Echoes Unveiled: Identifying Synthetic Voices

Daniel Pluth
Vail Systems, Inc.
Chicago, USA
dpluth@vailsys.com

Jordan Hosier
Vail Systems, Inc.
Chicago, USA
jhosier@vailsys.com

Yu Zhou
Vail Systems, Inc.
Chicago, USA
yzhou@vailsys.com

Vijay K. Gurbani
Vail Systems, Inc.
Chicago, USA
vgurbani@vailsys.com

***Abstract*—The advent of deep neural networks in natural language and speech processing has created a new attack vector in the form of synthetic voices cloned from less than 30-seconds of voice sample from a human counterpart. Effectively detecting spoofing attacks is critical for any speech application that uses voice for authentication, verification, and identification. With the rapid rise of highly effective speech synthesis, it is challenging to identify synthetic voices while generalizing to novel voices, synthesizers, and channel conditions. In this paper, we present a model aimed at identifying synthetic voices and demonstrate its effectiveness and generalizability. We further motivate the need for the research community to consider channel conditions when detecting voice spoofing. Our work demonstrates that channel conditions play an inordinate role in identifying a spoofed voice, and detection techniques that do not consider variable channel conditions will exhibit high error rates.**

***Index Terms*—synthetic speech detection, voice authentication, anti-spoofing**

I. Introduction

In November 2023, the US Federal Trade Commission (FTC) announced a voice cloning challenge to address the "present and emerging harms" of such technologies[1]. In January 2024 during the New Hampshire primary, a robocall cloned from President Biden's voice urged citizens not to cast ballots[2]. As the need for voice recognition grows (banks using voice as a biometric print, phones unlocking via the owner's voice), the ability to spoof and impersonate voices using deep learning-based speech synthesis systems has also significantly improved. Such high-quality text-to-speech (TTS) synthesis and voice cloning (VC) approaches can successfully deceive humans and automatic speaker verification systems. Commercial and open-source voice cloning systems can generate an audio sample reproducing a source voice saying any given prompt. This vulnerability creates the need for systems that can distinguish between authentic voices and spoofed voice clones to protect voice-based authentication systems from adversarial attacks.

Contributions: Liveness detection identifies a set of characteristics intrinsic to a real human voice [1]. In the context of voice authentication, this effort targets the identification of spoofed voices. Such features include articulatory gestures, vocal pitch and depth, etc. There are several approaches aimed at solving liveness detection; a primary concern, however, is the generalizability of a solution: existing countermeasures can only detect spoofed voices using prior knowledge of existing VC models, hindering generalization. Developing voice authentication systems that are resilient to a range of spoofing techniques hinges on robust feature extraction such that the model requires little re-calibration to detect novel spoofing attacks.

We present a model that generalizes better than existing solutions (cf. Section IV). Further, we call attention to the dependence on channel conditions during liveness detection (cf. Section VII); most detection models are trained on clean data in (near) optimum channel conditions (no background noise, high-fidelity codecs, etc.). Consequently, such models exhibit high error rates when faced with detecting voice spoofs that occur in the real world over noisy channel conditions, such as a telephone call using codecs that can alter the shape and relevance of extracted features used for speaker verification.

II. Related Work

Synthetic-based voice attacks using TTS typically consist of three modules: a text analysis model, an acoustic model, and a vocoder. To generate synthetic spoofed audio, raw audio is collected with an associated transcript. Next, the TTS model is trained using the collected data to build a synthetic speech model. WaveNet [2] was the first end-to-end speech synthesizer that directly used raw audio for training, and showed a mean opinion score (MOS) very close to human speech. Similar quality was shown by other TTS systems such as Deep Voice [3] and Tacotron [4]. These breakthroughs in TTS and VC technologies made spoofing attack detection more challenging.

While VC generation has improved considerably in recent years, the accuracy of liveness detection lingers around 80% on open-source datasets [5]. Though performance is strong from an academic perspective, it is insufficient for real-world applications. Given two major emerging issues, i.e., less-than-perfect accuracy of detection and widened target range, adaptability of liveness detection has become a critical consideration.

The most significant body of work on liveness detection is driven largely by the ASVspoof challenges and datasets [6]–[8]. The reported performance of these models is based on a limited set of TTS synthesis algorithms. ASVspoof 2019 is based on the VCTK dataset (version 0.92) [9], which features

[1] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/11/preventing-harms-ai-enabled-voice-cloning

[2] https://www.npr.org/2024/02/07/1229856682/what-a-robocall-of-bidens-ai-generated-voice-could-mean-for-the-2024-election

110 native English speakers with various accents recorded in a studio environment. The limitations of this ASVspoof challenge and resulting models are highlighted in [10]. In this work, researchers implement twelve of the most popular architectures and identify fundamental properties for well-performing voice clone detection. They further introduce a new voice spoof dataset[3] which consists of 17.2 hours of high-quality voice clones and 20.7 hours of authentic material from 58 politicians and celebrities. Results demonstrate that liveness models generally perform poorly on such real-world data underlining the discrepancy between reported and actual generalization ability.

Chaubey et al. 2023 proposed a speaker-specific threshold technique for liveness detection [11]. The approach uses enrollment samples to compute an adaptive intra-speaker threshold, computed using the equal error rate (EER) metric (defined in Appendix A). The threshold for a particular dataset depends on the inter-speaker and intra-speaker separability of the audio samples in the speaker embedding space. Further, this separability depends on the speaker distribution and recording conditions present in the dataset. Results showed a superior performance as compared to a fixed threshold approach. However, the datasets used were all composed of high-quality audio (VoxCeleb, VCTK, FFSVC). Chaubey et al.'s model also shows degraded performance under variable microphone distances, highlighting the challenge of differences in channel (recording) conditions. Indeed, Gupta et al., 2024 presented a survey on strategies and vulnerabilities in Automatic Speaker Verification (ASV) [12]. In this survey, authors point to the importance of evaluating over-the-air performance, over various channels, due to increased chances of successful attacks under these conditions.

Although liveness detection is an active area of study, further research is still needed to address the existing gaps, particularly as it relates to real-world, noisy environments and spoofed voices derived from different TTS systems.

III. Voice Cloning Models and Dataset Preparation

Three open-source synthesizers are used to generate voice clones: 1. **XTTS**[4], 2. **StyleTTS2** [13], and 3. **YourTTS** [14], all of which have been found to achieve state-of-the-art performance in voice-cloning tasks.

Dataset Generation The reference human speech is taken from three sources: 1. 110 real speakers from the **VCTK dataset**, 2. a single speaker extracted from the **Mozilla Common Voice data (MozVox)**, and 3. an anonymous, proprietary single-speaker dataset referred to as **AnonVox**, collected for TTS training and included for its unique studio recording environment. The MozVox speaker was selected as the leading English speaking contributor to the Movilla Common Voice dataset at the time of this study. Both MozVox and AnonVox are generated from male speakers. For each speaker in all the datasets, a 30-second reference audio sample is extracted by randomly selecting audio, trimming silence, and concatenating segments. At the end of this process, we obtain 112[5] 30-second reference audio files that can be used to clone new utterances across each of the three synthesizers. Spoofed audio samples are created using the reference audio and the desired output text transcript. The result is a spoofed audio file of the reference voice saying the words in the given transcript.

TABLE I
Dataset distribution

	MozVox	AnonVox	VCTK
Real	3,601	3,601	44,057
StyleTTS2	3,601	3,601	44,057
XTTS	3,601	3,601	44,057
YourTTS	3,601	3,601	44,057

These real and synthetic voice pairs are used in training and testing. In all, the dataset consists of 51,259 real voice samples and 109,720 voice spoof samples as shown in Table I.

IV. Titanet Model for Liveness Detection

The liveness detection model presented in this work is based on Titanet [15]. Titanet is a convolutional network-based model, largely derived from ContextNet with an adaptive pooling layer. It is trained directly to classify speakers based on audio samples of their speech. To apply the model to novel speakers, a speaker embedding is extracted from the penultimate layer. Titanet uses additive angular margin (AAM) loss to optimize the distance between speaker embeddings [16]. The extracted embeddings can then be compared using cosine similarity to determine a distance measure, and a threshold can be chosen to determine whether samples belong to the same speaker. However, the cosine similarity method itself is insufficient for the task of liveness detection as it struggles to fully disambiguate between the real speaker and the synthetic voice, as seen in Figure IV. In the figure, all of the synthetic voices overlap with the real voice: StyleTTS2, despite being separated from the real distribution, still shares a long tail with the real voice. The synthetic voices generated by the other two TTS engines show a more pronounced overlap.

Table II further demonstrates this failing by presenting the EER for each synthesizer on the AnonVox dataset using the cosine similarity of Titanet embeddings. StyleTTS2 has the lowest EER, but the EER is high enough to make it unsuitable for real-world deployment. Equal rate is described in Appendix A.

TABLE II
The EER (%) for each synthesizer on the AnonVox dataset using cosine similarity of Titanet embeddings.

YourTTS	XTTS	StyleTTS2
12.6	20.7	4.8

Motivated by this limitation, this work proposes extending the speaker embedding method by treating the embeddings as

[3] https://deepfake-demo.aisec.fraunhofer.de/in_the_wild

[4] https://github.com/coqui-ai/TTS

[5] 110 unique VCTK speakers, plus the two single speakers.

a feature set for additional classification models. In our model pipeline, the Titanet embeddings, composed of 192 elements, are used as input to a Support Vector Machine (SVM) model which classifies a data point as either a spoofed or real voice.[6]

The advantage of the Titanet-SVM model is that it allows a smaller model (SVM) to be quickly trained using limited training data derived from the output of a large pre-trained model (Titanet). Using the embeddings from Titanet to train an SVM model allows for rapid prototyping and exploration. When a novel synthesizer appears in the wild, it is easier to train an SVM model using embeddings than it would be to train larger, dedicated spoof-detection models that require thousands or even millions of observations to adapt to the novel synthesizer. The Scikit-learn implementation of SVM was used with the default parameters and trained with various datasets as described in the following sections [17].

We test the efficacy of our augmented model across diverse conditions by isolating three axes of variation: 1. distinct voices; 2. multiple synthesizers; and 3. channel conditions.

V. Voice Diversity Experiment

Titanet is designed as a speaker recognition model. Thus, it is expected that the characteristics of a person's speech are reflected in the corresponding speaker embedding. However, it is not clear that the embedding is capable of distinguishing synthetic voices in general from real voices; it may be that the embedding may encode the synthetic variant of a speaker simply as a distinct speaker. If the synthetic features themselves are not distinguishable, the SVM model will fail to generalize for a held-out set of speakers with real and synthetic samples.

A. Data and Design

To test the extent to which spoofed features can be identified and spoofed voices separated, we use the multi-speaker VCTK dataset and derive spoofed voices using the StyleTTS2 synthesizer[7]. Each VCTK speaker has roughly 400 utterances with corresponding transcripts from which another 400 synthetic utterances are generated for each speaker using StyleTTS2. This process yields about 88,000 utterances or 44,000 (real, synthetic) voice pairs. Early experimentation on this dataset indicated that the SVM model was able to separate real voices from synthetic voices with a small training set — both real and synthetic — of about 10% of the speakers was enough to separate the real and synthetic voices of the 90% remaining speakers.

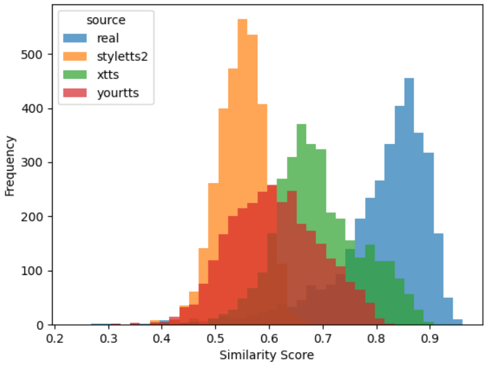

Fig. 1. Cosine similarity between a held-out subset of real recordings and synthetic and real sources. This comparison used the AnonVox dataset.

B. Results and Discussion

Figure 2 shows the classification accuracy of the SVM model as it determines which embeddings correspond to the real speaker and which are synthetic. The classification accuracy is asymptotic to 1.00 at about 6 speakers in the training set; adding more speakers in the training set does not improve the accuracy. To gain confidence in the results, and to avoid picking a particularly good (or bad) set of speakers by chance, we ran the above experiment 200 times, randomly shuffling the speakers in the train/test sets to mitigate the effect of differing numbers of samples, quality of particular voice spoof samples, and other sample idiosyncrasies. The 90% confidence interval is shown in the shaded portion of Figure 2.

This experiment demonstrates our contention that limited training data — only a handful of speakers — is required to accurately train an SVM model. Figure 2 shows that Titanet-SVM trained on a few speakers generalizes well to a much larger set of held-out speakers. It is unlikely that a single speaker can represent 88% to 97% of the speaker diversity of the VCTK dataset, as demonstrated in Figure 2. This result instead indicates that the Titanet embeddings contain identifying characteristics of the StyleTTS2 model and the SVM can find a hyperplane separating the real and fake voices.

VI. Synthesizer Diversity Experiment

The results of Section V illustrate that Titanet-SVM is capable of distinguishing spoofed voices based on the characteristics of the synthesizer. However, since this experiment used a single synthesizer, StyleTTS2, it remains unclear whether different synthesizers have common signatures reflected in their Titanet embeddings such that knowledge of the features of one synthesizer enables the detection of another. ASVSpoof 2019 [6] in particular has attempted to address this problem through experiment design, where the training set had 6 synthesizers and the evaluation set had another 11 held-out synthesizers. However, a zero-day vulnerability is always present as there is no guarantee that future synthesizers will share characteristics with current ones. This experiment aims to evaluate how well the Titanet-SVM is capable of adapting to a handful of novel synthesizers.

[6] Random Forest models were also trained to similar results, which are omitted for brevity.

[7] The choice of using StyleTTS2 was primarily driven by its low latency in generating synthetic voices from a large corpus like VCTK.

Fig. 2. Classification accuracy by number of training speakers for Titanet-SVM. The selected speakers were chosen randomly 200 times per datapoint to estimate the confidence interval.

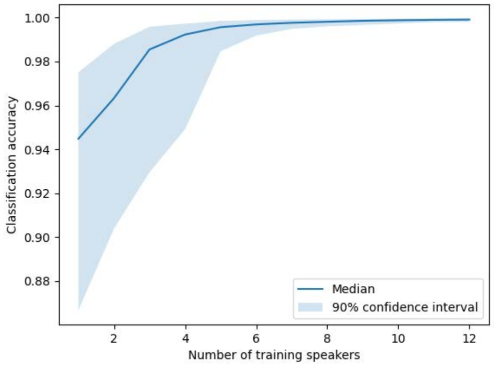

TABLE III
IN-DOMAIN AND OUT-OF-DOMAIN PRECISION (IN %).

	Evaluation		
Train	YourTTS	XTTS	StyleTTS2
YourTTS	100	79.3	27.0
XTTS	85.1	100	34.1
StyleTTS2	51.2	21.9	100

A. *Data and Design*

The MozVox, AnonVox, and VCTK datasets were combined and utilized for this experiment, each contributing 3601 utterances. The VCTK dataset was subset randomly in order to balance the contributions from the various sources. Thus, in all, we curate 10,803 human utterances. Each human utterance will be used to generate a synthetic utterance from each of the YourTTS, XTTS, and StyleTTS2 synthesizers. In total we obtain 43,212 utterances. With this data, a set of three SVM models were trained, each is trained on a single, distinct synthesizer with its real pairs and tested on the two held-out synthesizers in addition to a held-out test set of the in-domain synthesizer. When used as the training set, each synthesizer+real dataset was split 90/10 into training and test sets.

B. *Results and Discussion*

Table III shows that the cross-synthesizer performance is highly dependent on the synthesizer used to train the model. Training on certain synthesizer embeddings, like XTTS, results in models that are highly effective for out-of-domain data, whereas the other two are much less effective in general. From analyzing the model architectures, it would appear that all three synthesizers use the HifiGan vocoder [18]. However, if there were artifacts present in the speaker embeddings as a result of this vocoder specifically, it would be expected that there would be a mutual classification benefit, which is not apparent in the results.

For comparison, and to demonstrate that loss of accuracy on out-of-domain data is not a quirk of the Titanet-SVM model, RawGAT-ST [19], a model trained on ASVSpoof 2019, was also evaluated on the VCTK[8] real vs StyleTTS2 samples; its overall accuracy was 75%. The analogous experiment with Titanet-SVM yields a comparable accuracy of 77.2%. Likewise SSLAntiSpoofing [20], a model built for ASVSpoof 2021 [8], records an accuracy of 53.2% on this same dataset [21]. These models were chosen because they are some of the best-of-breed models at the date of this study.

VII. CHANNEL DIVERSITY EXPERIMENT

Titanet was trained on data collected from a variety of environments, microphones, codecs, etc. Recording conditions ideally would have minimal influence on extracted embeddings. In a voice-authentication setting, a model should be able to identify a speaker calling in from different sources under different conditions as the same speaker. This would require that the model embeddings minimize these variances and focus on the characteristics of the speaker. This experiment aims to test the impact of channel conditions on the extracted speaker embeddings.

A. *Data and Design*

The three speech datasets were collected under different recording conditions. The VCTK dataset was recorded in a hemi-anechoic chamber. The recording environment of MozVox is unknown but is seemingly consistent. Due to the volunteer nature of the dataset and the quality of the recordings, it was likely recorded on a laptop microphone. AnonVox is composed of recordings collected for commercial use and is not publicly available. The utterances were recorded in a private recording studio. While it is already understood that noisy environments can have a significant detrimental effect on liveness detection [22], it should be noted that none of these samples were recorded in what could be considered 'noisy' environments. All of the samples are near pristine in terms of recording conditions.

Using the human-generated and YourTTS cloned audios generated from VCTK, MozVox and AnonVox, Titanet embeddings were generated and used to train various SVM models. A subset of 3,000 samples from each of the three datasets were used in turn as the training set and the remaining datasets were held out as test sets. For comparison, three other spoof-detection models that were trained partially or entirely on the ASVSpoof2019 dataset were also tested on these datasets[9].

B. *Results and Discussion*

Table IV shows the accuracy of the four liveness models on out-of-domain channels, where out-of-domain here refers to data collected under different recording conditions (across different datasets). Generally, the in-domain performance is quite good across all models. However, performance on out-of-domain datasets drops significantly, with a few exceptions. Titanet-SVM trained on AnonVox performs surprisingly well

[8] The training dataset of ASVSpoof 2019 is descended from VCTK. Thus VCTK is an appropriate comparison.

[9] See footnote 8, ASVSpoof will be labeled here as VCTK

TABLE IV
DEPENDENCE OF PERFORMANCE (IN %) ON ENVIRONMENTAL CONDITIONS. SAME DATASET PERFORMANCE IS ON A HELD OUT EVALUATION SET (90/10 SPLIT).

		Evaluation		
Model	Train	VCTK	MozVox	AnonVox
Titanet-SVM	VCTK	100	63.0	54.5
	MozVox	58.3	100	68.7
	AnonVox	83.0	93.0	100
RawGAT-ST	VCTK	96.7	47.2	60.6
SSLAntiSpoofing	VCTK	98.7	91.5	66.0
SASV [23]	VCTK	99.6	50.0	52.2

across all datasets. The reason for this generalizability is not immediately clear. SSLAntiSpoofing also performs quite well on MozVox, but this may be due to the model's usage of wav2vec 2.0 XLSR [24] as a backbone, which is trained on the Common Voice dataset. MozVox is a single voice contained within the Common Voice dataset.

It is possible that a training dataset with sufficient channel variation would ameliorate these channel effects, but this remains to be seen. The external models compared here are trained on tens of thousands of examples. Beyond this already cumbersome training size, to offset the lack of variation present in the training data, these models are often supplemented with additional noise or temporal augmentation. What's more, the poor performance of state-of-the-art models trained on channel-diverse datasets, like InTheWild and MLAAD, suggests that the problem is not easily solved [10], [20]. The Titanet-SVM models in Table IV on the other hand, are trained on 5,400 total samples. The need for less training data makes Titanet-SVM much more adaptable to novel channel input.

VIII. CONCLUSION

This work highlights the difficulty of liveness detection across the main axes of variability in the wild - speaker, synthesizer, and channel. Speaker voice variability, keeping aside synthesizer and channel diversity, can be achieved easily with modern liveness models. Robustness against unseen synthesizers was examined in ASVSpoof 2019, however, leading models from the competition fare poorly against novel sythesizers. Likewise, channel variation was examined in ASVSpoof 2021, but novel datasets confound the leading models from that competition.

While these competitions are useful, this work demonstrates that solutions that excel with in-domain data struggle to generalize to common sources of variability. We demonstrate that spoof-detection which is resilient to channel conditions and robust against unseen synthesizers is an unsolved problem and emphasize the importance of synthesizer and channel diversity in future liveness detection endeavors. Though Titanet-SVM suffers these same limitations, it is able to uniquely deal with these challenges through its ease of adaptation to novel scenarios.

REFERENCES

[1] S. Gupta, K. Khoria, A. T. Patil, and H. A. Patil, "Deep convolutional neural network for voice liveness detection," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 775–779.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International conference on machine learning*. PMLR, 2017, pp. 195–204.

[4] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[5] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[7] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[8] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 2507–2522, 2023. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2023.3285283

[9] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[10] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[11] A. Chaubey, S. Sinha, and S. Ghose, "Speaker-specific thresholding for robust imposter identification in unseen speaker recognition," *arXiv preprint arXiv:2306.00952*, 2023.

[12] P. Gupta, H. A. Patil, and R. C. Guido, "Vulnerability issues in automatic speaker verification (asv) systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–14, 2024.

[13] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 19 594–19 621. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf

[14] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[15] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8102–8106, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238582695

[16] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, oct 2022. [Online]. Available: https://doi.org/10.1109/tpami.2021.3087709

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch-

esnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[19] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. 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*, 2021, pp. 1–8.

[20] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, “Mlaad: The multi-language audio anti-spoofing dataset,” 2024.

[21] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” 2022.

[22] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, “Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward,” 2022. [Online]. Available: https://arxiv.org/abs/2210.00417

[23] S. H. Mun, H. jin Shim, H. Tak, X. Wang, X. Liu, M. Sahidullah, M. Jeong, M. H. Han, M. Todisco, K. A. Lee, J. Yamagishi, N. Evans, T. Kinnunen, N. S. Kim, and J. weon Jung, “Towards single integrated spoofing-aware speaker verification embeddings,” 2023.

[24] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” 2020.

Appendix A
Equal Error Rate

EER is a common metric for comparing models, especially in biometric systems. It is calculated by setting a threshold such that the false positive rate is equal to the false negative rate. For a model such as Titanet the metric being calculated between any two particular samples is a cosine similarity with values mostly between 0 and 1. The similarity value can be used as a discriminant and then the threshold can be chosen such that the false positive and false negative rates are equal. The value of the error at that threshold is reported as the EER. Lower values of EER are preferred.