One-class classification for Speaker-Specific Audio Spoof Detection

Anonymous ICCV submission

Paper ID *****

Abstract

Advancements in text-to-speech (TTS) and voice conversion (VC) technologies have significantly increased the threat posed by audio spoofing attacks, particularly in high-profile applications such as political or public figure impersonation. Existing binary classification-based Audio Spoof Detection (ASD) methods face critical limitations in generalizing to novel and unseen spoofing techniques due to the growing diversity and sophistication of synthetic speech. This paper presents a speaker-specific framework for detecting audio deepfakes, leveraging Self-Supervised Learning (SSL) embeddings and one-class classification to address these challenges. The proposed methodology employs a one-class Support Vector Machine (SVM) trained exclusively on genuine speech samples from individual speakers to identify deviations indicative of synthetic speech. We conducted extensive evaluations on controlled datasets (ASVSpoof 2019 and DFADD) and real-world scenarios (In-The-Wild and political figures dataset) to demonstrate the effectiveness of this approach. We achieved robust and consistent performance across diverse synthesis methods and acoustic conditions. The results highlight the practical viability of speaker-specific ASD models in safeguarding against audio spoofing, particularly in applications requiring targeted protection of known individuals.

1. Introduction

The increasing sophistication of synthesized speech [23, 31, 35] and its easy dissemination through social media platforms present a growing threat to the integrity of the information. According to the World Economic Forum Global Risk Report, misinformation and disinformation represent the most serious anticipated threats over the next two years [25], potentially undermining democratic processes for approximately three billion people expected to participate in upcoming electoral polls across multiple countries. This threat has already manifested in high-profile incidents: from the Cambridge Analytica scandal in the 2016 US presidential elections [16] to more recent AI-generated deepfakes, including a viral video of President Zelenskyy purportedly asking his troops to surrender [14], a fake audio of President Biden misleading New Hampshire voters [15], and fabricated recordings of London Mayor Sadiq Khan making inflammatory remarks [27]. These incidents demonstrate how deepfakes can potentially trigger political unrest, from protests to civil confrontation, while enabling government censorship and propaganda that erodes press freedom and access to information. As synthetic media quality improves, distinguishing truth from falsehood becomes increasingly challenging for humans, making it imperative to develop strategies that can effectively differentiate genuine content from sophisticated falsifications.

Speech antispoofing research has evolved from traditional hand-crafted features (LFCC, CQCC) [2, 26, 41] and CQCC [32, 33] to end-to-end models using raw waveforms [20, 21, 28, 29], achieving state-of-the-art performance. Despite innovations in data augmentation, multitask learning, and attention mechanisms, generalization to unseen attacks remains a critical challenge as detection systems significantly degrade when facing novel spoofing techniques. Recent approaches employing acoustic perturbations [3, 4, 7, 13, 30] and pre-trained speech foundation models [10, 36] have shown promise, yet developing robust capabilities for detecting unseen synthesis systems continues to be a paramount challenge, demanding more sophisticated and adaptable methodologies.

Recent research has identified fundamental limitations in approaching Audio Spoof Detection as a binary classification problem between genuine and synthetic speech [22, 40]. As speech synthesis technologies rapidly evolve, the assumption of a unified synthetic speech distribution becomes increasingly problematic. In response, several studies have proposed one-class classification methods [1, 22, 34, 40] that model only the distribution of genuine speech, categorizing samples outside these boundaries as synthetic. Approaches like OC-Softmax [40] and ACS [22] enforce compact clustering of genuine speech from multiple speakers into a single cluster, while SAMO [11] adopts a multi-center approach that preserves speaker-specific characteristics, acknowledging the inherent acoustic diversity among

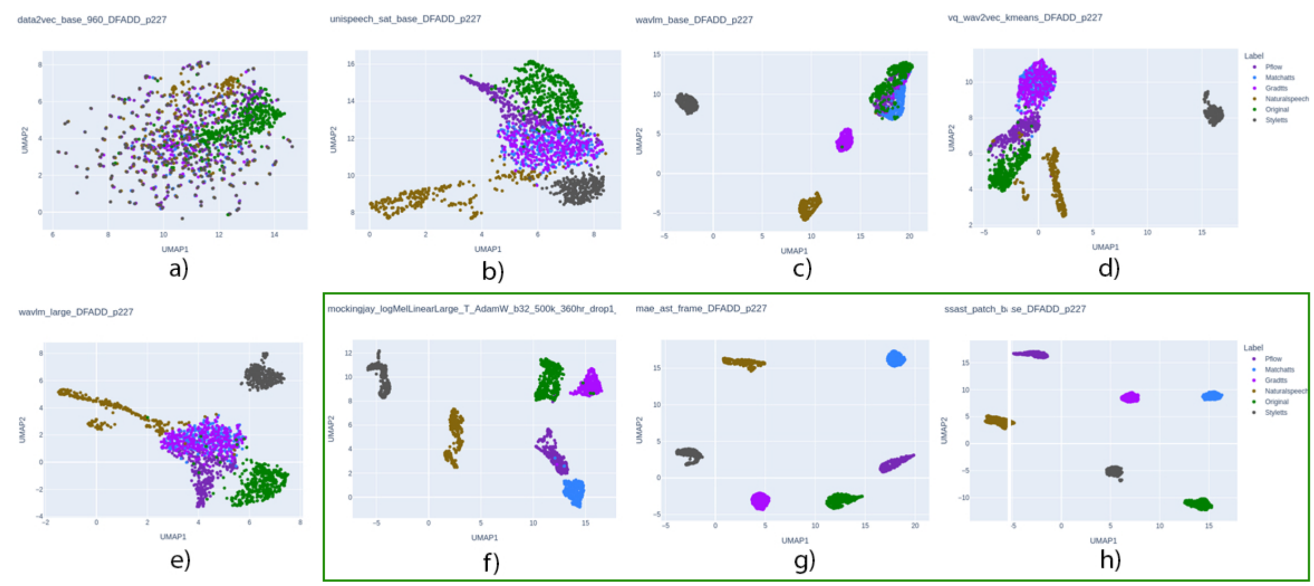

Figure 1. UMAP-based visualization of self-supervised learning embeddings for speaker p227, arranged to highlight the formation of distinct clusters. Here original(green) labels are bonafide audios and other colored labels are different deepfake types

speakers that naturally results in multiple clusters within the embedding space.

While these approaches represent significant advances, we posit that training one-class classifiers on multi-speaker datasets may still lead to suboptimal performance due to the fundamental challenge of modeling diverse speaker characteristics simultaneously. Our empirical analysis shows that UMAP [24] embeddings of multi-speaker scenarios exhibit significant overlap between genuine and synthetic speech, whereas single-speaker embeddings show clear cluster separation. This insight is particularly pronounced when considered within the context of contemporary real-world applications. Recent incidents, such as the Biden deepfake robocall in New Hampshire, demonstrate scenarios where the target identity is known a priori. This convergence of empirical evidence and practical requirements provides a compelling basis for the development of speaker-specific approaches to the detection of audio spoofing.

This paper presents a novel speaker-specific approach to audio deepfake detection, focusing on protecting individual speakers rather than developing universal detectors. Our methodology is particularly relevant for high-profile individuals who are frequent targets of voice spoofing attacks. Using a one-class SVM framework and self-supervised learning embeddings, we train models exclusively on genuine speech samples from target speakers, eliminating the need for synthetic examples during development. This approach offers adaptability to emerging spoofing techniques and leverages the readily available genuine speech data of public figures. Through comprehensive evaluation across diverse datasets, we demonstrate the effectiveness of our speaker-specific approach.

Table 1. Dataset organization for speaker-specific audio spoof detection. For 19LA, DFADD, and CodecFake, the training data comes from VCTK, while synthetic speech from these datasets serve as test data. S&I: Speeches/Interviews

Dataset	# Speakers	Training Data	Test Data
19LA DFADD	107 109	VCTK corpus (∼400 utterances per speaker)	108,978 163,500
ITW	54	19,963	11,816
Pol. Figures	2	S&I	750

The primary contributions of this work are as follows.

1. We propose a novel speaker-specific approach to audio spoof detection that protects individual speakers rather than using universal detectors, particularly beneficial for high-profile individuals who are frequent targets of voice spoofing attacks.
2. We introduce a one-class classification framework in which models are trained exclusively on genuine speech samples using self-supervised learning embeddings, eliminating the dependency on synthetic examples while capturing distinctive vocal characteristics.
3. Our evaluation demonstrates that vision transformer-based embeddings significantly outperform other representations, achieving near-perfect detection across diverse acoustic conditions with EERs of 1.30%, 1.82%, and 1.99% for controlled, in-the-wild, and political

Table 2. Statistical overview of datasets used in this study. For each dataset, we provide the number of speakers, total utterances (both bonafide and synthetic), and synthesis methods used.

Dataset	# Speakers	Bonafide Utterances	Synthetic Utterances	Duration (hrs)	Synthesis Methods
VCTK	110	44,000	–	44	–
ASVSpoof19 LA	107	12,483	108,978	∼100	A01-A19 (TTS & VC)
DFADD	109	44,455	163,500	∼50	GradTTS, YourTTS, StyleTTS2, NaturalSpeech2, matcha, pflow
In-the-Wild	54	19,963	11,816	∼38	Various
Trump Dataset	1	740	500	∼1	E2TTS, F5TTS, FishSpeech, MaskGCT, SSRSpeech, StyleTTS, and XTTS.
JD Vance Dataset	1	800	250	∼0.5	

datasets, respectively.

4. Through comprehensive evaluation across multiple datasets (ASVSpoof 2019, DFADD, In-The-Wild, and FakeXpose), we provide empirical evidence that speaker-specific modeling offers superior discrimination between genuine and synthetic speech compared to universal approaches, even when both use identical feature representations.

2. Benchmarking pretrained SSL model representations

We investigate how well pre-trained self-supervised learning (SSL) audio representations transfer to the task of audio deepfake detection. For that purpose, we extracted 80 SSL embeddings from 31 different SSL models using the S3PRL toolkit [38, 39] for the DFADD dataset [12]. These models span a diverse range of architectures and pre-training approaches, including transformer-based models (e.g., HuBERT [19], WavLM [8], wav2vec 2.0 [6]), convolutional models (e.g., VGGish [18]), LSTM-based models (e.g., APC [9]), and vision transformer adaptations (e.g., AST [17] and MAE-AST [5]). This comprehensive selection of SSL models allows us to thoroughly evaluate different speech representations for our speaker-specific spoofing detection system.

After extracting the SSL embeddings, we visualized them for the DFADD dataset using uniform manifold approximation and projection (UMAP), as shown in Fig. 1, to analyze their effectiveness in separating original from synthetic speech. Our analysis revealed that the MAE-AST (mae_ast_large), SSAST (ssast_patch_base) and Mockingjay (mockingjay_logMelLinearLarge) models provided the best separation between genuine and fake speech samples. To ensure comprehensive evaluation, we also included three widely adopted SSL models in the speech processing community, which include wav2vec 2.0 (large variant), WavLM (large variant), and HuBERT (large variant) for our further analysis. This selection balances models with optimal separation capabilities and those with established performance across diverse speech processing tasks, allowing us to evaluate different representation approaches for speaker-specific spoofing detection.

3. Datasets

To thoroughly evaluate our proposed speaker-specific audio spoof detection technique, we used diverse datasets spanning both controlled laboratory conditions and real-world scenarios. As shown in Tables 2 and 1, our framework incorporates four distinct types of datasets:

VCTK-based Datasets: The Voice Cloning Toolkit (VCTK) corpus forms the foundation for ASVSpoof 2019 LA [37] and DFADD [12] datasets. Instead of following standard partitioning, we used all VCTK recordings (400 utterances per speaker) to train speaker-specific models, while using the respective synthetic speech for testing. This approach maximizes available training data while enabling evaluation across diverse synthesis techniques.

In-The-Wild (ITW): This dataset bridges laboratory testing and practical applications with 38 hours of speech from online platforms. ITW presents more challenging evaluation scenarios with significant acoustic variations and commercially generated synthetic samples. For ITW, we used its own bonafide recordings for training, mimicking real-world scenarios where pristine recordings are unavailable.

FakeXpose Political Figures: Motivated by recent incidents like the Biden robocall controversy, we curated a specialized dataset for Donald Trump and JD Vance using bonafide samples from diverse speaking contexts. The synthetic portion comprises samples generated with seven state-of-the-art TTS systems, selected for their high-fidelity output and practical accessibility, representing real-world scenarios.

4. Modeling

Our detection framework employs One-Class Support Vector Machine (OC-SVM), a machine learning algorithm designed for scenarios where training data are only available from one class. In the context of audio spoofing detection,

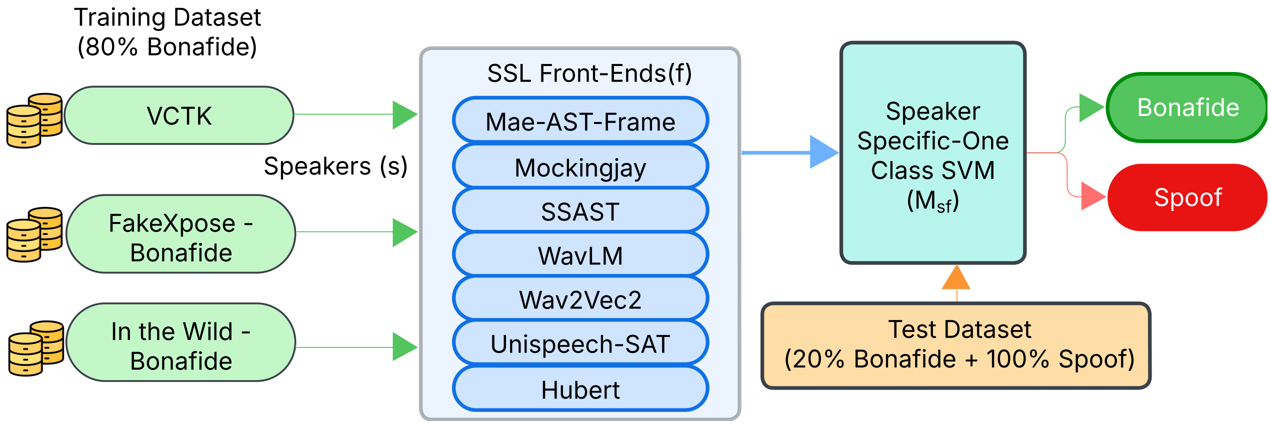

Figure 2. **Overview of the proposed speaker-specific audio deepfake detection framework.** The system trains separate one-class SVM models for each speaker using only genuine speech samples (80% for training, 20% for testing) from three datasets: VCTK, FakeXpose, and In-The-Wild. Multiple SSL front-end embeddings (Mae-AST-Frame, Mockingjay, SSAST, WavLM, Wav2Vec2, Unispeech-SAT, and Hubert) are extracted from the genuine speech data and fed into speaker-specific one-class SVMs (M_{sf}). During testing, the trained models classify audio samples as either bonafide (genuine) or spoof (synthetic) by detecting deviations from the learned genuine speech distribution.

this characteristic is particularly advantageous as it allows the model to learn the distribution of genuine speech without requiring examples of synthetic speech during training. The OC-SVM learns to construct a decision boundary that encapsulates the genuine speech characteristics in the feature space, enabling the detection of synthetic speech as deviations from this learned distribution.

The SVM hyperparameters γ and ν that control the Gaussian kernel width and outlier percentage are optimized using the VCTK dataset. Specifically, we performed a grid search over γ and ν and selected the parameters that yielded the highest discrimination between the speaker of interest and the remaining speakers. These hyperparameters were trained for each speaker. An SVM model is trained for each speaker using the SSL embeddings extracted from 85% of their genuine speech samples. During inference, the signed distance from the decision boundary serves as the detection score, where positive values indicate genuine speech and negative values indicate synthetic speech. The magnitude of this score provides a measure of confidence in the classification decision.

5. Experimental Setup

5.1. Implementation Details

The experimental framework is implemented on a high performance computing infrastructure comprising three NVIDIA A100 GPUs, each with 12 GB of memory, utilizing Python 3.10.16. All audio samples are standardized to a 16 kHz sampling rate prior to processing.

5.2. Model Training and Evaluation

In section 2, we found that the pretrained *mae_ast*, *wav2vec2_large*, *ssast_patch*, *mockingjay_logMelLinearLarge*, *hubert_large* and *wavlm_large* models exhibited good separation between the fake and real speech. We used these SSL features as front-ends and trained a one-class svm (OC-SVM) classifier for each speaker, which requires only genuine speech samples for model training.

Following established protocols in ASD research, we employ Equal Error Rate (EER) as our primary evaluation metric. EER represents the operating point where the false acceptance rate is equal to the false rejection rate, providing a balanced measure of system performance.

6. Results and Analysis

Our evaluation examines the effectiveness of speaker-specific models across three distinct scenarios: (1) cross-dataset generalization using VCTK-derived datasets, (2) robustness in real-world conditions using the ITW dataset, and (3) practical application for protecting high-profile individuals using the FakeXpose Political Figures dataset. We also compare our approach with established baselines including AASIST [21], RawNet2 [28], and wav2vec2-AASIST [30], all of which are trained on the ASVSpoof 2019 LA dataset using both real and synthetic speech samples.

ICCV
#*****

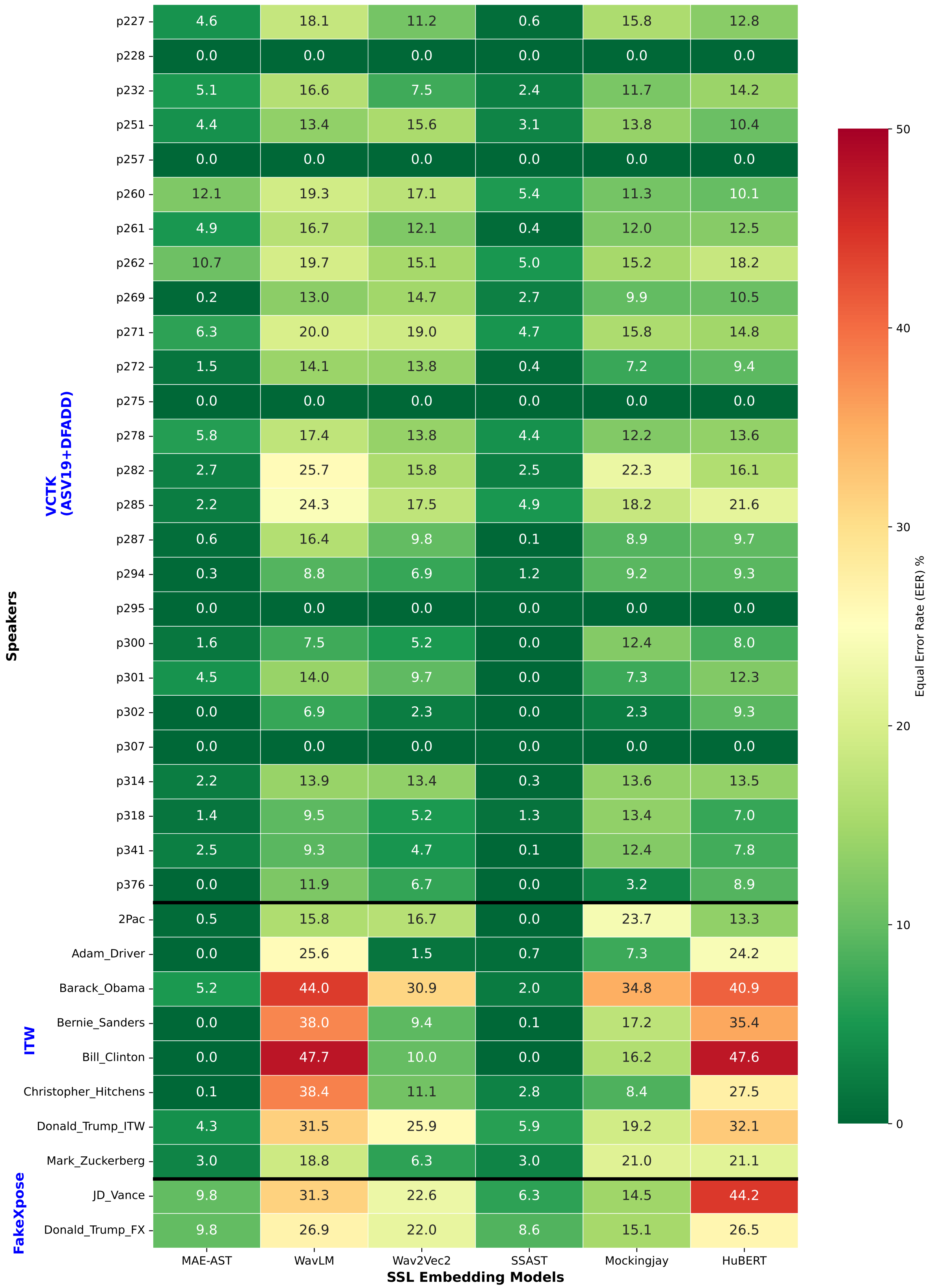

Figure 3. **EER Performance Heatmap across speakers and SSL embeddings.** The heatmap visualizes Equal Error Rate (EER) percentages for speaker-specific audio deepfake detection across three datasets: VCTK (ASV19+DFADD), In-The-Wild (ITW), and FakeXpose. Each cell shows the EER value for a specific speaker-SSL embedding combination, with color coding ranging from dark green (excellent performance) to dark red (poor performance).

6.1. Speaker-Specific SSL Performance

Figure 3 presents the EER performance across speakers from three datasets using six different SSL embeddings. The *ssast_patch_base* embedding consistently achieves the lowest EER values across most speakers (average EER of 1.30% for VCTK, 1.82% for ITW, and 7.45% for FakeXpose), with many instances of perfect detection (EER = 0), followed by *mae_ast_frame* (average EER of 2.65% for VCTK, 1.63% for ITW, and 9.80% for FakeXpose). In contrast, *wavlm_large* and *hubert_large* typically yield significantly higher EER values, particularly for ITW speakers (32.49% and 30.26%, respectively) and FakeXpose speakers (29.10% and 35.35%, respectively).

We observe notable variations in performance across individual speakers, with certain speakers like p260 (5.4%) and p262 (5.0%) showing consistently higher EER values even with the best-performing embedding. Among political figures, Donald Trump and JD Vance exhibit substantially higher EERs (8.6% and 6.3% respectively) compared to the VCTK average. These variations likely stem from several factors: the distinctive vocal characteristics of certain speakers may be easier to synthesize convincingly; speakers with more varied speaking styles or emotional range present greater challenges for detection; and the quality and acoustic diversity of available training samples significantly impact model performance.

The speaker-to-speaker variability in detection performance strongly validates our speaker-specific approach. Using a single universal model would inevitably compromise performance, as the model would need to make trade-offs between speakers that are easier to detect (e.g., p228, p257, p275) and those that present greater challenges (e.g., p260, political figures). By modeling each speaker individually, we can optimize the detection boundary specifically for each speaker's unique vocal characteristics, ensuring optimal performance regardless of whether a speaker is inherently easy or difficult to protect. This speaker-specific optimization is particularly critical for high-profile individuals, more vulnerable to deepfake attacks.

6.2. Comparison with Baseline Systems

Table 3 demonstrates the substantial performance advantage of our speaker-specific approach over established baselines across all datasets. Our ssast-OCSVM system achieves remarkably low EER values of 1.30%, 1.82%, and 1.99% on ASV19+DFADD, ITW, and FakeXpose datasets respectively, significantly outperforming all baselines. A particularly insightful comparison can be made between wav2vec2-AASIST and w2v2-OCSVM, as both utilize the same wav2vec2 feature space but differ in their modeling approach, universal versus speaker-specific. Although both systems use identical front-end features, our speaker-specific w2v2-OCSVM (8.38%, 13.97%, 10.92%) outperforms the universal wav2vec2-AASIST (5.59%, 12.34%, 29.23%) on FakeXpose by a substantial margin, although performs slightly worse on ASV19+DFADD and comparably on ITW. This mixed result suggests that speaker-specific modeling offers practical advantages when protecting high-profile individuals, even when using the same feature representation. The slightly worse performance on ASV19+DFADD may be attributed to certain speakers with challenging vocal characteristics that disproportionately impact the average performance, as observed in Figure 3 where speakers like p260 and p262 consistently show higher EER values.

Critically, it is important to emphasize that our speaker-specific models are trained exclusively on genuine speech samples without exposure to any synthetic data, while all baseline systems are trained on both genuine and synthetic speech samples. This training methodology makes our system inherently more generalizable and robust to novel attacks, as it does not rely on characteristics of specific synthetic speech techniques seen during training. Instead, by learning the genuine speech distribution for each speaker independently, our approach can detect any deviation from this distribution regardless of the synthesis method used to

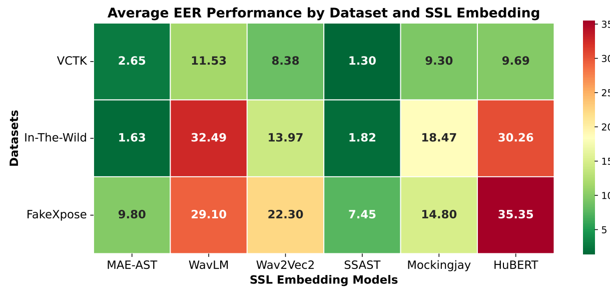

Figure 4. **Average EER Performance Summary by Dataset and SSL Embedding.** This summary heatmap presents the mean Equal Error Rate (EER) percentages across three datasets (VCTK, In-The-Wild, and FakeXpose) for six different SSL embedding models. The SSAST embedding demonstrates consistently superior performance across all datasets with the lowest average EERs (1.30% for VCTK, 1.82% for ITW, 7.45% for FakeXpose), followed by MAE-AST.

Table 3. Performance comparison of baseline systems and our Speaker-Specific System (using *ssast_patch_base* and *wav2v2_large* embedding) across different datasets averaged over all speakers (EER %).

System	ASV19 + DFADD	ITW	FakeXpose
AASIST	30.268	25.44	9.62
wav2vec2-AASIST	5.58	12.34	29.23
RawNet2	36.32	48.563	50.09
ssast-OCSVM	1.30	1.820	7.45
w2v2-OCSVM	9.12	13.970	22.30

create the deepfake.

The exceptional performance of ssast-OCSVM can be primarily attributed to the superior quality of the ssast feature space, as evidenced by the significant performance gap between ssast-OCSVM and w2v2-OCSVM across all datasets. These results highlight the importance of both selecting appropriate SSL embeddings and employing speaker-specific modeling approaches, particularly for protecting high-profile individuals against increasingly sophisticated audio deepfake attacks.

6.3. Key Findings

Our investigation reveals several significant findings:

- Our speaker-specific approach using *ssast_patch_base* embeddings consistently outperforms all baselines across datasets, achieving near-perfect detection, despite being trained exclusively on genuine speech samples.
- Vision transformer-based SSL embeddings (particularly *ssast_patch_base*) demonstrate superior effectiveness for speaker-specific spoofing detection compared to other architectures, with average EERs of 1.30%, 1.82%, and 7.45% across the three datasets, significantly outperforming wav2vec2 (8.38%, 13.97%, 22.30%) and other embeddings.
- Individual speakers exhibit substantial variability in detection difficulty, with some speakers and political figures consistently showing higher EER values across all embeddings, validating the need for speaker-specific modeling that can optimize detection boundaries for each unique vocal profile.

Despite these promising results, our approach requires training and maintaining separate models for each protected speaker, which may introduce scaling challenges for large-scale deployments. However, significant performance advantages, combined with the enhanced generalizability to unseen attacks due to training exclusively on genuine speech, make this approach particularly valuable for high-stakes applications where protecting specific individuals is critical.

7. Conclusion

This paper presents a novel speaker-specific approach to audio deepfake detection that addresses the growing threat of voice spoofing attacks, particularly against high-profile individuals. Our methodology leverages one-class Support Vector Machines trained exclusively on genuine speech samples, eliminating the dependency on synthetic examples during model development while achieving superior performance compared to traditional universal detection systems.

Our comprehensive evaluation across diverse datasets, from controlled laboratory conditions (ASVSpoof 2019, DFADD), to challenging real-world scenarios (In-The-Wild) and practical applications (FakeXpose Political Figures), demonstrates the effectiveness of our approach. The use of vision transformer-based SSL embeddings, particularly ssast_patch_base, consistently achieves remarkably low EER values of 1.30%, 1.82%, and 1.99% across the three dataset types, substantially outperforming established baselines including AASIST, RawNet2, and wav2vec2-AASIST.

A key strength of our approach lies in its training methodology. By learning exclusively from genuine speech patterns for each individual speaker, our system becomes inherently more generalizable and robust to novel attacks, as it does not rely on characteristics of specific synthetic speech techniques. This speaker-specific optimization ensures optimal performance regardless of whether a speaker is inherently easy or difficult to protect, making it particularly valuable for safeguarding high-profile individuals who are frequent targets of sophisticated deepfake attacks.

Our findings reveal significant speaker-to-speaker variability in detection performance, strongly validating the need for individualized modeling approaches. The substantial performance advantages of our method, combined with its enhanced generalizability to unseen synthesis techniques, make this approach particularly valuable for high-stakes applications where protecting specific individuals is critical, such as political communications, financial security systems, and other security-critical environments.

Although our approach requires training separate models for each protected speaker, the demonstrated performance benefits and adaptability to emerging spoofing techniques position speaker-specific detection as a promising direction for practical deepfake defense systems in an era of increasingly sophisticated audio synthesis technologies.

References

[1] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2013. 1

[2] Hashim Ali, Dhimant Khuttan, Rafi Ud Daula Refat, and Hafiz Malik. Protecting voice-controlled devices against laser injection attacks. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2023. 1

[3] Hashim Ali, Surya Subramani, and Hafiz Malik. Augmentation through laundering attacks for audio spoof detection. In *Proc. ASVspoof 2024*, pages 181–187, 2024. 1

[4] Hashim Ali, Surya Subramani, Shefali Sudhir, Raksha Varahamurthy, and Hafiz Malik. Is audio spoof detection robust to laundering attacks? In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, pages 283–288, 2024. 1

[5] Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022. 3

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 3

[7] Weicheng Cai, Danwei Cai, Wenbo Liu, Gang Li, and Ming Li. Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion. In *Interspeech*, pages 17–21, 2017. 1

[8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. 2021. 3

[9] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*, 2019. 3

[10] David Combei, Adriana Stan, Dan Oneata, and Horia Cucu. Wavlm model ensemble for audio deepfake detection. *arXiv preprint arXiv:2408.07414*, 2024. 1

[11] Siwen Ding, You Zhang, and Zhiyao Duan. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[12] Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-yi Lee, and Jyh-Shing Roger Jang. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 921–928. IEEE, 2024. 3

[13] Raphaël Duroselle, Olivier Boeffard, Adrien Courtois, Hubert Nourtel, Pierre Champion, Heiko Agnoli, and Jean-François Bonastre. Data augmentations for audio deepfake detection for the asvspoof5 closed condition. In *ASVspoof Workshop 2024*, pages 16–23. ISCA, 2024. 1

[14] ebaker. Russian War Report: Hacked news program and deepfake video spread false Zelenskyy claims, 2022. 1

[15] Vittoria Elliott. The Biden Deepfake Robocall Is Only the Beginning. *Wired*, 2024. Section: tags. 1

[16] Christina Georgacopoulos and Grayce Mores. How fake news affected the 2016 presidential election, 2020. 1

[17] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 3

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 3

[20] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *arXiv preprint arXiv:2004.00526*, 2020. 1

[21] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371. IEEE, 2022. 1, 4

[22] Hyun Myung Kim, Kangwook Jang, and Hoirin Kim. One-class learning with adaptive centroid shift for audio deepfake detection. *arXiv preprint arXiv:2406.16716*, 2024. 1

[23] Jiaxin Li and Lianhai Zhang. Zse-vits: A zero-shot expressive voice cloning method based on vits. *Electronics*, 12(4): 820, 2023. 1

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2

[25] Marsh McLennan et al. Global Risks Report 2024. World Economic Forum, 2024. 1

[26] Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Interspeech*, pages 2062–2066, 2015. 1

[27] Marianna Spring. Sadiq Khan says fake AI audio of him nearly led to serious disorder. 2024. 1

[28] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021. 1, 4

[29] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 1–8, 2021. 1

[30] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE, 2022. 1, 4

[31] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021. 1

[32] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, pages 283–290, 2016. 1

[33] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant q cepstral coefficients: A spoofing countermeasure

for automatic speaker verification. *Computer Speech & Language*, 45:516–535, 2017. 1

[34] Jesus Villalba, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge. In *Proc. Interspeech*, pages 2067–2071, 2015. 1

[35] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers, 2023. *URL: https://arxiv.org/abs/2301.02111. doi: doi*, 10. 1

[36] Xin Wang and Junichi Yamagishi. Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv:2111.07725*, 2021. 1

[37] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020. 3

[38] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. 3

[39] Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, et al. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 3

[40] You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021. 1

[41] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE workshop on automatic speech recognition & understanding*, pages 559–564. IEEE, 2011. 1