Journal of Information Systems Engineering and Management
2025, 10(31s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing the Security of Speaker Verification: A Hybrid Feature and Xception-Based Method for Spoof Detection

Selin M*[1], Preetha Mathew K[2]

[1*]Department of Computer Applications, Cochin University of Science and Technology, Kerala, India.

2Cochin University College of Engineering, Kerala, India.

Email ID's: - 1*selin.m.a@gmail.com, 2preetha.mathew.k@gmail.com.

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Even though Automatic Speaker Verification (ASV) systems are an essential part of biometric authentication, they are nevertheless vulnerable to spoofing attacks, particularly logical access attacks such as voice conversion and text-to-speech (TTS) synthesis. In order to increase ASV security, an effective spoof detection system is suggested that integrates the complementary data from Mel-Frequency Cepstral Coefficients (MFCC) and Constant Q Cepstral Coefficients (CQCC). The Xception model, the most advanced deep learning (DL) architecture created for high-dimensional extraction of feature, handles these characteristics, because capture both short-term and long-term spectrum properties. With the ASVspoof 2019 Logical Access dataset, the suggested approach achieves 92.11% accuracy, 92% precision, 93% recall, and a 92% F1-score on average. Outperforming traditional GMM-based and deep learning-based approaches, the system also achieves a low Tandem Detection Cost Function (t-DCF) score of 0.0464 and an Equal Error Rate (EER) of 0.0511. These findings show that the suggested approach, which offers high verification reliability and enhanced resistance to spoofing attacks, has potential in real-world ASV applications.

Keywords: Spoof Detection, Speaker Verification, MFCC, CQCC, Xception Model.

INTRODUCTION

Automatic Speaker Verification (ASV) acts as a flexible and economical biometric approach for personal authentication that has been widely employed in many security and authentication applications, including secure account access and smartphone unlocking [1]. ASV systems function by capturing the unique attributes of an individual's voice including their pitch, accent, and speaking style. These characteristics are then compared to a reference model of the speaker's voice that has been pre-registered in the system [2]. The speaker is verified and granted access if the speaker's voice characteristics match with the reference model in an acceptable threshold. However, spoofing attacks, in which malicious attackers attempt to circumvent security measures by posing as reliable users, might affect ASV systems. Spoofing attacks in speaker verification can range from simple tasks like playing back pre-recorded audio to more intricate ones like voice conversion (VC) [3] as well as speech synthesis (SS) [4], which replicate the target speaker's tone by changing the speaker's voice. These attacks cause significant risk to ASV systems, which lead to security breaches and unauthorized access to confidential data. With the exponential development of social networks people are sharing their audio and video recordings on online platforms. As the target speaker's voiceprint information is readily available on the internet, an imposter can use it to create high-quality speech signals that closely resemble the target voice. The ASV systems can be manipulated using these spoofed speech signals.

Four main categories can be used to classify spoofing attacks: replay attacks using pre-recorded audio, impersonation attacks involving vocal mimicry, with twin impersonation posing a unique challenge, SS using VC, and TTS technology [5] to alter the voice of one speaker to sound like that of another. These attacks can be further classified based on their mode of execution: physical access, where the spoofed audio is introduced through the microphone, and logical access, where the attack bypasses the sensor and directly targets the ASV system, by the application of TTS or VC techniques. Spoofing detection, often referred to as presentation attack detection, is becoming more and more important in ASV, as seen by the emergence of multiple unique evaluation challenges. Logical access attacks with TTS and VC are addressed by ASVspoof 2015 [6], logical and physical access attacks are addressed by BTAS 2016 [7], real replay attacks in noisy environments are addressed by ASV spoof 2017 [8],and ASVspoof 2019 [9] addresses logical access attacks using sophisticated TTS as well as VC technologies as well as simulating replay attacks under various acoustic settings.

Several features have been studied for spoofing detection, including magnitude-based characteristics like log-magnitude spectrum [10] along with residual LMS, phase-based features [11] like group delay and improved group delay, and cepstral coefficients obtained through both spectral and phase information, that include Linear Frequency Cepstral Coefficients (LFCC), MFCC, [12], and their variants. Researchers have also investigated other features like local binary patterns, pitch patterns, i-Vectors, and modulation features for their potential in spoofing detection.

In recent years, audio spoofing detection has advanced significantly. Unfortunately, the majority of current techniques are only designed to identify particular kinds of attacks, which reduces their usefulness in practical situations. Additionally, these methods frequently have issues with robustness against hostile environments, generalization, and feature extraction. To overcome these limitations, a DL-based spoof detection framework has been brought ahead, which combines complementary feature representations of characteristics with an advanced classification model to improve the accuracy and efficacy of autonomous speaker verification systems. The following are the key elements of this work:

1. **Hybrid feature extraction for enhanced spoofing detection:** Both short-term and long-term spectrum features of speech are captured by combining MFCCs and constant Q cepstral coefficients (CQCCs). These characteristics offer a more thorough depiction of the speech signal, making it possible to identify various spoofing techniques, such as VC and TTS synthesis.
2. **Deep learning-based classification using the Xception model:** The Xception model, a deep learning framework specialized for high- dimensional feature extraction, is used to process the extracted features. Xception's depthwise separable convolutions increase classification accuracy and processing efficiency.
3. **Comprehensive performance evaluation and benchmarking:** With an accuracy of **92.11%**, an EER of **0.0511**, and a t-DCF score of **0.0464**, the suggested framework exhibits strong detection performance when tested on the ASVspoof 2019 Logical Access dataset.

The following is how this document is structured: Section 2 discusses current studies in the areas of spoofing audio detection. A thorough discussion of the suggested methodology is given in Section 3. The dataset and evaluation metrics used for performance assessment are described in detail in Section 4. Section 5 discusses the findings and an in-depth analysis of them. The paper is finally concluded in Section 6.

RELATED WORKS

During Interspeech 2013, the first symposium on spoofing countermeasures em- phasized the necessity of a consistent dataset, methods, and metrics for Au- tomatic Speaker Verification systems [13]. This led to the development of the ASV Spoofing and Countermeasures (ASVspoof). After that, the Interspeech 2015 ASVspoof challenge was organized, and its second iteration was held in 2017. The replay attack detection, one of the most prevalent and easily accessi- ble types of spoofing in ASV systems, was the main goal of the ASVspoof 2017 competition.

ASVspoof 2019 was the first competition to use spoofing methods such as VC, voice synthesis, and replay assaults [14]. It featured sep- arate scenarios for logical and physical access, allowing a more comprehensive evaluation of ASV systems under different types of attacks. ASVspoof 2021 [15] presented a more comprehensive assessment approach that encompassed novel spoofing attack types, including adversarial attacks.

For speaker verification and spoofing detection, a multitask Conformer model based on Conformer blocks [16] and X-vector model is proposed. This method was the very first to apply Conformer for combined tasks in speaker verifica- tion and anti-spoofing, achieving a 70% improvement in SASV-EER on the ASVspoof2019 LA dataset. A method for detecting playback spoofing in ASV systems by combining temporal and spectral features with ML and DL tech- niques using Recursive Feature Elimination (RFE) and XGBoost was introduced in [17], the approach achieved significant performance improvements, with ac- curacy rising to 99.86% and EER dropping to 0.69%. The CQCCs are used in ASV systems to identify spoofing attacks, using the constant Q transform for improved time-frequency analysis was proposed in [18]. Studies employing the ASVspoof 2015 database show that CQCCs with a Gaussian mixture model outperform other methods by 72% in detecting unknown spoofing attempts. The MFCC, Constant Q Transform (CQT), CQCC, and LFCC are commonly used speech features and extraction techniques for spoofing detection. A common baseline classification technique is the Gaussian Mixture Model (GMM) [19]. A few more DL methods that are used include Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Siamese neural networks (Siamese CNN), and Deep Residual Networks (ResNet). On the ASVspoof 2019 dataset, this study [20] suggests a replay attack detection technique for ASV utilizing an 18-layer ResNet algorithm containing LFCCs, producing findings with an EER of 0.29%.

METHODOLOGY

The proposed method combines MFCCs and CQCCs [21] features for enhanced spoof detection. MFCCs uses a logarithmic frequency scale based on the mel scale, which effectively captures broad perceptual features of human speech, they offer uniform frequency resolution and are primarily sensitive to low-frequency components. This can make them less effective in identifying fine-grained spectral details in higher frequencies where certain spoofing artifacts may reside. However, by using the constant-Q transform (CQT), CQCCs provide changing resolution of frequency with more accuracy at the lower frequencies, enabling a more thorough depiction of minute spectral changes over the whole frequency spectrum. Combining the advantages of MFCCs and CQCCs makes it more resistant to various spoofing strategies, improving detection accuracy and dependability. The suggested approach is shown in Figure 1.

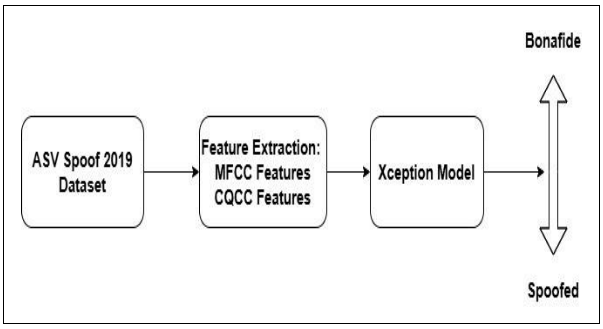

Figure 1: Block Diagram of Proposed Method

1.1 MFCC

By inversely transforming the logarithm of the spectrum, the cepstrum of the time series *y(n)* may be determined. The spectrum in voice processing is often calculated using the “discrete Fourier transform” (DFT), whereas the inverse transform is computed using the “discrete cosine transform” (DCT). The cepstrum offers a modified spectrum representation, capturing the spectral information in a more condensed and frequently decorrelated fashion. In particular, it reduces duplication and efficiently summarizes the most important spectrum properties by converting K Fourier coefficients with $q \ll K$ independent cepstral coefficients. Prior to cepstral analysis, the Mel-cepstrum employs a frequency scale based on auditory significant bands. The resulting features are usually extracted and shown in the Equation (1) and (2) as Mel frequency cepstral coefficients.

$$MFCC(q) = \sum_{n=1}^{N} \log[T(n)] \cos\left[\frac{q\left(n-\frac{1}{2}\right)\pi}{N}\right] \tag{1}$$

And the Mel frequency spectrum is as follow:

$$T(n) = \sum_{k=1}^{K} |Y^{DFT}(k)|^2 G_n(k) \tag{2}$$

$G_n(k)$ indicates the triangle filter function for the n^{th} Mel-scaled bandpass filter, and k represents the DFT index. Using the function $MFCC(q)$, the number of coefficients usually smaller than the number of Mel-filters, and N, is retrieved. q, often falls between 13 and 20. In Figure 2, the MFCC extraction of features process is shown.

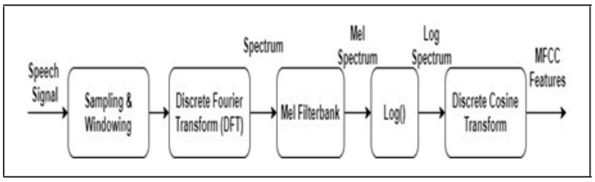

Figure 2: Schematic Diagram of MFCC Feature Extraction.

1.2 CQCC

Youngberg and Boll first introduced the CQT, a time-frequency analysis method, in 1978. [21]. The CQT scale's core frequencies are dispersed geometrically, as opposed to using traditional Fourier-based techniques. When comparedto the Short-time Fourier transform (STFT), the CQT provides higher frequency resolution for lower frequencies and better temporal resolution for higher frequency bands. The procedure of CQCC feature extraction is depicted in Figure 3.

Equation (3) describes the CQCC characteristics that may be retrieved from the audio signals utilizing the Constant Q Transform:

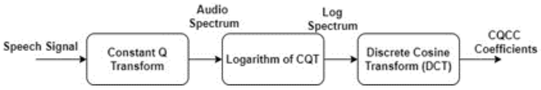

Figure 3: Schematic Diagram of CQCC feature Extraction. The CQT is calculated as follows:

$$C(q,t) = \sum_{r=0}^{K-1} s[r] * w[r-t] * e^{-j2\pi\frac{rq}{Q}} \qquad (3)$$

Where s[r] represent the input speech signal, w[r-t] is the window function, q is the frequency bin in the CQT domain, and $e^{-j2\pi\frac{rq}{Q}}$ represents the frequency-based transformation. After obtaining the CQT, the power spectrum is derived by taking the magnitude of the CQT coefficients. The dynamic range is then compressed by applying a logarithmic function to the power spectrum, as shown in Equation (4):

$$S(q,t) = \log(\,|C(q,t)|)$$

The Discrete Cosine Transformation (DCT) is applied to this log-scaled signal to obtain the CQCC features, as shown in Equation (5):

$$CQCC(n) = \sum_{r=0}^{K-1} S(r)\cos\left(\frac{\pi}{K}\left(r+\frac{1}{2}\right)n\right), for\ n = 0,1,2,\ldots..,N-1 \qquad (5)$$

1.3 Xception Model for Classification

One prominent deep CNN architecture is the Xception model (Extreme Inception), which uses depth-wise separable convolutions [22]. It replaces traditional Inception modules with depth-wise separable convolutions, significantly improving efficiency and performance. Depth-wise separable convolutions, in contrast to conventional convolutional layers, divide the entire process into two separate stages: a pointwise convolution (combining the output) and a depth-wise convolution (spatial filtering). This approach reduces the computational complexity while maintaining accuracy.

MFCCs are highly effective in capturing phonetic information because they mimic how humans perceive sound, focusing on the formant structure and spectral envelope. CQCCs are highly sensitive to pitch-related features due to the constant-Q transform, which offers better frequency resolution in lower-frequency bands. By combining these features, the model becomes more resilient to speech variations. The hybrid features were obtained by concatenating 128 MFCC features and 98 CQCC features, which were then provided as input to the fed into the Xception architecture. The Xception model acts as a feature extractor, and the following Dense layers perform the classification task. The architecture of Xception [22] is shown in Figure 4.

The Xception model is highly effective for tasks like spoof detection in speaker verification systems due to its ability to learn complex patterns efficiently through depthwise separable convolutions. Unlike standard convolutions, which simultaneously capture spatial and channel-wise correlations, Xception decouples these processes. In the context of speaker verification, the model can better distinguish between authentic and spoofed audio features by extracting relevant patterns more precisely, such as voice characteristics across different channels.

The Xception model is divided into Entry Flow, Middle Flow, and Exit Flow. For spoof detection, the Entry Flow helps capture lower-level audio features from inputs, such as spectral properties or feature maps derived from combined MFCC and CQCC features. The Middle Flow, repeated multiple times, progressively extracts higher-level patterns that may indicate spoofing attempts. Global Average Pooling and completely connected layers come after the Exit Flow, which gathers the learned features and makes categorization easier.

To enhance generalization and prevent overfitting, techniques like weight decay and dropout are employed. Additionally, residual connections help maintain gradient flow in the network, ensuring efficient learning even in deep architecture. Because of this, Xception is an effective tool for speaker verification systems to identify minute distinctions between bonafide and spoofed sounds.

In training phase, the model is fine-tuned using an appropriate loss function, such as cross-entropy loss, which is particularly effective for multi-class classification tasks. The Adam optimizer adjusts the learning rate during training to enhance convergence. To guide the training process, a number of hyperparameters are set, such as learning rate, batch size, and epoch count. The model is evaluated using validation data and trained using the training dataset to reduce the likelihood of overfitting.

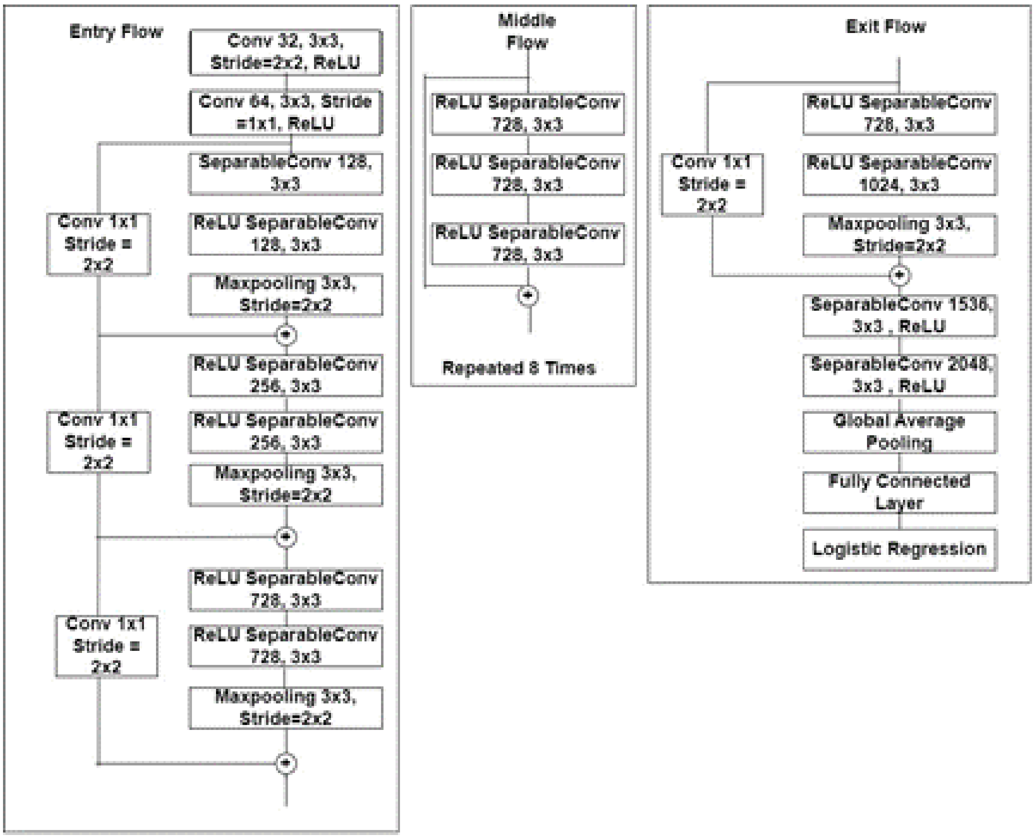
Entry Flow
Conv 32, 3x3, Stride=2x2, ReLU
Conv 64, 3x3, Stride =1x1, ReLU
SeparableConv 128, 3x3
Conv 1x1 Stride = 2x2
ReLU SeparableConv 128, 3x3
Maxpooling 3x3, Stride=2x2
ReLU SeparableConv 256, 3x3
Conv 1x1 Stride = 2x2
ReLU SeparableConv 256, 3x3
Maxpooling 3x3, Stride=2x2
ReLU SeparableConv 728, 3x3
Conv 1x1 Stride = 2x2
ReLU SeparableConv 728, 3x3
Maxpooling 3x3, Stride=2x2
Middle Flow
ReLU SeparableConv 728, 3x3
ReLU SeparableConv 728, 3x3
ReLU SeparableConv 728, 3x3
Repeated 8 Times
Exit Flow
ReLU SeparableConv 728, 3x3
Conv 1x1 Stride = 2x2
ReLU SeparableConv 1024, 3x3
Maxpooling 3x3, Stride=2x2
SeparableConv 1536, 3x3 , ReLU
SeparableConv 2048, 3x3 , ReLU
Global Average Pooling
Fully Connected Layer
Logistic Regression

Figure 4: Architecture of Xception Model [22]

2 Dataset and Evaluation Metrics

The suggested approach was assessed using the ASVspoof 2019 dataset, which comprises two scenarios: logical access (LA) and physical access (PA), as indicated in Table 1. The ASVspoof 2019 LA data collection was utilized. The 2019 edition is the first to concentrate on defenses against the three primary attack types of replay spoofing, VC, and TTS. EER and t-DCF serve as the evaluation metrics.

Table 1: ASVspoof 2019 LA Dataset [23]

Partition	Male	Female	Bonafide	Spoof
Train	8	12	2580	22800
Development	8	12	2548	22296
Evaluation	30	37	7355	63882

4.0.1 EER and tDCF

EER and tDCF are the evaluation measures. The False Rejection Rate (FRR) and the False Acceptance Rate (FAR) are comparable at the EER. In mathematical terms, it can be expressed as:

$$EER = FAR(\theta_{EER}) = FRR(\theta_{EER}) \tag{6}$$

where θ_{EER} is the threshold where the FAR and FRR are equal, and:

$$FAR(\theta) = \frac{Number\ of\ false\ acceptances\ at\ \theta}{Total\ number\ of\ impostor\ attempts}$$

$$FRR(\theta) = \frac{Number\ of\ false\ rejections\ at\ \theta}{Total\ number\ of\ genuine\ attempts}$$

The EER is a commonly used metric in biometric and authentication systems to evaluate the system's accuracy, with a lower EER indicating better performance.

tDCF, introduced as the primary metric for the 2019 challenge, evaluates the tandem performance of ASV and countermeasure (CM) systems by incorporating real-world costs and prior probabilities, reflecting their combined effectiveness in mitigating spoofing attacks which is calculated as:

the normalized minimum t-DCF is defined as:

$$t - DCF_{norm}^{min} = \min\{\alpha P_{miss}^{cm}(s) + P_{fa}^{cm}(s)\} \qquad (7)$$

Here, α is determined by ASV performance (miss, false alarm, and spoof failure rates) and application factors (priors, cost), while $P_{miss}^{cm}(s) + P_{fa}^{cm}(s)$ are the countermeasure miss and false alarm rates at threshold s.

False acceptance arises when an imposter is mistakenly identified as the target speaker during speaker verification, and false rejection occurs when a real speaker is incorrectly identified as an impostor. When it comes to spoof detection, false acceptance happens when a spoofed speech is incorrectly identified as authentic, whereas false rejection happens when a genuine statement is wrongly identified as spoofed. For the combined ASV and anti-spoofing system, an utterance is accepted only if it is identified as both the target speaker and bonafide. Consequently, false rejection occurs when a bonafide utterance from the target speaker is misclassified, and false acceptance happens when an utterance from an impostor or spoof is mistakenly accepted as a bonafide target speaker. EER provides a balanced evaluation point for analyzing system performance across these error types.

RESULTS AND DISCUSSION

The research addressed the challenge of detecting spoofing attempts in automatic speaker-verification systems, utilizing the ASVspoof 2019 dataset. The focus was on extracting two widely recognized audio features: MFCC and CQCC. These features were selected due to their demonstrated efficacy in capturing both spectral and temporal aspects of audio signals. By integrating these two features, the study aimed to exploit their complementary strengths and enhance the model's capacity to differentiate between authentic and artificially generated audio samples.

Table 2: Performance of Proposed System

Feature+Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MFCC and CQCC features, Xception model	92.11	92	93	92

The integrated set of features was input into the Xception model, a sophisti- cated CNN prominent for its well-organized and effective feature learning abilities. A highly effective 92.11% accuracy, 92% precision, 93% recall, and 92% F1-score were all achieved by the model. These metrics show a robust equilibrium between the capacity of the model to recognize authentic samples (high precision) and identify spoofed samples (high recall), which is crucial in this field. Additional assessments using the EER and t-DCF offered more comprehensive insights into the model's resilience as depicted in Table 2. The obtained t-DCF of.0464 and EER of 0.0511 demonstrate that the recommended approach successfully lowers the rates of false rejection and false acceptance, satisfying the exacting requirements of actual speaker verification systems. These outcomes highlight the efficacy of merging MFCC and CQCC features, which capture the complementary aspects of audio signals, and the capability of the Xception model to process high-dimensional input data. Ta- ble 3 shows that this method yields competitive results compared with current techniques, suggesting its potential for practical applications. Figure 5 shows the suggested system's model accuracy and loss.

Table 3: Performance Comparison of the Proposed Approach with Existing Approaches.

Model	t-DCF	EER (%)
LFCC and GMM [24]	0.2116	8.09
CQCC and GMM [24]	0.2366	9.57
DNNS and 1024D [25]	-	15.3
MFCC and GMM [26]	0.1826	7.56
MFCC+ CQCC and Xception (Our Model)	0.0511	0.0464

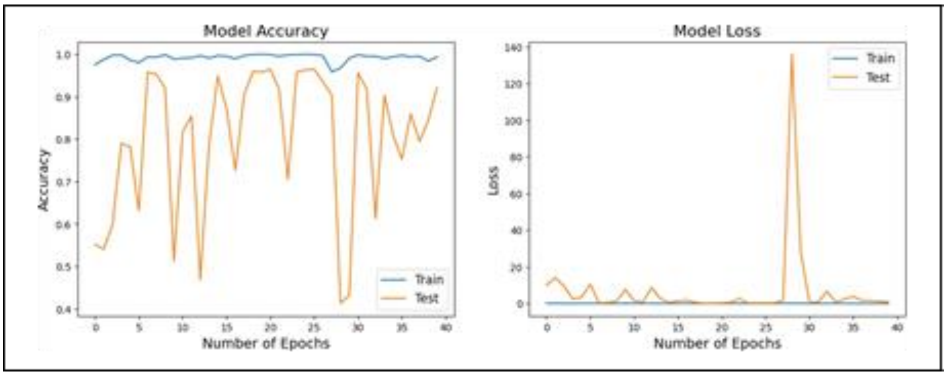

Figure 5: Model Accuracy and Loss

CONCLUSION

A deep convolutional architecture and hybrid cepstral features enable the system to demonstrate great generalization against a range of logical access spoofing attacks, such as VC and TTS synthesis. Using the complementing advantages of MFCC and CQCC as input characteristics to the Xception DL model, a strong spoof detection solution for ASV was suggested in this study. The experimental assessment conducted on the ASVspoof 2019 Logical Access dataset showed the efficacy of the proposed technique with a t-DCF score of 0.0464, an EER of 0.0511, and a high detection accuracy of 92.11%. Compared to previous deep learning-based countermeasures and conventional Gaussian Mixture Model-based countermeasures, the results show a notable improvement.

REFERENCES

[1] D. A. Reynolds, “An overview of automatic speaker recognition technol- ogy,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, May 2002, pp. IV-4072–IV-4075.

[2] A. K. Sarkar and Z.-H. Tan, “Text dependent speaker verification using un- supervised HMM-UBM and temporal GMM-UBM,” in *Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2016, pp. 425–429.

[3] Z. Wu, E. S. Chng, and H. Li, “Conditional restricted Boltzmann machine for voice conversion,” in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2013, pp. 104–108.

[4] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, “Voice conversion using dynamic kernel partial least squares regression,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[5] Z. Wu et al., “Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human perfor- mance,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4,pp. 768–783, Apr. 2016.

[6] Z. Wu et al., “ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge,” in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2037–2041.

[7] P. Korshunov et al., “Overview of BTAS 2016 speaker anti-spoofing com- petition,” in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Niagara Falls, NY, USA, Sep. 2016, pp. 1–6.

[8] T. Kinnunen et al., “The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2–6.

[9] M. Todisco et al., “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1008–1012.

[10] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge,” in *Proc. Conf. Int. Speech Com- mun. Assoc. (INTERSPEECH)*, 2015, pp. 2052–2056.

[11] T. B. Patel and H. A. Patil, “Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech,” in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 2062–2066.

[12] Z. Wu et al., “Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance,” *IEEE/ACM Trans. Audio, Speech, Language Process.*,

vol. 24, no. 4, pp. 768–783, Apr. 2016.

[13] Z. Wu, Zhizheng, et al., “ASVspoof: the automatic speaker verification spoofing and countermeasures challenge,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588-604, 2017.

[14] A. Nautsch et al., “ASVspoof 2019: spoofing countermeasures for the de- tection of synthesized, converted and replayed speech,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[15] X. Liu et al., “Asvspoof 2021: Towards spoofed and deepfake speech de- tection in the wild,” *IEEE/ACM Transactions on Audio, Speech, and Lan- guage Processing*, vol. 31, pp. 2507–2522, 2023.

[16] T. B. Ta et al., “A multi-task conformer for spoofing aware speaker veri- fication,” *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, IEEE, 2022.

[17] H. Sanghvi and S. H. Mankad, “XGB-RFE: An XGBoost Approach for Improved Playback Spoofing Detection in Automatic Speaker Verification Systems Using Recursive Feature Elimination,” *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*, IEEE, 2023.

[18] M. Todisco, H. Delgado, and N. W. D. Evans, “A New Feature for Au- tomatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coeffi- cients,” *Odyssey*, 2016.

[19] D. A. Reynolds, “Gaussian mixture models,” *Encyclopedia of biometrics*,pp. 741.659–663, 2009.

[20] A. Chaiwongyen et al., “Replay attack detection in automatic speaker ver- ification based on resnewt18 with linear frequency cepstral coefficients,” in *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, IEEE, 2021.

[21] M. Todisco, H. Delgado, and N. W. D. Evans, “Constant Q cepstral co- efficients: A spoofing countermeasure for automatic speaker verification,” *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[22] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolu- tions,” in *2017 IEEE Conference on Computer Vision and Pattern Recog- nition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1800–1807.

[23] J. Yang, J. Yamagishi, I. Echizen, and S. Kajita, “Discriminative features based on modified log magnitude spectrum for playback speech detection,” in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, 2020, pp. 1–14.

[24] B. Chettri et al., “Ensemble models for spoofing detection in automatic speaker verification,” *arXiv preprint arXiv:1904.04589*, 2019.

[25] M. Y. Faisal and S. Suyanto, “SpecAugment Impact on Automatic Speaker Verification System,” in *2019 International Seminar on Research of Infor- mation Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indone- sia, 2019, pp. 305-308.

[26] R. K. Bhukya and A. Raj, “Automatic speaker verification spoof detection and countermeasures using gaussian mixture model,” in *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, 2022.