A Comparative Evaluation on Data Transformation Approach for Artificial Speech Detection

Choon Beng Tan[1] *and Mohd Hanafi Ahmad Hijazi*[1*]

[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Abstract. The rise of voice biometrics has transformed user authentication and offered enhanced security and convenience while phasing out less secure methods. Despite these advancements, Automatic Speaker Verification (ASV) systems remain vulnerable to spoofing, particularly with artificial speech generated swiftly using advanced speech synthesis and voice conversion algorithms. A recent data transformation technique achieved an impressive Equal Error Rate (EER) of 1.42% on the ASVspoof 2019 Logical Access Dataset. While this approach predominantly relies on Support Vector Machine (SVM) as the backend classifier for artificial speech detection, it is vital to explore a broader range of classifiers to enhance resilience. This paper addresses this research gap by systematically assessing classifier efficacy in artificial speech detection. The objectives are twofold: first, to evaluate various classifiers, not limited to SVM, and identify those best suited for artificial speech detection; second, to compare this approach's performance with existing methods. The evaluation demonstrated SVM-Polynomial as the top-performing classifier, surpassing the end-to-end learning approach. This work contributes to a deeper understanding of classifier efficacy and equips researchers and practitioners with a diversified toolkit for building robust ASV spoofing detection systems.

1 Introduction

Speaker recognition, including Automatic Speaker Verification (ASV) system, has diverse applications, from voice-controlled devices to forensics [1]. However, it faces vulnerabilities, notably from spoofing attacks, including replay and artificial speech attacks. Artificial speech, produced using advanced technology, can deceive ASV systems, presenting a serious security risk. These attacks are particularly concerning because they can mimic human speech with remarkable accuracy, challenging traditional speaker recognition systems.

One notable approach that has garnered attention in addressing this vulnerability employs data transformation techniques for artificial speech detection [2]. This approach has exhibited impressive results, achieving a remarkably low Equal Error Rate (EER) of 1.42% on the ASVspoof 2019 Logical Access Dataset [3]. However, it is important to underscore that this

[*] Corresponding author: hanafi@ums.edu.my

method predominantly relies on a single classifier, Support Vector Machine (SVM), as the backend classifier for artificial speech detection. Despite its success, this approach has not undergone extensive exploration of alternative classifiers or demonstrated the same level of effectiveness when faced with different ASV spoofing scenarios.

This paper addresses the need to better understand classifier effectiveness in artificial speech detection. It has two main goals: to evaluate various classifiers, including SVM, for artificial speech detection, and to assess the performance of a data transformation approach compared to existing methods. The motivation behind this work is the demand for stronger defenses against artificial speech attacks and the importance of exploring a variety of classifiers that can adapt to evolving threats. The study offers a comprehensive analysis of classifier performance, highlighting their strengths and weaknesses. The paper's structure includes Sections 2 (related works), 3 (comparative classifier evaluation), and 4 (conclusion).

2 Related Works

Artificial speech detection techniques have witnessed significant evolution, falling into three primary categories: classic machine learning, end-to-end learning, and hybrid methodologies. Classic machine learning entails the manual crafting and extraction of predetermined features from data samples, which are then subject to separate classification modules [4,5]. In contrast, end-to-end learning orchestrates the automatic and joint identification and learning of all data sample features to determine their class labels. Unlike classic machine learning, where feature extraction and classification are distinct, end-to-end learning encompasses the entire learning process from input data to output prediction.

On the other hand, a hybrid approach harnesses the strengths of classic machine learning and end-to-end learning techniques [6,7]. This fusion aims to enhance artificial speech detection by capitalizing on both paradigms' advantages, providing a comprehensive approach to combat artificial speech attacks [7]. Existing research underscores the hybrid approach's propensity to outperform traditional machine learning or end-to-end learning [1].

Recent advancements in artificial speech detection have significantly contributed to fortifying the security of ASV systems, which are susceptible to spoofing attacks involving artificial speech. Notably, one recent study has explored different Support Vector Machine (SVM) kernels for artificial speech detection [2]. This recent work focused on evaluating SVM kernels in conjunction with a diverse set of fused features. Additionally, various SVM settings such as normalization and probability estimates were compared. This recent work aimed to identify the optimal SVM kernels for artificial speech detection by conducting an in-depth analysis of feature-kernel combinations. The findings highlighted the exceptional performance of the polynomial kernel, achieving an impressively low Equal Error Rate (EER) of 1.42%, underscoring the importance of handcrafted features in the investigation. This study significantly enhances our understanding of SVM kernel effectiveness in artificial speech detection and contribute to the ongoing efforts to fortify voice biometric applications and strengthen ASV systems against emerging threats.

3 Comparative Evaluation

In Section 3, we conduct a comparative evaluation of artificial speech detection using features optimized in our previous SVM-focused work. These features include Mel-Frequency Cepstral Coefficients (MFCC), Hexadecimal Frequencies, and image-based features. We employ three core classifiers: SVM, Random Forest (RF), and the Multi-Layer Perceptron (MLP). SVM is a continuity choice from our prior work, RF is selected for its ASVspoof

2015 success, and MLP represents a deep learning approach. While consistent with our previous work, we remain open to exploring additional classifiers in future research.

3.1 Feature Engineering Using Data Transformation Approach

MFCCs are effective for speech analysis and a crucial part of this study. The MFCC extraction process involves several steps: the input signal is divided into short frames, the discrete Fourier transform (DFT) calculates the power spectrum, the amplitude is logged to create the log-amplitude spectrum, mel-scaling is applied using a mel filterbank to generate the mel spectrum, and DCT produces the MFCC coefficients. The first 13 coefficients are especially informative about formants and spectral characteristics [8].

Hexadecimal representation for audio data was included as it provides a more human-friendly format compared to binary. This approach extracts text-based features from the hexadecimal representation of audio data to create a feature space. Drawing inspiration from previous studies [5,9], which effectively counted opcode occurrences in executable files for malware classification, we adapted this method for artificial speech detection. Leveraging the hexadecimal representation of speech, our study aimed to extract features capable of distinguishing between bonafide and spoofed speech. Anomalies in the occurrences of specific hexadecimal values, ranging from 00 to FF, in artificial speech data were used as indicators to differentiate authentic from counterfeit voices.

In addition, spectrogram and MFCC representations as images was used in this study for artificial speech detection. While these are typically used for audio analysis, we treated them as images [10,11]. We created these images using Python libraries like pyplot and librosa. From these images, we extracted two types of image-based features: color layout filter (CLF) and local binary patterns (LBP). We used Weka's CLF implementation to get 33 CLF features [12]. For LBP, we followed the original LBP settings [13], using a neighborhood radius (r) of 1 and considering 8 neighboring pixels in a 3×3-pixel window. This process generated 256 LBP features.

In this evaluation, features including MFCC coefficients, hexadecimal-based features, and image-based features (CLF and LBP) were combined into a unified input for artificial speech detection. This fusion strategy aims to provide a rich set of information for classifiers, enhancing their ability to differentiate between genuine and spoofed speech.

3.2 Classifiers

Like in our prior work, we considered four SVM kernels: radial basis function (RBF), linear, polynomial, and sigmoid for this evaluation. The RBF kernel, commonly used as the default kernel in machine learning tools like Weka and sklearn, is a real-valued function for building function estimates. The linear kernel is well-suited for linearly separable data, allowing a straight-line separation in two-dimensional graphs. The polynomial kernel introduces non-linearity by representing feature vectors as polynomials, a common practice in image processing. The sigmoid kernel resembles a two-layer perceptron model and functions as an activation function in neural networks. More in-depth information on these kernels can be found in references [14].

RF was considered in this evaluation due to its demonstrated strong performance in the ASVspoof 2015 challenge [4]. RF is an ensemble learning model that employs decision trees for classification and regression. It enhances prediction accuracy and stability by combining multiple decision trees trained on random subsets of the data through a voting mechanism. Unlike traditional decision trees, RF disrupts the greedy splitting algorithm during tree creation, restricting the choice of split points to random subsets of input features. This reduces similarity between trees, resulting in predictions with lower bias and higher variance. For more details, please refer to reference [15].

Lastly, MLP was selected as one of the classifiers in this evaluation as this versatile classifier excels at capturing complex patterns within the engineered features. By including the MLP alongside other classifiers like the SVM and RF, this study offers a comprehensive exploration of classification approaches for artificial speech detection, enriching the insights available to researchers and practitioners in this field.

3.3 Performance Evaluation

In this study, we set up an experiment to compare the performance of classifiers for artificial speech detection, which included SVM, RF, and MLP. In the experiment, the ASVspoof 2019 Logical Access (LA) dataset, which contains various speech synthesis and voice conversion attacks was used [3,16]. This dataset is ideal for testing artificial speech detection approaches because it covers a wide range of challenges that ASV systems might face. It's divided into three parts: training, development, and evaluation, with 107 speakers, including 46 males and 61 females. Six spoof algorithms (A01-A06) were part of the training and development data. Our experiments ran on a machine with an Intel(R) Xeon(R) CPU E5-2620 v4 processor, 2.10 GHz, 16 GB of RAM, and the Windows 10 (64-bit) operating system. We used Weka's classifier implementations to ensure consistent and reliable results.

In this study, we employ the EER as the evaluation metric to gauge the performance of the compared models. The EER corresponds to the threshold point where the false rejection rate (FRR) and false acceptance rate (FAR) are equal, and a lower EER signifies superior performance. The performance of each experimented model is presented in Table 1. In addition, Table 1 also provides a comparison with recent works for reference.

Table 1. Comparative evaluation of model performance in artificial speech detection

Model	Performance (EER %)
Model 1 : Fused Features + MLP	12.59
Model 2 : Fused Features + RF	13.24
Model 3 : Fused Features + SVM-RBF	10.84
Model 4 : Fused Features + SVM-Linear	3.55
Model 5 : Fused Features + SVM-Sigmoid	14.00
Model 6 : Fused Features + SVM-Polynomial	1.42
Model 7 : LFCC + AIR-ASVspoof [17]	2.19
Model 8 : LFCC + HAIR-ASVspoof [18]	0.57

In the comparative evaluation of the models, we observe notable differences in their performance, measured by the Equal Error Rate (EER %). Model 4, which combines Fused Features with an SVM using a Linear kernel, stands out with an impressively low EER of 3.55%. This result demonstrates the effectiveness of this combination in distinguishing between bonafide and spoof samples. Model 6, employing Fused Features with an SVM using a Polynomial kernel, also yields a remarkable performance with an EER of 1.42%. The polynomial kernel SVM outperformed other compared SVM kernels due to the prevalence of image-based features in the dataset. Notably, the polynomial kernel, commonly used in image processing, proved effective.

In contrast, Models 2 and 5, which incorporate Fused Features with RF and SVM using a Sigmoid kernel, respectively, show higher EER values of 13.24% and 14.00%. These findings indicate that RF and the Sigmoid kernel may not be the most suitable choices for artificial speech detection within this context. Additionally, the evaluation extends to recent works, Model 7 and Model 8, which serve as points of reference for comparison.

Model 7, incorporating Linear Frequency Cepstral Coefficients (LFCC) with an attentive ResNet18 model, achieves a competitive Equal Error Rate (EER) of 2.19%. This approach, distinct from Models 1-6, focuses solely on LFCC features and utilizes an attention mechanism within the ResNet18 architecture. In contrast, Model 8, inspired by Model 7, enhances performance by replacing the final fully connected layer with a linear SVM classifier, resulting in a significantly lower EER of 0.57%. These recent works underscore the efficacy of LFCC features when combined with machine learning classifiers and deep learning models, demonstrating promising results in the field of artificial speech detection.

Overall, the results indicate the significance of model choice and feature engineering in artificial speech detection, providing valuable insights for enhancing the security of voice biometric applications and ASV systems against evolving spoofing attacks. Comparisons with existing methods highlight the distinct practical advantages of the hybrid approach embodied in Model 8. The low EER of 0.57% positions this hybrid model as a standout choice for artificial speech detection, surpassing recent works that leverage different techniques. This underscores the immediate practical relevance and superior performance of the hybrid approach in the evolving landscape of speaker verification.

4 Conclusion

This study explores the vulnerability of ASV systems to spoofing attacks, especially in the context of advanced artificial speech generation. While prior research achieved promising results with an EER of 1.42% using SVM-based methods, this paper systematically assesses various classifiers for artificial speech detection. Our goal was to enhance our understanding and evaluate the performance of classifiers beyond SVM, aiming to improve ASV spoofing detection systems. Results show that SVM-Polynomial outperforms other classifiers, surpassing end-to-end learning, underscoring the importance of classifier diversity. Future research should focus on noise resilience in artificial speech detection to enhance ASV system reliability in real-world scenarios, ensuring security and convenience [19,20].

References

1. C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. binti Nohuddin, Z. Zainol, F. Coenen, and A. Gani, Multimed Tools Appl **80**, 32725 (2021)
2. C. B. Tan, M. H. A. Hijazi, and P. N. E. Nohuddin, TELKOMNIKA (Telecommunication Computing Electronics and Control) **21**, 97 (2023)
3. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf,

J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, Comput Speech Lang **64**, 101114 (2020)

4. C. B. Tan, M. H. Ahmad Hijazi, F. Kok, M. S. Mohamad, and P. N. Ellyza Nohuddin, IAES International Journal of Artificial Intelligence (IJ-AI) **11**, 161 (2022)
5. M. H. A. Hijazi, T. C. Beng, J. Mountstephens, Y. Lim, and K. Nisar, Adv Sci Lett **24**, 1172 (2018)
6. N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (IEEE, 2020), pp. 459–465
7. C.-X. Qin, D. Qu, and L.-H. Zhang, EURASIP J Audio Speech Music Process **2018**, 18 (2018)
8. C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, in *2006 International Conference on Computing & Informatics* (IEEE, 2006), pp. 1–5
9. L. Li, Y. Ding, B. Li, M. Qiao, and B. Ye, Alexandria Engineering Journal **61**, 91 (2022)
10. Y. Jia, X. Chen, J. Yu, L. Wang, Y. Xu, S. Liu, and Y. Wang, Complex & Intelligent Systems **7**, 1749 (2021)
11. K. S. Rao, V. R. Reddy, and S. Maity, in (2015), pp. 55–81
12. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, ACM SIGKDD Explorations Newsletter **11**, 10 (2009)
13. T. Ojala, M. Pietikäinen, and D. Harwood, Pattern Recognit **29**, 51 (1996)
14. N. Kalcheva, M. Karova, and I. Penev, in *2020 International Conference on Biomedical Innovations and Applications (BIA)* (IEEE, 2020), pp. 141–145
15. O. Sagi and L. Rokach, WIREs Data Mining and Knowledge Discovery **8**, (2018)
16. M. Todisco, X. Wang, V. Vestman, Md. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, in *Interspeech 2019* (ISCA, ISCA, 2019), pp. 1008–1012
17. Y. Zhang, F. Jiang, and Z. Duan, IEEE Signal Process Lett **28**, 937 (2021)
18. C. B. Tan, Mohd. H. Ahmad Hijazi, and P. N. Ellyza Nohuddin, in *2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (IEEE, 2023), pp. 236–240
19. S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, Multimed Tools Appl **79**, 24013 (2020)
20. A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, Appl Soft Comput **103**, 107141 (2021)