Trident of Poseidon: A Generalized Approach for Detecting Deepfake Voices

Thien-Phuc Doan
phucdt@soongsil.ac.kr
Soongsil University
Seoul, South Korea

Hung Dinh-Xuan
hungdx@soongsil.ac.kr
Soongsil University
Seoul, South Korea

Taewon Ryu
taewonryu@soongsil.ac.kr
Soongsil University
Seoul, South Korea

Inho Kim
inho107@soongsil.ac.kr
Soongsil University
Seoul, South Korea

Woongjae Lee
woongjae_l@ssu.ac.kr
Soongsil University
Seoul, South Korea

Kihun Hong
khong@ssu.ac.kr
Soongsil University
Seoul, South Korea

Souhwan Jung*
souhwanj@ssu.ac.kr
Soongsil University
Seoul, South Korea

Abstract

Deepfakes, an increasingly prevalent form of information attack, pose serious threats to security and privacy. Deepfake voice attacks, in particular, have the potential to cause widespread disruption, creating an urgent need for an effective detection system. In this research, we propose the **Trident of Poseidon** - a novel set of triad training strategies aimed at enhancing the generalizability of deepfake voice detection models. Our solution comprises three key components: (1) Supervised Contrastive Learning, (2) Hard Negative Mining by Audio Re-synthesizing, and (3) Effective Proactive Batch Sampling. Together, these enable the model to learn more robust features. Our extensive experiments demonstrate that our approach outperforms existing methods in both in-domain and out-of-domain testing scenarios, making significant strides toward securing digital media against deepfake voice attacks.

Furthermore, we conducted a deeper analysis to explore whether deepfake voices can be categorized into *families*. By identifying the factors that contribute to the formation of a *deepfake voice family*, we can better organize a deepfake voice corpus, thereby reducing the effort needed to combat the arms race challenge. Finally, to promote practical utility and community-wide adoption, we have made our solution publicly available as a web application available on *deepfake.aisrc.technology*[1], where users can utilize this tool to test for potential deepfake voices.

Keywords

Domain Generalization; Speech Synthesis; DeepFake Voice Detection

ACM Reference Format:
Thien-Phuc Doan, Hung Dinh-Xuan, Taewon Ryu, Inho Kim, Woongjae Lee, Kihun Hong, and Souhwan Jung. 2024. Trident of Poseidon: A Generalized Approach for Detecting Deepfake Voices. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3658644.3690311

1 Introduction

Audio deepfake (or deepfake voice[2]) technology, employing cutting-edge Deep Learning (DL) techniques like Text-To-Speech (TTS) and Voice Conversion (VC), is revolutionizing the way we interact with digital media [53]. These technologies leverage deep neural networks, particularly Generative Adversarial Networks (GANs), Diffusion models, and other advanced DL algorithms, to generate synthetic yet highly realistic speech content. The benefits of deepfake voice are numerous, providing significant advancements in fields such as entertainment, accessibility, and communication [3]. For instance, they can be used to generate voice overs in various languages[68], assist individuals with speech impairments[22], provide personalized virtual assistants[4], create an AI singer[13], and even recreate voice from a face image[18].

Threat model: The misuse of audio deepfake technology poses significant challenges, highlighting the gravity of the threat, especially in real-world. A notable example is the 2019 case where the CEO of a UK-based energy firm was duped into transferring €220,000 to a fraudulent account. The CEO was tricked by a phone call from an attacker who used audio deepfake technology to impersonate the voice of the company's chief executive [3]. In another

*Corresponding author
[1] Please send email to aisrc1@ssu.ac.kr for accessing our website.

[2] In this work, the terms *audio deepfake, deepfake audio*, and *deepfake voice* are used interchangeably.
[3] https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=54259f562241

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690311

instance, the popular podcast 'The Joe Rogan Experience' was targeted with deepfake audio clips that convincingly mimicked Rogan's voice and speech patterns, creating potential for misinformation and reputation damage [4]. These cases highlight the pressing need for robust detection and verification methods. However, the task is daunting due to the rapid advancements in deepfake technology, making the fakes increasingly difficult to distinguish from real ones.

To measure how people react to the audio deepfake, we made *Deepfake voice detection game*, attracted more than 250 participants during the World IT Show, located in Seoul, Korea. In this game, participants need to classify 5 random speech segments, 4 Korean and 1 English samples, into AI-generated voice and real one. Participants can listen the samples more than one time. We analyzed the answer sheets and found that: (1) achieving a perfect score (correctly classifying all 5 out of 5 samples) proves to be challenging, with fewer than 10% of participants winning; (2) the misclassification rate between fake and real classes is similar, indicating that the quality of real and fake voices is closely matched. In particular, 1.1% of participants answered all 5 samples incorrectly.

The post-game analysis reveals that deepfake voices are hard for the human auditory system to detect, even in situations without negative constraints. This highlights the danger that individuals could be more likely to fall for impersonation attacks, particularly in emergency situations.

Challenges of detecting deepfake: Current investigations into identifying audio deepfakes show potential but are hindered by substantial challenges[2]. Primarily, these models struggle with generalizability, indicating that they perform well on the specific types of data they were trained on but fail to maintain accuracy across unseen datasets. This limitation not only restricts the practical applicability of these detection systems in real-world scenarios, where audio deepfakes can vary greatly in quality and style, but also underscores the need for developing more generalized detector's architecture.

An additional significant obstacle in the fight against deepfake voice attacks is the reliance on outdated training datasets. As generative AI rapidly advances, detectors trained on older data find increasingly difficult to identify new deepfake samples. That evolution poses an arms race between detectors and adversaries who misuse deepfake technologies for harmful purposes.

Proposed system: In this work, we focus on enhancing the generalization of deepfake voice detection. We have introduced **Trident of Poseidon** - a novel training framework combined three techniques: (1) Supervised Contrastive Learning (SCL) training with (2) Hard negative mining by Audio Re-synthesizing (HAR) aimed at guiding the model to learn *in-variant features*. Furthermore, we proposed a novel (3) Proactive Batch Sampling (PBS) that shapes the training mini-batch in a manner that ensures a balanced and stable training mini-batch. The effectiveness of our training strategy lies in its ability to address the issue of an unbalanced training set while simultaneously conducting representation learning. We also introduced the implementation of the Variational Information Bottleneck (VIB) network to condense the feature space in order to ignore unnecessary features while preserving the model's generalizability. Evaluation results on both *in-domain* and *out-of-domain* settings showed outstanding performance of our system with the Equal Error Rate (EER) at **2.07%** on ASVSpoof Deepfake Track 2021 and **3.78%** In-the-wild dataset.

To meet the need for a newer dataset, we have compiled a large-scale corpus specifically for deep voice detection with diverse synthesizers. The utility of our **Diverse Synthesizer for Deepfake voice detection Corpus (DSD-corpus)** was compared with three other training sets, examining both intra-dataset and cross-dataset performance. Benchmarking results show an improvement of two different detectors' performance up to 20 times across evaluation sets. However, this surprising result is strong proof of the inevitable arms race issue.

The fact that the innovation of new deepfake technology frequently builds upon previous technologies. Inspired by the benefits of classifying malware into families, we wonder, **Could deepfake voices also be categorized into families?** The systematic categorization of deepfake voices into defined families could facilitate the process of analyzing and detecting unknown deepfake audio clips, particularly when the characteristics of these samples bear resemblance to those of a well-known family. This approach potentially reduces the resources and effort required to develop more sophisticated detection tools. We conducted **cross-synthesizer detection** experiment to identify the factors that could be used to categorize deepfake voices into families. Despite being in the early stages of this research, our findings showed potential for the existence of deepfake voice families. Based on our findings, we also provided several propositions for collecting a better deepfake voice dataset.

In short, the contribution of this work is listed as follows:

- We proposed the **Trident of Poseidon** — a novel deepfake voice detection system that combines three training techniques in collaboration with the VIB network. Experimental results showed that our system has greater generalizability than state-of-the-art solutions.
- We argue that the currently available datasets are *left behind* by the rapid development of audio deepfake generative technologies. To address this, we collected and published a new dataset — the **DSD-corpus** — to facilitate the development of deepfake voice detection. Utilizing the DSD-corpus significantly improves detection accuracy across a range of datasets.
- We carried out a detailed analysis to determine the common factors shared among different voice synthesizers, subsequently demonstrating the high possibility of the existence of **deepfake voice families**. Employing representative members from a well-known family could assist in addressing the arms race issue and facilitate the analysis of unknown deepfake samples.
- We published our detection system through a web application and asked end users for their usage experience. The feedback shows the necessity and reliability of our system in the real world to combat AI-based voice scamming/phishing attacks.

[4] https://www.linkedin.com/pulse/how-deepfakes-can-used-spread-misinformation-damage-reputations-vp/

2 System Design

2.1 Overview

The challenge is that current detection models do not generalize well. Existing systems have limitation to capture robust features. Suitable regularization techniques, which prevent overfitting and enhance model generalization, are often not adequately applied or are missing altogether. To address this challenge, we propose the "Trident of Poseidon" - an innovative training approach that integrates three distinct methodologies: Supervised Contrastive Learning (SCL) (**Sect. 2.2**), Hard Negative Mining through Audio Re-synthesis (HAR) (**Sect. 2.3**), and Proactive Batch Sampling (PBS) (**Sect. 2.4**). Additionally, we cooperate our triad strategies with a self-supervised learning framework based on Variational Information Bottleneck (VIB) specifically tailored for deepfake audio detection (**Sect. 2.5**).

2.2 Model Generalization Enhancement using Supervised Contrastive Learning (SCL)

Supervised Contrastive Learning (SCL)[26] is a learning technique that has been gaining attention in the field of machine learning due to its effectiveness in representation learning and classification tasks. Unlike traditional contrastive learning, which relies on unsupervised learning to learn representations by contrasting positive (similar) and negative (dissimilar) pairs, SCL incorporates class labels into the learning process, making it a supervised learning method. The supervised contrastive loss can be defined as follows:

$$\mathcal{L}_{SCL} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \quad (1)$$

where $i \in I \equiv \{1, \ldots, N\}$ is served as the indicator of a randomly selected anchor sample, I is usually called *multi-viewed batch*, $A(i) \equiv I \setminus \{i\}$, $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ is the of all positive in multi-viewed batch. z indicates the latent feature, and τ is the temperature value.

This loss function combines the principles of contrastive learning with the benefits of supervised learning. It encourages the model to learn representations that are close for instances from the same class and far apart for instances from different classes. This leads to more discriminative representations, which can significantly improve the model's performance on classification tasks.

2.3 Hard Negative Mining by Audio Re-synthesizing (HAR)

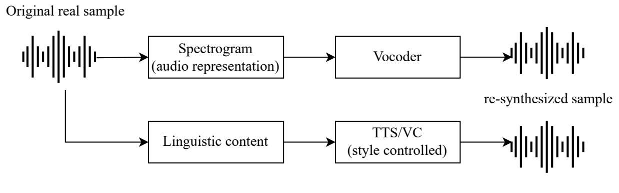


Figure 1: Audio re-synthesizing methods flow-graph.

Hard negative mining is indeed an important concept in representation learning, especially Supervised Contrastive Learning (SCL)[47]. In the context of contrastive learning, *negative* instances are those that are different from a given instance, while *positive* instances are similar or identical to the given instance - the *anchor*. Hard negatives are instances that are different from the *anchor* but could be easily mistaken as similar due to certain shared characteristics.

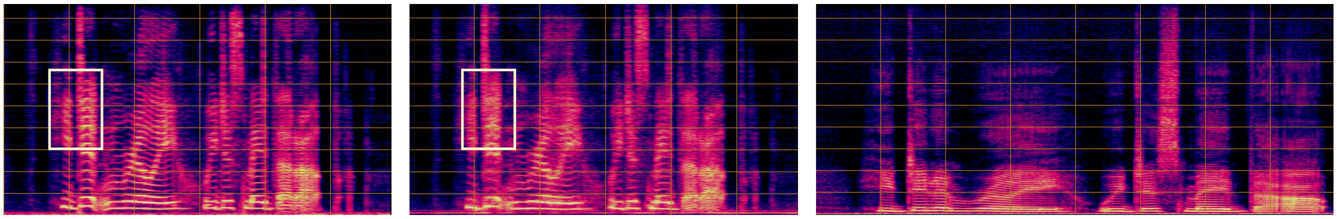
Look back to the **Formula. 1**, when the model is trained, it's encouraged to make $z_i \cdot z_p$ large for positive pairs (to increase the numerator) and $z_i \cdot z_a$ small for negative pairs (to decrease the denominator). This would make the overall fraction large, and since the loss is the negative log of this fraction, it would make the loss small. Hard negative examples are challenging for the model to distinguish from the anchor. In terms of the equation, this means that for hard negative samples, $z_i \cdot z_a$ might be large. By including these hard negatives in the set $A(i)$, we force the model to learn representations z_i and z_a that make $z_i \cdot z_a$ small, even for hard negative examples. This can lead to more discriminative representations that can better distinguish between different classes.

Indeed, finding hard negative samples in the domain of deepfake voice detection is challenging. Speech signal encompasses two factors: the **linguistic content** and the **vocal style**. In the context of deepfake detection, the goal is to identify discrepancies in the vocal style that indicate a voice has been artificially generated, while ignoring variations in the linguistic content. Ideally, we aim for pairs of real and synthetic samples that share identical linguistic content but differ in vocal style due to the potential presence of AI-generated artifacts in the synthetic voice, even if they sound similar. However, collecting such samples can be difficult because it requires having both real and fake utterances of the same linguistic content. For effective hard negative mining, the fake sample should have three principles:

- **Principle 1 (Same Content):** The pair of real and fake should speak the same content.
- **Principle 2 (Similar Vocal Style):** The fake speech should generated to mimic the real utterance's vocal style, ideally with the same speed and volume.
- **Principle 3 (AI generated sample):** The deepfake sample should be generated by a generative AI model to distinguish it from other forms of presentation attacks[49], such as replay attack.

Following these principles, we utilize **audio re-synthesizing** technique to generate hard negative samples for the anchor x_i. Audio re-synthesizing, as illustrated in **Fig. 1**, has two common ways to re-synthesize speech samples:

- **Neural Vocoder-Based Re-Synthesizing (Upper Branch):** A neural vocoder, such as WaveNet or WaveGlow, can convert high-dimensional representations like spectrograms back into raw audio waveforms. For generating hard negative samples, you could start with a real audio sample, extract its spectrogram, and then use neural vocoders to convert the spectrogram back into an audio waveform.
- **TTS/VC Based Re-Synthesizing (Lower Branch):** TTS and VC models can generate synthetic speech from provided



a. Original sample b. Vocoder-based re-synthesized sample c. TTS/VC-based re-synthesized sample

Figure 2: Spectrogram of the original sample (VCTK p310_034) and its two different re-synthesized samples.

linguistic content with a target speaker's voice style. To generate a hard negative sample, we could extract the source speech's linguistic content, and then together with the original speech as a style reference sample, feed it to the TTS/VC model to generate a new sample with the same speaker style and source content.

The examples of two audio re-synthesizing approaches are shown in **Fig. 2**. It is evident that the vocoder-based re-synthesizing approach (**Fig. 2.b**) maintains time and frequency domain similarities to the original sample, as the TTS/VC-based approach (**Fig. 2.c**) faces difficulties in accurately recreating the signal while adhering to all defined principles. Re-synthesized samples in both cases contain AI-related artifacts that could potentially influence the training procedure. Therefore, in this work, we employ the neural vocoder-based audio re-synthesizing method to enhance the development of our model.

2.4 Proactive Batch Sampling (PBS)

The problem of class imbalance is a common issue in training machine learning models, where the dominant class may bias the model's learning. In Supervised Contrastive Learning (SCL), maintaining a balanced number of samples from different classes in each mini-batch is critical. To tackle this issue, we propose a method called Proactive Batch Sampling for setting up a balanced training mini-batch in SCL, which we believe will effectively enhance the performance of audio deepfake detection models.

Given an anchor sample $x_i \in D_{Real}$ in the i^{th} training batch:

- Randomly select p other real samples from D_{Real}.
- Choose k augmentation methods uniformly at random. These augmentation methods are applied exclusively to the anchor x_i for the current batch i^{th}.
- Randomly select v vocoder systems to re-synthesize v *vocoded samples* from x_i. In this work, we use *vocoded samples* and *re-synthesized samples* interchangeably since we only use the Neural Vocoder-based re-synthesizing technique.
- Randomly select q other fake samples from D_{Fake}.
- Apply v augmentation methods, chosen randomly, exclusively to the *vocoded samples.* Note that each vocoded sample is used to generate *augmented vocoded sample* once.

In this setup, k, k', and v serve as hyperparameters because of limited available augmentation and vocoder algorithms, while p and q are dynamically adjusted to maintain balance in the mini-batch, following the formula:

$$1 + k + p \approx 2v + q \tag{2}$$

This proactive approach ensures a balanced representation of real and fake samples in each mini-batch, which is crucial for effective learning in SCL. Moreover, by applying different augmentation methods and vocoder systems to various samples, we increase the diversity of the samples in the mini-batch, potentially leading to a more robust and generalizable model. The detail of utilized vocoder and augmentation methods is shown in **Table. 1**

Table 1: Description of Vocoder and Augmentation Methods

	Name	Description
Vocoder	HN-Sinc-NSF	Digital signal processing (DSP) based vocoder
	HIFI-GAN	Generative adversarial network (GAN) based vocoder
	WaveGlow	Flow-based Deep Generative vocoder
Augmentation Method	Reverberation	Add an echo effect to the audio, simulating the sound reflections in a room or space.
	Background noise	Add different types of noise to the audio to simulate real-world environments.
	Speed	Alter the speed of the audio playback, which can change the perceived pitch and length of the audio.
	Volume	Adjust the loudness of the audio.
	Pitch	Change the perceived pitch of the audio without affecting the speed.
	Telephone	Apply ALAW, ULAW, and other telephone codecs with bandpass filtering.

2.5 VIB-based Audio Deepfake Detection Architecture

Our proposed architecture, as shown in **Fig. 3**, for audio deepfake detection comprises two primary components: a front-end feature extractor based on the Wav2Vec 2.0 model, and a back-end classifier that leverages the Variational Information Bottleneck (VIB) approach.

The front-end feature extractor utilizes a self-supervised learning framework called Wav2Vec 2.0[6], which has demonstrated its efficacy in converting raw audio waveforms into high-dimensional

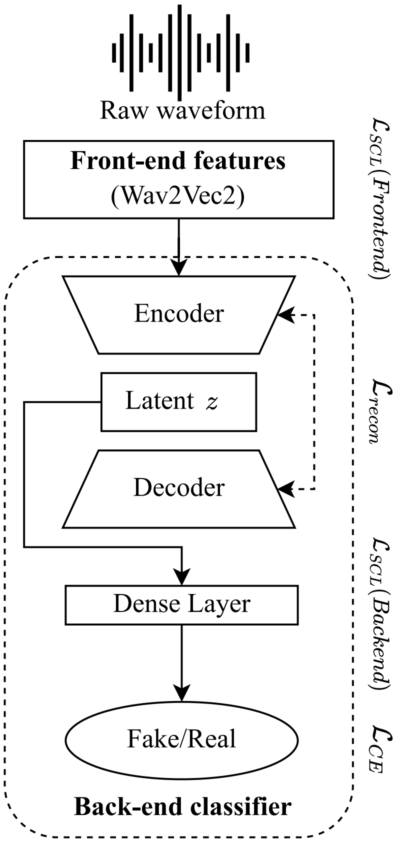


Figure 3: Overview of the proposed system's network architecture.

representations. This model has proven to excel in various downstream tasks, such as speech recognition. However, these high-dimensional features often encapsulate both linguistic and voice style dependencies. While these dependencies can be beneficial for certain tasks, they may not be necessary for audio deepfake detection and could potentially introduce noise into the system. To address this, the high-dimensional features are compressed into a smaller dimensional latent space via the information bottleneck (IB) back-end classifier. This compression process aims to reduce the unnecessary features and retain the most discriminative ones. However, a potential pitfall of this compression process is the risk of overfitting, where the model becomes excessively specialized on the training data and performs poorly on unseen data.

To mitigate this risk, the VIB-based network incorporates a reconstruction loss with an additional decoder as a regularization technique. This encourages the model to retain as much information as possible during the compression, thereby preventing the model from discarding too much information, leading to overfitting. The reconstruction loss is written as:

$$\mathcal{L}_{recon} = \mathbb{E}_{q(z|x)}(-\log p(y|z)) + \beta D_{KL}(q(z|x), r(z)) \qquad (3)$$

where $q(z|x)$ denotes the estimate of the posterior distribution of latent z (i.e., encoder output) $p(y|z)$ means a variational approximation of $q(z|x)$ (i.e., decoder output), and $r(z)$ means an estimate of the prior distribution of z treated as standard normal distribution [1] while computating Kullback-Leibler (KL) divergence. The value β for D_{KL} is used as a regularization term.

Table 2: Training and Evaluation datasets' description

Dataset Name	Description	# fake	# real
LA19 (train/dev)[41]	ASVSpoof 2019 training and development partition, including real and 6 different types of spoofing samples	22800	2580
LA19 eval[41]	ASVSpoof 2019 evaluation partition, 11 new types of spoofing samples	63882	7355
DF21[61]	ASVSpoof 2021 deepfake audio track evaluation partition, much large evaluation set comprising 100 different types of fake samples	589212	22617
InWild[38]	In-the-Wild dataset collected from publicly available video and audio files of English-speaking celebrities and politicians	11815	19963
FoR[46]	A normalized version of the Fake or Real dataset, a collection of real (Arctic, LJSpeech, VoxForge) and TTS (Deep voice 3, Wavenet TTS) samples	34695	34605
Wavefake[17]	A collection of synthesized samples in English and Japanese from 6 different generative models, including MelGAN, ParallelWaveGAN, Multi-band MelGAN, Full-band MelGAN, HiFi-GAN, WaveGlow.	117985	-
MLAAD[39]	A multi-language audio dataset comprising 52 TTS models in 23 different languages	74,000	-

The loss function for the whole training progress is shown in **Formular. 4**.

$$\mathcal{L} = \mathcal{L}_{SCL}(front-end) + \mathcal{L}_{SCL}(back-end) + \lambda\mathcal{L}_{recon} + \mathcal{L}_{CE} \quad (4)$$

The weight $\lambda \in [0, 1]$ of the reconstruction loss plays a pivotal role in controlling the amount of information that we compress. A low reconstruction loss (smaller λ) implies that more general information is retained. Conversely, a higher reconstruction loss (larger λ) indicates that more domain-specific features are retained. In this work, we found that $\lambda = 10^{-5}$ and $\beta = 0.05$ give the best performance for our task.

3 System Evaluation

3.1 Dataset

For fair comparison to other state-of-the-art solutions, we trained our system using ASVSpoof 2019 dataset (LA19)[57]. We evaluated our system with 2 ASVSpoof challenge sets, LA19 eval and DF21 [34], and 4 out-of-distribution (OOD) datasets, In-the-Wild [40], Fake-or-Real [46], Wavefake [17], and MLAAD [39]. The dataset description can be found in **Table. 2**.

3.2 Evaluation Metrics

The Equal Error Rate (EER) is a commonly used metric in the evaluation of biometric systems, such as voice recognition or fingerprint identification systems. The EER is defined at the point where the system's False Acceptance Rate (FAR) equals its False Rejection Rate (FRR). Mathematically, the EER can be expressed as follows:

- Let $FAR(\theta)$ denote the False Acceptance Rate at a given threshold θ, which is the rate at which fake voices are incorrectly classified as real ones by the system.

- Let $FRR(\theta)$ denote the False Rejection Rate at the same threshold θ, which is the rate at which real voices are incorrectly classified as fake ones by the system.

The EER is found when:

$$EER = FAR(\theta) = FRR(\theta)$$

At this point, the system has balanced its security and usability. A lower EER indicates a more accurate system.

In addition to the EER, this study also employs detection accuracy (ACC) when the benchmarking datasets lack real samples (for instance, the MLAAD and WaveFake datasets). Detection accuracy is defined as follows:

$$\text{ACC} = \frac{TP + TN}{N}$$

Where TP indicates True Positive and TN indicates True Negative. A higher accuracy indicates that the model is more effective.

3.3 Experimental Setup

We trained our model using NVIDIA RTX A5000-16G, and limited the batch size to less than 16 because we chose XLS-R-300M [5] as the front-end feature extractor, which has around 300 million parameters. The front-end extractor consists of a linear projection layer designed to map the final output of XLS-R into a smaller dimensional space, $R^{n\times1024} \rightarrow R^{n\times128}$, following the success from[52]. The VIB-based back-end classifier is constructed using `2xLinear` layers for both the Encoder and Decoder, each with a hidden size of 64. The dense layer begins with average pooling, followed by a linear projection layer. The dense layer initiates with an average pooling step, subsequently followed by a linear projection layer. We experimented with multiple configurations to identify the optimal PBS setting within our hardware constraints. The best PBS setting identified in this study is $p = q = 1; k = 5; v = 3$. We used ADAM [28] as the optimizer, with a learning rate from 1×10^{-8} to 1×10^{-4} scheduled by the CyclicLR[50] algorithm. We applied early stopping when the validation accuracy on the development set showed no more improvement for 10 epochs.

3.4 Experiment Results

We benchmarked our system in two scenarios: in-domain and out-of-domain settings. To ensure an equitable comparison, we trained our model using the ASVSpoof 2019 (LA19) training partition as other state-of-the-art deepfake voice detectors. For **in-domain** testing, we evaluated on ASVSpoof 2021 Deepfake partition (DF21) as it from the same challenge to the LA19. For **out-of-domain** testing, which aims to show the superior generalization ability of our model, we evaluated on In-the-Wild dataset (InWild), which consists of a vast collection of samples primarily sourced from social networking sites, effectively simulating real-world conditions.

Table 3: Performance comparison, in EER (%), of ours and other deepfake voice detection solutions, training with the LA19 dataset and testing on the in-domain dataset - DF21, and out-of-domain dataset - InWild

	Network architecture	DF21	InWild
[64]	Wav2Vec + WavLM + HuBERT	11.78	24.27
[59]	Wav2Vec + ASDG	8.07	24.50
[56]	Distillation Wav2vec	5.67	6.10
[52]	Wav2Vec + AASIST	2.84	12.19
[48]	Wav2Vec + Conformer	2.58	12.23
[35]	One class Distillation Wav2Vec	2.27	7.68
(Our)	**Wac2Vec + Linear**	2.17	4.51
(Our)	**Wac2Vec + VIB**	**2.07**	**3.78**

The **Table. 3** presents the evaluation results of various deepfake voice detection solutions in EER (%). In the in-domain evaluation on DF21, notable approaches showcase varying levels of success, with EER values ranging from 11.78% to 2.07%. Despite the use of the simple back-end network with `3xLinear layers`, we still achieve superior performance at EER of **2.17%**. Notably, **VIB-based** back-end system has the lowest EER at **2.07%**, indicating its robustness in distinguishing between real and deepfake voices within the same dataset domain.

On the other hand, the out-of-domain evaluation on the InWild dataset presents the significant challenge of distribution shift. While some systems demonstrate strong performance in the in-domain setting, their effectiveness diminishes when tested on out-of-domain data, as seen in the EER values ranging from 24.27% to 3.78%. For instance, the solution proposed by[52] achieves the competitive EER of 2.84% on DF21 but 4 times worse at EER of 12.19% on InWild. Our system maintains its competitive edge with an EER of **4.51%** and **3.78%**, for the *linear-based* and *VIB-based* systems, showcasing its ability to generalize well across different dataset domains. This suggests that our approach offers a level of adaptability and robustness crucial for detecting deepfake voices in real-world scenarios where the manipulation techniques may vary, emphasizing the significance of developing versatile detection systems capable of handling distribution shift challenges more effectively.

3.5 Ablation Study

We conducted an ablation study to clarify which components have the most impact on enhancing generalization and model robustness. The experiment results, as shown in **Table. 4**, emerge from sequentially omitting one strategy at a time:

- Without Hard negative mining by Audio Re-synthesizing (w/o HAR), the system is trained with batches sampled through PBS. However, instead of utilizing re-synthesized samples, existing fake samples from the dataset are used.
- Without Supervised Contrastive Learning (w/o $\mathcal{L}_{\text{SCL}}$), the model is trained solely with Cross-entropy loss. The training batches are sampled via PBS, including re-synthesized samples as per the HAR configuration.

- Without Proactive Batch Sampling (w/o PBS), the model undergoes training within the SCL framework, incorporating re-synthesized samples. However, the proportion of fake to real samples is not managed and is randomly selected during the training process.

Table 4: Ablation study on three different evaluation sets. We removed one of three methodologies of the triad of our training strategies (i.e., HAR, SCL, and PBS). We also compare two different back-end networks, the simple 3xlinear layers, and the Variational Information Bottleneck (VIB). The performance of these systems is shown in EER (%).

	Training Strategy			**Back-end Network**	
Eval set	w/o HAR	w/o $\mathcal{L}_{\text{SCL}}$	w/o *PBS*	3xLinear	VIB
LA19 eval	**0.57**	2.88	10.73	2.88	2.68
DF21	3.38	5.49	8.66	2.17	**2.07**
InWild	9.78	8.78	11.88	4.51	**3.78**
Pooled	4.58	5.72	10.42	3.18	**2.84**

Proactive Batch Sampling (PBS) emerged as the most pivotal component of our triad training strategy. Its omission led to a significant performance decline, with the model's error rate soaring to 10.42%. The Hard negative mining by Audio Re-synthesizing (HAR) method, although less impactful compared to the other two strategies, still contributed to enhancing the model's overall performance by reducing the error rate by 1.4%. The HAR approach creates real-fake pairs, enabling the model to disregard irrelevant linguistic content and concentrate on more robust features for detection. Interestingly, the absence of HAR yielded the best results on the LA19 evaluation set. This phenomenon is reasonable since LA19 comprises not only neural-based synthesized speech but also more waveform concatenation-based spoofing samples. Utilizing re-synthesized samples in training as hard negative samples might cause the model to become biased towards detecting neural-based synthesized samples. Consequently, our proposed system's effectiveness is limited to neural-based synthesized voice. Given our focus on deepfake voice detection, this outcome is deemed acceptable. Integrating all three training strategies significantly lowers the error rate, demonstrating effectiveness across different back-end networks. This experiment underscores that the traditional Supervised Contrastive Learning (SCL) framework can be enhanced with a well-designed data sampling approach, such as our proposed PBS. Furthermore, in the realm of deepfake voice detection, employing hard negative samples generated through the re-synthesizing method not only boosted performance but also addressed the challenge of limited fake data availability.

The VIB-based network outperformed the simple back-end network of 3xLinear layers, as expected, primarily due to its capability to condense the information essential for the task at hand, especially when leveraging a large self-supervised learning front-end feature extractor like XLSR-300M. This efficiency in information compression allows the model to focus on the most relevant features for deepfake voice detection, enhancing its performance. When conducting in-domain evaluations, such as those on the ASVSpoof dataset (including both LA19 and DF21), the performance of the VIB-based network remained consistent. However, the most notable performance enhancement was observed during evaluations on out-of-domain datasets, such as InWild. This improvement underscores the VIB-based network's superior generalization capabilities, making it a promising approach for enhancing the robustness and reliability of deepfake voice detection systems.

4 DSD-Corpus: A New Audio Deepfake Detection Dataset

The challenge of detecting deepfake voices is significantly compounded by the gap between ideal laboratory conditions and the unpredictable nature of real-world environments. In controlled settings, models might demonstrate high accuracy rates, but these successes often do not translate when the models are applied outside the lab. This discrepancy arises because real-world scenarios introduce a plethora of variables and complexities that are typically absent or controlled for in lab settings. To bridge this gap and enhance the practical applicability of deep voice detection technologies, it's crucial to train models on data that reflect the diversity and complexity of the real world. In response to this need, we argue that the collected dataset should follow these four main rules:

- **Diversity of Sources:** Collect data from a wide range of sources to capture various scenarios and contexts. Ensure the inclusion of both controlled (studio-recorded) and uncontrolled (real-world) audio samples.
- **Temporal Coverage:** Incorporate data from different time periods to address changes in recording technology and deepfake techniques over time. Aim for a historical span that reflects both past and contemporary audio characteristics.
- **Linguistic Variety:** Include audio samples in multiple languages to ensure the model's effectiveness across different linguistic contexts. Prioritize languages based on geographical diversity and the prevalence of the language among potential users.
- **Balance Among Categories:** Ensure a balanced representation of various types of audio deepfakes and genuine audio samples. Aim for an equitable distribution across different categories to avoid model bias and improve detection accuracy across all types of deepfakes.

4.1 Dataset Collecting Strategy

Adhering to the fundamental principles outlined previously, we have gathered both authentic and counterfeit samples up until the end of 2023. These samples have been meticulously categorized into several groups, as illustrated in Table 5, hereby designated as the **DSD-corpus**.

Real Sample Collection: Our methodology for assembling a robust collection of controlled audio samples involved the random selection of 100 samples per speaker from renowned databases such as LibriSpeech[43], VCTK[60], and AIHUB. This approach was designed to ensure a comprehensive representation of the diverse audio characteristics these databases offer. We also extended our collection to include uncontrolled samples. These were crowdsourced from various social networking sites, such as YouTube, Facebook, and Instagram. This strategy was aimed at capturing a wide spectrum of acoustic environments, and speaking styles.

Table 5: The detailed description of our Diverse Synthesizer for Deepfake Voice Detection - DSD-Corpus

Fake samples						
Synthesizer	**TTS/VC**	**Acoustic**	**Vocoder**	**Training set**	**Description**	**#samples**
StarGANv2-VC[32]	VC	Auto-Encoder, GAN	Parallel-WaveGAN[62] (VCTK)	VCTK	Multi-speaker English, cloning random 20 VCTK speakers.	3000
DiffGAN[33]	TTS	Diffusion	HifiGAN (LJSpeech)	LJSpeech.	English, cloning LJSpeech voice.	500
Diff-HierVC[11]	VC	Diffusion	HifiGAN[29] (VCTK)	LibriTTS	Multi-speaker English, cloning random 20 VCTK speakers.	2000
DDDM-VC[12]	VC	Diffusion	HifiGAN (VCTK)	LibriTTS	Multi-speaker English, cloning random 20 VCTK speakers.	2000
Tortoise-TTS[8]	TTS	Diffusion	Univnet[21] (LibriTTS)	LibriTTS	Multi-speaker English, cloning 30 English voices utilizing pre-trained model.	6000
QuickVC[19]	VC	Transformer, Flow	MS-iSTFT-VITS[25] (E2E)	VCTK	Multi-speaker English, cloning random 20 VCTK speakers.	3000
VITS-TTS[27]	TTS	Transformer, Flow	HifiGAN (E2E)	VCTK, LJSpeech	Multi-speaker English, cloning random 41 VCTK and LJSpeech speakers.	4100
VITS AIHUB[27]	TTS	Transformer, Flow	HifiGAN (E2E)	AIHUB[1]	Utilized VITS, re-trained on AIHUB dataset to generate 163 Korean voices.	10595
StyleTTS2[31]	TTS	BiLSTM, Style Diffusion	HifiGAN (E2E)	VCTK, LJSpeech	Multi-speaker English, cloning random 13 VCTK and LJSpeech speakers.	1300
OpenVoice-TTS[45]	TTS	VITS-based, Auto-Encoder	HifiGAN (E2E)	Chinese-MSML[45]	Multi-speaker Chinese, cloning random 13 Chinese voices utilizing pre-trained model.	2600
MeloTTS[69]	TTS	VITS-based, Auto-Encoder	HifiGAN (E2E)	MSML[69]	Multilingual, including Japanese, Spanish, Korean, Chinese, and French, cloining 5 voices utilizing pre-trained model on HuggingFace[3].	1000
MMSTTS[44]	TTS	VITS	HifiGAN (E2E)	MMS	Multilingual, cloning 9 different voices on 9 different languages utilizing pre-trained model on HunggingFace[3].	1800
SeamlessM4T-TTS[7]	TTS	Transformer	HifiGAN (E2E)	Seamless-M4T	Multilingual, cloning 11 different voices on 9 different languages utilizing pre-trained model on HunggingFace[3].	2200
Elevenlabs[2]	TTS	Unknown	Unknown	Unknown	A commercial AI voice generator website, cloning 50 English voices utilizing the default voice provided by Elevenlabs.	5000
Social Network Sites	Both	Unknown	Unknown	Unknown	Collected samples from social networking sites (SNS), mainly in Korean, English, and Japanese.	960
					Total	**46055**

Real samples		
Real Dataset	**Description**	**# samples**
LibriSpeech[43]	Large English corpus with 767 speakers. We drew 166 speakers for our dataset.	16600
AIHUB[1]	Large Korean corpus with 163 speakers	16300
VCTK[60]	English corpus with 109 speakers.	10900
Social Network Sites	Collected from social networking sites (SNS), mainly in Korean, English and Japanese	2530
	Total	**46330**

*Training set of the pre-trained vocoders is listed in parentheses, except end-to-end (E2E) system. [1]https://www.aihub.or.kr, [2]http://elevenlabs.io, [3]https://huggingface.co

Fake Sample Collection: The initiation of our fake audio sample collection involved identifying and acquiring voice synthesis systems from repositories such as GitHub, Hugging Face, and other open-source repositories. Leveraging publicly available pre-trained models, we embarked on the generation of fake voices employing both Text-to-Speech (TTS) and Voice Conversion (VC) methodologies. For the TTS systems, we utilized ChatGPT (developed by OpenAI) to generate a corpus of 2000 sentences across varied contexts. From this GPT-based corpus, thousands of samples were synthesized randomly, ensuring a rich tapestry of audio diversity. For VC systems, source content speech samples from a different dataset were utilized. Both the TTS and VC systems were equipped with mechanisms for controlling speaker style, though through different approaches. Systems with predefined speaker IDs facilitated the generation of audio in specific, predetermined voice styles. Conversely, systems that support speaker embedding allow for the use of reference speech. Based on this reference speech, the system calculates a speaker embedding that captures the unique characteristics of the speaker's style. This enables the creation of audio that closely mimics the vocal traits of the reference sample. To enrich the diversity of our fake sample collection, we employed reference samples from various voice corpora. Additionally, we crowdsourced fake voices from social networking sites, treating these as uncontrolled samples.

We partitioned our collected dataset into three subsets: Training, Developing (Dev), and Evaluating (Eval) with the ratio of `10:10:80`.

4.2 DSD-corpus Evaluation

Table 6: Evaluation result on cross data benchmarking

Train	Model →	Our system		SSL AASIST[52]	
set	Eval set ↓	EER(%)	ACC(%)	EER(%)	ACC(%)
DSD-corpus	**InWild**	**1.99**	-	12.93	-
	DF21	3.72	-	10.83	-
	FoR	**3.93**	-	17.64	-
	WaveFake	-	**99.79**	-	**83.31**
	MLAAD	-	**98.53**	-	**84.01**
	DSD-Corpus	0.11	-	0.94	-
ASVSpoof 2019	**InWild**	3.78	-	**10.48**	-
	DF21	**2.07**	-	**2.85**	-
	FoR	10.32	-	**12.19**	-
	WaveFake	-	86.67	-	70.08
	MLAAD	-	72.44	-	60.27
	DSD-Corpus	32.27	-	26.49	-
	ASVSpoof 2019	2.68	-	0.22	-
Fake or Real	**InWild**	7.27	-	35.72	-
	DF21	4.07	-	10.55	-
	WaveFake	-	15.33		1.52
	MLAAD	-	34.50	-	26.83
	DSD-Corpus	28.58	-	34.71	-
	Fake or Real	0.14	-	0.39	-
In the wild	**DF21**	8.01	-	5.23	-
	FoR	13.91	-	18.21	-
	WaveFake	-	93.68	-	81.71
	MLAAD	-	63.81	-	53.56
	DSD-Corpus	26.31	-	26.41	-
	In the wild	0.58	-	0.25	-

To rigorously evaluate the effectiveness of the DSD-corpus in facilitating the development of robust audio deepfake detection models, we undertook a comprehensive validation process. We trained our model, along with SSL-AASIST[52], using the DSD-corpus and conducted evaluations across six diverse sets. These sets were carefully designed to cover a broad spectrum of scenarios, challenges, and deepfake generation techniques, as detailed in **Table 2**.

The detailed results shown in **Table. 6** compellingly demonstrate the effectiveness of the **DSD-corpus** in advancing deepfake voice detection systems. This corpus has been instrumental in training systems that perform exceptionally well across a variety of datasets, covering both controlled environment domains (e.g., DF21, FoR, Wavefake, MLAAD) and uncontrolled environment domains (e.g., InWild). Our system trained on the DSD-corpus achieved the best performance on the majority of these datasets, with the notable exception of DF21, where the error rate was 1.7% higher compared to training with the LA19 dataset. This discrepancy is understandable given that LA19 and DF21 originate from the same challenge and can be considered to have a closed distribution corpus.

Particularly impressive was the performance on the InWild dataset, where the DSD-corpus-trained models significantly outperformed the others, achieving an Equal Error Rate (EER) of 1.99%. This trend continued with the newer Wavefake (2020) and MLAAD (2024) datasets, further underscoring the DSD-corpus's utility in training state-of-the-art deepfake voice detection systems.

Conversely, when models not trained on our DSD-corpus were tested against it, they exhibited significantly higher error rates, ranging from 26.31% to 34.71%. This stark contrast, with error rates more than double those of models trained on the DSD-corpus, underscores the corpus's value in enhancing the performance of deepfake voice detection systems, suggesting its superior capability.

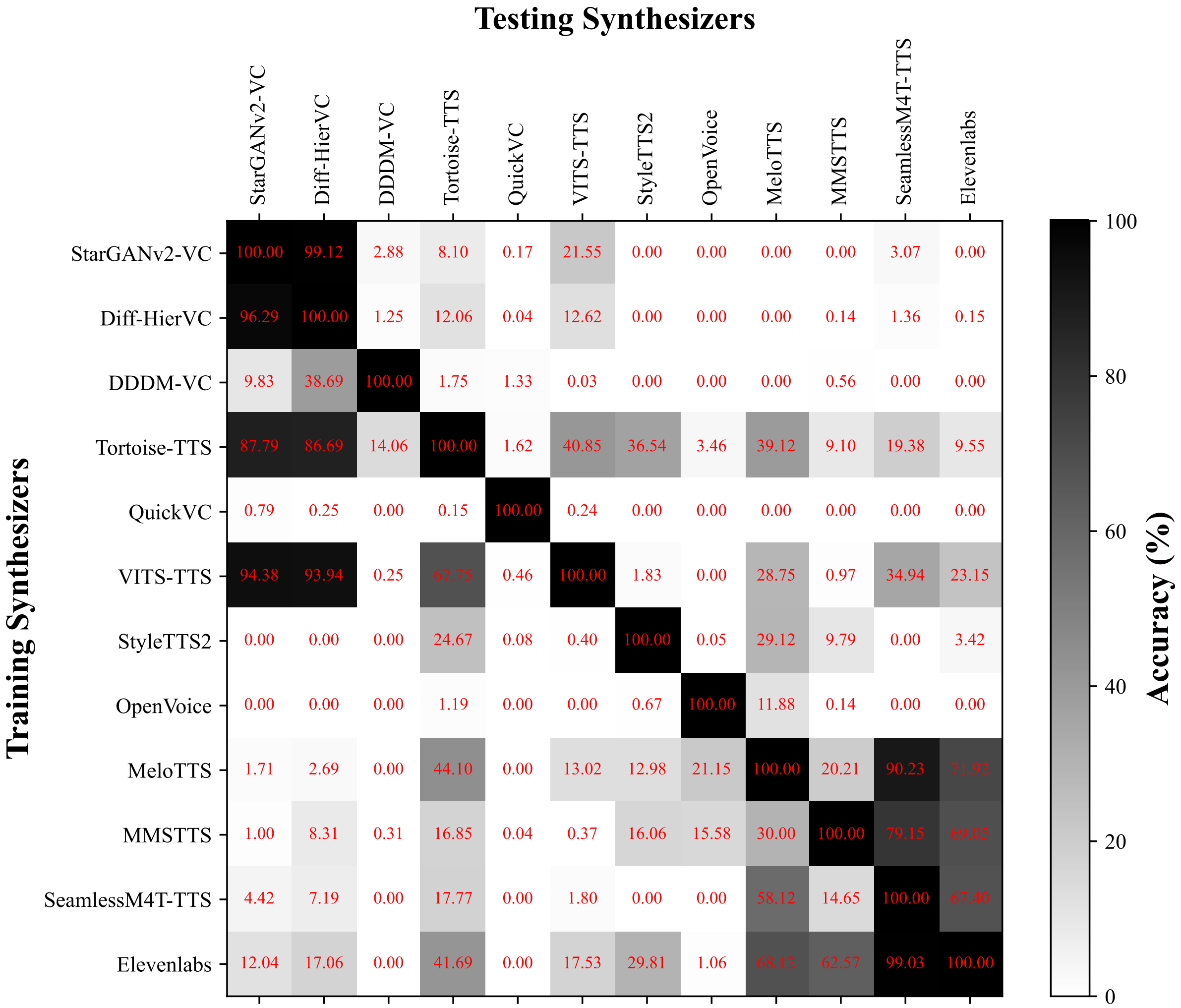
The use of outdated deepfake voice datasets, such as FoR, underscores the persistent challenge of distribution shift. Our model encountered significant difficulties in accurately identifying fake samples from the WaveFake and MLAAD datasets, achieving accuracy rates of only 15.33% and 34.50%, respectively. In short, the experimental results show the effectiveness of using our new **DSD-corpus** in developing a better deepfake voice detection system.

On the other hand, the SSL AASIST system, when trained on our new DSD-corpus, could not generalize well to older benchmarking sets (e.g., For, DF21, and InWild) compared with other versions. This underscores that the generalization capability has to be achieved by two factors: *(1) the well-designed model* and *(2) the large and diverse training set.* Furthermore, this situation accentuates the ongoing arms race in deepfake technology, prompting us to pursue a more generalized solution to effectively address this challenge.

5 Do Deepfake Voices Have Families?

The rapid advancement of generative AI across both academic and industrial landscapes has significantly intensified the arms race between deepfake voice detection (blue team) and generation (red team). In other adversarial detection systems, such as malware and network intrusion detection, harmful examples are analyzed and categorized into **families**. This categorization provides insights into the adversaries, facilitating the analysis of unknown attack samples in the future. Inspired by this approach, we pose an intriguing research question: *"Do deepfake voices have families? If yes, what factors contribute to the formation of a family?"*

This question goes beyond the rhetorical, laying the foundation for our strategy to enhance generalization capabilities. In this



	StarGANv2-VC	Diff-HierVC	DDDM-VC	Tortoise-TTS	QuickVC	VITS-TTS	StyleTTS2	OpenVoice	MeloTTS	MMSTTS	SeamlessM4T-TTS	Elevenlabs
StarGANv2-VC	100.00	99.12	2.88	8.10	0.17	21.55	0.00	0.00	0.00	0.00	3.07	0.00
Diff-HierVC	96.29	100.00	1.25	12.06	0.04	12.62	0.00	0.00	0.00	0.14	1.36	0.15
DDDM-VC	9.83	38.69	100.00	1.75	1.33	0.03	0.00	0.00	0.00	0.56	0.00	0.00
Tortoise-TTS	87.79	86.69	14.06	100.00	1.62	40.85	36.54	3.46	39.12	9.10	19.38	9.55
QuickVC	0.79	0.25	0.00	0.15	100.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00
VITS-TTS	94.38	93.94	0.25	67.75	0.46	100.00	1.83	0.00	28.75	0.97	34.94	23.15
StyleTTS2	0.00	0.00	0.00	24.67	0.08	0.40	100.00	0.05	29.12	9.79	0.00	3.42
OpenVoice	0.00	0.00	0.00	1.19	0.00	0.00	0.67	100.00	11.88	0.14	0.00	0.00
MeloTTS	1.71	2.69	0.00	44.10	0.00	13.02	12.98	21.15	100.00	20.21	90.23	71.92
MMSTTS	1.00	8.31	0.31	16.85	0.04	0.37	16.06	15.58	30.00	100.00	79.15	69.05
SeamlessM4T-TTS	4.42	7.19	0.00	17.77	0.00	1.80	0.00	0.00	58.12	14.65	100.00	67.40
Elevenlabs	12.04	17.06	0.00	41.69	0.00	17.53	29.81	1.06	68.12	62.57	99.03	100.00

Figure 4: Cross-Synthesizer detection accuracy: The X-axis represents the testing sets, and the Y-axis represents the training sets. Each set contains samples exclusive to its corresponding synthesizer, alongside a common list of real samples. The darker the cell, the higher the accuracy.

context, the term *family* refers to a group of synthesizers whose synthetic speech shares several characteristics or properties. By identifying and categorizing deepfake voices into distinct families, we argue that the development of deepfake detection systems could be significantly streamlined. Specifically, having a comprehensive collection of representative samples from each identified family could simplify the detection of new variants, particularly those emerging from known families, making the process more straightforward.

The foundation of this approach is based on the premise that new voice synthesizers, despite their innovations, typically evolve from existing technologies rather than emerging as revolutionary breakthroughs. They often enhance existing frameworks with modifications, rather than constructing completely new architectures from the ground up. By comprehensively understanding and classifying *deepfake voice families*, our goal is to forge a more proactive and effective strategy for the early detection and mitigation of these evolving synthetic voices.

To address the research question of whether deepfake voices can be categorized into identifiable *families*, we implemented an experimental setup termed **cross-synthesizer checking**. This method involves training with a dataset comprising 1000 samples from the target synthesizer and an additional 1000 random real voice samples from the DSD-corpus. Testing was conducted using samples from all other synthesizers. Throughout this discussion, we refer to the SSL AASIST detector trained with samples from a specific target synthesizer using the names of the synthesizers themselves. Our investigation covered 12 different synthesizers, with the exclusion of AIHUB-VITS due to its overlap with the VITS-TTS synthesizer, and DiffGAN, because its open-source version offers only a single-speaker model (LJSpeech), rendering it unsuitable as a standalone training set.

We analyzed six anticipated factors that might contribute to the formation of the *deepfake voice family*, including the system's purpose (TTS or VC), acoustic architecture, vocoder architecture, the synthesizers' training dataset, the system's training/inference type (i.e., two-stage with separate optimized acoustic and vocoder, or jointly optimized in end-to-end (E2E) setting), and the target speakers set for voice style control (described in **Table. 5**).

Fig. 4 presents the outcomes of the cross-synthesizer detection accuracy analysis. These results substantiate the existence of deepfake voice families, as evidenced by certain pairs of synthesizers demonstrating a high rate of cross-detection, notably between StarGANv2-VC & Diff-HierVC, and SeamlessM4T-TTS & MeloTTS. However, the phenomenon does not appear to be governed by a consistent set of factors, making it challenging to fully elucidate. In the preliminary phase of analyzing deepfake voice families, we highlight several intriguing findings:

- **Finding 1: Enhancing detector's generalization through training with end-to-end GAN-based synthesizers' samples.** Training detectors with samples from end-to-end (E2E) GAN-based synthesizers improves their generalization capabilities. Apart from synthesizers that are challenging to detect (e.g., DDDM-VC, QuickVC), other synthesizers are more easily identified by detectors trained with E2E samples. For instance, VITS-TTS, MeloTTS, and MMSTTS can achieve a cross-detection rate of over 20% with more than three other synthesizers.
- **Finding 2: Enhancing Detector Generalization through training with high-quality vocoder samples.** This finding is an extension of **Finding 1**, highlighting that the shared AI-related properties in deepfake voice samples predominantly originate from the vocoders. Indeed, end-to-end (E2E) synthesizers incorporate the vocoding component, with the high quality of generated speech stemming from their unified, jointly trained process that converts text to raw waveforms directly. A similar level of generative quality can be achieved by employing a large-scale universal vocoder. For example, the Tortoise-TTS detector can achieve a cross-detection rate of over 20% with five other synthesizers. In contrast, using samples from lower-quality vocoder synthesizers, such as StarGANv2-VC, Diff-HierVC, or DDDM-VC, results in lower generalization.
- **Finding 3: The special case of QuickVC.** This synthesizer utilizes the Inverse Short-Time Fourier Transform (iSTFT) to convert the predicted spectrogram into a raw waveform, resulting in no shared properties with other neural-based vocoders.
- **Finding 4: AI-related factors produced by the Acoustic network have a minimal effect in two-stage synthesizer** This finding, while still supported by limited evidence, can be inferred by examining two opposing pairs: StarGANv2-VC & Diff-HierVC (A) versus DDDM-VC & Diff-HierVC (B). Pair (A) does not share a similar architecture yet exhibits a high cross-detection rate. Conversely, pair (B), despite having the same architecture and being proposed by the same authors, displays a low cross-detection rate.
- **Finding 5: Elevelabs, a commercial product, likely employs a GAN-based end-to-end TTS/VC approach..** It demonstrates a significant cross-detection rate when compared to MeloTTS, MMSTTS, and SeamlessM4T-TTS. Building on Finding 2, it's inferred that Elevenlabs does not utilize a separate, large-scale universal vocoder, as evidenced by the Tortoise-TTS detector's difficulty in classifying Elevenlabs samples. Additionally, the use of Elevenlabs samples yields the highest pooled cross-detection rate at 37.41%.

Based on the findings discussed above, we can outline several key propositions aimed at building a more generalized dataset for deepfake voice detection systems:

- **Proposition 1:** Vocoders play significant roles in deepfake voice detection. As demonstrated by Findings 1 and 2, the presence of high-quality vocoder samples enhances generalization.
- **Proposition 2:** Diversity in the training dataset, particularly regarding the samples from different synthesizers, is crucial. Assessing whether the samples from a new synthesizer are unique and truly contribute to enhancing generalization based on the model architecture or training set alone is challenging, as evidenced by Findings 3 and 4. Through cross-synthesizer checking akin to this experiment, we could identify which synthesizers are redundant and which are essential. This allows us to create a more balanced dataset that includes only representative synthesizers.
- **Proposition 3:** The cross-synthesizer checking experiment elucidates why our DSD-corpus is more generalized than others, as detailed in **Table. 2**. Obviously, there are not many high cross-detection rates across our DSD-corpus, proving that our dataset possesses high level of diversity.

Despite the very first stage of our analysis, we believe that deepfake voice families have possibly existed. More analysis should be done to better understand the characteristics of different deepfake voices. For example, explainable artificial intelligence (XAI) might be utilized to visualize the model detection behavior. We open this issue for future work.

6 Real-World User Experience Assessment

To assess the real-world user experience and to increase awareness about the threats posed by deepfake voices, we deployed and published our web application, available at deepfake.aisrc.technology. We also invited end users to review our system during the World IT Show[5], a three-day technology exhibition in Seoul, Korea. We asked users to answer four questions that required yes/no or rating responses without further information. No personal information was collected (an anonymous survey); therefore, there are no ethical concerns regarding the survey. The survey results shown in **Table 7** in statistical form provide insights into users' experiences, highlighting the perceived necessity and reliability of our Deepfake Voice Detection application.

[5] https://www.worlditshow.co.kr/

Table 7: Survey results from World IT Show visitors after using our Deepfake voice detection system

Question	Answer
Have you ever heard of AI voice phishing/scamming? (yes/no)	84.6% yes
Have you or someone close to you ever been a victim of AI voice phishing/scamming? (yes/no)	12.3% yes
Do you think a Deepfake voice detection app is necessary to prevent voice phishing/scamming? (1-5)	4.63
Was the accuracy of our Deepfake voice detection app reliable? (1-5)	4.29

7 Related Works

7.1 Deepfake Voice Technology

In the rapidly evolving landscape of artificial intelligence (AI), the advent of deep learning, particularly deep neural networks (DNNs), has revolutionized numerous fields. Deepfake voice is a technology that can synthesize realistic-sounding speech that mimics a specific person's voice by utilizing DNNs. This technology has two primary forms: Text-to-Speech (TTS) synthesis, which generates speech directly from text, and Voice Conversion (VC), which modifies source utterances to mimic the reference voice of a target speaker.

Deepfake voice can be realized through two primary architectures: the traditional two-stage process and the end-to-end (E2E) jointly trained method [53]. The traditional approach divides the task into acoustic modeling, which generates intermediate speech representation like a spectrogram, and vocoding, which synthesizes actual audio from this representation. In contrast, the end-to-end method streamlines this by directly converting text/speech to speech in a single step, leveraging deep learning models trained on extensive datasets. While the traditional method offers granular control over speech synthesis, allowing for detailed adjustments, it can be complex and may result in less cohesive outputs. The end-to-end approach, meanwhile, aims for a more integrated and natural-sounding result but demands significant computational power and may offer less flexibility for fine-tuning. Despite its integrated approach, the E2E approach implicitly encompasses aspects of both acoustic modeling and vocoding within a unified framework.

The development and application of deep learning techniques have enabled the creation of highly realistic synthesized speech, leading to the emergence of audio deepfakes. These are artificially generated audio clips that convincingly imitate the voice and speech patterns of a particular individual [42]. The realism of these audio deepfakes can be so high that it becomes challenging to distinguish them from authentic speech, thus raising both exciting possibilities and significant ethical concerns [2].

7.2 Audio Deepfake Detection Systems

The ease and speed of creating audio deepfakes raise significant ethical and security concerns. As these technologies become more advanced and accessible, the potential for misuse also increases. This underscores the importance of developing robust detection methods and ethical guidelines to govern the use of these technologies.

Acknowledging the potential misuse of deepfake voice technologies, the community has organized competitions like ASVSpoof[41, 58, 61] and the ADD challenges[65] to address these concerns. The ASVSpoof challenge focuses on the issue of spoofing attacks aimed at deceiving Automatic Speaker Verification (ASV) systems through synthesized speech. Notably, the most recent ASVSpoof 2021[61] introduced a deepfake track, highlighting the urgent need to confront deepfake threats in real-world scenarios. Meanwhile, the ADD challenges have brought attention to more sophisticated and stealthy techniques employed by attackers, such as the introduction of noise or the creation of partially fake samples, which mimic the practical challenges faced in developing deepfake detection systems.

The development of systems for detecting deepfake voices generally falls into two primary methodologies: Feature Learning[15, 20, 36, 51] and back-end-Classifier Modeling[9, 23, 24, 54]. Deep Neural Networks (DNNs) are highly effective in tackling the challenges associated with both these approaches. While targeting similar ends, the methodologies diverge in their feature extraction techniques, opting either for hand-crafted features like the Linear Frequency Cepstral Coefficient (LFCC) or direct analysis of raw waveforms. Additionally, certain studies have focused on characteristics predominantly found in real human voices, such as emotional expression[14, 30] or breathing patterns[16, 37]. Recent advancements have seen the successful application of Self-supervised Learning (SSL) pre-trained models. These pre-trained SSL models, adept at representing speech signals, offer more pertinent information for downstream tasks than traditional hand-crafted features, leading to superior performance in detecting deepfake voices[52, 55]. Nonetheless, ensuring the reliability of deepfake detection remains a challenging issue. While some research has suggested the addition of more defensive layers to bolster security[63, 66], other studies continue to focus on improving model generalization to enhance performance across varied scenarios[10, 67].

8 Conclusion

The rapid advancement of audio deepfake technology has ushered in an era where the authenticity of audio content can no longer be taken at face value. This technological leap forward has, unfortunately, paved the way for nefarious applications, including sophisticated voice phishing attacks that pose significant security threats. Recognizing the urgent need to mitigate these risks, our research has focused on enhancing the robustness of detection mechanisms against such deceptive practices. Our innovative approach, dubbed the "Trident of Poseidon," integrates three advanced training strategies: Supervised Contrastive Learning (SCL), Hard Negative Mining by Audio Re-synthesizing (HAR), and Proactive Batch Sampling (PBS). When combined with a Variational Information Bottleneck (VIB) back-end network, this triad of strategies significantly improves the generalizability of our model, setting a new benchmark in the fight against audio deepfake misuse. However, the proposed system fails to detect non-neural-based synthesized speech (not deepfake voice) due to the use of a neural-based vocoder audio re-synthesizing method.

Despite these advancements, the issue of data distribution shift remains a formidable challenge in ensuring the effectiveness of deepfake detection models. To address this, we introduced the Diverse Synthesizers for Deepfake Voice Detection dataset (**DSD-corpus**), a pioneering resource designed to enhance the generalization capabilities of detection algorithms. Our experiment results demonstrate that leveraging the DSD-corpus leads to a marked improvement in the model's ability to classify deepfake audio across varied and previously unseen data distributions.

Looking ahead, our research has begun to explore the underlying factors and properties that bind them. Preliminary findings suggest the existence of *deepfake voice families*, a concept that could gain our understanding and categorization of synthesizers. The initiation of our *cross-synthesizer detection* experiments has opened up a new avenue of research focused on deepfake voice family analysis which helps to develop more effective strategies to stay ahead in the perpetual arms race against the evolving threats posed by audio deepfake technology.

Acknowledgments

This research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2024-2020-0-01602) supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP); and was partially supported by IITP grant funded by the Korea government (MSIT) (No.RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes); and was partially supported by IITP grant funded by the Korea government (MSIT) (No.RS-2023-00263037, Robust deepfake audio detection development against adversarial attacks).

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).

[2] Zaynab Almutairi and Hebah Elgibreen. 2022. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms* 15, 5 (2022), 155.

[3] Naroa Amezaga and Jeremy Hajek. 2022. Availability of voice deepfake technology and its impact for good and evil. In *Proceedings of the 23rd Annual Conference on Information Technology Education.* 23–28.

[4] Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing* 11, 2 (2017), 361–372.

[5] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* (2021).

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[7] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596* (2023).

[8] James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243* (2023).

[9] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. 2020. Generalization of Audio Deepfake Detection. In *Proc. Odyssey).* https://doi.org/10.21437/Odyssey.2020-19

[10] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. 2020. Generalization of Audio Deepfake Detection.. In *Odyssey.* 132–137.

[11] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation. In *Proc. INTERSPEECH 2023.* 2283–2287. https://doi.org/10.21437/Interspeech.2023-817

[12] Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17862–17870.

[13] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. 2020. Korean singing voice synthesis based on auto-regressive boundary equilibrium gan. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 7234–7238.

[14] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. 2022. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 8962–8966.

[15] Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, and Agha Ali Raza. 2021. Fake Audio Detection in Resource-Constrained Settings Using Microfeatures. In *Proc. Interspeech.* https://doi.org/10.21437/Interspeech.2021-524

[16] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. 2023. BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 1–5.

[17] Joel Frank and Lea Schönherr. 2021. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* https://openreview.net/forum?id=74TZg9gsO8W

[18] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. 2020. Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image.. In *INTERSPEECH.* 1321–1325.

[19] Houjian Guo, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2023. QuickVC: Any-to-many Voice Conversion Using Inverse Short-time Fourier Transform for Faster Conversion. *arXiv preprint arXiv:2302.08296* (2023).

[20] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. 2021. Towards End-to-End Synthetic Speech Detection. *IEEE Signal Processing Letters* 28 (2021). https://doi.org/10.1109/LSP.2021.3089437

[21] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. Interspeech 2021.* 2207–2211. https://doi.org/10.21437/Interspeech.2021-1016

[22] Camil Jreige, Rupal Patel, and H Timothy Bunnell. 2009. VocaliD: Personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility.* 259–260.

[23] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan. 2021. CRIM's System Description for the ASVSpoof2021 Challenge. In *Proc. 2021 Edition of the ASVSpoof Challenge.* https://doi.org/10.21437/ASVSPOOF.2021-16

[24] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan. 2021. Investigation on activation functions for robust end-to-end spoofing attack detection system. In *Proc. 2021 Edition of the ASVSpoof Challenge.* https://doi.org/10.21437/ASVSPOOF.2021-13

[25] Masaya Kawamura, Yuma Shirahata, Ryuichi Yamamoto, and Kentaro Tachibana. 2023. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 1–5.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[27] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning.* PMLR, 5530–5540.

[28] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[29] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* 33 (2020), 17022–17033.

[30] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2022. A comparative study on physical and perceptual features for deepfake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia.* 35–41.

[31] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems* 36 (2024).

[32] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. 2021. StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion. In *Proc. Interspeech 2021.* 1349–1353. https://doi.org/10.21437/Interspeech.2021-319

[33] Songxiang Liu, Dan Su, and Dong Yu. 2022. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972* (2022).

[34] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2507–2522. https://doi.org/10.1109/TASLP.2023.3285283

[35] Jingze Lu, Yuxiang Zhang, Wenchao Wang, Zengqiang Shang, and Pengyuan Zhang. 2024. One-Class Knowledge Distillation for Spoofing Speech Detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11251–11255.

[36] Youxuan Ma, Zongze Ren, and Shugong Xu. 2021. RW-Resnet: A Novel Speech Anti-Spoofing Model Using Raw Waveform. arXiv 2108.05684. arXiv:2108.05684 [cs, eess] http://arxiv.org/abs/2108.05684

[37] Zohreh Mostaani and Mathew Magimai-Doss. 2022. On Breathing Pattern Information in Synthetic Speech.. In *INTERSPEECH*. 2768–2772.

[38] Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Interspeech 2022*. ISCA, 2783–2787. https://doi.org/10.21437/Interspeech.2022-108

[39] Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. *arXiv preprint arXiv:2401.09512* (2024).

[40] Nicolas M. Müller, Philip Sperl, and Konstantin Böttinger. 2023. Complex-valued neural networks for voice anti-spoofing. In *Proc. INTERSPEECH 2023*. 3814–3818. https://doi.org/10.21437/Interspeech.2023-901

[41] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. 2021. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 2 (2021), 252–265.

[42] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences* 9, 19 (2019), 4050.

[43] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[44] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research* 25, 97 (2024), 1–52.

[45] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. OpenVoice: Versatile Instant Voice Cloning. *arXiv preprint arXiv:2312.01479* (2023).

[46] Ricardo Reimao and Vassilios Tzerpos. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 1–10.

[47] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592* (2020).

[48] Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. 2023. A Conformer-Based Classifier for Variable-Length Utterance Processing in Anti-Spoofing. In *INTERSPEECH 2023*. ISCA, 5281–5285. https://doi.org/10.21437/Interspeech.2023-1820

[49] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Andreas Nautsch, Xin Wang, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, and Kong-Aik Lee. 2023. Introduction to voice presentation attack detection and recent advances. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment* (2023), 339–385.

[50] Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 464–472.

[51] Hemlata Tak, Jee-weon Jung, Jose Patino, et al. 2021. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proc. 2021 Edition of the ASVSpoof Challenge*. https://doi.org/10.21437/ASVSPOOF.2021-1

[52] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA, 112–119. https://doi.org/10.21437/Odyssey.2022-16

[53] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A Survey on Neural Speech Synthesis. *arXiv:2106.15561 [cs, eess]* (July 2021). arXiv:2106.15561 [cs, eess]

[54] Zhongwei Teng, Quchen Fu, Jules White, Maria E. Powell, and Douglas C. Schmidt. 2022. SA-SASV: An End-to-End Spoof-Aggregated Spoofing-Aware Speaker Verification System. arXiv 2203.06517. arXiv:2203.06517 [cs, eess] http://arxiv.org/abs/2203.06517

[55] Xin Wang and Junichi Yamagishi. 2022. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA, 100–106. https://doi.org/10.21437/Odyssey.2022-14

[56] Xin Wang and Junichi Yamagishi. 2024. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10311–10315.

[57] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech. *Computer Speech & Language* 64 (Nov. 2020), 101114. https://doi.org/10.1016/j.csl.2020.101114

[58] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Hector Delgado. 2017. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing* 11, 4 (2017), 588–604.

[59] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. In *Proc. INTERSPEECH 2023*. 2808–2812. https://doi.org/10.21437/Interspeech.2023-1383

[60] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). (2019).

[61] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*.

[62] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6199–6203.

[63] Chen-Zhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung Liew. 2020. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security* 16 (2020), 1841–1854.

[64] Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, and Yunhe Wang. 2024. A robust audio deepfake detection system via multi-view feature. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 13131–13135.

[65] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9216–9220.

[66] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. 2023. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 460–474.

[67] Yibo Zhang, Weiguo Lin, and Junfeng Xu. 2024. Joint Audio-Visual Attention with Contrastive Learning for More General Deepfake Detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 5 (2024), 1–23.

[68] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In *Interspeech*.

[69] Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. *MeloTTS: High-quality Multilingual Multi-accent Text-to-Speech*. https://github.com/myshell-ai/MeloTTS