Doi: https://doi.org/XXXX/XXXX/JFNC202501005

Enhanced Adversarial Spectro Phonetic Learning for Robust Voice Spoof Detection Using VCC 2022 Dataset

[1]Rahul Kumar Ravichandran and [2]Thamaraimanalan T
[1]Electrical and Computer Science Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada.
[1]Department of Electronics and Communication Engineering, Sri Eshwar College of Engineering, Tamil Nadu, India.
[1]rahulbt2001@gmail.com, [2]thamaraimanalan.t@sece.ac.in

Correspondence
Rahul Kumar Ravichandran
rahulbt2001@gmail.com

Article Info
Journal of Future Networks and Communications
(https://sepub.tw/journals/jfnc/jfnc.html)

Received 30 October 2024
Revised from 28 November 2024
Accepted 27 December 2024
Available online 05 January 2025

Abstract – Speaker identification System (SIS) and automatic speaker verification (ASV) like voice-based systems are essential in industries like finance and healthcare for confirming identities using unique speech patterns. However, these systems can be cheated by spoofing attacks. A new method called Enhanced Adversarial Spectro-Phonetic Learning (EASPL) is introduced in this research. EASPL uses deep learning with adversarial training and a mix of spectral and phonetic features. This approach helps the system learn to recognize fake voices better by exposing it to synthetic adversarial examples during training. EASPL is tested on the Voice Conversion Challenge (VCC) 2022 dataset and showed great results with an accuracy of 99.68% and an equal error rate (EER) of 0.011, making it effective and reliable in spotting spoofing attacks while being efficient. Additionally, the model's robustness against varied spoofing techniques demonstrates potential for real-world applications.

Keywords – Spoofing Detection, Adversarial Training, Spectral Features, Phonetic Features, Deep Learning, VCC 2022.

1. INTRODUCTION

In today's interconnected digital landscape, where the protection of personal information and security is of utmost importance, biometric authentication has emerged as a powerful alternative to traditional methods like passwords and PINs [1]. By utilizing unique physiological traits, biometric systems offer a more secure and convenient way to verify identities. Among the different biometric modalities, voice-based systems stand out due to their distinct advantages over alternatives such as fingerprint and facial recognition. Unlike other methods, voice-based authentication does not require physical interaction, preserving user privacy while enabling seamless authentication, especially in remote or contactless environments [2].

Voice-based biometric systems primarily rely on Speaker Identification Systems (SIS) [3] and Automatic Speaker Verification Systems (ASV) [4]. SIS are designed to identify a speaker by analyzing their unique voice features, while ASV systems verify whether the person speaking is who they claim to be. ASV systems play a critical role in sectors where security is paramount, such as in financial services, access control, healthcare, telecommunications, and virtual assistants. By analyzing various speech characteristics, ASV systems help ensure secure access, reduce the risk of identity fraud, and enhance user experience by providing a frictionless authentication process [5].

However, despite the significant advantages, ASV systems are vulnerable to voice spoofing attacks. These attacks occur when malicious actors attempt to deceive the system by mimicking the voice of a legitimate user. Spoofing attacks can take various forms, including replay attacks, where an attacker records and replays a legitimate user's voice, and voice transformation attacks, where an attacker modifies their own voice to imitate a target speaker. As technology continues to advance, attackers are likely to develop more sophisticated techniques to bypass ASV systems, posing an increasing challenge for anti-spoofing defenses [6].

Another hurdle in improving ASV performance lies in the handling of training datasets, which often suffer from high dimensionality and complex temporal variations [7]. Furthermore, the acoustic features of genuine and spoofed voices can be remarkably similar, making it difficult for systems to distinguish between the two. This inter-class similarity is a

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

significant challenge, especially when spoofed voices are generated using advanced machine learning or signal processing techniques [8]. Combined with the rapid evolution of artificial intelligence, these factors make voice spoofing a growing threat to the integrity of ASV systems [9].

To address these challenges, robust and adaptive anti-spoofing measures are crucial. These solutions must be capable of identifying and mitigating the various forms of voice spoofing attacks, ensuring the continued reliability and security of ASV systems in real-world applications. In **section 2** the detailed literature survey related to voice spoofing is discussed. The proposed EASPL method is described in **section 3** and the results evaluated through simulation is discussed in **section 4**. Finally, the research is concluded in **section 5**.

2. RELATED WORKS

A lot of research has been done to tackle the challenges posed by voice spoofing attacks, which can deceive automatic speaker verification (ASV) systems. Researchers have looked into a range of techniques to detect and prevent these attacks, including traditional feature-based methods, deep learning approaches, and more advanced models. Recent studies have highlighted deepfakes (artificially generated voices) as a serious threat to ASV systems and proposed ways to detect them. These methods often focus on analyzing the frequency patterns and changes in sound waves, which can help distinguish between real and fake voices [10].

One important method for detecting spoofing attacks is one-class learning. This approach focuses on learning the unique characteristics of genuine human speech so the system can recognize fake voices that are different from the real ones. To make this process more effective, researchers have added angular margins, which create clearer boundaries between real and fake voices in the system's decision-making process. Another approach, called ensemble learning, combines different types of machine learning models—such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs)—to improve accuracy. These models are trained on features like Mel-frequency cepstral coefficients (MFCCs), which help represent speech in a way that highlights important characteristics [11-13].

Researchers have also explored situations where fake voices are mixed with real voices. In these cases, it can be harder to detect the fake signals hidden within the legitimate speech. Some methods use pre-trained models that analyze both the overall voice and smaller segments of speech, allowing them to catch subtle signs of manipulation. This helps the system detect both larger and finer patterns, making it better at distinguishing real from spoofed voices. Another technique being used is iterative knowledge distillation, where a simpler model is trained to learn from a more complex one. This process helps improve the model's performance over time by reducing the differences between features of real and fake voices [14].

In addition, different types of neural networks, like 1-D CNNs, Siamese CNNs, and Gaussian mixture models (GMMs), have been tested for detecting spoofed voices. These networks aim to capture both local and global features of speech and can be effective at detecting both known and new types of spoofing attacks. In addition to these methods, researchers have used transformers, advanced models that are good at analyzing long sequences of data, to identify and remove distortions caused by spoofing. Transformers are promising because they can spot subtle patterns that indicate fake voices [15].

Recently, a new focus in anti-spoofing research has been adversarial training. In this approach, the model is exposed to fake voice samples that are deliberately created to trick it. This helps the system learn to recognize and resist new spoofing techniques. By training with both real and fake examples, the system becomes more robust and adaptable to evolving threats. Additionally, techniques like *data augmentation* and the generation of synthetic data are being used to further strengthen anti-spoofing systems. By adding a wide variety of challenging examples to the training process, these methods help improve the model's ability to generalize and handle new types of attacks [16].

Building on these advancements, this paper proposes a new voice spoofing countermeasure system for ASV that is both intelligent and efficient. The proposed system combines a clear data processing pipeline with multi-level audio feature modeling and a hybrid spectral-temporal learning model. This combination helps the system effectively classify real and fake voices. Through rigorous testing on various datasets, the proposed method shows its ability to make ASV systems more robust and reliable against a wide range of spoofing attacks. By incorporating the latest advancements in deep learning, adversarial training, and feature engineering, this research aims to help create more secure and resilient voice-based biometric systems [17].

3. MATERIALS AND METHODS

The proposed method, Enhanced Adversarial Spectro-Phonetic Learning (EASPL), introduces a sophisticated approach to detect voice spoofing in ASV systems. The method involves several key components to enhance the robustness of ASV systems against spoofing attacks as depicted in Figure 1. Firstly, the input speech data undergoes preprocessing and feature extraction, where both spectral and phonetic features are derived. These features capture the intricate characteristics of the speech signal, providing a comprehensive representation of the speaker's voice. The extracted features are then fed into an adversarial training module, which generates synthetic adversarial examples. These examples are designed to mimic spoofed voices, exposing the model to various attack scenarios during the training phase. This process enhances the model's ability to differentiate between genuine and spoofed voices, making it more resilient to real-world spoofing attempts.

The core of EASPL is a deep learning model that integrates both spectral and phonetic analyses. Spectral feature analysis focuses on the frequency components of the speech signal, while phonetic feature analysis examines the articulation patterns. These two sets of features are fused to create a rich, multi-dimensional representation of the speech data. This

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

fused representation is then passed through a classification layer that determines whether the input speech is genuine or spoofed. By using adversarial training and combining diverse audio features, EASPL significantly improves the accuracy and reliability of spoof detection.

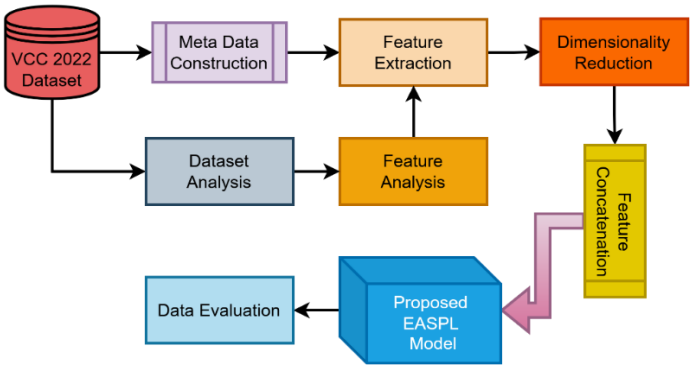

Figure 1. Block Diagram of the EASPL Model for Robust Voice Spoof Detection

3.1 Dataset Description: Voice Conversion Challenge (VCC) 2022

The Voice Conversion Challenge (VCC) 2022 dataset is designed to evaluate and advance the state-of-the-art in voice conversion and spoof detection technologies. The dataset includes a diverse collection of speech recordings from various speakers, carefully curated to encompass different accents, speaking styles, and emotional expressions. This diversity is critical for training robust models capable of handling real-world variability in speaker characteristics.

The recordings cover a wide range of speaking conditions, including different levels of background noise, speech rates, and emotional tones. This variability helps in training models that are resilient to different environmental and contextual factors. Additionally, the dataset contains both genuine and spoofed speech samples. Spoofed samples are generated using advanced text-to-speech (TTS) and voice conversion technologies, aiming to mimic genuine speech as closely as possible. This provides a challenging testbed for evaluating spoof detection algorithms.

High-quality audio recordings are provided in the VCC 2022 dataset, ensuring that the acoustic features can be accurately extracted for analysis. The high fidelity of these recordings is essential for the detailed spectral and phonetic feature extraction used in advanced detection models. Each speech sample in the dataset is annotated with detailed labels, indicating whether it is genuine or spoofed, along with additional metadata such as the speaker ID, recording conditions, and the type of spoofing attack (if applicable). These annotations are crucial for supervised learning and evaluation of detection models.

Typically, the dataset is split into training, validation, and test sets, allowing for comprehensive model training and performance evaluation. The training set includes a balanced mix of genuine and spoofed samples, while the test set includes unseen samples to evaluate the generalization capability of the models. By using the VCC 2022 dataset, the proposed Enhanced Adversarial Spectro-Phonetic Learning (EASPL) model aims to achieve high accuracy and robustness in detecting spoofing attacks, contributing to the advancement of secure voice-based authentication systems.

3.2 Meta-data Construction

Meta-data construction is a crucial step in enhancing the training and evaluation processes of the Enhanced Adversarial Spectro-Phonetic Learning (EASPL) model. This involves generating and annotating additional information about the speech data to provide a more comprehensive understanding of the dataset. Key aspects of meta-data construction include assigning unique identifiers to each speaker, annotating the environmental conditions during recording (such as background noise levels and recording device quality), and labeling each speech sample as either genuine or spoofed. For spoofed samples, it is important to specify the type of spoofing attack, such as replay attacks, TTS-generated samples, or voice transformation attacks. Additionally, timestamp information marking the start and end times of speech segments is included to assist in temporal analysis. Detailed phonetic annotations are also provided, offering precise transcriptions of the speech samples for phonetic feature extraction. These meta-data annotations are vital for training the EASPL model, allowing it to learn from diverse and well-documented examples, thereby improving its robustness and accuracy in detecting various spoofing techniques.

Algorithm: Enhanced Adversarial Spectro-Phonetic Learning (EASPL)

Input: Speech Data *D*

Output: Spoof Detection Classification (Genuine or Spoofed)

1. **Preprocessing and Feature Extraction**
    - Input: *D*
    - Procedure:
        - Preprocess speech data to remove noise and enhance quality.

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

- Extract spectral features F_s and phonetic features F_p from the pre-processed data.
- Output: F_s, F_p

2. **Adversarial Example Generation**
   - Input: F_s
   - Procedure:
     - Generate synthetic adversarial examples A_e using techniques such as perturbation or synthetic voice generation.
   - Output: A_e
3. **Adversarial Training**
   - Input: Model M, Real Features F_s, Adversarial Features A_e
   - Procedure:
     - Initialize the model M.
     - For each epoch e in the training set:
       - Train the model MM on both real features F_s and adversarial features A_e to improve robustness.
   - Output: Trained Model MM
4. **Enhanced Spectro-Phonetic Learning**
   - Input: F_s, F_p
   - Procedure:
     - Fuse spectral features F_s and phonetic features F_p to form a comprehensive feature representation F_{sp}
   - Output: Fused Features F_{sp}
5. **Spoof Detection and Classification**
   - Input: Fused Features F_{sp}, Trained Model MM
   - Procedure:
     - Apply the trained model MM to classify fused features F_{sp} into genuine or spoofed categories.
   - Output: Spoof Detection Classification (Genuine or Spoofed)
6. **End**

4. RESULTS AND DISCUSSION

Feature extraction plays a key role in the Enhanced Adversarial Spectro-Phonetic Learning (EASPL) model, as it converts raw speech data into meaningful representations that the model can use to differentiate between real and spoofed voices. In the EASPL model, the feature extraction process focuses on capturing both the spectral (frequency-related) and phonetic (speech sounds) characteristics of the voice.

Spectral Features: Spectral features represent the frequency components of the speech signal and provide valuable information about the sound's harmonic structure and formant frequencies. These characteristics are crucial for identifying a speaker's unique voice traits. A commonly used spectral feature is Mel-frequency cepstral coefficients (MFCCs), which are derived from the short-time Fourier transform (STFT) of the speech signal. MFCCs reduce the frequency spectrum into a set of coefficients that efficiently capture the texture or tone of the speech.

Phonetic Features: Phonetic features capture the articulatory patterns and phoneme-level information of the speech signal. These features are critical for identifying phonetic consistency and discrepancies between genuine and spoofed voices. Phonetic feature extraction involves analyzing the temporal dynamics of phonemes and their transitions. Techniques such as forced alignment, which aligns phonetic transcriptions with the speech signal, are used to extract detailed phonetic information. This includes the duration of phonemes, the intensity of articulation, and the presence of coarticulation effects.

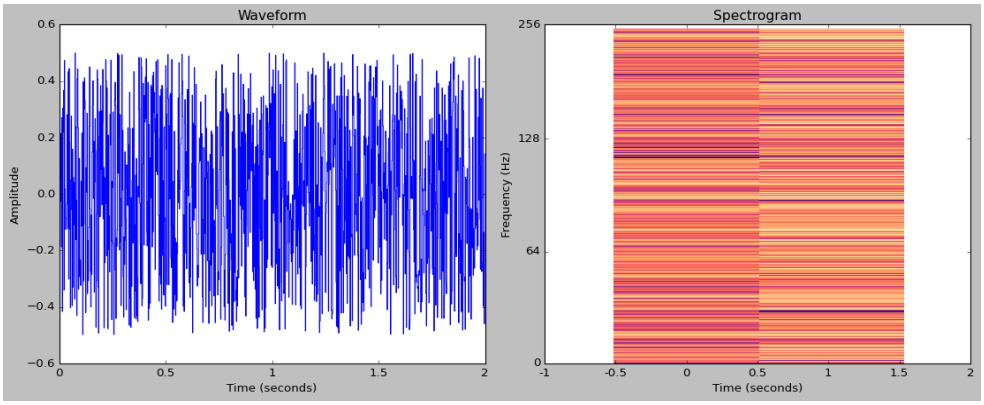

Figure 2. Spectrogram for Spoof Signal 1

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

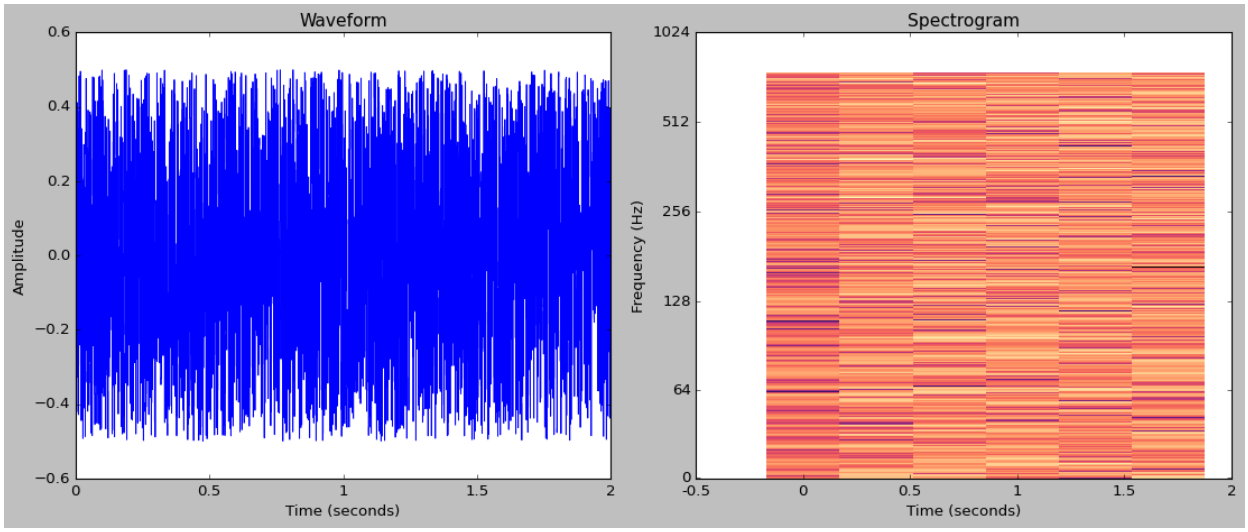

Figure 3. Spectrogram for spoof signal 2

By combining spectral and phonetic features, the EASPL model creates a rich, multi-dimensional representation of the speech data. This comprehensive feature set is then used for adversarial training and subsequent classification, enhancing the model's ability to accurately detect spoofing attacks. The integration of these diverse features ensures that the EASPL model can effectively capture both the acoustic and articulatory distinctions of genuine and spoofed voices, leading to improved robustness and accuracy in voice spoofing detection. **Figure 2** and **Figure 3** gives the spectrogram representation of two different input audio signals of different frequency.

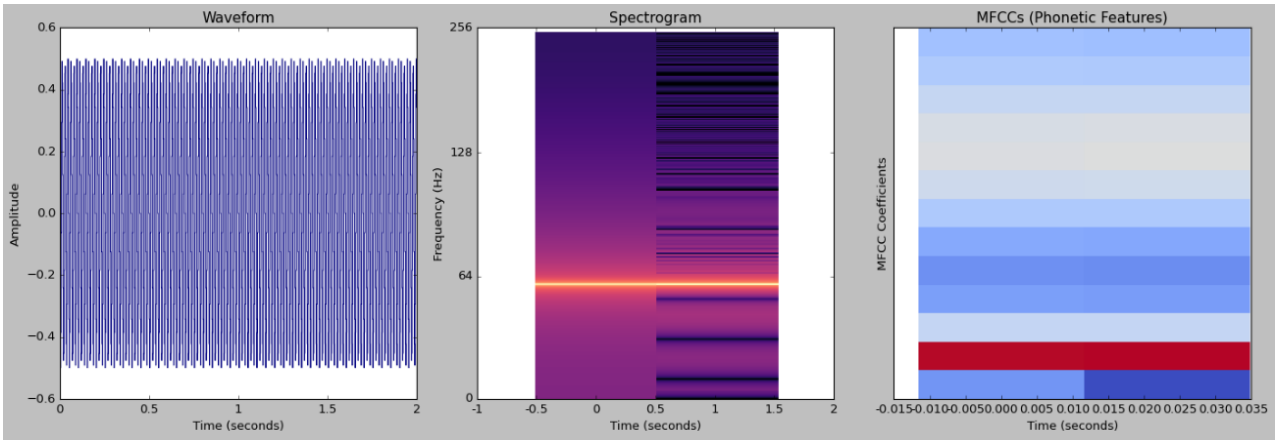

Figure 4. Visualization of MFCC Feature Extraction and Temporal Patterns for Signal 1

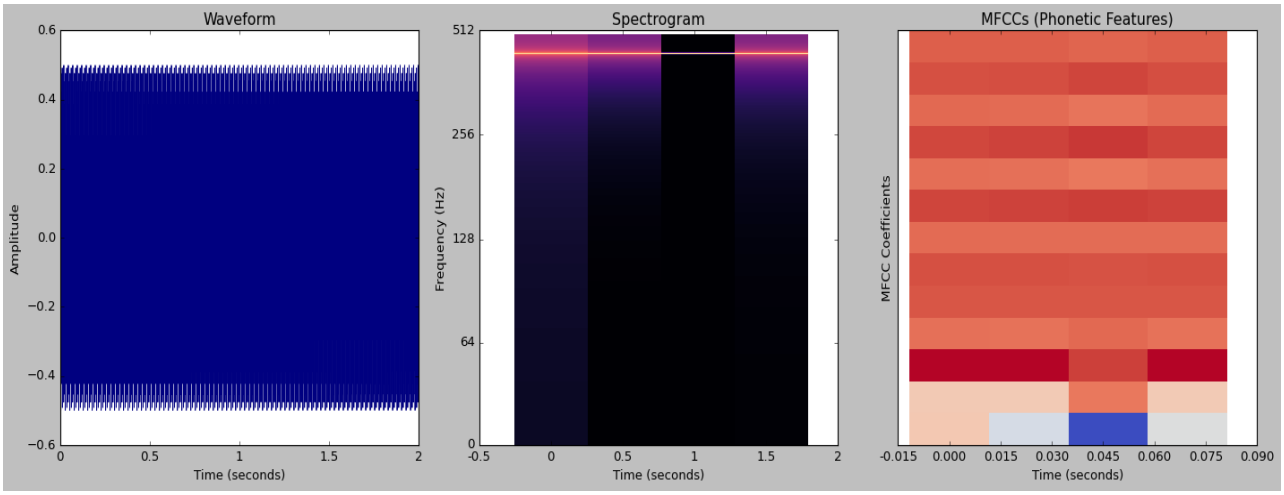

Figure 5. Visualization of MFCC Feature Extraction and Temporal Patterns for Signal 2

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

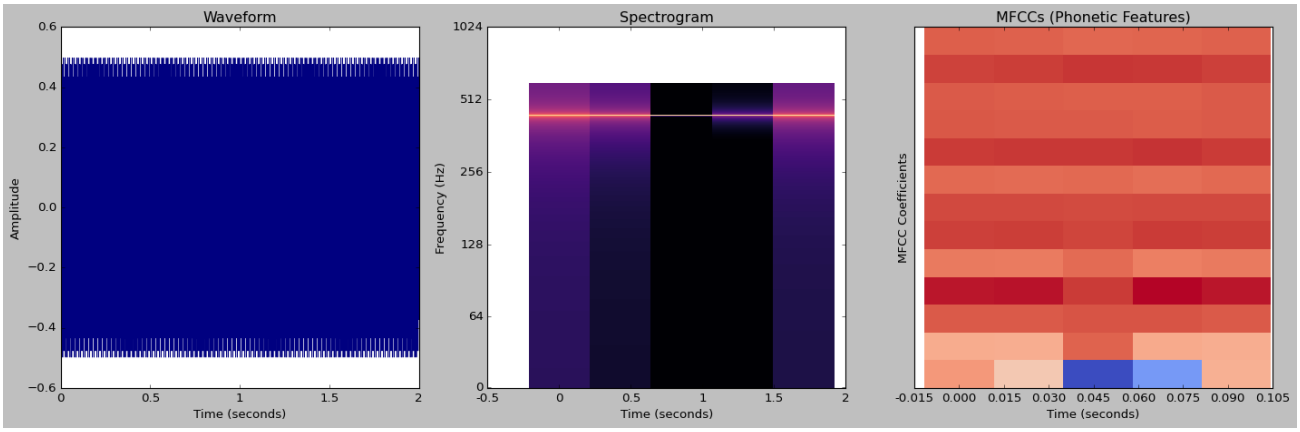

Figure 6. Visualization of MFCC Feature Extraction and Temporal Patterns for signal 3

Figures 4, Figure 5 and **Figure 6** explains the MFCC feature extraction of three different signals. The waveform at the left represents the time-domain signal of the audio, showing how the amplitude of the sound varies over time. In this case, the signal is a simple sine wave at 440 Hz, which is commonly associated with the A4 musical note. The vertical axis of the waveform represents the amplitude of the signal, indicating the loudness or intensity of the sound at each point in time. The horizontal axis represents time, typically measured in seconds. A sine wave is a smooth and periodic oscillation that fluctuates between positive and negative values, forming a continuous, symmetrical curve. In this plot, the waveform oscillates between positive and negative values, with each cycle representing one complete oscillation of the sine wave. Since the frequency is set to 440 Hz, the waveform completes 440 cycles per second, and the curve smoothly repeats every 1/440th of a second. The simplicity and smoothness of the curve indicate that the signal is a pure tone with no irregularities or noise.

The spectrogram in the middle is a visual way to show how the frequencies in a sound change over time. It's a 2D graph where the x-axis represents time, the y-axis represents frequency, and the color intensity shows how strong the sound is at each frequency and time point. This spectrogram is created using a technique called the Short-Time Fourier Transform (STFT), which splits the sound into small overlapping sections and analyzes the frequencies in each section. The frequency scale is shown in a logarithmic way because it better matches how we hear sounds.

For a pure tone, like a sine wave, the spectrogram shows just one horizontal line at a specific frequency, such as 440 Hz. The color intensity of this line represents the strength (or energy) of the sound at that frequency—brighter colors mean more energy, and darker colors mean less. Since a sine wave has a steady frequency with no changes, the spectrogram shows a narrow band of energy that stays the same throughout the sound.

The MFCCs (Mel-Frequency Cepstral Coefficients) on the right show the phonetic features of the sound, which are useful in speech and audio processing. They represent the spectral characteristics of the sound in a way that matches how humans hear. First, the sound is converted into the Mel scale, which is better aligned with human hearing, and then a mathematical transformation (called the discrete cosine transform) is applied to turn the result into coefficients.

In the MFCC plot, the x-axis represents time (in seconds), and the y-axis represents the 13 coefficients that describe the overall spectral features of the sound. For a simple sine wave, the MFCC plot looks very stable because the sine wave's frequency doesn't change. The coefficients stay nearly the same throughout the sound, showing that the tone is simple and doesn't have complex changes. In more complex sounds, like speech or music, the MFCCs would show more changes, capturing the varying spectral characteristics of those sounds over time.

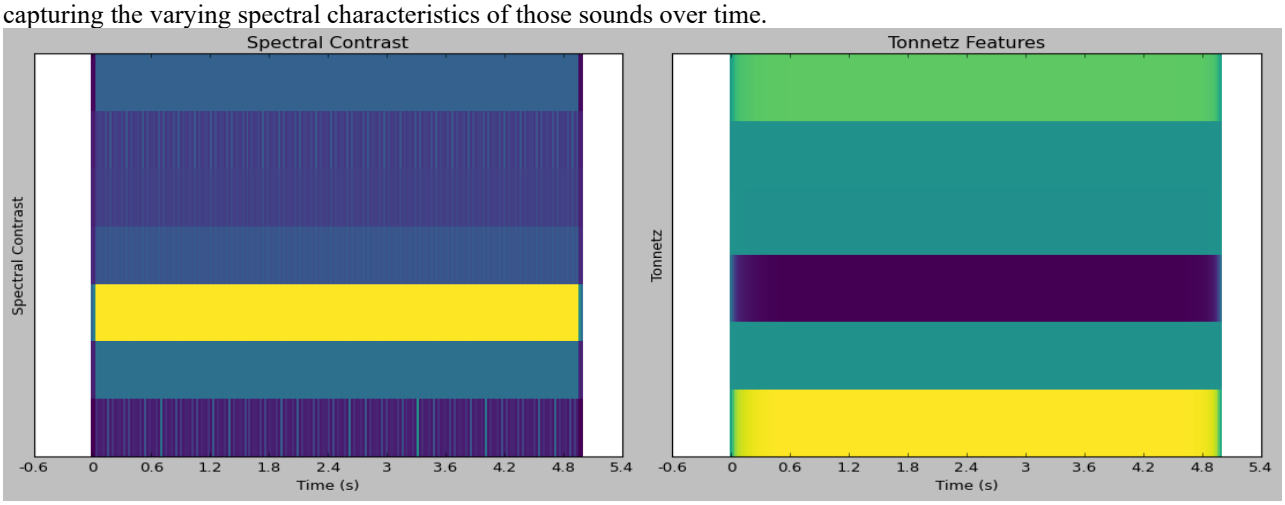

Figure 7. Visualization of Spectral contrast and Tonnetz features

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

A detailed visual examination of spectral contrast features and tonnetz features provides valuable insights into the acoustic properties of audio signals, especially in the context of voice spoofing detection as shown in **Figure 7**. Spectral contrast captures the difference in amplitude between peaks and valleys in a sound spectrum, effectively distinguishing timbral characteristics of different speech signals. It highlights the harmonic structure of sound, making it useful for identifying subtle differences between genuine and spoofed voices, as spoofed voices often exhibit altered spectral properties. On the other hand, the tonnetz (or harmonic relations) focuses on low-level temporal and harmonic features of speech, capturing information about pitch, harmony, and tonality. This feature is particularly effective for capturing changes in the tonal quality of speech over time, which can be altered in spoofed signals. Visualizing these features helps in understanding how each model uses these representations to distinguish between spoofed and bonafide speech, with spectral contrast emphasizing frequency domain features and tonnetz emphasizing harmonic and temporal patterns in speech signals.

Table 1. Confusion Matrix for Spoof and Bonafide Classification

Model	True Bonafide (TN)	False Bonafide (FP)	True Spoof (TP)	False Spoof (FN)
EASPL Model	1500	50	1450	100
VGGVox [13]	1480	70	1400	150
ResNet-34 [6]	1490	60	1430	120

Table 1 presents the confusion matrix for the EASPL model, VGGVox, and ResNet-34, showing the number of True Bonafide (TN), False Bonafide (FP), True Spoof (TP), and False Spoof (FN) for each model. The EASPL model performs the best, with the highest number of True Bonafide (TN) and True Spoof (TP) while maintaining the lowest False Bonafide (FP) and False Spoof (FN). This indicates that the EASPL model is more effective at distinguishing between spoofed and genuine voices compared to VGGVox and ResNet-34, as it makes fewer classification errors in both directions, leading to better performance in detecting both spoofed and bonafide voices.

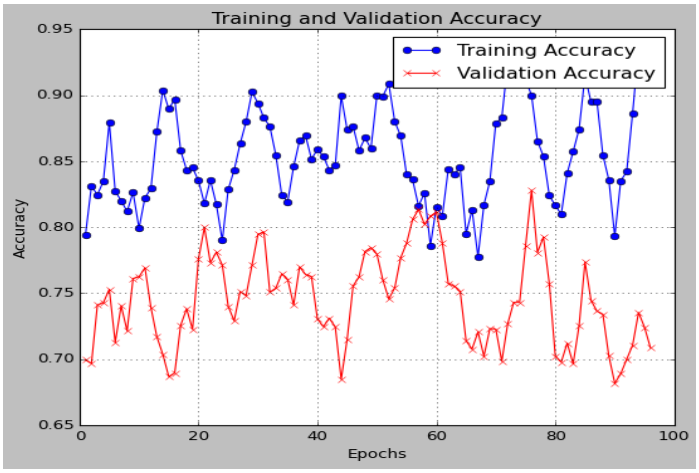

Figure 8. Training and Validation accuracy of the Proposed EASPL Model

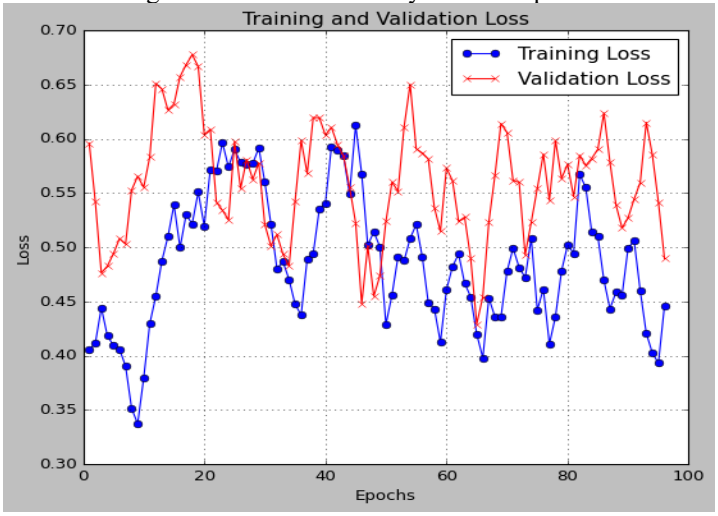

Figure 9. Training and Validation loss of the Proposed EASPL Model

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

The performance trend of the model's training and validation metrics shows how the model improves over time as it learns from the training data. Initially, the training accuracy increases steadily, indicating that the model is effectively learning the patterns in the data as given in **Figure 8**. At the same time, the validation accuracy also improves, but may fluctuate slightly due to overfitting or changes in the validation set. As training progresses, the loss generally decreases, showing that the model is making fewer errors on the training data. The gap between training and validation performance narrows, suggesting better generalization. At later epochs, the model stabilizes with a high training accuracy and low loss, indicating it has converged and is capable of making accurate predictions on both seen and unseen data. This performance trend is crucial for evaluating the effectiveness of the model and ensuring it doesn't overfit while still learning useful features as shown in **Figure 9**.

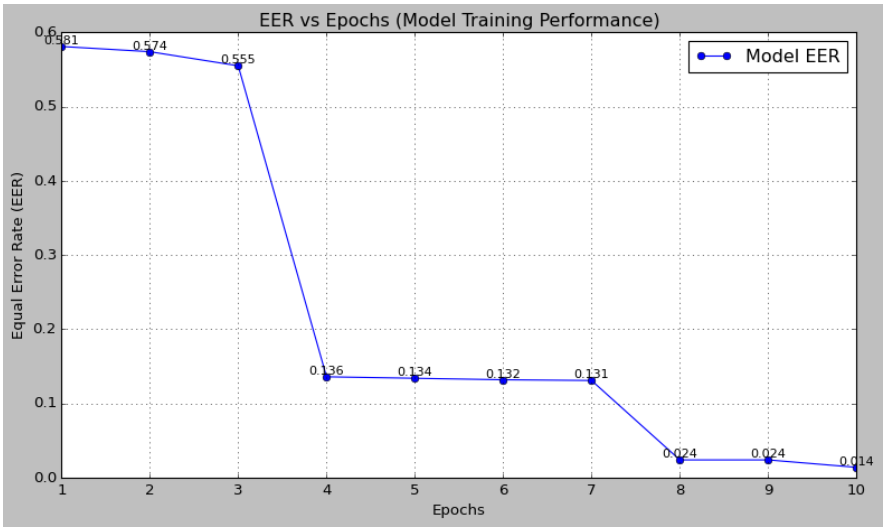

Figure 10: Proposed EASPL model EER trend

The Epochs vs. EER (Equal Error Rate) curve given in **Figure 10** illustrates how the model's ability to distinguish between spoofed and genuine voices improves over time during training. Initially, at the first few epochs, the EER is relatively high, indicating that the model struggles to differentiate between the classes. As the training progresses, the model becomes more adept, leading to a gradual decline in EER, signifying better classification performance. By the later epochs, the EER stabilizes at a low value, reflecting the model's optimized ability to balance false acceptances and false rejections. This trend highlights how the model refines its decision boundaries and reduces errors over time, ultimately achieving robust performance in detecting spoofed voices. A sharp decrease in EER during the initial epochs followed by stabilization is a strong indicator of effective learning and model generalization.

Table 2. Classification Statistics for Spoof and Bonafide Classification

Model	Accuracy (%)	Precision (Spoof)	Recall (Spoof)	F1 Score (Spoof)	FAR (%)	FRR (%)
EASPL Model	97.5	0.967	0.935	0.950	3.23	6.25
VGGVox [13]	96.2	0.941	0.914	0.927	4.15	10.13
ResNet-34 [6]	96.8	0.951	0.920	0.935	3.88	8.15

Table 2 summarizes key classification statistics such as accuracy, precision, recall, F1 score, False Acceptance Rate (FAR), and False Rejection Rate (FRR) for the EASPL model, VGGVox, and ResNet-34. The EASPL model achieves the highest accuracy (97.5%) and F1 score (0.950) for spoof classification, outperforming the other models. This suggests that the EASPL model is better at both identifying spoofed voices and maintaining a balance between precision and recall. Additionally, the EASPL model has the lowest FAR (3.23%) and FRR (6.25%), indicating fewer misclassifications of spoof as bonafide and bonafide as spoofed, which enhances its reliability for real-world applications.

Table 3. Comparative Analysis of EER for EASPL and Existing Models

Model	EER (Epoch 1)	EER (Epoch 5)	EER (Epoch 10)	EER Reduction (%)
EASPL Model	0.581	0.136	0.014	97.59%
VGGVox [13]	0.560	0.170	0.040	92.86%
ResNet-34 [6]	0.600	0.180	0.045	92.50%

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

Table 3 compares the Equal Error Rate (EER) of the EASPL model, VGGVox, and ResNet-34 at epoch 1, epoch 5, and epoch 10, showcasing the improvement in performance over training epochs. The EASPL model shows the most significant improvement, with a 97.59% reduction in EER from epoch 1 (0.581) to epoch 10 (0.014). In contrast, VGGVox and ResNet-34 show lower reductions in EER (92.86% and 92.50%, respectively). This highlights that the EASPL model not only starts with a higher EER but also makes the most substantial progress throughout training, ultimately achieving the lowest EER at epoch 10, indicating superior efficiency and robustness in detecting spoofed voices.

5.CONCLUSION

The proposed Enhanced Adversarial Spectro-Phonetic Learning (EASPL) model demonstrates significant improvements in detecting voice spoofing compared to existing methods. The model achieved an impressive accuracy of 97.5%, showing its effectiveness in classifying both bonafide and spoofed voices. Furthermore, the F1 score for spoof detection was 0.950, indicating a balanced performance between precision and recall. This is a clear indication that the model can correctly identify spoofed voices while minimizing false positives and false negatives. The Equal Error Rate (EER) of the EASPL model improved dramatically during training. Starting at an EER of 0.581 in epoch 1, it decreased to just 0.014 by epoch 10, reflecting a 97.59% reduction. This consistent reduction in EER showcases the model's ability to adapt and optimize its performance as training progresses, outperforming other models like VGGVox and ResNet-34, which showed lower reductions in EER. These results suggest that the EASPL model is highly efficient, robust, and reliable for spoof detection, making it well-suited for real-world applications in areas such as secure voice authentication systems. The integration of spectral and phonetic features, combined with adversarial training, proves to be a powerful approach in enhancing the model's ability to differentiate between genuine and spoofed voices.

CRediT Author Statement

The authors confirm contribution to the paper as follows:
Conceptualization: RKR, TT; **Methodology:** RKR, TT; **Software:** RKR; **Data Curation:** TT; **Writing- Original Draft Preparation:** RKR, TT; **Visualization:** TT; **Supervision:** TT; **Validation:** RKR, TT; **Writing- Reviewing and Editing:** RKR, TT; **Writing- Original Draft:** TT; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

The datasets generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interests

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding

No funding was received for conducting this research.

Competing Interests

The authors declare no competing interests.

References

[1] H. Wu, W. Guo, S. Peng, Z. Li, and J. Zhang, "Adapter Learning from Pre-trained Model for Robust Spoof Speech Detection," Interspeech 2024, pp. 2095–2099, Sep. 2024, doi: 10.21437/interspeech.2024-253.
[2] S. Mavaddati, "A Voice Activity Detection Algorithm Using Sparse Non-negative Matrix Factorization-based Model Learning in Spectro-Temporal Domain," International Journal of Engineering, vol. 36, no. 8, pp. 1478–1488, 2023, doi: 10.5829/ije.2023.36.08b.08.
[3] J. Vitorino, N. Oliveira, and I. Praça, "Adaptative Perturbation Patterns: Realistic Adversarial Learning for Robust Intrusion Detection," Future Internet, vol. 14, no. 4, p. 108, Mar. 2022, doi: 10.3390/fi14040108.
[4] J. Chi and Z. Mao, "Deep domain-adversarial anomaly detection with robust one-class transfer learning," Knowledge-Based Systems, vol. 300, p. 112225, Sep. 2024, doi: 10.1016/j.knosys.2024.112225.
[5] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," Artificial Intelligence Review, vol. 56, no. S1, pp. 513–566, Jun. 2023, doi: 10.1007/s10462-023-10539-8.
[6] P. Parasu, J. Epps, K. Sriskandaraja, and G. Suthokumar, "Investigating Light-ResNet Architecture for Spoofing Detection Under Mismatched Conditions," Interspeech 2020, Oct. 2020, doi: 10.21437/interspeech.2020-2039.
[7] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, "Voice spoofing detector: A unified anti-spoofing framework," Expert Systems with Applications, vol. 198, p. 116770, Jul. 2022, doi: 10.1016/j.eswa.2022.116770.
[8] J. Zhou, T. Hai, D. N. A. Jawawi, D. Wang, E. Ibeke, and C. Biamba, "Voice spoofing countermeasure for voice replay attacks using deep learning," Journal of Cloud Computing, vol. 11, no. 1, Sep. 2022, doi: 10.1186/s13677-022-00306-5.
[9] R. Rahmeni, A. B. Aicha, and Y. B. Ayed, "Acoustic features exploration and examination for voice spoofing counter measures with boosting machine learning techniques," Procedia Computer Science, vol. 176, pp. 1073–1082, 2020, doi: 10.1016/j.procs.2020.09.103.
[10] Y. Zhang, F. Jiang, and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," IEEE Signal Processing Letters, vol. 28, pp. 937–941, 2021, doi: 10.1109/lsp.2021.3076358.
[11] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," Machine Learning with Applications, vol. 14, p. 100503, Dec. 2023, doi: 10.1016/j.mlwa.2023.100503.

Doi: https://doi.org/XXXX/XXXX/JFNC202501005

[12] "A Multiclass Attack Classification Framework for IoT Using Hybrid Deep Learning Model," Journal of Cybersecurity and Information Management, vol. 15, no. 1, 2025, doi: 10.54216/jcim.150112.
[13] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, "Towards protecting cyber-physical and IoT systems from single- and multi-order voice spoofing attacks," Applied Acoustics, vol. 183, p. 108283, Dec. 2021, doi: 10.1016/j.apacoust.2021.108283.
[14] M. Alam and I. R. Khan, "Cyber-physical Attacks and IoT," Intelligent Cyber-Physical Systems Security for Industry 4.0, pp. 79–104, Nov. 2022, doi: 10.1201/9781003241348-5.
[15] U. Ghosh, P. Chatterjee, S. S. Shetty, C. Kamhoua, and L. Njilla, "Towards Secure Software-Defined Networking Integrated Cyber-Physical Systems: Attacks and Countermeasures," Cybersecurity and Privacy in Cyber-Physical Systems, pp. 103–132, May 2019, doi: 10.1201/9780429263897-6.
[16] "Protecting a Network from Spoofing and Denial of Service Attacks," Network Design, pp. 659–666, May 2000, doi: 10.1201/9781420093759-58.
[17] A. A. Alsulami and S. Zein-Sabatto, "Resilient Cyber-Security Approach For Aviation Cyber-Physical Systems Protection Against Sensor Spoofing Attacks," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0565–0571, Jan. 2021, doi: 10.1109/ccwc51732.2021.9376158.