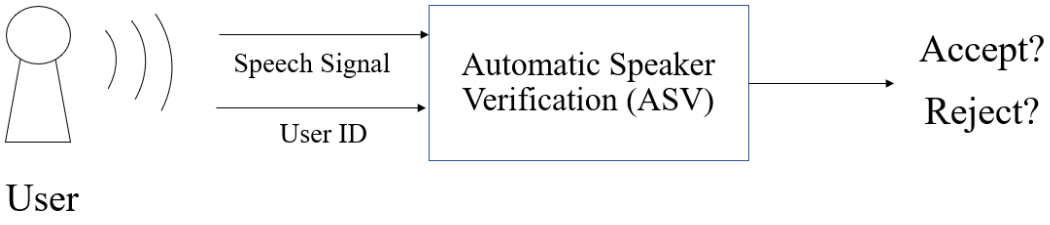

Figure 1.1 Diagram for automatic speaker verification system.

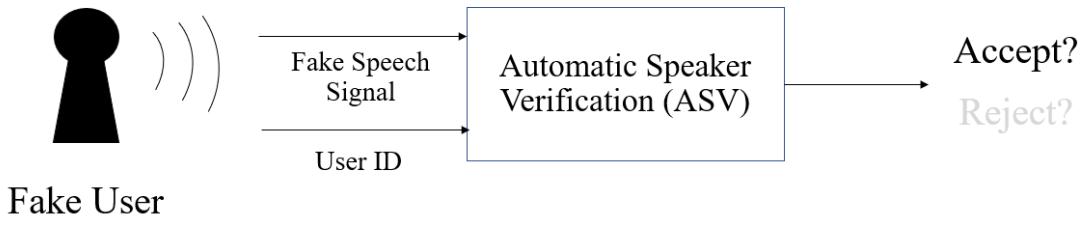

Figure 1.2 Diagram for automatic speaker verification spoofing.

However, despite the growth of technology, human voices can still be easily spoofed by machines or through various types of attacks with the purpose of mimicking or impersonating legitimate users in order to bypass the security mechanisms of ASV systems (Sahidullah et al., 2019). Spoof attacks can take multiple forms, including replay attacks, voice conversion, and voice synthesis (Li et al., 2024). In these cases, an attacker can send a manipulated speech signal together with a legitimate user's registration ID to the system, deceiving it to accept those spoofed inputs, as shown in Figure 1.2. These attacks represent one of the most critical challenges in the field of ASV and are the primary focus of this research.

Spoofing attacks in ASV systems can be broadly categorized into two main types: Physical Access (PA) and Logical Access (LA). Physical Access attacks involve the use of recorded or replayed speech. For example, an attacker might use a device to play back a genuine recording of a speaker's voice to gain unauthorized access. These attacks exploit the physical transmission channel between the speaker and the microphone. Logical Access attacks, on the other hand, are more advanced and are carried out using artificially generated speech. This includes voice conversion (VC), where an attacker modifies their voice to sound like someone else, and text-to-speech (TTS) or speech synthesis, where a machine generates an entirely fake speech signal using AI-based models trained to imitate a target speaker's voice.

In this study, we focus specifically on Logical Access (LA) attacks, which present greater challenges due to their ability to generate high-quality synthetic voices that are nearly indistinguishable from human speech. The increasing availability and sophistication of speech synthesis technologies have made LA attacks more prevalent and difficult to detect, thus posing a serious threat to ASV systems. The ASVspoof 2019 Logical Access (LA) dataset, which simulates such attacks using a variety of TTS and VC systems, will be used to evaluate and validate our proposed spoof detection framework.

In response to these concerns, numerous studies and efforts have been undertaken

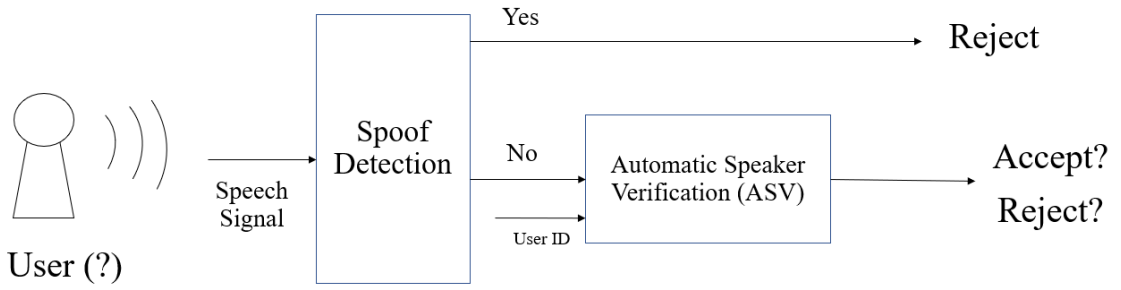

Figure 1.3 Diagram for automatic speaker verification with spoof detection.

to develop anti-spoofing systems, particularly through initiatives such as the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenge, which began in 2015 (Wu et al., 2017). This biennial challenge aims to advance research and promote the development of more methods to detect and mitigate spoof attacks, improving the security and reliability of ASV systems (Wu, Kinnunen, et al., 2015). To support these goals, the challenge also provides standardized datasets that cover a broad range of spoofing techniques. These datasets are periodically updated and released under two main tracks, Physical Access (PA) and Logical Access (LA). The PA dataset typically involves replay attacks captured through real-world recording and playback devices, while the LA dataset focuses on synthetic and voice conversion attacks generated using speech synthesis and modification algorithms. Over time, these datasets have become key benchmarks in the field, enabling fair comparisons across systems and facilitating the advancement of anti-spoofing research.

Based on the ASVspoof challenge, various spoofing countermeasure techniques have been proposed and developed. One of the commonly used methods is a spoof detection system, which acts as a front-end component between the user and the core ASV system, as shown in Figure 1.3. The purpose of this system is to detect whether an input speech signal is spoofed. If the signal is classified as spoofed, it will be immediately rejected; otherwise, it will be forwarded to the ASV system for further authentication. This additional detection step significantly enhances the overall security of ASV systems by filtering out suspicious inputs and reducing the risk of successful spoofing attempts.

Generally, spoof detection consists of two parts: feature extraction and classification (BT et al., 2019). Current methodologies for feature extraction are categorized into three groups, which are hand-crafted spectral features, deep-learning features, and other analysis-oriented approaches (Li et al., 2024). In many works, hand-crafted spectral features extraction method have been used, such as linear frequency cepstral coefficient (LFCC)

(Zhou et al., 2011), mel-frequency cepstral coefficient (MFCC) (Davis & Mermelstein, 1980a), rectangular filter cepstral coefficient (RFCC) (Hasan et al., 2013), and constant-Q cepstral coefficient (CQCC) (Todisco et al., 2017). These features are particularly effective in capturing spectral and temporal patterns associated with both bona fide and spoofed speech signals, and are often used as a baseline model in recent anti-spoofing studies (Liu et al., 2023).

In contrast, deep learning-based approaches such as deep neural networks (DNN) (BT et al., 2019), residual networks (ResNet) (Chakravarty & Dua, 2024), and recurrent neural networks (RNN) (Khan et al., 2023) have also been widely explored for spoof detection tasks. Compared to hand-crafted spectral features, deep learning models are capable of automatically learning complex, discriminative features relevant to spoofing, often leading to superior performance. While the use of deep learning-based methods in this domain is relatively recent, the number of studies continues to grow as research advances (Das et al., 2019; Zhiqiang et al., 2022).

In addition to these methods, analysis-oriented approaches have also been investigated. These include techniques such as analyzing the interaction between vocal folds and the vocal tract (Blue et al., 2022), exploring the role of silence in speech (Zhang et al., 2023), and examining energy loss in pauses between words (Deng et al., 2022). These approaches aim to capture subtle acoustic anomalies that distinguish spoofed speech from genuine speech. By incorporating insights from speech production and prosodic behavior, they provide complementary information that can enhance the robustness and accuracy of spoof detection systems.

As spoofing techniques continue to advance, models based on hand-crafted features, which rely on fixed spectral patterns, are increasingly vulnerable to being deceived. Deep learning-based approaches offer the flexibility to automatically learn discriminative features that adapt to various spoofing scenarios. Therefore, in this study, we focus on a deep learning-based model for spoof detection. We adopt the residual network (ResNet) architecture, which is originally proposed for image recognition tasks, due to its strong performance in extracting hierarchical features and its ability to mitigate the vanishing gradient problem in deep networks. ResNet has demonstrated success in many speech-related applications by treating time-frequency representations, such as spectrograms. In our work, we modified the ResNet50 model and integrated it into the feature extraction process, comparing the results with those obtained from the original ResNet50 architecture. This comparison enabled us to assess the effectiveness of our modifications in enhancing the performance of the spoof detection system.

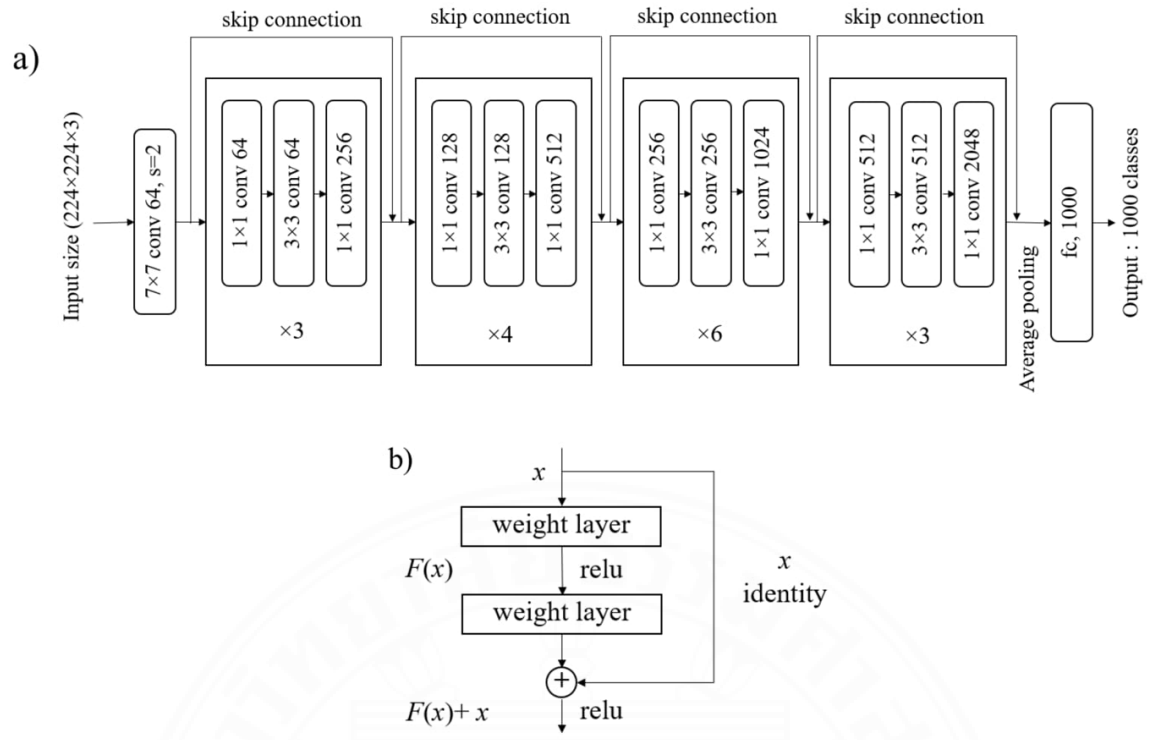

Figure 1.4 (a) ResNet50 architecture and (b) an example of skip connection.

Residual networks (ResNet) are classic neural network models mostly used in computer vision tasks, which were first introduced by He et al. (2015). There are many types of ResNet model based on the number of layers, e.g., ResNet18, ResNet34, ResNet50, where the number indicates the total number of layers in the network. Each variant of the model offers a different trade-off between complexity and performance. In our work, we select the ResNet50 model, which contains 50 layers, because it strikes an effective balance between depth and computational efficiency (Shin et al., 2021). It is sufficiently deep to capture complex patterns and extract high-level features from the data, making it optimal for our experiment.

According to Figure 1.4 (a), the ResNet50 architecture is composed of the following components: convolutional layers, residual blocks, and fully connected layer. ResNet50 consists of a series of residual blocks, each consisting of three convolutional layers of specific sizes, with batch normalization and ReLU activation. These blocks implement skip connections, where the original input is added to the output of the convolutional blocks as shown in Figure 1.4 (b), helping mitigate the vanishing gradient problem and improve training stability (Adnan et al., 2023). After passing through those residual blocks, the feature maps are processed by a global average pooling, and then fed into the fully connected layer, which outputs the class predictions. In the default model, the number of classes is 1,000, but this can be adapted for specific use cases, such as binary classification or other

multiclass problems, making ResNet50 highly versatile for various tasks.

In order to provide suitable inputs for the ResNet model, two-dimensional time-frequency features are required. In this work, we employed mel-spectrograms and gammatone spectrograms extracted from speech signals. The details of both features are as follows:

Mel-spectrogram is a temporal-frequency representation that maps speech signals to the mel scale, which approximates human auditory perception by focusing on lower frequencies (Lambamo et al., 2023). This transformation captures perceptually relevant features of the signal, making the mel-spectrogram one of the popular choices in speech and audio processing tasks (Tak et al., 2017).

As shown in Figure 1.5 (a), mel-spectrograms are created by first breaking down an speech signal into short, overlapping frames. This process is called framing and windowing. Then, a fast Fourier transform (FFT) is applied to convert each frame from the time domain to the frequency domain. The magnitude specrum is then mapped to the mel scale by the mel filterbank, which is better aligned with human hearing perception. The frequency response of the mel filterbank is shown in Figure 1.5 (b). Finally, the result is converted to a decibel scale to improve interpretability. An example of a mel-spectrogram, which is in the form of a two-dimensional image, as shown in Figure 1.5 (c).

Mel-spectrograms were generated using the `Librosa` library in Python. The conversion to the mel scale is achieved by applying the mel filterbank $H_m(f)$ to the magnitude spectrum of each frame:

$$S_{\text{mel}}(m,t) = \sum_{f} |X(f,t)|^2 \cdot H_m(f), \tag{1.1}$$

where f is the frequency components of the audio signal in each frame, t is the time indices corresponding to that frame, $S_{\text{mel}}(m,t)$ is the mel-spectrogram, $X(f,t)$ is the magnitude spectrum, and $H_m(f)$ is the mel filterbank response for the m-th filter.

Gammatone spectrogram is also a temporal-frequency representation used in speech processing similar to mel-spectrogram, but it is based on the gammatone filterbank, which models the auditory filters of the human cochlea more closely than the mel filterbank. The gammatone spectrogram emphasizes frequency components in a way that mimics human auditory perception, making it useful for capturing perceptually relevant features in audio signals.

As shown in Figure 1.6 (a), the speech signal will be windowed in a specific time.

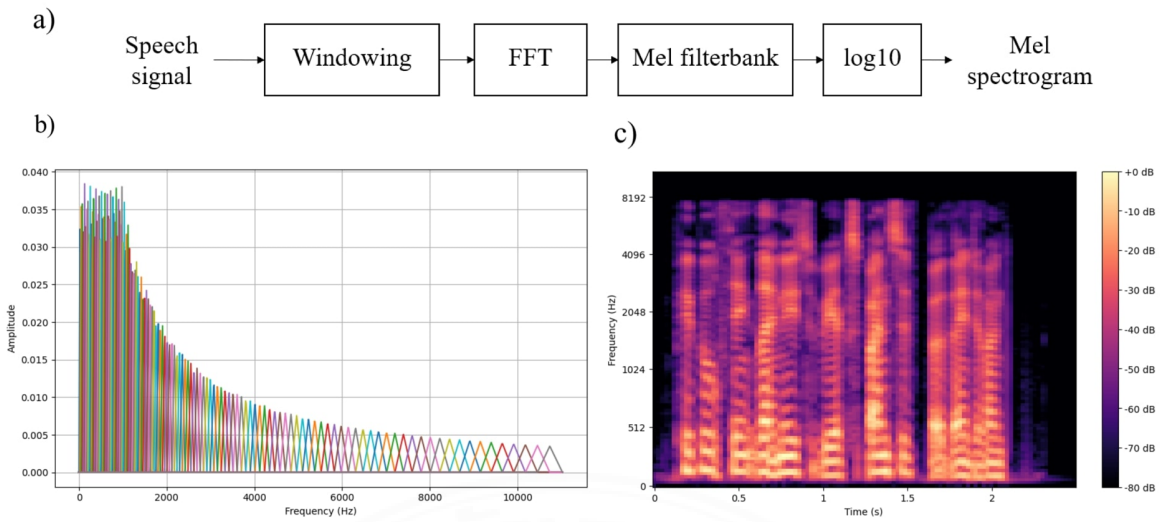

Figure 1.5 (a) Block diagram of mel-spectrogram extraction, (b) frequency response of mel-spectrogram filterbank, and (c) an example of mel-spectrogram.

Then, it will be passed through the fast Fourier transform (FFT) to convert to the frequency domain. The filter frequency response is shown in Figure 1.6 (b). Then, it will be processed through a gammatone filterbank. Key parameters, such as the number of channels, window length, and hop length, are set to control the temporal and frequency resolution of the analysis. The output is a spectrogram that represents the energy or magnitude of each frequency band over time, capturing how different parts of the frequency spectrum change. The spectrogram will be converted to a decibel scale to improve visualization and highlight subtle variations in the audio signal. An example of a gammatone-based spectrogram shown in Figure 1.6 (c).

In our work, we utilized the gammatone spectrogram as a feature representation. The gammatone filter is defined as:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \tag{1.2}$$

where a is the amplitude, n is the filter order, b is the bandwidth, f_c is the center frequency, and ϕ is the phase.

Another important component in the spoof detection pipeline is dimensionality reduction, which is often applied after feature extraction to reduce the complexity of the input and enhance class separability. In this study, we employed Linear discriminant analysis (LDA) as the main dimensionality reduction technique. LDA is a statistical technique used to reduce the number of dimensions in data and improve classification (Sorzano et al., 2014). Its goal is to project the data onto a lower-dimensional space that enhances the separability

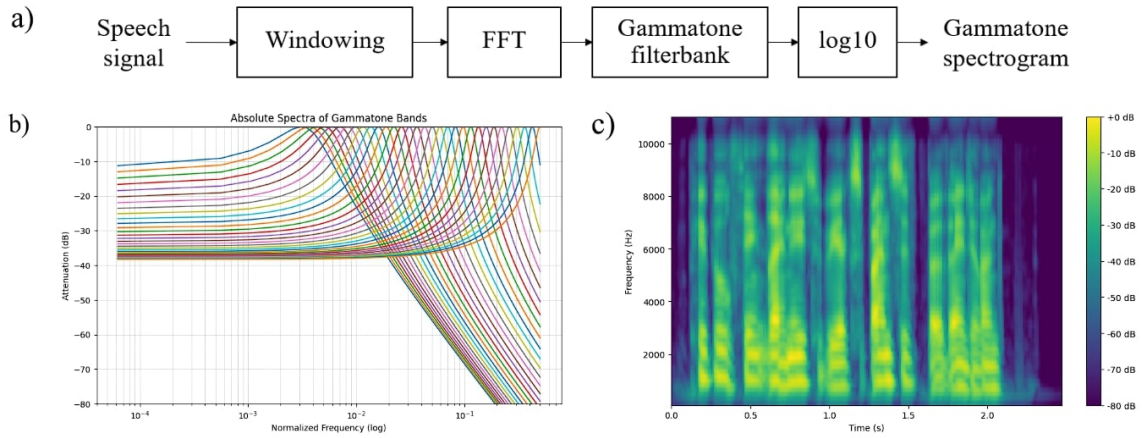

Figure 1.6 (a) Block diagram of gammatone extraction, (b) frequency response of gammatone spectrogram filterbank, and (c) an example of gammatone spectrogram.

between different classes. Unlike principal component analysis (PCA), which focuses on maximizing variance without considering class labels, LDA takes class information into account, making it particularly effective for classification problems. The resulting transformation seeks to maximize the ratio of the between-class variance to the within-class variance, ensuring that the classes are as distinct as possible in the reduced-dimensional space.

Finally, we used a simple decision-based classifier to determine whether the input signal is bona fide or spoofed. The output from the dimensionality reduction process will be compared against a decision threshold, which was optimized using the training data to minimize classification error. If the projected value exceeded the threshold, the input was classified as bona fide, otherwise, it was considered spoofed. This threshold-based approach allows for a lightweight and interpretable classification stage, which is particularly useful for evaluating the effectiveness of different feature extraction and dimensionality reduction strategies within the spoof detection pipeline.

1.2 Problem statement

Despite recent advances in automatic speaker verification (ASV) systems, they remain vulnerable to prevent various types of spoofing attack, such as voice conversion and speech synthesis. These attacks can deceive ASV systems into accepting faked speech signals as genuine, posing serious security threats in applications that rely on voice-based authentication. While traditional countermeasures based on hand-crafted features have shown some practical result, they often struggle to generalize across advanced spoofing methods due to their limited adaptability.

Deep learning-based models, particularly convolutional neural networks such as ResNet50, offer greater flexibility by automatically learning discriminative features from time-frequency representations. However, the effectiveness of these models depends on the choice of input features, preprocessing techniques, and the complexity of the extracted features. High-dimensional feature spaces can lead to overfitting or poor generalization. Therefore, an efficient mechanism to reduce dimensionality while preserving class separability is crucial for enhancing spoof detection performance.

1.3 Objectives

Our objective of this research is to propose a new method to improve the performance of the spoof detection system and to expand the knowledge of deep-learning-based spoof detection methodologies using ResNet50 based model, exploring the impact of different spectrogram-based inputs which are mel-spectrogram and gammatone, and investigated the role of the dimensionality reduction methods, e.g., linear discriminant analysis (LDA) and Principal Component Analysis (PCA). Furthermore, we propose and evaluate the effect of different input tensor types, using padding and resized tensor, along with the duplication and differentiation along both direction techniques, and determine which configuration offers better discriminative performance. Through these investigations, we aim to identify an effective and lightweight approach that enhances the reliability of spoof detection systems.

1.4 Contribution and impact of our research

This research contributes to the field of voice anti-spoofing by proposing a novel framework that integrates the deep feature extraction capabilities of ResNet50 with the dimensionality reduction strengths of Linear Discriminant Analysis (LDA). This combination not only enhances class separability but also improves the overall performance in detecting spoofed speech.

To further investigate the influence of input representation, we conduct a systematic comparison between two widely used time-frequency representations which are mel-spectrograms and gammatone spectrograms, as inputs to the ResNet50 model. This comparative analysis provides empirical insights into the strengths and limitations of each representation in the context of spoof detection.

In addition, we explore and evaluate multiple tensor preparation strategies, including duplication and differentiation along the time and frequency axes, as well as resizing and zero-padding approaches. These preprocessing techniques were shown to significantly influence model performance and are therefore critical considerations for system design.

Our proposed approach is rigorously evaluated on the Logical Access (LA) subset of the ASVspoof 2019 dataset, a widely recognized benchmark in the field. Experimental results demonstrate substantial improvements over baseline systems, reflected in reduced equal error rates (EER) and increased balanced accuracy. Furthermore, the use of a simple threshold-based classifier following the LDA stage allows the system to remain computationally efficient and lightweight, making it practical for real-time or resource-constrained environments.

1.5 Thesis Structure

This thesis explains previous work and related studies on topics similar to our research, including the literature review in Chapter 2. Next, we present the overview of our work, proposed methodology, a reference model, and the experimental setting in Chapter 3. Then, we show evaluation result and discussion in Chapter 4. Finally, we conclude everything we have learned from this research and mention our future work direction in Chapter 5.

CHAPTER 2
LITERATURE REVIEW

This section presents a review of recent literature relevant to spoof detection in automatic speaker verification (ASV) systems, including the ASVspoof countermeasure competition, which provides benchmarking datasets and evaluation protocols. In addition, we review various feature extraction techniques, focusing particularly on mel-spectrogram and gammatone spectrogram representations. Furthermore, we explore the Residual Network (ResNet50) which is a main model we used in our works, and examine the use of dimensionality reduction methods such as Linear Discriminant Analysis (LDA), which are key components in our proposed framework.

2.1 ASVspoof countermeasure competition and dataset

As mentioned in Section 1, ASVspoof countermeasure competition was launched to address the lack of standardized evaluation protocols in the study of spoofing and countermeasures for automatic speaker verification (ASV).

2.1.1 ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan (Wu, Kinnunen, et al., 2015)

This paper focused on spoofing detection systems and aimed to promote the development of generalized countermeasures that could detect various spoofing attacks without relying on prior knowledge of specific attack algorithms. A key objective of the challenge was to provide a level playing field by releasing a standard dataset, consisting of both genuine and spoofed speech generated using multiple voice conversion and speech synthesis techniques, along with a clearly defined evaluation protocol. The evaluation metric used was the Equal Error Rate (EER), and the challenge included training, development, and evaluation subsets with increasing diversity, particularly in the spoofing methods used. ASVspoof 2015 played a critical role in shaping research direction by introducing a benchmarking framework that continues to influence subsequent works in the field.

The standard dataset provided by this paper is divided into bona fide and spoofed speech. Bona fide speech is collected from 106 speakers (45 male, 61 female) and with no significant channel or background noise effects. Spoofed speech is generated from the

Table 2.1 Number of non-overlapping target speakers and utterances in the training, development and evaluation sets in ASVspoof2015 dataset.

Subset	Speakers		Utterances	
	Male	Female	Bona fide	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

genuine data using a number of different spoofing algorithms. The full dataset is partitioned into three subsets, the first for training, the second for development and the third for evaluation. The number of speakers in each subset is illustrated in Table 2.1. There is no speaker overlap across the three subsets regarding target speakers used in voice conversion or text-to-speech (TTS) adaptation.

2.1.2 ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan (Kinnunen et al., 2017)

The ASVspoof 2017 challenge was introduced as a follow-up to the 2015 edition, aiming to address several limitations of the previous dataset. While ASVspoof 2015 focused solely on synthetic spoofing methods such as text-to-speech (TTS) and voice conversion (VC), ASVspoof 2017 shifted its focus entirely to replay attacks, which are more accessible to attackers and realistic in practical scenarios. The 2017 dataset was built upon the RedDots corpus (Lee et al., 2015) and included both bona fide and replayed utterances recorded across a wide variety of playback and recording devices. The replay data was collected through a crowd-sourcing effort to simulate diverse in-the-wild conditions.

The dataset was divided into training, development, and evaluation subsets, with no overlap in replay configurations between them. Unlike the 2015 edition, which averaged equal error rates (EER) per spoofing method, ASVspoof2017 used pooled EER across all trials to encourage the development of generalizable countermeasures. This shift marked an important step toward more realistic and robust evaluation settings for spoofing detection systems.

2.1.3 ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection (Consortium, 2019)

The ASVspoof 2019 dataset represents a significant advancement from previous editions by incorporating all three major spoofing attack types: text-to-speech (TTS), voice conversion (VC), and replay attacks. It is the first challenge to address both logical access

Table 2.2 Dataset statistical information in the training, development and evaluation sets in Logical Access (LA) and Physical Access (PA).

Subset	Speakers		Logical Access (LA)		Physical Access (PA)	
	Male	Female	Bona fide	Spoofed	Bona fide	Spoofed
Training	8	12	2580	22800	5400	48600
Development	4	6	2548	22296	5400	24300
Evaluation	20	26	7355	63882	8000	135000

(LA) and physical access (PA) scenarios within a unified evaluation framework. Unlike ASVspoof 2015, which focused only on synthetic speech, and ASVspoof 2017, which emphasized replay attacks in uncontrolled environments, the 2019 edition expanded the threat model and improved the simulation of acoustic conditions for greater realism and diversity.

The LA subset includes bona fide speech and spoofed speech generated from 17 different TTS and VC systems. Six systems are used for training and development as known attacks, while the evaluation set introduces 11 unknown attacks constructed with cutting-edge neural vocoding and waveform synthesis techniques. The PA subset, on the other hand, simulates replay attacks in various room and device conditions. It includes 27 different acoustic configurations (e.g., room size, reverberation, talker-to-mic distance) and 9 replay configurations (e.g., loudspeaker quality, attacker-to-talker distance), making it more controlled and diverse than its 2017 counterpart.

The statistical information of this dataset in both logical access (LA) and physical access (PA) as shown in Table 2.2. In our work, we used the Logical Access (LA) portion of the ASVspoof 2019 dataset for training, validation, and evaluation (Liu et al., 2022). The reason for choosing the LA subset is that it focuses on synthetic spoofing attacks, such as TTS and VC, which are more relevant to our research objective of evaluating the effectiveness of deep learning-based models in detecting machine-generated speech. Furthermore, LA provides a diverse range of attack algorithms, including both known and unknown systems, making it suitable for assessing the generalization capability of our proposed model.

2.1.4 ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild (Liu et al., 2023)

Unlike previous versions that relied heavily on clean and controlled data, the 2021 edition focused on channel robustness and real-world distortions, including compression, telephony transmission, and social media-style deepfakes. In addition to the existing Logical

Table 2.3 Statistics of ASVspoof 2021 dataset across all three tasks: Logical Access (LA), Physical Access (PA), and Deepfake (DF).

Task	Subset	Bona fide	Spoofed	Female	Male
Logical Access (LA)	Progress	1,676	14,788	37	30
	Evaluation	14,816	133,360	37	30
Physical Access (PA)	Progress	14,472	72,576	37	30
	Evaluation	94,068	627,264	37	30
Deepfake (DF)	Progress	5,768	53,557	37	30
	Evaluation	14,869	519,059	50	43

Access (LA) and Physical Access (PA) tracks, this edition introduced a new Deepfake (DF) task, which aims to detect manipulated speech data compressed by various codecs and shared across online platforms. The DF track also included data from previously unseen corpora (e.g., VCC 2018 and VCC 2020), making the detection task more complex and domain-diverse.

According to Table 2.3, the ASVspoof 2021 dataset includes a total of 16 subsets across three tasks: Logical Access (LA), Physical Access (PA), and Deepfake (DF). The LA track contains 1,676 bona fide and 14,788 spoofed utterances in the progress set, and 14,816 bona fide and 133,360 spoofed utterances in the evaluation set. The PA track includes 14,472 bona fide and 72,576 spoofed utterances for progress, and 94,068 bona fide and 627,264 spoofed utterances for evaluation. The DF track provides 5,768 bona fide and 53,557 spoofed utterances in the progress set, and 14,869 bona fide and 519,059 spoofed utterances in the evaluation set. All subsets are gender-balanced, with 67 speakers (37 female, 30 male) for LA and PA progress sets, and 93 speakers (50 female, 43 male) for the DF evaluation set.

Despite its realistic conditions and broader scope, we chose to use the ASVspoof 2019 LA dataset in our study. This decision was driven by three primary reasons. First, the 2019 dataset provides clean and controlled conditions suitable for evaluating the core spoof detection performance of our deep learning-based model. Second, the training protocol of the 2021 challenge still relied on the 2019 LA subset, with no newly matched training data provided for its LA or DF tasks, which may result in overfitting or domain mismatch when evaluated on heavily compressed or unseen sources. Third, the LA dataset in ASVspoof 2019 includes a well-defined set of both known and unknown attacks from various VC and TTS systems, offering a balanced and interpretable benchmark for evaluating generalization performance under controlled in varied spoofing scenarios.

2.2 Feature Extraction for Spoof Detection

According to BT et al. (2019), features that used in spoof detection system were categorized into three main types: (1) hand-crafted spectral features, (2) deep-learning-based features, and (3) other analysis-oriented inspired approaches. Each subsection provides an overview of the corresponding technique and discusses previous works that have applied these features to the spoof detection task.

2.2.1 Hand-crafted Spectral Features

Hand-crafted spectral features were popular in spoof detection in ASV system. These methods rely on expert knowledge to design features, often based on spectral characteristics, that aim to capture distinctions between bona fide and spoofed speech. They are typically combined with classifiers like Gaussian Mixture Models (GMMs) or some simple neural networks.

2.2.1.1 Constant Q Cepstral Coefficients: A Robust Descriptor for General Audio Signals with Applications in Verification and Spoofing (Todisco et al., 2017)

This work introduced Constant Q Cepstral Coefficients (CQCCs) as a powerful alternative to MFCCs and LFCCs for anti-spoofing. CQCCs utilize a perceptually motivated Constant Q transform, providing geometrically spaced frequency bins, which proves effective in capturing artifacts across different spoofing types, especially those generated by synthesis and conversion algorithms. When combined with a GMM backend, CQCC-GMM systems demonstrated state-of-the-art performance on the ASVspoof 2015 dataset and became a strong baseline in subsequent challenges like ASVspoof 2017 and 2019, highlighting the robustness of well-designed hand-crafted features.

2.2.1.2 Spoofing Detection Goes Noisy: An Analysis of Synthetic Speech Detection in the Presence of Background Noise (Hanilci et al., 2016)

: This study investigates the impact of noise on spoofing detection using hand-crafted features like MFCC, LFCC, and CQCC with GMM classifiers. It demonstrated that while these features perform well in clean conditions, their robustness degrades significantly in noisy environments. The analysis revealed that CQCCs generally offered better robustness compared to MFCCs and LFCCs under various noise types and levels, likely due to the Constant Q transform's properties. This work highlights the challenges faced by hand-crafted features in realistic, noisy conditions and the importance of feature robustness.

While hand-crafted spectral features might seem relatively traditional compared to

deep learning approaches, they continue to prove themselves as solid baselines in spoof detection research. Techniques like LFCC and CQCC, when paired with simple classifiers such as GMMs, still competitive and usable in spoof detection.

2.2.2 Deep-learning-based Features

As technology keeps advancing, spoofing techniques are also getting better at copying not just the surface of human speech, but also deeper patterns like spectral and temporal details. Because of that, hand-crafted features alone sometimes aren't enough to catch these attacks. So, deep learning-based features were introduced to help detect spoofing more effectively, especially when the attack method is something the system hasn't seen before. These methods let the model learn useful features by itself, without needing to be manually designed. But the trade-off is that they usually require more time and computational power compared to using hand-crafted features.

Here are some examples of recent works that used deep learning based features to solve the problem:

2.2.2.1 An Empirical Study on Channel Effects for Synthetic Speech Detection (Zhang et al., 2021)

This paper explores the impact of channel variability on deep learning-based spoofing detection, primarily using light convolutional neural network (LCNN) architectures, which were highly successful in the ASVspoof 2019 challenge. The study showed that while powerful models like LCNNs achieve excellent performance, they are sensitive to channel mismatches between training and testing data. Techniques like channel-robust feature normalization (e.g., cepstral mean and variance normalization) and multi-condition training were investigated to improve the generalization capability of these deep learning systems across diverse acoustic channels.

2.2.2.2 RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies (Kim et al., 2022)

Architectures like RawNet operate directly on raw audio waveforms, bypassing traditional hand-crafted feature extraction. These deep 1D Convolutional Neural Networks based on ResNet-like blocks learn relevant filters directly from the waveform. Adapted for anti-spoofing model, these models have shown competitive performance, potentially capturing fine-grained phase and artifact information lost during spectral feature extraction.

Their success highlights the power of end-to-end deep learning in discovering discriminative patterns directly from the signal.

2.2.3 Other Analysis-oriented Features

This category includes methods focusing on aspects beyond standard spectral envelope features, such as characteristics of the vocal source, phase information, background noise analysis, or acoustic channel properties that might be altered by spoofing processes.

Here are some examples of work that use this kind of method:

2.2.3.1 Phase-Aware Features Based on Group Delay for Replay Spoofing Detection

This paper explored group delay–based analysis to detect replay attacks by focusing on high-frequency phase distortions and artifacts caused by recording and replay devices. The phase-sensitive features outperformed MFCC and CQCC under noisy and reverberant test conditions in the ASVspoof 2017 and 2019 PA tasks. The results emphasized that phase cues offer additional robustness, especially when replay conditions vary in the wild.

2.2.3.2 Modified Magnitude-Phase Spectrum Information for Spoofing Detection (Yang et al., 2021)

This work introduced modified magnitude-phase spectrum (MMPS), a joint spectral and phase feature derived using the constant-Q transform. MMPS captures phase artifacts and high-resolution spectral structure that may indicate voice conversion or synthesis. Despite relying solely on hand-crafted features, their system achieved competitive results with deep-learning-based models on the ASVspoof 2019 LA dataset, showcasing the potential of phase information in distinguishing spoofed speech.

2.2.3.3 Glottal Source Processing: from Analysis to Applications (Drugman et al., 2019)

This study proposed a detection system based on glottal flow parameter estimation from inverse filtering. The authors extracted glottal source features and combined them with spectral features like MFCC and CQCC. On the ASVspoof 2019 LA dataset, this fusion of glottal and spectral achieved an EER of 2.39%, indicating improved discriminative power over spectral features alone, particularly in neural vocoder scenarios.

2.3 Temporal-frequency Representation

Temporal-frequency representation is a way to transform a speech signal into a two-dimensional format that captures both time and frequency information, making it suitable for input to convolutional neural networks.

In our work, we focus on mel-spectrogram and gammatone spectrogram which are widely used because they preserve important acoustic characteristics while providing a visually structured input for deep learning models. Here are some related works that used those representation for spoof detection task.

This following section provides a brief explanation of two time-frequency representations used in this study, namely mel-spectrogram and gammatone spectrogram.

2.3.1 Mel-spectrogram

The mel-spectrogram is a time-frequency representation that maps the linear frequency scale into the mel scale, which is designed to match the nonlinear frequency sensitivity of human hearing. This concept was first introduced by (Stevens et al., 1937) through perceptual studies of pitch scaling, and later adopted in practical speech processing systems by (Davis & Mermelstein, 1980b) through the introduction of the mel filterbank in the MFCC framework. The mel scale approximates human auditory perception by emphasizing lower-frequency details, which are more perceptually relevant in speech.

2.3.2 Gammatone spectrogram

The gammatone spectrogram was developed to more closely to the filtering behavior of the human cochlea by using a filterbank based on gammatone filters, first proposed by (Patterson et al., 1988). These filters are characterized by impulse responses shaped like gamma distribution functions modulated by sinusoids, reflecting the auditory system's response to incoming sound. Holdsworth *et al.*(1991) later provided an efficient implementation of the gammatone filterbank, making it feasible for audio processing applications.

2.4 Residual Network (ResNet) model

Residual Network (ResNet) is a convolutional neural network architecture that introduces shortcut connections to help gradients flow through deep layers more effectively. While it is commonly used in image recognition tasks, its ability to learn complex feature representations without performance degradation makes it adaptable to other domains as

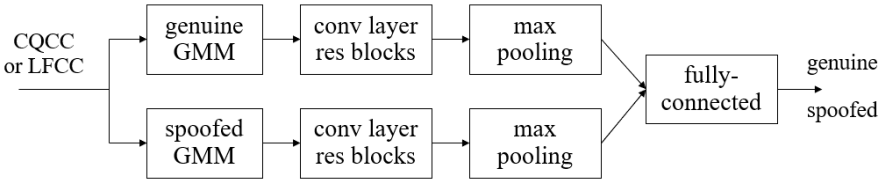

Figure 2.1 Diagram of the two-path GMM-ResNet or GMM-SENet.

well. In spoof detection, ResNet can be used in both the feature extraction and classification stages, depending on the overall system design, which makes it a flexible and effective backbone for deep learning–based countermeasure systems. Because of these advantages, ResNet has been chosen for further investigation in this research. The following works are examples of how ResNet has been applied in ASV spoofing tasks.

2.4.1 Two-path GMM-ResNet and GMM-SeNet for ASV Spoofing detection (Lei et al., 2022)

This paper proposed a novel spoof detection framework that integrates classical Gaussian Mixture Models (GMMs) with deep neural architectures, specifically ResNet and SENet, to overcome the limitations of conventional GMM-based countermeasures. In traditional systems, GMMs treat each speech frame independently, ignoring the correlation between adjacent frames and the contribution of individual Gaussian components. To address this, the authors introduced the GMM-ResNet and GMM-SENet models, where the input features are log Gaussian probabilities derived from two GMMs trained separately on bona fide and spoofed speech. These features are then passed through convolutional layers and six residual blocks (with or without squeeze-excitation units), followed by max pooling and a fully connected layer for classification. To further enhance robustness, a two-step training scheme is applied: first training the convolutional layers with softmax outputs, then freezing them and training the final classifier.

According to Figure 2.1, the two-path architecture separates the input CQCC or LFCC features into two branches processed by GMMs trained on bona fide and spoofed speech, respectively. Each GMM outputs log-probability features, which are then passed through shared convolutional layers and residual blocks in each path. After max pooling is applied to reduce spatial dimensions, the resulting feature maps from both paths are concatenated and passed through a fully connected layer to make the final classification between genuine and spoofed speech. This structure allows the model to learn from both distributions separately while preserving discriminative cues from each class.

Experimental evaluations were conducted on the ASVspoof 2019 dataset under both Logical Access (LA) and Physical Access (PA) scenarios. The results demonstrate that the proposed systems significantly outperform the baseline GMM and ResNet-only models. Specifically, the LFCC+GMM-ResNet (2P2S) system achieved an EER of 1.80% in the LA evaluation set—representing a 76.3% relative improvement over the baseline GMM. In the PA condition, the LFCC+GMM-SENet (2P2S) model achieved an EER of just 0.59%, outperforming several state-of-the-art single-system baselines. The authors also performed score fusion across multiple subsystems, achieving competitive performance close to the top-ranked systems in the ASVspoof 2019 challenge. These findings suggest that incorporating GMM-derived probability features with deep ResNet-based networks can offer both interpretable structure and high accuracy in detecting spoofed speech.

2.4.2 A lightweight feature extraction technique for deepfake audio detection (Chakravarty & Dua, 2024)

This work is one of the main reference that give us an idea to do the further research about ResNet model for ASV spoof detection. Chakravarty N. *et al.* proposed a lightweight audio deepfake detection system that leverages mel-spectrograms as input and employs a modified ResNet50 architecture for deep feature extraction, alongside with the Linear Discriminant Analysis (LDA) for dimensionality reduction, reducing each sample to a one-dimensional discriminative feature, then they used to train several traditional machine learning classifiers, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Naive Bayes (NB).

The system is trained on the ASVspoof2019 LA dataset and evaluated using the deepfake partition of ASVspoof2021, as well as a noisy unseen dataset like deepfake cross-lingual (DECRO) to assess robustness. Among various configurations, the combination of ResNet50 and LDA followed by RF classification achieved the best performance, yielding an Equal Error Rate (EER) of just 0.4% and an accuracy of 99.7%. These results demonstrate the effectiveness of combining deep CNN-based feature extraction with traditional machine learning classifiers for detecting audio spoofing attacks, particularly when using compact and discriminative features optimized through LDA.

2.4.3 Spoof Detection using Voice Contribution on LFCC features and ResNet34 (Mon et al., 2023)

This paper proposed a spoof detection method combining Linear Frequency Cepstral Coefficients (LFCCs) and a ResNet-34 model to identify various spoofing attacks, including replay, speech synthesis, and voice conversion, within automatic speaker verification systems. Rather than using full utterances alone, the study explores the impact of extracting LFCC features from specific portions of speech, such as the initial silence and fixed percentages from both the beginning (head) and end (tail) of each utterance. This segmented input aims to highlight distinguishing patterns between genuine and spoofed speech that are often found in the margins of the signal. These features are passed through a ResNet-34 architecture, which was selected for its proven effectiveness in extracting hierarchical representations for classification tasks. The model is trained on the ASVspoof 2019 dataset and evaluated on both Physical Access (PA) and Logical Access (LA) conditions.

The experimental results demonstrate that the proposed method significantly outperforms traditional LFCC-GMM and CQCC-GMM baselines. For the PA scenario, the lowest Equal Error Rates (EERs) achieved were 3.11% on the development set and 3.49% on the evaluation set using 15% of both head and tail segments. For the LA condition, the best performance was obtained with 40% of the segmented voice, yielding 0.16% EER on the development set and 6.89% on the evaluation set. These results, along with high accuracy and F1 scores, suggest that focusing on key voice segments improves model sensitivity to spoofed patterns. The study also highlights that this segment-based feature extraction helps optimize computation without sacrificing performance. Overall, the combination of LFCC and ResNet34, particularly when coupled with voice segment analysis, proves to be a promising direction for robust and efficient spoof detection.

2.4.4 Replay Attack Detection in Automatic Speaker Verification Using GTCCs and ResNet-based Model (Chaiwongyen et al., 2022)

This paper propose a replay attack countermeasure for speaker verification using biologically inspired Gammatone Cepstral Coefficients (GTCCs) as features and a deep ResNet-based classifier. GTCCs are an auditory filterbank variant of MFCC that allocates more resolution to lower-frequency bands, so it can capture the subtle spectral artifacts of replayed speech better.

In their method, the speech signal with no voice activity detection is divided into short frames and passed through a 60-channel Gammatone filter bank spanning between

10 Hz and 11 kHz, the log energy of each sub-band is then taken and a Discrete Cosine Transform (DCT) is applied to yield the cepstral coefficients. This produces a 60×128 time-frequency feature map for each utterance, which is fed into a convolutional neural network. The authors use a ResNet-based architecture, comparing a standard ResNet-34 to a deeper "Deep ResNet" model that employs multi-branch (grouped) convolutional blocks (32 groups) to improve accuracy without excessive complexity.

This work evaluated on the ASVspoof2019 Physical Access benchmark, the GTCC + ResNet approach showed substantially improved performance over the challenge's baseline systems. The proposed single-feature system achieved an equal error rate (EER) of about 8.5%, compared to 15% EER for the best baseline (CQCC with GMM) on the same dataset. It also outperformed the authors' earlier MFCC/LFCC-based ResNet models and yielded higher accuracy, balanced accuracy, and F1-score than those counterparts. These results demonstrate that the GTCC front-end combined with a ResNet-based deep model can more effectively detect replay attacks in speaker verification, outperforming conventional features on the benchmark evaluation. Like mentioned earlier, ResNet model was designed

2.5 Dimensionality Reduction

As the complexity and dimensionality of extracted features increase, dimensionality reduction becomes a useful technique to reduce computational cost and memory usage without decreasing data quality. Although this technique is not commonly used in ASV spoofing tasks, it becomes relevant in our case due to the high-dimensional outputs produced by the ResNet feature extractor.

In our work, we focus on using the dimensionality reduction method called Linear Discriminant Analysis (LDA). Also, we try to use another technique called Principal Component Analysis (PCA) to compare the result. This following section is the literature review of such a paper that introduced and the implementation of those techniques:

2.5.1 Linear Discriminant Analysis (LDA)

LDA was originally introduced by R. A. Fisher in 1936 as a classification method in statistics, as a statistical classification method that projects high-dimensional data onto a lower-dimensional space by maximizing the separability between classes. It achieves this by finding a linear combination of features that best separates two or more classes by maximizing the between-class variance while minimizing the within-class variance. In the context of spoof detection, LDA is used after feature extraction to reduce dimensionality while enhancing the discrimination between bona fide and spoofed speech samples. In our

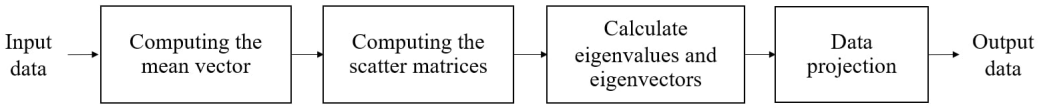

Figure 2.2 Block diagram of LDA.

study, LDA was applied to deep features extracted by ResNet50, compressing them into a single discriminative dimension. This not only reduced computational complexity but also improved the performance of the downstream classifier, as the projection emphasized differences relevant to spoof detection rather than speaker identity or channel variation.

As shown in Figure 2.2, applying LDA to the data involves four steps. First, the mean vectors for each class and the overall mean of the data are calculated. Second, the within-class scatter matrix and the between-class scatter matrix are computed to show how data points spread within each class and how class means differ from each other. Third, the generalized eigenvalue problem for these matrices is solved to find the eigenvectors that maximize class separability. The data is then projected onto these eigenvectors to create the reduced feature space. The number of LDA components is limited by the number of classes, with the maximum number of components being one less than the total number of classes. The resulting output data consists of transformed features that emphasize the distinctions between classes, optimizing them for efficient classification.

When applying LDA, we aim to find an optimal projection matrix $\mathbf{W}$ that maximizes the class separability, which is achieved by solving the following optimization equation as following:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} \tag{2.1}$$

where $\mathbf{W}$ is the linear projection matrix that maps the original high-dimensional feature space to a lower-dimensional subspace. $\mathbf{S}_B$ denotes the between-class scatter matrix (equation 2.2), which captures the dispersion between class means, while $\mathbf{S}_W$ denotes the within-class scatter matrix (equation 2.3), which reflects the variation of samples within each class. These matrices are defined as:

$$\mathbf{S}_B = \sum_{i=1}^{C} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{2.2}$$

$$\mathbf{S}_W = \sum_{i=1}^{C} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \tag{2.3}$$

where C is the number of classes, N_i is the number of samples in class i, $\boldsymbol{\mu}_i$ is the mean vector of class i, and $\boldsymbol{\mu}$ is the global mean of all samples.

To apply LDA for dimensionality reduction, the process begins by grouping the high-dimensional input features according to their class labels, namely bona fide and spoofed speech in our case. Then, the class-wise mean vectors and global mean vector are computed in order to construct the between-class scatter matrix ($\mathbf{S}_B$) and the within-class scatter matrix ($\mathbf{S}_W$). The objective of LDA is to find a linear transformation matrix $\mathbf{W}$ that maximizes the ratio of the between-class variance to the within-class variance, as shown in Equation 2.1. This optimization ensures that the projected data in the lower-dimensional space remains well-separated across classes while preserving as much discriminatory information as possible. In our implementation, the final output from LDA is a one-dimensional representation for each input utterance, which serves as a highly discriminative feature used in the final classification stage.

2.5.2 Principal Component Analysis (PCA)

PCA is a classic technique introduced over a century ago by Karl Pearson (1901) and later formalized by Harold Hotelling (1933), is an unsupervised dimensionality reduction technique that identifies orthogonal directions (principal components) in the feature space along which the data varies the most. Unlike LDA, PCA does not use class label information. Instead, it projects the data into a lower-dimensional space based on directions that retain the most variance. In spoof detection tasks, PCA can be used to remove noise and redundancy from high-dimensional input features, although it may not always retain class-separating information. In our work, PCA was used as a comparative baseline to evaluate the effectiveness of LDA. The result showed that LDA outperformed PCA in classification performance, which is expected given LDA's supervised nature and its ability to focus on spoof-related distinctions.

According to Figure 2.3, the eigenvectors of the data's covariance matrix can be calculated by giving a centered dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, calculated as:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X} \tag{2.4}$$

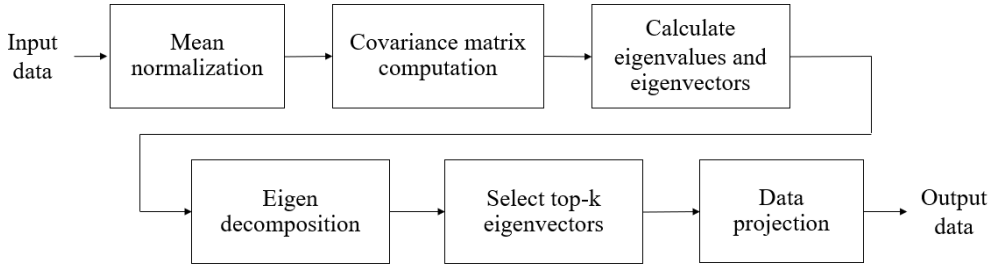

Figure 2.3 Block diagram of PCA.

where n is the number of samples, and $\mathbf{X}$ is the zero-mean data matrix. The covariance matrix captures how much each feature varies with respect to the others, serving as a foundation to identify meaningful patterns of variation in the dataset.

Next, PCA solves the following eigenvalue problem:

$$\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{2.5}$$

where $\mathbf{v}_i$ and λ_i are the eigenvectors and eigenvalues of the covariance matrix, respectively. Each eigenvector $\mathbf{v}_i$ represents a direction in the feature space, and the corresponding eigenvalue λ_i quantifies the amount of variance captured in that direction. By sorting the eigenvectors in descending order of their eigenvalues, PCA identifies the principal directions (or components) that explain the most variance in the data.

The top k eigenvectors $\mathbf{v}_i$ corresponding to the largest eigenvalues λ_i form the projection matrix. The transformed data is then:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}_k \tag{2.6}$$

where $\mathbf{Z}$ is the reduced-dimensional representation of the data. This transformation retains the most informative components of the original dataset while reducing its dimensionality, helping to simplify subsequent processing and reduce computational costs.

In our study, PCA was used as a dimensionality reduction method to compare with the supervised LDA approach. However, since PCA does not utilize class label information during the projection process, it may not effectively preserve discriminative information between bona fide and spoofed speech samples. Therefore, while PCA can reduce noise and redundancy, it may not always be ideal for classification-oriented tasks such as spoof detection.

CHAPTER 3
METHODOLOGY

In this chapter, we explain our proposed method and contribution to improving the performance of the overall ASV system. As mentioned in the previous section, our objective is to explore and expand the use of deep learning–based feature extraction. We focus specifically on using the ResNet model, which is well-known for its ability to extract deep features in computer vision and has been increasingly applied in speech-based spoof detection.

In most cases, ResNet is mainly used as a classifier. However, in our approach, we aim to utilize ResNet50 as a feature extractor, removing its final classification layers and using the deep embeddings it generates for further dimensionality reduction and decision-making. To study the effectiveness of this method, we explore different input configurations, spectrogram types, and tensor arrangements, then evaluate how these variations impact the spoof detection performance.

According to Figure 3.1, our baseline model will include only ResNet50 feature extraction with its own fully-connected classification. The input speech signal is first converted into a temporal-frequency representation. This representation is then passed into a deep feature extractor based on the ResNet50 architecture, which is responsible for learning discriminative patterns from the spectrogram. Finally, the extracted features are fed into the built-in fully connected layer of ResNet50 for binary classification between spoofed and bona fide speech. This simple yet effective framework serves as our starting point for further improvement and experimentation.

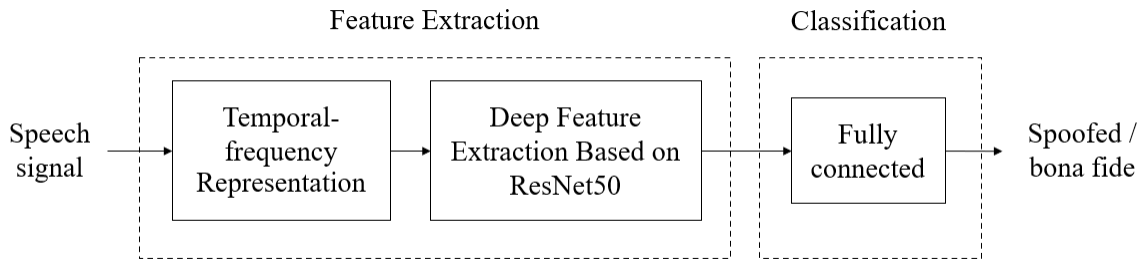

Figure 3.1 Baseline framework.

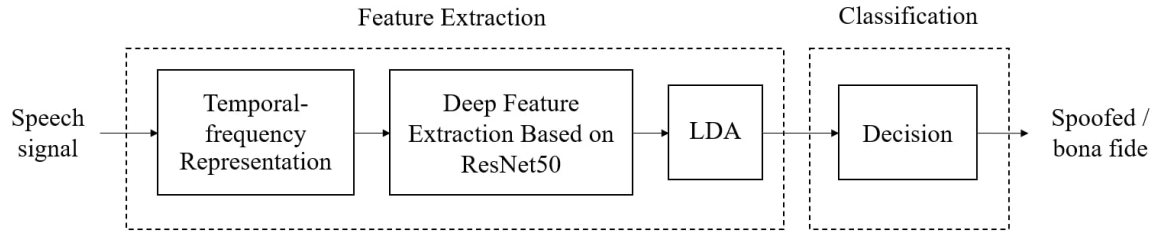

Figure 3.2 Proposed framework.

3.1 Proposed framework

The proposed method analyzes input speech signals, determines their authenticity, and classifies them as bona fide or spoofed. It consists of four steps: temporal frequency representation, deep feature extraction based on ResNet50, linear discriminant analysis (LDA) and decision, as shown in Figure 3.2.

First, the input speech signal is transformed into a temporal-frequency representation. In this work, we explored two representations, which are mel-spectrogram and gammatone spectrogram, as described in the previous section.

Second, the signal representation is fed into a ResNet50 model trained to differentiate between spoofed and bona fide signals. The conventional ResNet50 model accepts a three-channel image input, i.e., the input shape is $224 \times 224 \times 3$. The dense layers of ResNet50 were dropped to retain only the convolutional feature extraction capability, ensuring that the extracted features remained focused on capturing spatial and temporal patterns in the spectrogram representation. The model produces a deep feature vector with 2,048 dimensions as output.

The dimensionality of the deep feature is reduced by using linear discriminant analysis (LDA) in the third step, which is the final stage of the feature extraction process. Since there are two classes, which are spoofed and bona fide signals, LDA can produce only one discriminant component.

Finally, the discriminant component of the previous step is used as input for classification, where a classifier determines whether the signal is spoofed or bona fide. The classification decision is based on an optimal threshold, which is determined by identifying the point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) intersect from the training dataset. The result of this classification is a prediction that indicates whether the input signal is spoofed or bona fide.

The detail of each step will be described as follow:

3.1.1 ResNet50 Configuration

The ResNet50 architecture used in our work consists of 50 layers, including convolutional, batch normalization, ReLU activation, and skip connections grouped into residual blocks. The model starts with a 7×7 convolutional layer with 64 filters and a stride of 2, followed by a 3×3 max-pooling layer. The core of the architecture comprises four sequential stages of residual blocks: Stage 1 contains 3 blocks with 64, 64, and 256 filters, stage 2 contains 4 blocks with 128, 128, and 512 filters, stage 3 contains 6 blocks with 256, 256, and 1024 filters, and stage 4 contains 3 blocks with 512, 512, and 2048 filters.

In the baseline model, the ResNet50 is used end-to-end with its default architecture, including the final global average pooling layer and the fully connected dense layer for binary classification. The model takes an input tensor of shape $224 \times 224 \times 3$ and outputs the prediction indicating whether the input is spoofed or bona fide.

In the proposed model, the same ResNet50 architecture is used, but the final fully connected classification layer is removed. Only the convolutional layers and the global average pooling layer are retained. The resulting output is a 2048-dimensional feature vector, which is then passed to a Linear discriminant analysis (LDA) module for dimensionality reduction and subsequent classification. This modification allows us to decouple feature extraction from classification, making it possible to analyze the behavior of deep embeddings more explicitly and flexibly combine them with other classifiers.

3.1.2 Spectrogram Preparation

To prepare the input for the ResNet50 model, the raw speech signal is first segmented and converted into a time-frequency representation. In our work, we used two different types of spectrograms: mel-spectrogram and gammatone spectrogram. Each spectrogram was computed using the Fast Fourier Transform with appropriate window size and hop length to maintain temporal resolution while capturing spectral characteristics.

For the mel-spectrogram experiment, we set the hop length of 512, window size of 2048 samples, and sampling rate at 22,050 Hz. The mel-spectrogram is generated using a 128-channel mel filterbank, which maps the linear frequency scale into the perceptually motivated mel scale that better reflects human auditory perception. After applying the filterbank to the magnitude spectrum, the resulting power spectrogram is then converted to decibel (dB) scale using logarithmic compression, which improves the visibility of both strong and weak components in the signal.

For the gammatone experiment, we used the *gtgram* function to compute the

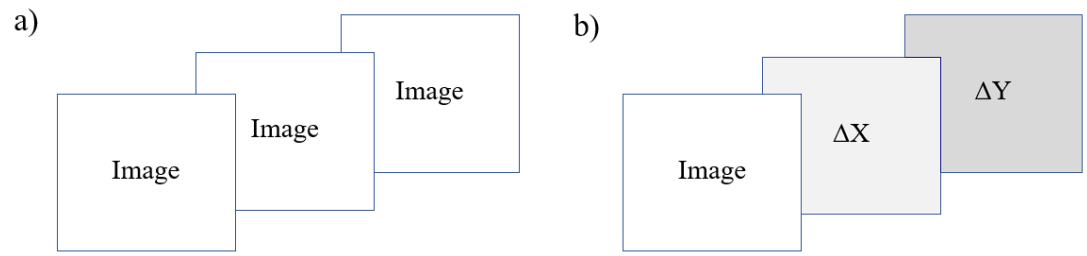

Figure 3.3 (a) Duplicated tensors (type-1), (b) Differentiated tensors (type-2).

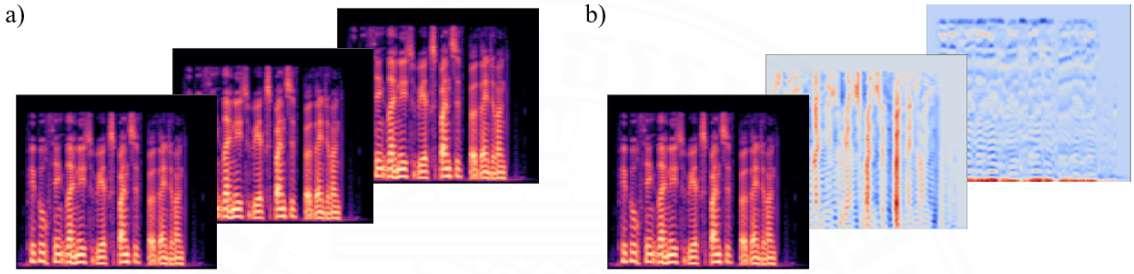

Figure 3.4 Example of mel-spectrogram for both types of tensors.

gammatone spectrogram, simulating the auditory filterbank responses of the human cochlea. The speech signal is zero-padded to 88,200 samples to ensure uniform input length. The spectrogram is then generated using a window size of 25 milliseconds and a hop size of 10 milliseconds, which correspond to 551 and 220 samples respectively at a standard sampling rate. A 128-channel gammatone filterbank is applied, starting from 20 Hz, to decompose the signal into critical bands with non-linear spacing that reflects the human ear's sensitivity across frequency.

3.1.3 Tensor type-1 and tensor type-2

In addition to a normal transformation from a signal representation to a three-channel image input, we also proposed two method to form the input tensor. In the first method, all three channels are duplicates of the representation obtained from the first step. In the second strategy, three channels consist of the representation obtained in the first step, its first-order derivative along the horizontal axis, indicated by ΔX, and its first-order derivative along the vertical axis, denoted by ΔY, respectively. The inclusion of derivatives aims to capture additional temporal and frequency dynamics that may enhance the model's ability to distinguish between bona fide and spoofed signals.

To compute the first-order derivatives (ΔX, ΔY) of the spectrogram, we use the

`librosa.feature.delta` function, which estimates the local derivative using linear regression over a specified window. The default setting uses `width=9`, meaning each derivative at a given frame is computed by fitting a straight line across a 9-frame window centered at that point. The resulting derivative is calculated using the formula:

$$\Delta X_t = \frac{\sum_{n=-4}^{4} n \cdot X_{t+n}}{\sum_{n=-4}^{4} n^2} \tag{3.1}$$

$$\Delta Y_t = \frac{\sum_{n=-4}^{4} n \cdot Y_{t+n}}{\sum_{n=-4}^{4} n^2} \tag{3.2}$$

where ΔX_t and ΔY_t are the first-order derivatives of the spectrogram at time (or frequency) index t, computed along the horizontal and vertical axes respectively. X_{t+n} and Y_{t+n} are the values of the spectrogram at neighboring frames (for time axis) or neighboring frequency bins (for frequency axis), relative to position t.

In our case, the variable n ranges from -4 to $+4$, corresponding to the half-window size when `width = 9`. The numerator $\sum n \cdot X_{t+n}$ computes the weighted sum of surrounding values to approximate the local slope. The denominator $\sum n^2$ serves as a normalization factor to stabilize the derivative value.

This operation captures the local slope of spectral energy changes along time (for ΔX) or frequency (for ΔY) while smoothing out high-frequency noise. Using a wider window like 9 frames provides a more stable and robust estimation of variation, which may be helpful in identifying subtle temporal and spectral cues useful for spoof detection. In our implementation, these derivatives are used to construct a type-2 input tensor where the three channels represent the original spectrogram, its horizontal derivative, and its vertical derivative respectively.

3.1.4 Dimensionality with LDA

The dimensionality reduction step is applied only in our proposed experiment after passing through the ResNet-based feature extraction model. The original ResNet50 model outputs a 2,048-dimensional feature vector, which may contain redundancy or irrelevant variations. To address this, we apply linear discriminant analysis (LDA) to reduce the feature vector to a one-dimensional scalar that emphasizes class separability between bona fide and spoofed speech.

Alternatively, we also apply Principal Component Analysis (PCA) with the number of components set to 10 in order to explore the effectiveness of unsupervised dimensionality reduction. PCA transforms the original 2,048-dimensional feature vector into a lower-dimensional space by projecting the data onto directions (principal components) that capture the maximum variance. Unlike LDA, PCA does not utilize class label information, and the resulting components are optimized for information preservation rather than class separability.

3.1.5 Classification Threshold

For the classification method, we use the fully connected layer in the ResNet baseline model and use the classification threshold decision for the proposed model with dimensionality reduction.

In the baseline configuration, the final prediction is made directly from the ResNet50 model through its built-in fully connected layer and softmax activation. The output provides a confidence score for each class (bona fide or spoofed), and the highest-scoring class is selected as the prediction.

For our proposed model, where features are reduced to a single scalar value using LDA or PCA, a simple thresholding approach is applied for classification. Specifically, we determine a decision threshold from the training set by identifying the point at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) intersect, this point is commonly referred to as the Equal Error Rate (EER) threshold. If the scalar feature value is greater than or equal to this threshold, the signal is classified as bona fide. Otherwise, it is classified as spoofed.

3.2 Dataset

For the experiment, we used the Logical Access (LA) subset of the ASVspoof 2019 dataset for training, validation, and evaluation (Liu et al., 2022). This dataset was specifically designed to address the problem of detecting spoofed speech generated by advanced synthesis techniques, such as text-to-speech (TTS) and voice conversion (VC). The LA subset contains both bona fide and spoofed utterances, where the latter were generated using a total of 17 different algorithms. Among these, 6 systems were made available in the training and development sets (known attacks), while 11 additional systems were reserved exclusively for the unknown attack in the evaluation set, making the task more challenging and suitable for evaluating generalization capability.

The dataset is widely adopted in the spoof detection community due to its realistic

scenarios, balanced class distribution, and availability of evaluation protocols. In our study, we used the training set for model training, the development set for parameter tuning and threshold optimization, and the evaluation set for final performance testing. The statistical information of the dataset used is given in Table 2.2.

3.3 Experiment setup

In this study, we set up experiments and aim to answer two questions. First, inspired by the work of Chakravarty and Dua (2024), can the concept of extraction deep features by ResNet50 be applicable and improved when the temporal-frequency representations are changed? To answer this question, we implemented and compared two models: one with mel-spectrograms and another with gammatone spectrogram. Second, how does different input tensor formation affect the performance of the model? To answer the second question, we also compared two input tensor types: (1) all three channels are the same (i.e., they are duplicated), and (2) three channels are the temporal-frequency representation, its ΔX, and its ΔY, respectively.

To set up the experiment, the model with mel-spectrogram is denoted by 'mel' and that with gammatone spectrogram is denoted by 'gt.' The model, of which its ResNet takes an input tensor forming by replication of the temporal-frequency representation, is marked with 'type-1 tensor,' and the model, where the tensor with its ΔX and ΔY, is marked with 'type-2 tensor.'

This setup results in a total of six experiments, namely: (1) **ResNet50/mel/type-1**, which is the baseline ResNet50 model using mel-spectrogram representation with duplicated tensor; (2) **ResNet50/mel/type-2**, which is the same baseline model but using a tensor consisting of the original spectrogram along with ΔX and ΔY channels; (3) **Proposed/mel/type-1**, which applies LDA on features extracted by ResNet50 from mel-spectrograms with duplicated tensor format; (4) **Proposed/mel/type-2**, which also applies LDA but uses the mel-spectrogram tensor with ΔX and ΔY; (5) **ResNet50/gt/type-2**, the baseline ResNet50 model using gammatone spectrogram with *original*, ΔX, and ΔY channels; and (6) **Proposed/gt/type-2**, which is the proposed LDA-enhanced framework using gammatone spectrogram and tensor type-2.

3.4 Further Experiment

In addition to the primary experiments described in the previous sections, we conducted several supplementary experiments to explore how different preprocessing steps and modeling variations affect the performance of the spoof detection system.

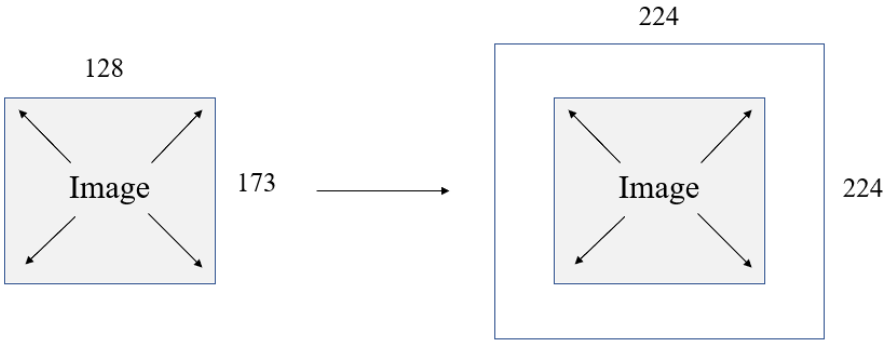

Figure 3.5 Padding spectrogram tensor.

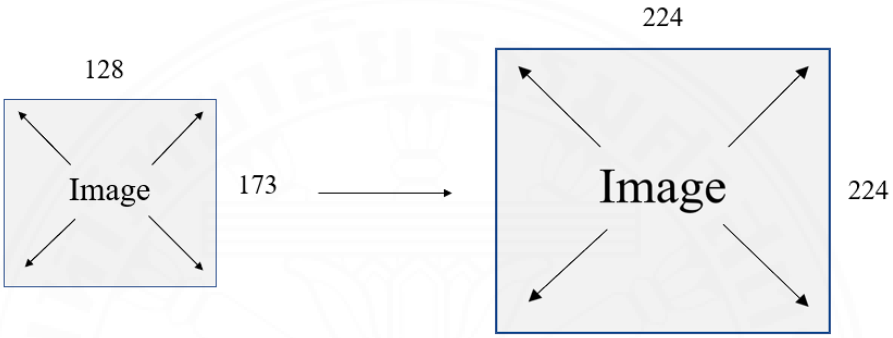

Figure 3.6 Resized spectrogram tensor.

In further experiment, we used mel-spectrogram in every model with some additional changes as following:

3.4.1 Padding and Resizing the Spectrogram Tensor

The spectrograms generated from raw speech signals in the ASVspoof2019 Logical Access dataset have a fixed shape of 128×173, corresponding to 128 mel or gammatone filter channels and 173 time frames. However, the ResNet50 model expects a 224×224 input tensor with three channels. Therefore, before applying the duplication or differentiation strategy mentioned in Section 3.1.3, we need to resize each spectrogram into the appropriate shape.

As illustrated in Figure 3.5 and Figure 3.6, we explore two methods to achieve this transformation. The first method is zero-padding, where zeros are symmetrically added around the edges of the spectrogram to increase its size to the target dimension. This method preserves the original data values but introduces empty regions around the input. The second method is resizing, where the spectrogram is interpolated and rescaled using image resizing techniques to fit into a 224×224 matrix, treating the spectrogram as a standard grayscale

image. This approach maintains the full tensor size and density but may smooth or distort fine-grained features.

These two preprocessing strategies were compared in the extended experiment to evaluate their impact on the performance of the spoof detection model.

3.4.2 Dimensionality with PCA

Apart from using LDA for dimensionality reduction, we also used principal component analysis (PCA) as an alternative technique. PCA is an unsupervised method that transforms high-dimensional data into a lower-dimensional space by identifying the directions (principal components) along which the data varies the most. This approach is particularly useful for reducing redundancy and compressing the feature space while retaining as much information as possible.

In our experiment, we applied PCA to the 2,048-dimensional deep feature vectors extracted from the ResNet50 model. These features, although rich in information, contain some degree of correlation and noise that may not contribute meaningfully to the classification task. By computing the covariance matrix of the feature set and solving its eigenvalue decomposition, we identified the principal components ranked by their corresponding eigenvalues. The top 10 eigenvectors ($k = 10$ associated with the largest eigenvalues were selected to form the projection matrix, and these values will be used to classify whether the speech are bona fide or spoofed in the classification part.

3.4.3 Classifier

After applying dimensionality reduction techniques, particularly PCA in our extended experiments, the resulting lower-dimensional features were further evaluated using several standard classification algorithms. Specifically, we employed three widely used classifiers: Random Forest (RF), k-Nearest Neighbors (KNN), and Naïve Bayes (NB), each selected for their distinct advantages in handling different data characteristics.

The KNN algorithm classifies each input by majority vote among its k closest neighbors in the feature space. In most cases, KNN can perform well in such settings without suffering from the curse of dimensionality. In our work, we have only 2 classes of data, so the number of k is equal to 2.

Random Forest is an ensemble learning method that builds multiple decision trees and merges their outputs to improve overall accuracy and reduce the risk of overfitting. It is well-suited for handling noisy or imbalanced data and provides robust decision boundaries, making it a strong candidate for classifying spoofed versus bona fide speech features.

Lastly, Naïve Bayes, a probabilistic classifier based on Bayes' theorem, assumes conditional independence among features. Despite this simplification, NB is computationally efficient and performs reasonably well on small datasets or when feature distributions are sufficiently separable.

In conclusion, our extended experiments can be grouped by classifier type and configuration as follows: (1) **Random Forest (RF)**: A total of 12 experiments were conducted. These cover all combinations of 4 tensor configurations: Resized + type-1, Resized + type-2, Padding + type-1, Padding + type-2, and 3 model types: Baseline ResNet50, Proposed with LDA, and Proposed with PCA. (2) **K-Nearest Neighbors (KNN)** and (3) **Naïve Bayes (NB)**: Each classifier was tested in 6 experiments, using Resized tensors only: both type-1 and type-2 and 3 model types as above (ResNet50, LDA, PCA).

CHAPTER 4
RESULT AND DISCUSSION

4.1 Simulation and Evaluation

For evaluating our proposed model, a testing dataset consisting of 7,355 bona fide and 63,882 spoof signals was used. We preprocessed the data in the same manner as the training set and predicted whether each speech sample was spoofed or bona fide, and then calculated the accuracy, F1-score, recall, precision, balanced accuracy (Bal), and equal error rate (EER) from each experiment. The comparison of model evaluation is shown in Table 4.1.

As a result, our proposed method demonstrates substantial improvements over the ResNet50 baseline in both EER and balanced accuracy. For the type-1 tensor configuration, the proposed method shows a 43.55% reduction in EER and a 48.59% increase in balanced accuracy compared to ResNet50 alone. For the type-2 tensor configuration, the improvements include an 8.95% reduction in EER and a 15.52% increase in balanced accuracy for the mel-spectrogram experiment, and a 44.14% reduction in EER and a 44.77% increase in balanced accuracy for the gammatone experiment. These results underscore the effectiveness of our enhancements in improving detection performance across different tensor types and spectrogram representations.

The second comparison focuses on the type-1 and type-2 tensor configurations. For the ResNet50 model, the type-2 configuration achieves a 36.85% reduction in EER and a 31.17% increase in balanced accuracy over type-1. In our proposed model, type-2 achieves a slightly lower EER of 1.55% compared to type-1, while the balanced accuracy is lower by 1.90%. These results suggest that the type-2 configuration improves precision, but may slightly reduce recall, which could be advantageous in tasks where a higher confidence in

Table 4.1 Performance comparison among different models.

Model	Accuracy	F1	Precision	Recall	Bal	EER (%)
ResNet50/mel/type-1	0.8586	0.9238	0.8931	0.9567	0.4544	51.86
ResNet50/mel/type-2	0.9137	0.9508	0.9722	0.9303	0.7661	15.01
Proposed/mel/type-1	0.9741	0.9856	0.9822	0.9890	0.9403	8.31
Proposed/mel/type-2	0.9736	0.9852	0.9880	0.9824	0.9213	6.06
ResNet50/gt/type-2	0.8967	0.9455	0.8968	1.0000	0.4484	50.00
Proposed/gt/type-2	0.9674	0.9817	0.9893	0.9743	0.8961	5.86

Table 4.2 Additional experiment with Random Forest (RF) classifier.

Model	Accuracy	F1	Precision	Recall	Bal	EER(%)
ResNet50/resized/type-1	0.8958	0.8960	0.8964	0.8958	0.7215	10.42
Proposed(PCA)/resized/type-1	0.8753	0.8530	0.8382	0.8753	0.5393	12.47
Proposed(LDA)/resized/type-1	0.9593	0.9551	0.9601	0.9593	0.8087	4.07
ResNet50/resized/type-2	0.8983	0.8965	0.8949	0.8983	0.7086	10.17
Proposed(PCA)/resized/type-2	0.8294	0.8134	0.7981	0.8293	0.4637	17.06
Proposed(LDA)/resized/type-2	0.9633	0.9602	0.9635	0.9633	0.8318	3.67
ResNet50/padding/type-1	0.8944	0.8989	0.9047	0.8943	0.7628	10.56
Proposed(PCA)/padding/type-1	0.8845	0.8543	0.8387	0.8845	0.5287	11.55
Proposed(LDA)/padding/type-1	0.9610	0.9577	0.9610	0.9610	0.8241	3.90
ResNet50/padding/type-2	0.8979	0.8995	0.9012	0.8979	0.7406	10.21
Proposed(PCA)/padding/type-2	0.8358	0.8527	0.8773	0.8358	0.6999	16.42
Proposed(LDA)/padding/type-2	0.9614	0.9582	0.9613	0.9614	0.8260	3.86

Table 4.3 Additional experiment with k-Nearest Neighbors (KNN) classifier.

Model	Accuracy	F1	Precision	Recall	Bal	EER(%)
ResNet50/resized/type-1	0.8705	0.8815	0.8982	0.8705	0.7606	12.95
Proposed(PCA)/resized/type-1	0.8676	0.8631	0.8590	0.8676	0.6111	13.24
Proposed(LDA)/resized/type-1	0.9566	0.9517	0.9578	0.9567	0.7946	4.34
ResNet50/resized/type-2	0.8694	0.8787	0.8918	0.8694	0.7359	13.06
Proposed(PCA)/resized/type-2	0.8625	0.8321	0.8047	0.8625	0.4855	13.75
Proposed(LDA)/resized/type-2	0.9627	0.9593	0.9631	0.9627	0.8266	3.73

positive detections is critical, although it may be less optimal in scenarios where balanced detection across all classes is necessary.

By comparing all six experiments, the model using our proposed method with mel-spectrogram features and the type-1 configuration achieves the highest balanced accuracy at 94.03%, while the model with our proposed method using gammatone spectrogram and the type-2 configuration achieves the lowest EER at 5.86%. These results highlight that, depending on the priority metric, either mel-spectrogram with type-1 for balanced accuracy or a gammatone spectrogram with type-2 for EER could be optimal configurations.

4.2 Further Experiment Result

In this section, we showed the result of the further experiment that are conducted from various tensor formats, dimensionality reduction techniques, and classifier types. We recorded the result in a different table separated by its classifier, which are Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB) classifiers. In each table, the models were trained on deep features extracted from the ResNet50-based model with a different types of tensors and processed with either LDA or PCA for dimensionality reduction with a specific classifier.

Table 4.4 Additional experiment with Naïve Bayes (NB) classifier.

Model	Accuracy	F1	Precision	Recall	Bal	EER(%)
ResNet50/resized/type-1	0.7336	0.7851	0.9170	0.7336	0.8274	26.64
Proposed(PCA)/resized/type-1	0.8923	0.8563	0.8467	0.8924	0.5247	10.77
Proposed(LDA)/resized/type-1	0.9667	0.9642	0.9667	0.9667	0.8493	3.33
ResNet50/resized/type-2	0.6671	0.7321	0.9105	0.6671	0.7872	33.29
Proposed(PCA)/resized/type-2	0.8875	0.8433	0.8035	0.8874	0.4949	11.25
Proposed(LDA)/resized/type-2	0.9720	0.9707	0.9716	0.9720	0.8818	2.80

In Table 4.2, the Random Forest (RF) classifier has been applied. We conducted the experiment for every type of tensor to see which one can perform the highest performace in term of balance accuracy (Bal) and equal error rate (EER). As a result, the highest balanced accuracy and the lowest equal error rate was achieved by the proposed method using LDA and resized type-2 tensor, reaching 83.18% and an equal error rate (EER) of 3.67%, which is an appropriate result.

However, we can see from the table that every result from the PCA will give us a worse ERR than the baseline ResNet50. Since we fixed the value of k equal to 10 in this experiment, we can conduct the further experiment with the PCA with different k in the future, the result of PCA would be different.

Similarly to the result while we used k-Nearest Neighbors (KNN) that shown in Table 4.3 and Naïve Bayes (NB) classifier that shown in Table 4.4, the best performance was still achieved by the proposed method using LDA with resized type-2 tensor, showing consistent superiority in both Bal and EER across all three classifiers. The KNN model achieved a balanced accuracy of 82.66% with an EER of 3.73%, while the NB classifier achieved 88.18% balanced accuracy and an EER of 2.80% under the same configuration.

4.3 Discussion

Despite these results, it is evident that our proposed method outperforms the ResNet50 baseline model for spoof detection, showing substantial improvements across key metrics such as EER and balanced accuracy. However, this work still has several limitations and there are numerous avenues to explore to improve deep learning-based feature extraction. One potential direction is to experiment with different types of spectrograms, which provide two-dimensional features, to identify which features are most effective for this model. Additionally, investigating hybrid techniques that combine both time-domain and frequency-domain features could offer a more comprehensive representation of the audio signal. Exploring alternative dimension reduction methods beyond LDA may further enhance performance by preserving key information while reducing redundancy. Moreover,

advanced neural network architectures, such as attention mechanisms or transformer-based models, could improve the model's ability to capture relevant temporal or spectral patterns in spoof detection. Future research should also consider larger and more diverse datasets to better generalize the findings across various types of audio spoofing techniques.

According to the further experiment result, applying some kind of classifier may give us a higher result comparing with using the decision threshold. The overall performance of the NB classifier was higher compared to RF and KNN due to its strong assumptions about feature independence. These observations indicate that both the input tensor structure and the choice of dimensionality reduction technique play crucial roles in optimizing downstream classifier performance. In particular, the combination of LDA and type-2 tensor constructed from resized spectrograms appears to provide a strong generalization ability across different classification strategies.

However, the results from this PCA-based setup were significantly worse compared to those obtained using LDA. The primary reason appears to be the unsupervised nature of PCA, which ignores class label information when computing its projection matrix. As a result, the reduced features may retain high variance but not necessarily the dimensions most useful for distinguishing bona fide and spoofed speech. This limitation was particularly evident in our binary classification task, where class separation is essential. Consequently, the PCA approach was deemed ineffective in our context, and we decided to proceed with LDA as the main dimensionality reduction method in our proposed framework.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

This work aimed to expand the exploration of deep learning-based feature extraction models. ResNet50 was used to extract deep features from temporal-frequency, e.g., mel-spectrogram and gammatone spectrogram, representation of a speech signal. Then, LDA was applied to reduce the feature dimensionality, and a decision is mode by setting a separation plan in the reduced-dimensional space. The experimental results show that the proposed method outperforms the model based on ResNet50 alone. Compared to ResNet50, our proposed method can reduce the EER by 43.55% and improve the balanced accuracy by 48.59% for type-1 tensor, and reduce the EER by 8.95% and improve the balanced accuracy by 15.52% for the type-2 tensor mel spectrogram and reduce the EER by 44.14% and improve the balanced accuracy by 44.77% for the type-2 tensor gammatone spectrogram.

In addition, the experimental results show that when the input tensor of ResNet50 is formed by the gammatone spectrogram with its derivatives, the performance of the model can also be improved.

To further validate the flexibility of the proposed feature extraction method, we extended our study to include traditional classification methods. Specifically, we investigated how deep features extracted from the ResNet50 model (and reduced by LDA or PCA) would perform under classifiers such as Random Forest (RF), k-Nearest Neighbors (KNN), and Naïve Bayes (NB). From the results, it was evident that LDA-based dimensionality reduction yielded the best performance across all classifiers, particularly when used with a resized type-2 tensor. Notably, the NB classifier achieved the highest balanced accuracy of 88.18% and the lowest EER of 2.80% under this configuration. These findings indicate that, beyond threshold-based classification, traditional classifiers can also effectively leverage deep features—especially when paired with supervised dimensionality reduction techniques.

On the contrary, the use of PCA as an unsupervised dimensionality reduction technique did not improve performance and, in most cases, performed worse than even the baseline ResNet50 model. This supports the hypothesis that supervised dimensionality reduction methods, such as LDA, which consider class separability, are more appropriate

for this binary classification task. This insight could guide future work in choosing feature projection methods for similar problems.

5.2 Future work

In the future, we plan to extend our model by experimenting with other types of time-frequency representations apart from mel-spectrogram and gammatone spectrogram, e.g. chroma spectrograms, constant-Q transform, or even wavelet-based representations, which may provide different perspectives for capturing spoofing cues. The baseline model we used, ResNet50, can also be replaced or compared with other vision-based architectures like DenseNet, or MobileNet to see whether lighter or deeper models offer any performance advantage.

In terms of dimensionality reduction, techniques such as Principal Component Analysis (PCA) have been used but in the fixed value of k, we can do the further experiment by focusing only on this method but with different k in order to find the optimal result. Or even learned methods like autoencoders could be applied in place of or alongside LDA to investigate how the structure of the feature space affects classification. Another possible direction is to explore end-to-end learning approaches where the feature extractor and classifier are trained jointly.

Finally, further testing with other datasets in different languages or recording conditions, such as ASVspoof 2021 or multilingual corpora, could provide a different result. Comparing performance across datasets and attack scenarios would help evaluate how effectiveness of the proposed method is in real-world applications.

REFERENCES

Adnan, M. M., Rahim, M. S. M., Khan, A. R., Alkhayyat, A., Alamri, F. S., Saba, T., & Bahaj, S. A. (2023). Automated image annotation with novel features based on deep resnet50-slt. *IEEE Access*, *11*, 40258–40277. https://doi.org/10.1109/ACCESS.2023.3266296

Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., O'Dell, J., Butler, K., & Traynor, P. (2022). Who are you (i really wanna know)? detecting audio DeepFakes through vocal tract reconstruction. *31st USENIX Security Symposium (USENIX Security 22)*, 2691–2708. https://www.usenix.org/conference/usenixsecurity22/presentation/blue

BT, B., Lin, K. W. E., Lui, S., Chen, J.-M., & Herremans, D. (2019). Towards robust audio spoofing detection: A detailed comparison of traditional and learned features. https://arxiv.org/abs/1905.12439

Chaiwongyen, A., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M. (2022). Replay attack detection in automatic speaker verification using gammatone cepstral coefficients and resnet-based model. *Journal of Signal Processing*, *26*(6), 171–175. https://doi.org/10.2299/jsp.26.171

Chakravarty, N., & Dua, M. (2024). A lightweight feature extraction technique for deepfake audio detection. *Multimedia Tools and Applications*, *83*. https://doi.org/10.1007/s11042-024-18217-9

Consortium, A. (2019). Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan [Accessed: 2024-04-30].

Das, R., Yang, J., & Li, H. (2019). Long range acoustic and deep features perspective on asvspoof 2019. https://doi.org/10.7488/ds/1994

Davis, S., & Mermelstein, P. (1980a). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Davis, S., & Mermelstein, P. (1980b). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Deng, J., Mao, T., Yan, D., Dong, L., & Dong, M. (2022). Detection of synthetic speech based on spectrum defects. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 3–8. https://doi.org/10.1145/3552466.3556529

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. (2019). Glottal source processing: From analysis to applications. https://arxiv.org/abs/1912.12604

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. https://arxiv.org/abs/1603.03947

Hasan, T., Sadjadi, S. O., Liu, G., Shokouhi, N., Bořil, H., & Hansen, J. H. (2013). Crss systems for 2012 nist speaker recognition evaluation. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6783–6787. https://doi.org/10.1109/ICASSP.2013.6638975

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385

Khan, A., Malik, K. M., & Nawaz, S. (2023). Frame-to-utterance convergence: A spectra-temporal approach for unified spoofing detection. https://arxiv.org/abs/2309.09837

Kim, J.-h., Shim, H.-j., Heo, J., & Yu, H.-J. (2022). Rawnext: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies. https://arxiv.org/abs/2112.07935

Kinnunen, T., Evans, N., Yamagishi, J., Lee, K. A., Sahidullah, M., Todisco, M., & Delgado, H. (2017). Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan [Accessed: 2024-04-30].

Lambamo, W., Srinivasagan, R., & Jifara, W. (2023). Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition. *Applied Sciences*, *13*(1). https://doi.org/10.3390/app13010569

Lee, K. A., Larcher, A., Wang, G., Kenny, P., Brümmer, N., van Leeuwen, D., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., Li, H., Stafylakis, T., Alam, J., Swart, A., & Perez, J. (2015). The RedDots Data Collection for Speaker Recognition. *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. https://hal.science/hal-01818427

Lei, Z., Yan, H., Liu, C., Ma, M., & Yang, Y. (2022). Two-path gmm-resnet and gmm-senet for asv spoofing detection. *ICASSP 2022 - 2022 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 6377–6381. https://doi.org/10.1109/icassp43922.2022.9746163

Li, M., Ahmadiadli, Y., & Zhang, X.-P. (2024). Audio anti-spoofing detection: A survey. https://arxiv.org/abs/2404.13914

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., & Lee, K. A. (2022). ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. https://doi.org/10.48550/ARXIV.2210.02437

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., & Lee, K. A. (2023). Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 2507–2522. https://doi.org/10.1109/taslp.2023.3285283

Mon, K. Z., Galajit, K., Mawalim, C. O., Karnjana, J., Isshiki, T., & Aimmanee, P. (2023). Spoof detection using voice contribution on lfcc features and resnet-34. *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1–6. https://doi.org/10.1109/iSAI-NLP60301.2023.10354625

Patterson, R., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1988). An efficient auditory filterbank based on the gammatone function.

Sahidullah, M., Delgado, H., Todisco, M., Kinnunen, T., Evans, N. W. D., Yamagishi, J., & Lee, K. (2019). Introduction to voice presentation attack detection and recent advances. *CoRR*, *abs/1901.01085*. http://arxiv.org/abs/1901.01085

Shin, S., Kim, J., Yu, Y., Lee, S., & Lee, K. (2021). Self-supervised transfer learning from natural images for sound classification. *Applied Sciences*, *11*(7). https://doi.org/10.3390/app11073043

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. https://arxiv.org/abs/1403.2877

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. https://doi.org/10.1121/1.1915893

Tak, R., Agrawal, D., & Patil, H. (2017). Novel phase encoded mel filterbank energies for environmental sound classification, 317–325. https://doi.org/10.1007/978-3-319-69900-4_40

Todisco, M., Delgado, H., & Evans, N. (2017). Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech Language*, *45*, 516–535. https://doi.org/https://doi.org/10.1016/j.csl.2017.01.001

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, *66*, 130–153. https://doi.org/https://doi.org/10.1016/j.specom.2014.10.005

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., & Sahidullah, M. (2015). Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. https://doi.org/10.21437/Interspeech.2015-462

Wu, Z., Yamagishi, J., Kinnunen, T., Hanilçi, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M., & Delgado, H. (2017). Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, *11*(4), 588–604. https://doi.org/10.1109/JSTSP.2017.2671435

Yang, J., Wang, H., Das, R., & Qian, Y. (2021). Modified magnitude-phase spectrum information for spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *PP*, 1–1. https://doi.org/10.1109/TASLP.2021.3060810

Zhang, Y., Zhu, G., Jiang, F., & Duan, Z. (2021). An empirical study on channel effects for synthetic voice spoofing countermeasure systems. https://arxiv.org/abs/2104.01320

Zhang, Y., Li, Z., Lu, J., Hua, H., Wang, W., & Zhang, P. (2023). The impact of silence on speech anti-spoofing. https://arxiv.org/abs/2309.11827

Zhiqiang, L., Zhang, S., Tang, K., & Hu, P. (2022). Fake audio detection based on unsupervised pretraining models, 9231–9235. https : / / doi . org / 10 . 1109 / ICASSP43922.2022.9747605

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., & Shamma, S. (2011). Linear versus mel frequency cepstral coefficients for speaker recognition. *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, 559–564. https://doi.org/10.1109/ASRU.2011.6163888

APPENDIX

APPENDIX A
IMPLEMENTATION OF RESNET50

In this appendix, we provide the implementation details of the ResNet50 architecture. The code follows the canonical structure of residual networks and is divided into several modular components, namely the convolutional block, the identity block, and the projection block. Each of these components plays a crucial role in enabling residual learning and alleviating the vanishing gradient problem commonly found in very deep neural networks.

The convolutional block serves as the basic feature extraction unit of the network. It performs a sequence of operations consisting of a two-dimensional convolution, batch normalization, and the ReLU activation function. By adjusting the filter size, kernel size, stride, and padding, the block is capable of capturing local spatial features at different scales and resolutions.

The identity block introduces the core concept of residual learning by incorporating a skip connection that directly adds the input tensor to the output of a series of convolutional layers. This addition allows the model to learn a residual mapping, $F(x) + x$, rather than a direct mapping $H(x)$, which significantly eases the training of deeper networks. Importantly, in the identity block, the input and output dimensions remain the same, making it possible to combine the two paths without any transformation.

The projection block, in contrast, is employed when the input and output dimensions are not aligned. In this case, a 1×1 convolution is applied in the shortcut pathway to match the dimensionality, followed by a 3×3 convolution consistent with the pattern used in the ResNet architecture. This block plays a critical role in transitioning between different stages of the network, where the number of filters increases, thereby ensuring that the residual connection remains valid and effective throughout the architecture.

Finally, all the components described above are integrated to manually construct a ResNet50 model, as illustrated in Figure A.4. This modular design allows flexibility, enabling modifications at any layer as required for specific experimental purposes.

```
def conv_block(x, filters, kernel_size, strides, padding='same'):
    x = Conv2D(filters=filters, kernel_size=kernel_size, strides=strides, padding=padding)(x)
    x = BatchNormalization()(x)
    x = Activation('relu')(x)
    return x
```

Figure A.1 Convolution block function

```
def identity_block(x, filters):
    shortcut = x
    x = conv_block(x, filters=filters, kernel_size=(1, 1), strides=(1, 1))
    x = conv_block(x, filters=filters, kernel_size=(3, 3), strides=(1, 1))
    x = Conv2D(filters=filters * 4, kernel_size=(1, 1))(x)
    x = BatchNormalization()(x)
    x = Add()([x, shortcut])
    x = Activation('relu')(x)
    return x
```

Figure A.2 Identity block function

```
def projection_block(x, filters, strides):
    shortcut = x
    x = conv_block(x, filters=filters, kernel_size=(1, 1), strides=strides)
    x = conv_block(x, filters=filters, kernel_size=(3, 3), strides=(1, 1))
    x = Conv2D(filters=filters * 4, kernel_size=(1, 1))(x)
    x = BatchNormalization()(x)
    shortcut = Conv2D(filters=filters * 4, kernel_size=(1, 1), strides=strides)(shortcut)
    shortcut = BatchNormalization()(shortcut)
    x = Add()([x, shortcut])
    x = Activation('relu')(x)
    return x
```

Figure A.3 Projection block function

```
def resnet_50_train(input_shape=(224, 224, 3), num_classes = 2):
    inputs = Input(shape=input_shape)
    # initial conv layer
    x = conv_block(inputs, filters=64, kernel_size=(7, 7), strides=(2, 2), padding='same')
    x = MaxPooling2D(pool_size=(3, 3), strides=(2, 2), padding='same')(x)
    # conv block 1
    x = projection_block(x, filters=64, strides=(1, 1))
    x = identity_block(x, filters=64)
    x = identity_block(x, filters=64)
    # conv block 2
    x = projection_block(x, filters=128, strides=(2, 2))
    x = identity_block(x, filters=128)
    x = identity_block(x, filters=128)
    x = identity_block(x, filters=128)
    # conv block 3
    x = projection_block(x, filters=256, strides=(2, 2))
    x = identity_block(x, filters=256)
    x = identity_block(x, filters=256)
    x = identity_block(x, filters=256)
    x = identity_block(x, filters=256)
    x = identity_block(x, filters=256)
    # conv block 4
    x = projection_block(x, filters=512, strides=(2, 2))
    x = identity_block(x, filters=512)
    x = identity_block(x, filters=512)
    # global average pooling and dense layer
    x = GlobalAveragePooling2D()(x)
    x = Dense(512, activation='relu')(x)
    outputs = Dense(1, activation='sigmoid')(x)
    model = Model(inputs=inputs, outputs=outputs)
    model.summary()
    return model
```

Figure A.4 Training model of ResNet50 with the convolution block, identity block and projection block

BIOGRAPHY

Name	Peemapot Uparakool
Education	2009: Bachelor of Engineering
	(Electronic and Communication Engineering)
	Sirindhorn International Institute of Technology
	Thammasat University

Publication

Uparakool, P. et al. (2025). Anti-spoofing Using ResNet50 with Linear Discriminant Analysis for Automatic Speaker Verification. In: Huynh, VN., Honda, K., Le, B., Inuiguchi, M., Huynh, H.T. (eds) *Integrated Uncertainty in Knowledge Modelling and Decision Making*. IUKM 2025. Lecture Notes in Computer Science(), vol 15585. Springer, Singapore. https://doi.org/10.1007/978-981-96-4606-7_24