TimbreAdv: Timbre Adversarial Attacks on Speaker Verification Systems

Ye Xiao$^{1[0009-0003-3611-7294]}$, Wenhan Yao$^{1[0000-0003-1014-9565]}$, Zexin Li$^{1[0009-0008-9231-8660]}$, Jinsu Yang$^{1[0009-0003-4897-5157]}$, Yuhao Chen[1], Xiandang Luo[1], Fen Xiao[1], and Weiping Wen[2]✉

[1] XiangTan University, Hunan Province, CN
xy@smail.xtu.edu.cn
[2] Peking University, Beijing, CN
weipingwen@pku.edu.cn

Abstract. In recent years, speaker verification (SV) systems have become ubiquitous across security-critical applications. While these systems encode speaker identities into high-dimensional embeddings, they remain vulnerable to adversarial attacks that manipulate these embeddings, so it is essential for us to expose as many "blind spots" of speaker verification systems as possible. Existing attacks predominantly inject additive noise, which often compromises speech naturalness and lacks semantic control. In this paper, we propose the Timbre Adversarial attack (TimbreAdv), a novel paradigm that exploits vocal tract characteristics to deceive SV systems. Our framework introduces hierarchical feature disentanglement, feature-level timbre blending, and multi-object adversarial optimization to generate adversarial samples under the setting of black-box. We use comprehensive metrics to evaluate our method, and the results show great attack effectiveness and stealthiness.

Keywords: Adversarial Attacks · Speaker Verification Systems · Multi-object Adversarial Optimization.

1 Introduction

Speaker verification systems [1], which authenticate individuals based on unique vocal characteristics, have become integral to modern security infrastructures. These systems are widely deployed in biometric authentication (e.g., smartphone unlocking, banking voiceprints), forensic analysis (e.g., courtroom evidence validation), and smart environments (e.g., personalized voice assistants). At their core, speaker verification pipelines involve two phases: enrollment and verification. During enrollment, a speaker's voice is converted into a high-dimensional embedding that captures vocal tract and prosodic traits. During verification, a similarity score is computed between the input voiceprint and stored references to accept or reject identity claims. The continuous and high-dimensional nature of voiceprint embeddings creates complex decision boundaries in feature space, which introduces fragility: minor perturbations(whether from environmental noise or adversarial manipulation) in the embedding space can shift samples across decision thresholds. This vulnerability renders speaker verification systems susceptible to *adversarial attacks* [5–10], where intentionally crafted perturbations induce misclassification. To avoid potential risks and further research the robustness of speaker verification systems, it is of great value to expose as many “blind spots” of speaker verification systems as possible at the current research stage.

Adversarial attacks aim at perturbing the system input in a purposefully designed way to make the system behave incorrectly. In general, adversarial attacks fall into two categories based on threat models: white-box attacks and black-box attacks. White-box attacks [5–7], such as the Fast Gradient Sign Method [2] (FGSM) and Projected Gradient Descent [3] (PGD), utilize gradient-based optimization to craft perturbations by leveraging full access to model parameters and training dynamics. In contrast, black-box attacks [8–10], exemplified by Zeroth-Order Optimization [4] (ZOO) and evolutionary strategies, rely on iterative queries to approximate gradients or heuristic optimization without direct model access.

The study of adversarial attacks originated in computer vision, where the differentiable nature of image pixels enabled systematic perturbation optimization. While adversarial attacks in computer vision have been widely studied, their application in the audio domain, particularly in speaker verification, remains underexplored for audio’s temporal nature and human perceptual sensitivity. A critical constraint is stealthiness: perturbations must remain undetectable by humans, while preserving semantic content and temporal coherence. Also, while many existing adversarial attack methods operate under the white-box assumptions, their real-world applicability is limited for attackers typically cannot access model internals. Black-box attacks relying solely on query interactions or transferability are more practical but inherently challenging. To address these challenges, we propose TimbreAdv, a timbre-based adversarial attack framework that fills the gap in speaker verification robustness research. Our contributions can be summarized as follows:

1. We propose a novel adversarial attack framework leveraging timbre characteristics to implement adversarial attacks while operating under a black-box setting.
2. We propose the framework introducing hierarchical feature disentanglement, feature-level timbre blending and multi-object adversarial optimization to generate adversarial samples.
3. Our method generates adversarial samples that effectively deceive SV systems while preserving linguistic content, achieving great stealthiness.

2 Background

2.1 Adversarial Attacks

Adversarial attacks refer to techniques that manipulate machine learning models by introducing small but carefully crafted perturbations to input data, leading to

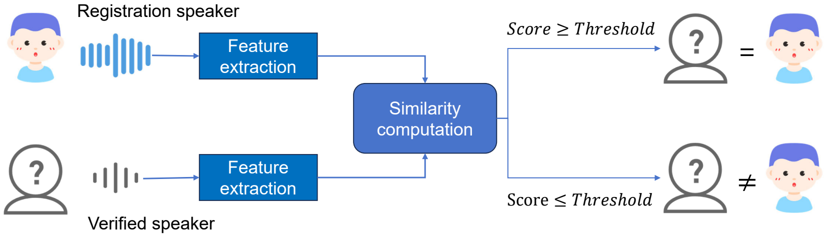

Fig. 1. Pipelines of speaker verification tasks.

incorrect predictions. These perturbations are often imperceptible to humans, yet significantly degrade model performance [2–4]. Formally, an adversarial attack aims to find a minimally modified input $x + \delta$ such that:

$$\mathcal{F}(x+\delta) \neq \mathcal{F}(x), \quad \text{s.t.} \quad |\delta|_p \leq \epsilon, \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the deep learning model, δ is the adversarial perturbation, and ϵ constrains perturbation magnitude under the ℓ_p-norm (typically $p = 2$ or ∞).

Adversarial example was first proposed in computer vision [2, 14, 15]. In recent years, researchers have proposed a variety of techniques in their attempts to challenge the robustness of SV systems [16–19]. These attacks can target a specific speaker or target multiple speakers, identify a speaker from a set of enrolled speakers, or identify a speaker if it is contained in this enrolled speakers set. Most attacks proposed in prior research are individual attacks [16, 20], in which the attacker must generate perturbations specific to each genuine sample, reducing the attack's efficiency. Less common are universal perturbations [23], which although usually considerably more costly and difficult to generate, are more efficient during test time. In the study performed by Li et al. [22], a generative adversarial network (GAN) was trained to serve as a universal approach. In this paper, we focus on the challenging task of generating adversarial samples to implement universal adversarial attacks and work effectively on an unknown set of speakers.

2.2 Speaker Verification Systems

The goal of the speaker verification task is to judge whether a given utterance is from a registered speaker, as illustrated in Fig. 1. Let $F(\cdot)$ denotes the speaker embedding extraction function, which receives an input audio x and gives the speaker embedding $v = F(x)$. For two utterances x_1 and x_2, we use cosine similarity as a score to measure the distance between their speaker embeddings. The score $s(F(x_1), F(x_2))$ is calculated by

$$s(F(x_1), F(x_2)) = \frac{F(x_1)F(x_2)}{\|F(x_1)\|_2 \, \|F(x_2)\|_2}, \tag{2}$$

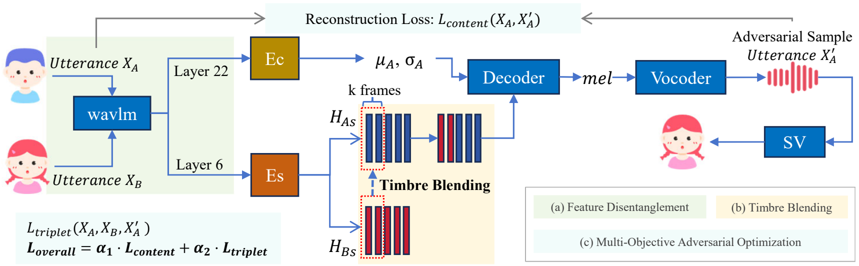
Reconstruction Loss: $L_{content}(X_A, X'_A)$
Utterance X_A
Utterance X_B
wavlm
Layer 22
Layer 6
Ec
Es
μ_A, σ_A
Decoder
mel
Vocoder
Adversarial Sample
Utterance X'_A
SV
k frames
H_{As}
H_{Bs}
Timbre Blending
$L_{triplet}(X_A, X_B, X'_A)$
$L_{overall} = \alpha_1 \cdot L_{content} + \alpha_2 \cdot L_{triplet}$
(a) Feature Disentanglement
(b) Timbre Blending
(c) Multi-Objective Adversarial Optimization

Fig. 2. Framework of TimbreAdv that adversarially misleads speaker verification models by confusing the timbre of speaker A and speaker B.

Table 1. Algorithm Symbol Definitions

Symbol	Description
X_A, X_B	Raw waveform inputs of source speaker A and target speaker B
f^n_{SSL}	Self-supervised learning encoder with n denoting layer index
E_s, E_c	Speaker encoder and content encoder
H_{As}, H_{Bs}	Speaker embeddings
μ_A, σ_A	Mean and standard deviation of source content features E_c
F	Speaker embedding extraction function
k	Critical frame length for timbre hybridization
$\mathcal{L}_{content}$	Waveform fidelity loss (L_1 distance)
$\mathcal{L}_{triplet}$	Triplet loss against SV model
α_i	Dynamic weights in MGDA optimization
η	Learning rate for gradient update

it will give a same speaker decision when the score satisfies $s(F(x_1), F(x_2)) \geq \theta$, where θ is a preset threshold, otherwise, it will give a different decision. This means that SV systems possess inherent defence capabilities and can resist weaker adversarial attacks, making it more difficult to craft suitable adversarial examples.

3 Methodology

3.1 Method Overview

Timbre serves as a fundamental attribute in speaker verification tasks, functioning as a unique auditory fingerprint determined by the physical characteristics of the vocal tract. The primary objective of SV systems is to distinguish between speakers based on timbre rather than the semantic content of speech. Consequently, altering timbre presents a direct and effective approach to disrupting SV systems, thereby increasing attack success rates. Furthermore, some speaker verification systems operate in a text-dependent manner, where authentication relies not only on the speaker's identity but also on the specific speech content.

This insight motivates a shift from traditional noise-based attacks to timbre-driven perturbations, which enable targeted speaker impersonation while preserving perceptual naturalness. We achieve this by altering the timbre of a source utterance X_A to resemble that of a target speaker X_B, generating adversarial examples that satisfy:

$$\mathrm{SV}(\mathrm{TimbreAdv}(X_A, X_B)) \neq \mathrm{SV}(X_A), \tag{3}$$

where $\mathrm{SV}(\cdot)$ denotes the output of a speaker verification model and $\mathrm{TimbreAdv}(\cdot)$ denotes our attack framework. The generated adversarial sample retains the linguistic content of X_A but adopts the speaker identity of X_B, so that it can deceive SV systems while preserving the content and perceptual naturalness. The framework comprises three core modules: hierarchical feature disentanglement, frame-level timbre blending, and multi-objective adversarial optimization, as illustrated in Fig. 2. We define some description symbols in Table 1. Below, we detail each component.

3.2 Hierarchical Feature Disentanglement

Self-supervised learning (SSL) trained on a large-scale unannotated speech corpus shows considerable potential for Voice Conversion(VC) tasks that require feature disentanglement. Also, the WavLM model [11] effectively addresses comprehensive downstream speech tasks, encompassing speaker verification, speaker diarization, speech separation and speech recognition [12]. Also, the research demonstrates that the upper layers of the model have greater semantic information, while the lower layers of the model contain more speaker-related information. So we employ WavLM pretrained model to extract self-supervised feature information from speech and choose layer 6 and layer 22 of the model to capture speaker characteristics and encode content information, respectively.

3.3 Frame-level Timbre Blending

After getting speaker embeddings H_{As}, H_{Bs} from E_s and getting content representation μ_B from E_c, we need to further deal with it to generate adversarial samples. As for speaker embedding, we employ frame-wise timbre blending to replace the k frames of H_{As} with those from H_{Bs} to inject target timbre features:

$$H_{As}[:, k :] = H_{Bs}[:, k :]. \tag{4}$$

Furthermore, for overall semantic consistency, we eliminate source timbre information from content features by whitening H^k_{Ac} with speaker A's statistics, then rescaling with speaker B's distribution:

$$H^k_{Ac} = \sigma_B \left(\frac{H^k_{Ac} - \mu_A}{\sigma_A} \right) + \mu_B, \tag{5}$$

Algorithm 1 Timbre Adversarial Samples Generation

```
Require: Utterance of speaker A $X_A$, utterance of speaker B $X_B$; SSL encoder $f_{\mathrm{SSL}}$;
    speaker verification model $SV$; Frame number $k$; speaker encoder $E_s$; content en-
    coder $E_c$; Decoder; Vocoder; learning rate $\eta$.
Ensure: Adversarial sample $X'_A$
Initialize model parameters $\theta$
while Not Converged do
    $H_{As} \leftarrow E_s(f^6_{\mathrm{SSL}}(X_A))$,$H_{Bs} \leftarrow E_s(f^6_{\mathrm{SSL}}(X_B))$          ▷ Extract speaker features
    $\mu_A, \sigma_A \leftarrow E_c(f^{22}_{\mathrm{SSL}}(X_A))$                              ▷ Content encoder
    $H_{As}[:, k:] \leftarrow H_{Bs}[:, k:]$                    ▷ Replace $k$ frames of speaker A with B
    mel $\leftarrow$ Decoder($H_{As}, \mu_A$)                     ▷ Generate Mel-spectrogram
    $X'_A \leftarrow$ Vocoder(mel)                                 ▷ Convert to waveform
    if $SV(X'_A) = SV(X_B)$ then                                   ▷ Check attack success
        break
    else
        $\mathcal{L}_{content} \leftarrow \|X'_A - X_A\|_1$
        $\mathcal{L}_{triplet} \leftarrow \max\left(\cos(F(X'_A), F(X_A)) - \cos(F(X'_A), F(X_B)) + m,\ 0\right)$
        Solve $\alpha_1, \alpha_2$ via: min $\|\alpha_1 \nabla_\theta \mathcal{L}_{content} + \alpha_2 \nabla_\theta \mathcal{L}_{triplet}\|^2$
        Update $\theta \leftarrow \theta - \eta(\alpha_1 \nabla_\theta \mathcal{L}_{content} + \alpha_2 \nabla_\theta \mathcal{L}_{triplet})$      ▷ Optimization
    end if
end while
return $X'_A$
```

where μ_A, σ_A denote the mean and standard deviation of source speaker A's content features computed by E_c, and μ_B, σ_B of each layer denote the linear transformation of the output of speaker encoder E_s. This alignment ensures spectral compatibility between hybridized timbre and original content.

To be more specific, the content encoder employs instance normalization (IN) to standardize input speech features by eliminating speaker-dependent spectral characteristics while preserving abstract linguistic attributes. The speaker encoder extracts a low-dimensional timbre embedding from target speaker B's utterance. To blend timbre information into content features, the decoder transforms it into channel-wise affine parameters via fully connected layers and output mel-spectrogram further synthesized into a waveform via a neural vocoder, producing an adversarial example X'_A that exhibits deceptiveness and imperceptibility.

3.4 Multi-Objective Adversarial Optimization

Generating effective adversarial samples for speaker verification requires simultaneously optimizing two competing objectives: (1) preserving the linguistic content and perceptual quality of the original utterance, and (2) altering the speaker identity by making the adversarial embedding resemble the target speaker rather than the source.

We formalize this as a multi-objective optimization problem with the following loss components: (1)$\mathcal{L}_{\mathrm{content}}$: ensures the adversarial sample retains the original semantic content and waveform quality. We use the L1 distance between the original and adversarial waveform:

$$\mathcal{L}_{\text{content}} = \|X'_A - X_A\|_1. \tag{6}$$

(2) $\mathcal{L}_{\text{triplet}}$: encourages the embedding of the adversarial sample X'_A to be close to the target speaker X_B, and distant from the source speaker X_A. This is formulated as a triplet margin loss:

$$\mathcal{L}_{\text{triplet}} = \max\left(\cos(F(X'_A), F(X_A)) - \cos(F(X'_A), F(X_B)) + m,\ 0\right), \tag{7}$$

where $F(\cdot)$ denotes the speaker embedding extractor, and $m > 0$ is the margin hyperparameter.

These objectives inherently conflict as preserving content may hinder the ability to alter identity, and vice versa. To dynamically balance them, we adopt the Multiple Gradient Descent Algorithm (MGDA) to compute dynamic weights α_1, α_2 at each iteration to minimize the total loss:

$$\mathcal{L}_{\text{overall}} = \alpha_1 \cdot \mathcal{L}_{\text{content}} + \alpha_2 \cdot \mathcal{L}_{\text{triplet}}. \tag{8}$$

The weights are optimized via the Frank-Wolfe algorithm. MGDA allows the model to emphasize underperforming objectives early in training, and gradually converge to balanced weights as both objectives improve. This optimization is performed iteratively as shown in Algorithm 1.

4 Experiments

4.1 Experiments Setup

Dataset. We use LibriSpeech [24], VoxCeleb1 [25], and VoxCeleb2 [26] to evaluate the effectiveness of our adversarial perturbations. These three datasets collectively span both clean, studio-quality recordings and noisy, in-the-wild utterances, enabling a comprehensive assessment of attack performance under controlled and unconstrained acoustic conditions. LibriSpeech provides paired transcripts, allowing precise evaluation of content preservation via word error rate (WER). For VoxCeleb1 and VoxCeleb2, which lack ground-truth transcripts, we employ the Whisper [27] automatic speech recognition (ASR) model to transcribe adversarial and original audio, enabling WER computation in real-world scenarios.

Model. We adopt WavLM [11] to extract self-supervised learning (SSL) features, which have demonstrated strong generalization across a range of downstream speech tasks, including speaker recognition and content modeling. We employ HiFi-GAN [28] as the vocoder due to its ability to generate high-fidelity speech with low computational cost. To guide adversarial sample generation, we use the state-of-the-art speaker verification model ERes2NetV2 [29] as a proxy model. During optimization, the adversarial perturbations are iteratively updated to deceive this proxy system. For black-box evaluation, we attack ECAPA-TDNN model [30], which is not involved in adversarial generation.

Metrics. To evaluate both the effectiveness and perceptual quality of our adversarial perturbations, we compute the cosine similarity (Similarity) between the original clean input and the adversarial example, as well as the attack success rate (ASR) to assess attack effectiveness. We also calculate the word error rate (WER) using an automatic speech recognition system to verify linguistic integrity. For perceptual quality, we report the Perceptual Evaluation of Speech Quality (PESQ) score, which ranges from 1.0 (poor) to 4.5 (excellent). In addition, we conduct a Mean Opinion Score (MOS) test, in which human listeners subjectively rate the naturalness of adversarial sample on a 1-5 scale.

Baseline methods. Our method was compared against with: (1) original clean utterances (Clean), (2) additive white Gaussian noise perturbations (Random), (3) white-box adversarial attacks based on projected gradient descent (PGD) [3], (4) GAN-generated perturbations (GAN) [22] and (5)universal adversarial perturbation (UAP) [32].

Training details. Raw audio waveforms X_A and X_B are resampled to 16 kHz for consistent processing. Utterance lengths are standardized to 4 seconds: shorter segments are zero-padded at the end, while longer segments are truncated at the end. The encoder and decoder are jointly optimized using the Adam optimizer with fixed hyperparameters: learning rate $\eta = 0.0005$, $\beta_1 = 0.5$, $\beta_2 = 0.9$, and batch size = 8. Training stops when any of these conditions is met: (1)The overall loss change $|\Delta\mathcal{L}_{\text{overall}}| < 10^{-4}$ for 10 consecutive iterations. (2)ASR $> 99.5\%$ against the proxy ERes2NetV2 model. (3) A maximum of 200 iterations is reached. The margin hyperparameter is set to $m = 0.4$ in our experiments.

Table 2. Cross-dataset evaluation of TimbreAdv

Dataset	**ASR(%)↑**	**Similarity(%)↓**	**WER(%)↓**
LibriSpeech [24]	96.91	46.86	13.3
VoxCeleb1 [25]	95.43	48.21	14.2
VoxCeleb2 [26]	93.87	51.05	15.8

4.2 Results

Results analysis. Table 2 presents the cross-dataset generalization performance of TimbreAdv on three datasets. Across all datasets, TimbreAdv maintains consistently high ASR, demonstrating robustness in both clean and in-the-wild acoustic conditions. On the clean and studio-quality LibriSpeech corpus, TimbreAdv achieves a near-perfect ASR of 96.91%, with a low speaker embedding similarity of 46.92% and minimal degradation in speech intelligibility. When

Table 3. Evaluation of different attack methods on LibriSpeech.

Method	**ASR(%)↑**	**Similarity(%)↓**	**WER(%)↓**	**PESQ↑**	**MOS↑**
Clean	0.00	84.71	12.3	4.01	4.5
Random	13.26	80.05	18.7	3.58	3.7
GAN [22]	72.65	82.47	21.8	3.26	3.5
UAP [32]	82.10	76.98	19.2	3.42	3.6
PGD [3]	94.80	74.56	28.6	2.78	2.9
TimbreAdv (Ours)	96.91	46.92	13.0	3.94	4.2

applied to the more challenging VoxCeleb1 and VoxCeleb2 datasets—characterized by background noise, varied recording channels, and cross-accent speech—TimbreAdv still achieves strong ASR values of 95.43% and 93.87%, respectively. As expected, the speaker similarity increases slightly, and WER rises moderately on VoxCeleb datasets due to greater acoustic variability.

As shown in Table 3, TimbreAdv achieves a high ASR of 96.91%, outperforming both GAN-based and UAP baselines, demonstrating its effectiveness in deceiving speaker verification systems under black-box conditions. The cosine similarity between adversarial and clean embeddings drops to 46.92%, confirming that timbre has been significantly altered. Meanwhile, the WER increases only marginally, indicating that linguistic content is largely preserved. For perceptual quality, TimbreAdv achieves a PESQ of 3.94, suggesting minimal audible artifacts. Furthermore, a small-scale listening study yields an average MOS of 4.2, confirming the naturalness and imperceptibility of the generated adversarial examples. Overall, TimbreAdv demonstrates a superior balance between attack effectiveness and perceptual stealth, outperforming both black-box and white-box attack baselines across all evaluation metrics.

Ablation analysis. We perform ablation studies to examine the impact of two key factors: the frame selection ratio k/T and the loss weighting strategy.

As illustrated in Fig. 3, we investigate the impact of the frame selection ratio k/T on speaker similarity between clean and adversarial samples. The results show a clear decreasing trend: as k increases, the similarity score drops accordingly. This indicates that incorporating a larger proportion of frames from the target speaker leads to more effective timbre transformation. In particular, higher k values inject more target-specific acoustic characteristics, thereby enhancing speaker identity manipulation. Conversely, smaller values of k result in more subtle, localized perturbations, which may better preserve perceptual naturalness while offering limited speaker obfuscation.

We compare fixed loss weight configurations (α_1, α_2) with our adaptive strategy based on Multiple Gradient Descent Algorithm (MGDA), as shown in Table 4. Fixed ratios, such as (0.7, 0.3) or (0.5, 0.5), are static throughout training and cannot adjust to shifting optimization dynamics. In contrast, MGDA dynamically allocates higher weight to the underperforming objective during early

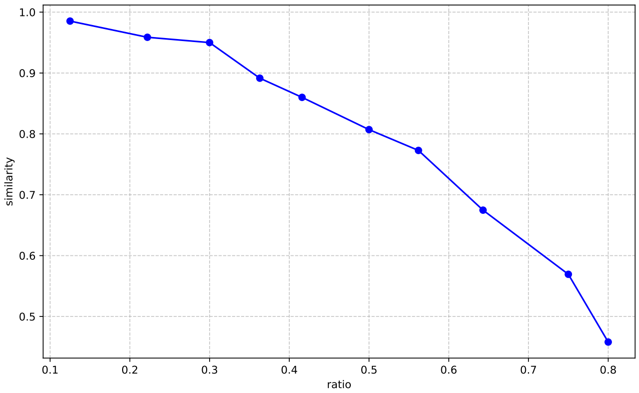

Fig. 3. Impact of frame selection ratio k/T on the same speaker similarity.

Table 4. Ablation Study on Loss Weighting Strategy (LibriSpeech)

Weighting Method	**ASR(%)↑**	**Similarity(%)↓**	**WER(%)↓**
Fixed ($\alpha_1 = 0.7$, $\alpha_2 = 0.3$)	91.82	53.48	13.7
Fixed ($\alpha_1 = 0.5$, $\alpha_2 = 0.5$)	94.17	50.13	14.2
MGDA (Ours)	96.45	45.63	13.1

training stages, allowing faster correction and more balanced convergence. Over time, MGDA weights tend to stabilize near (0.5, 0.5), reflecting equilibrium between adversarial strength and linguistic preservation.

5 Conclusion

In this paper, we present TimbreAdv, a novel adversarial attack framework targeting speaker verification systems by manipulating vocal timbre rather than introducing additive noise. Our method operates under black-box settings and comprises three key modules: hierarchical feature disentanglement, frame-level timbre blending, and multi-objective adversarial optimization. Extensive experiments across three datasets show that TimbreAdv achieves high attack success rates with low speaker similarity and minimal degradation in intelligibility and perceptual quality. Compared to both black-box and white-box baselines, our method consistently outperforms in terms of effectiveness and stealthiness.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bimbot, Frédéric, et al. "A tutorial on text-independent speaker verification." EURASIP Journal on Advances in Signal Processing 2004 (2004): 1-22.
2. Goodfellow, Ian J. et al. “Explaining and Harnessing Adversarial Examples.” CoRR abs/1412.6572 (2014): n. pag.
3. Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
4. Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017.
5. Li, Xu, et al. "Adversarial attacks on GMM i-vector based speaker verification systems." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
6. Villalba, Jesús, Yuekai Zhang, and Najim Dehak. "x-Vectors Meet Adversarial Attacks: Benchmarking Adversarial Robustness in Speaker Verification." Interspeech. 2020.
7. Luo, Hongwei, et al. "Spoofing speaker verification system by adversarial examples leveraging the generalized speaker difference." Security and Communication Networks 2021.1 (2021): 6664578.
8. Bi, Xin, et al. "Boosting question answering over knowledge graph with reward integration and policy evaluation under weak supervision." Information Processing & Management 60.2 (2023): 103242.
9. Liu, Jinbo, et al. "GNN-based long and short term preference modeling for next-location prediction." Information Sciences 629 (2023): 1-14.
10. Jia, Yan, et al. "Extrapolation over temporal knowledge graph via hyperbolic embedding." CAAI Transactions on Intelligence Technology 8.2 (2023): 418-429.
11. Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.
12. Lin, Guan-Ting, et al. "On the utility of self-supervised models for prosody-related tasks." 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023.
13. Xintao Zhao et al. "Adversarial Speaker Disentanglement Using Unannotated External Data for Self-supervised Representation-based Voice Conversion." 2023 IEEE International Conference on Multimedia and Expo (ICME) (2023): 1691-1696.
14. Feng, Yan, et al. "Adversarial attack on deep product quantization network for image retrieval." Proceedings of the AAAI conference on Artificial Intelligence. Vol. 34. No. 07.
15. Wang, Jiakai, et al. "Dual attention suppression attack: Generate adversarial camouflage in physical world." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
16. Shamsabadi, Ali Shahin, et al. "Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
17. Kreuk, Felix, et al. "Fooling end-to-end speaker verification with adversarial examples." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.
18. Xie, Yi, et al. "Real-time, universal, and robust adversarial attacks against speaker recognition systems." ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020.

19. Zhang, Weiyi, et al. "Attack on practical speaker verification system using universal adversarial perturbations." ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2021.
20. Chen, Guangke, et al. "AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems." IEEE Transactions on Dependable and Secure Computing (2022).
21. Xie, Yi, et al. "Enabling fast and universal audio adversarial attack using generative model." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 16. 2021.
22. Li, Jiguo, et al. "Universal adversarial perturbations generative network for speaker recognition." 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020.
23. Hanina, Shoham, et al. "Universal Adversarial Attack Against Speaker Recognition Models." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.
24. Panayotov, Vassil, et al. "Librispeech: an asr corpus based on public domain audio books." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.
25. Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset." arXiv preprint arXiv:1706.08612 (2017).
26. Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition."
27. Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International conference on machine learning. PMLR, 2023.
28. Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis." Advances in neural information processing systems 33 (2020): 17022-17033.
29. Chen, Yafeng, et al. "Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency." arXiv preprint arXiv:2406.02167 (2024).
30. Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." (2020).
31. Zolfi, Alon, et al. "Adversarial mask: Real-world universal adversarial attack on face recognition models." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland, 2022.
32. Hanina, Shoham, et al. "Universal Adversarial Attack Against Speaker Recognition Models." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.