

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

ORPHEUS: A Separation-Robust Proactive Defense for Singing Voice Conversion

Zhaolin Wei, *School of Cyber Science and Engineering, Wuhan University;* Dengpan Ye, *School of Cyber Science and Engineering, Wuhan University, and Cyberspace Institute of Advanced Technology, Guangzhou University;* Yanjiao Chen, *College of Electrical Engineering, Zhejiang University;* Jiacheng Deng, Ziyi Liu, Yuhan Lin, and Yunna Lv, *School of Cyber Science and Engineering, Wuhan University;* Yueyun Shang, *School of Mathematics and Information Science, Guangzhou University;* Zhihong Tian, *Cyberspace Institute of Advanced Technology, Guangzhou University*

https://www.usenix.org/conference/usenixsecurity26/presentation/wei-zhaolin

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

ORPHEUS: A Separation-Robust Proactive Defense for Singing Voice Conversion

Zhaolin Wei[1] Dengpan Ye[1,2,*] Yanjiao Chen[3] Jiacheng Deng[1] Ziyi Liu[1]
Yuhan Lin[1] Yunna Lv[1] Yueyun Shang[4,*] Zhihong Tian[2]

[1]*School of Cyber Science and Engineering, Wuhan University*
[2]*Cyberspace Institute of Advanced Technology, Guangzhou University*
[3]*College of Electrical Engineering, Zhejiang University*
[4]*School of Mathematics and Information Science, Guangzhou University*

Abstract

Recent advances in singing voice conversion enable realistic cloning of a singer's voice, raising concerns about unauthorized voice misuse. Existing proactive defenses inject imperceptible perturbations to disrupt such conversion, and they are designed for vocal signals rather than mixed music. In real-world scenarios, however, attackers usually obtain mixed music and apply source separation to decompose the mixture into vocal and backing tracks before conversion. Since separation models are trained to distinguish vocals from background components, perturbations are often treated as backing, causing the perturbations to be filtered out. To address this, we present ORPHEUS, the first proactive defense framework tailored for singing voice conversion involving source separation. Specifically, we jointly (i) interfere with the separation model via mask-misguiding and cross-track losses, (ii) disrupt identity information using an ensemble of heterogeneous speaker encoders to enhance transferability, and (iii) optimize perceptual quality through tonality harmony constraints and psychoacoustic masking. To evaluate ORPHEUS, we conduct experiments on two datasets with three source separation models and four singing voice conversion systems. Compared with the best-performing baseline, ORPHEUS further reduces identity similarity by 8.02% and 5.61% on two speaker verification models, respectively, demonstrating consistently stronger defense effectiveness and cross-model transferability.

1 Introduction

Recent breakthroughs in generative artificial intelligence have revolutionized the landscape of AI-driven artistic creation. Among these advances, singing voice conversion has emerged as one of the most prominent and impactful techniques [27]. By leveraging algorithm-driven acoustic modeling and timbre transfer, singing voice conversion enables high-fidelity vocal reconstruction, significantly lowering the barrier for non-professional creators to produce realistic artistic expressions. With continuous improvements in model architectures, AI music creation has rapidly gained global visibility, increasingly permeating mainstream cultural and commercial contexts. Real-world deployments illustrate the profound impact of singing voice conversion. For instance, the SYKI-SVC team successfully integrated AI vocals into the opera Turandot [9], garnering over 3.2 million views and diplomatic praise for 'redefining opera inheritance'. Similarly, in the commercial sector, the AI artist Sanja Monét achieved Billboard success and a USD 3 million recording contract [8]. However, the widespread adoption of singing voice conversion has also precipitated growing ethical concerns. Unauthorized voice cloning can directly infringe upon the rights to performers' vocal identities. On the legislative front, U.S. Congresswoman Madeleine Dean and others have proposed the 'No AI FRAUD Act' [5], aiming to establish a federal legal framework to prevent generative AI from utilizing individuals' voices without authorization. Within the industry, the Artist Rights Alliance released an open letter supported by over 200 artists calling on digital music and AI developers to cease utilizing AI technologies in ways that undermine and devalue human artists [7]. Developing effective technical protection methods is an urgent imperative for both the music industry and society.

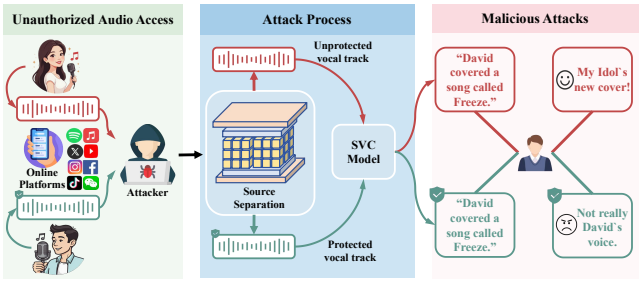


Figure 1: Threat Scenario. Red arrows indicate unprotected music, while green arrows indicate protected music.

To mitigate the threats posed by voice conversion, prior

*Corresponding authors: yedp@whu.edu.cn, 36214001@qq.com.

research [26, 33, 64] typically introduce adversarial perturbations to force synthesized speech to deviate from the original speaker in terms of timbre or within the speaker embedding space. However, most existing methods are tailored for spoken speech and fail to adequately account for the unique characteristics of music signals, such as backing, harmonic structures, and multi-track compositions [43]. Crucially, existing proactive defense methods generally overlook the actual attack pipeline employed in real-world singing voice conversion misuse scenarios. In practice, attackers rarely perform cloning directly on mixed music. Instead, they first employ source separation models [51] to decompose a song into vocal and backing tracks, subsequently using the clean vocal track as input for the singing voice conversion system [13]. This preprocessing step is essential for high-quality cloning, as converting mixed music directly inevitably introduces backing noise into the synthesized vocals [19]. Source separation models are trained to distinguish vocals from backing using time–frequency cues, and perturbations that are not strongly aligned with vocal structures are often interpreted as backing components and suppressed during separation. As a result, adversarial perturbations injected by existing defenses are largely removed before reaching the singing voice conversion model, causing their protection effectiveness to degrade significantly under the realistic 'separation + conversion' setting.

To address these limitations, we propose ORPHEUS, a novel proactive defense method designed for practical singing voice conversion attack scenarios. ORPHEUS is resilient to source separation and effectively protects the privacy of a singer's vocal identity in music. This protection can be embedded into music tracks by copyright holders prior to release. When protected music is maliciously exploited for singing voice conversion, the downstream singing voice conversion system produces distorted and identity-inconsistent outputs, thereby preventing unauthorized voice cloning. Unlike prior work, ORPHEUS explicitly models the complete attack pipeline that consists of source separation followed by singing voice conversion. To the best of our knowledge, it is the first proactive music protection framework that is specifically designed to defend against the realistic 'separation + conversion' attack setting. In designing ORPHEUS, we identify and address the following key technical challenges:

Challenge 1: Unknown source separation models and transferability. In practice, defenders have no access to the specific separation models used by attackers [47], requiring perturbations to generalize across diverse architectures [57]. To this end, we target model-agnostic mechanisms shared by separation systems rather than overfitting to a particular model. We introduce two strategies. First, we disrupt the prediction matrices that characterize vocal–backing assignment, a common intermediate representation across separation models. Second, we propose a cross-track loss that enforces interference between the separated vocal and backing tracks, directly blurring their semantic boundary. By operating on universal separation cues instead of model-specific parameters, these designs enable perturbations to remain effective across unseen architectures.

Challenge 2: The singing voice conversion model used by the attacker is unknown. Since the defense objective is to protect vocal timbre features, the adversarial perturbations must remain effective across different singing voice conversion architectures. To this end, ORPHEUS adopts an ensemble-learning strategy [14, 38, 61, 64], jointly optimizing against multiple speaker representation models with diverse architectures serving as surrogates, thereby improving transferability to unseen singing voice conversion systems.

Challenge 3: Music scenarios impose stricter requirements on perceptual quality. Unlike spoken speech, music is an artistic medium demanding significantly higher auditory harmony and imperceptibility. Existing approaches primarily rely on psychoacoustic models [14, 33] or frequency-domain penalties to limit perceptibility [64], often neglecting the interaction with the backing. In this work, we introduce a harmonic alignment loss that measures tonality consistency [23] between the perturbation and both the vocal and backing tracks, ensuring that perturbations align naturally with the musical structure. Combined with a psychoacoustic loss, this design further optimizes overall perceptual quality.

We conduct comprehensive experiments on two public music datasets (MUSDB18-HQ [49] and DSD100 [40]), covering three representative source separation models and four state-of-the-art zero-shot singing voice conversion systems. Compared with three representative proactive defense methods, ORPHEUS consistently achieves lower speaker identity similarity under a realistic attack pipeline that combines source separation and singing voice conversion, demonstrating its effectiveness in practical scenarios.

In summary, our main contributions are as follows:

- We propose ORPHEUS, the first proactive music protection framework explicitly designed for real-world singing voice conversion pipelines that include source separation. To achieve strong black-box transferability across unknown separation and singing voice conversion models, ORPHEUS introduces mask-misguiding loss and cross-track loss to perturb the universal decision boundaries of separation networks, together with an ensemble-based identity disruption objective that simultaneously biases multiple heterogeneous speaker encoders.

- We exploit the multi-track and harmonic characteristics of music to design adversarial perturbations with tonality harmony constraints. Combined with a psychoacoustic loss, this design significantly improves imperceptibility and musical naturalness while preserving defensive effectiveness.

- We conduct extensive evaluations on multiple source separation models, zero-shot singing voice conversion

systems, and adaptive attack settings, including over-the-air recording, as well as human studies. The results show that ORPHEUS consistently outperforms existing defenses, reducing speaker identity similarity by up to 8.02% and 5.61% on two speaker verification models, and demonstrating superior effectiveness and robustness.

2 Related Work

2.1 Singing Voice Conversion

Singing voice conversion aims to transform the voice of one singer into that of another while preserving the original lyrical content and melodic structure. Recent advances have enabled zero-shot synthesis, allowing the generation of singing voices for target singers whose voices are not included in the training data. Grad-SVC [4] is an open-source singing voice conversion framework derived from Grad-TTS [46] that extends diffusion-based modeling to the singing voice conversion task. It employs a diffusion probabilistic model to iteratively optimize the noise injection and denoising processes at each diffusion step, thereby achieving high-quality singing voice synthesis. Sha et al. proposed NeuCoSVC [53], a neural concatenative singing voice conversion framework that leverages self-supervised learning (SSL) features extracted from reference audio to effectively mitigate the color leakage problem commonly observed in prior separation-based approaches. Liu et al. introduced Seed-VC [37], which incorporates an external timbre perturber during training to further alleviate color leakage and adopts a diffusion-based converter for fine-grained timbre modeling. KNN-SVC [54] utilizes a k-nearest neighbor (kNN) framework to enhance concatenative synthesis quality. By leveraging nearest-neighbor matching within a self-supervised feature space—augmented by additive synthesis and temporal smoothing—it achieves robust tonality consistency in zero-shot scenarios. HQ-SVC [12] employs a unified decoupled encoding mechanism to simultaneously extract content and speaker representations via a decoupled audio codec. By integrating an enhanced voice adaptation module, pitch/range modeling, and differentiable signal processing with diffusion techniques, it facilitates high-fidelity zero-shot conversion even under low-resource constraints.

While these zero-shot singing voice conversion methods substantially advance synthesis quality and flexibility, they also significantly lower the barrier for misuse. In particular, the widespread availability of open-source zero-shot singing voice conversion models enables attackers to perform high-fidelity singing voice cloning with minimal technical expertise, thereby exacerbating the risk of voice identity theft [14].

2.2 Source Separation

Source separation is a core technique in audio signal processing aiming to decompose mixed musical signals into independent source tracks. This technology holds significant value across music production, audio editing, and downstream tasks such as singing voice conversion and music analysis. SIGSEP proposed Open-Unmix [56], an architecture combining CNNs and three-layer BiLSTMs that trains independently for each target source, employing spectral mask prediction to extract specific tracks from audio of arbitrary duration. Facebook introduced Demucs [51], an end-to-end waveform model featuring a hybrid U-Net convolutional encoder-decoder and BiLSTM architecture; by modeling directly in the time domain, it effectively mitigates the phase reconstruction and authenticity limitations of traditional spectral methods. Luo et al. developed Band-Split RNN [41], which introduces a 'band-split' mechanism to partition the spectrum into sub-bands of varying bandwidths, utilizing dual-path RNNs to alternately model sequence and band dimensions for fine-grained separation. Finally, the Deezer team presented Spleeter [25], a widely adopted model based on a 12-layer U-Net that leverages skip connections to integrate multi-scale features via spectral soft-masking, while additionally supporting customized training for diverse processing needs.

2.3 Proactive Defense of Adversarial Examples

Proactive defense methods mitigate the threat of synthesized speech by introducing adversarial perturbations that hinder the generation of speech clones. Huang et al. proposed AttackVC [26], which represents the first attempt to employ adversarial examples to disrupt voice conversion systems. However, its reliance on the L_∞ norm to constrain perturbation imperceptibility does not adequately reflect human perceptual perception. Li et al. introduced VoiceGuard [33], which extends adversarial perturbations to the time domain and incorporates psychoacoustic models to enhance perceptual concealment. Nevertheless, VoiceGuard exhibits limited transferability to singing voice conversion systems. Yu et al. developed AntiFake [64], which improves transferability under black-box settings through encoder ensemble strategies and weighted loss functions, albeit at the cost of degraded speech quality. Zhang et al. proposed SafeSpeech [65], which leverages a surrogate text-to-speech model to inject imperceptible perturbations, effectively disrupting unauthorized speech synthesis. While achieving strong transferability and robustness within text-to-speech threat models, SafeSpeech is ill-suited for the singing voice conversion setting. Crucially, it fails to account for the source separation stage, which is ubiquitous in realistic music-based attack pipelines. Chen et al. proposed SongBsAb [14], the first method designed to protect both timbre and content in singing voice conversion. However, SongBsAb shows limited effectiveness in realistic attack scenarios, as it lacks robustness against source separation models. In Table 1, we describe the comparison between ORPHEUS and related works.

Although these approaches have demonstrated effective-

Table 1: Comparison between ORPHEUS and related works.

Methods	Target Model	Purpose	Imperceptibility	Separation Robustness
AttackVC	Voice conversion	identity disruption	L_∞ norm	✗
VoiceGuard			psychoacoustics model	✗
AntiFake	Voice conversion & synthesis		frequency penalty & SNR	✗
SafeSpeech	Text-to-speech		perceptual fidelity optimization	✗
Songbsab	Singing voice conversion	identity & lyric disruption	psychoacoustics model (with backing tracks)	✗
ORPHEUS **(ours)**		identity disruption & separation interference	harmonic alignment loss & psychoacoustics model	✓

ness in mitigating voice conversion, text-to-speech, and singing voice conversion attacks, they do not fully capture the distinctive challenges of music protection. Unlike ordinary speech, music is a structured mixture that contains long-range temporal dependencies, rich semantic content, and strong coupling between vocals and accompaniment [50]. These properties make transferability in music substantially more challenging, especially when perturbations must remain effective after source separation. In contrast, ORPHEUS is explicitly designed for this music-specific setting and the realistic separation-plus-conversion attack pipeline.

3 Threat Model

3.1 Attacker Capability and Motivation

Attacker Capabilities. We assume that attackers can acquire singing data of the target singer through publicly available channels, such as social media platforms, videos, and live performances, as well as through illicit means, including unauthorized recordings. Consequently, attackers can construct a sufficiently large and diverse dataset for model training. Furthermore, attackers are assumed to possess substantial computational resources and domain expertise, enabling them to train or fine-tune advanced singing voice conversion and source separation models, and to perform zero-shot or few-shot singing voice conversion using the victim's vocal data.

Figure 1 illustrates the realistic threat scenario considered in this work. After obtaining music recordings from online platforms, an attacker first applies a source separation model to extract the vocal track, and then feeds the separated vocals into a singing voice conversion system to synthesize a fake song that mimics the target singer. The red arrows in Figure 1 indicate the conventional pipeline operating on unprotected audio, where the clean separated vocals enable high-fidelity identity cloning. In contrast, the green arrows denote the case where the released music has been protected, in which the separation–conversion pipeline no longer yields clean and usable vocal inputs for reliable identity cloning. This two-stage pipeline reflects the dominant workflow in practical attacks, as directly applying singing voice conversion to mixed music would introduce backing leakage and severely degrade synthesis quality.

Attacker Motivation. With the rapid advancement of singing voice conversion models, attackers are now capable of generating highly realistic singing voices that closely resemble specific target singers. In practical application scenarios, we identify two primary categories of threats.

(1) *Privacy infringement and reputational harm.* Attackers may exploit singing voice conversion technology to generate malicious or deceptive content, such as manipulated songs or fabricated audio recordings, thereby infringing upon the privacy rights of the target individual. Such misuse can result in long-term damage to the victim's reputation, public image, and social standing.

(2) *Artistic copyright infringement.* Attackers may also misuse a singer's vocal characteristics for unauthorized creative activities. These include the illegal production of songs for commercial gain, tampering with the original vocals in classic works to provoke copyright disputes, and the unlicensed commercial exploitation of a singer's voice identity.

3.2 Defender Goals

The defender's objective is to protect a singer's vocal identity by embedding adversarial perturbations into music recordings, thereby preventing malicious exploitation of the singer's voice on public platforms such as video-sharing websites and music streaming services. To effectively counter real-world attackers, the defense mechanism must simultaneously satisfy the following three goals.

(1) *Robustness to source separation.* The adversarial perturbations injected into the music must remain effective under source separation. Specifically, even after an attacker applies a source separation model to extract the vocal track from the protected music, the perturbations should persist within the separated vocals.

(2) *Voice identity protection.* The embedded perturbations must effectively protect the singer's vocal characteristics. That is, when the separated vocals are used as input to a singing voice conversion system, the synthesized output should not resemble the victim singer's voice.

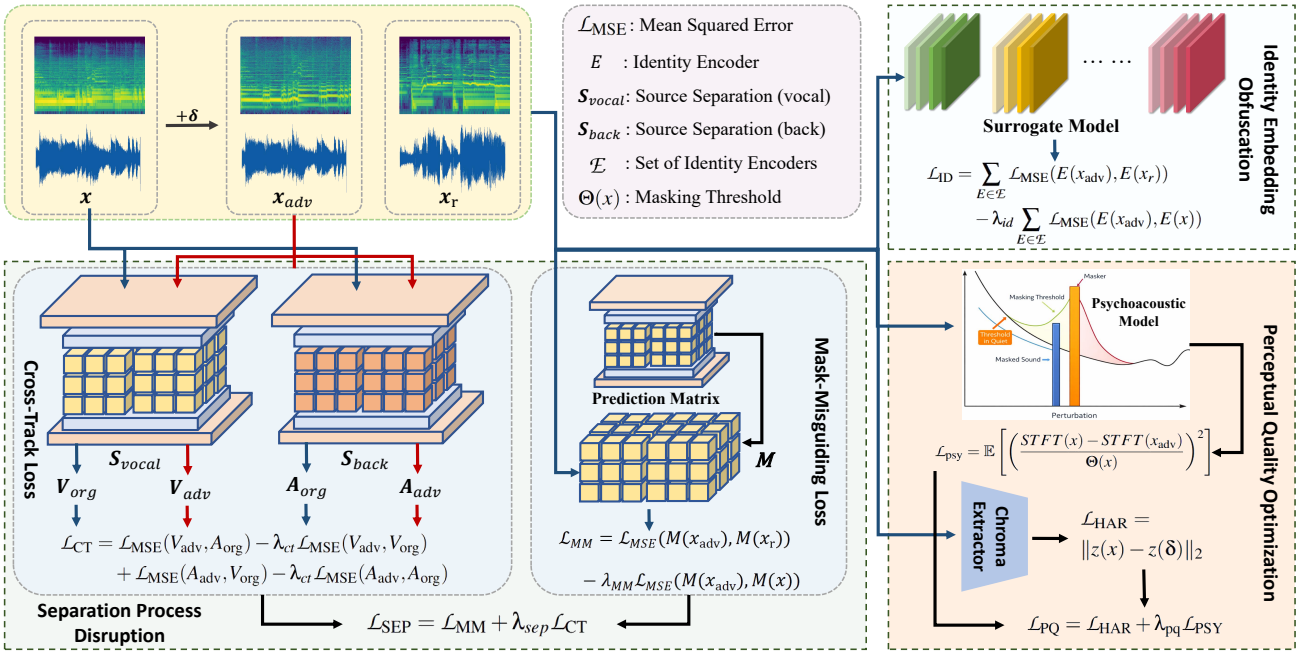


Figure 2: Overview of the methodology of ORPHEUS.

(3) *High perceptual quality.* The introduction of adversarial perturbations should minimally degrade the perceptual quality of the music, preserving both the naturalness of the singing voice and the overall musical experience.

Among these objectives, robustness to source separation is fundamental. If perturbations are removed during separation, they cannot reach the downstream singing voice conversion model, rendering identity protection and perceptual optimization ineffective. However, most existing defenses overlook this stage and assume that perturbations are directly applied to clean vocals. In contrast, practical singing voice conversion attacks require a defense that remains effective after source separation.

4 ORPHEUS: Detailed Construction

In this section, we first formalize the design objective of ORPHEUS and then describe the three components.

Problem Formulation. Let x denote the original music sample of a target singer, and let $x_{\text{adv}} = x + \delta$ denote the protected version produced by ORPHEUS, where δ is an imperceptible adversarial perturbation. The goal of ORPHEUS is to optimize δ such that the protected sample simultaneously satisfies three requirements: (1) it disrupts the source separation stage exploited in realistic attacks, (2) it obfuscates the singer's identity in the downstream singing voice conversion pipeline, and (3) it preserves the perceptual quality of the original music. Accordingly, we formulate the overall optimization objective as

$$\arg\min_{\delta} \lambda_{\text{SEP}}\mathcal{L}_{\text{SEP}} + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}} + \lambda_{\text{PQ}}\mathcal{L}_{\text{PQ}}, \quad (1)$$

where $\mathcal{L}_{\text{SEP}}$ denotes the separation disruption loss, $\mathcal{L}_{\text{ID}}$ denotes the identity obfuscation loss, and $\mathcal{L}_{\text{PQ}}$ denotes the perceptual quality loss. The three coefficients λ_{SEP}, λ_{ID}, and λ_{PQ} balance the tradeoff among robustness to source separation, resistance to identity extraction, and preservation of auditory quality.

As illustrated in Fig. 2, ORPHEUS comprises three corresponding components:

- *Separation process disruption*. This component targets the source separation model exploited in realistic attacks. It contributes $\mathcal{L}_{\text{SEP}}$ through two objectives: (1) misguiding the predicted masks that delineate vocal and accompaniment regions, and (2) confounding the semantic distinction between the separated vocal and backing tracks.
- *Identity embedding obfuscation*. This component targets the singer-specific embeddings extracted from the separated audio. It contributes $\mathcal{L}_{\text{ID}}$ through both untargeted and targeted obfuscation objectives, causing the extracted identity features to deviate from those of the protected singer.
- *Perceptual quality optimization*. This component ensures that the protected music remains perceptually faithful to the original. It contributes $\mathcal{L}_{\text{PQ}}$ through two objectives: (1) a harmonic alignment term that improves the

compatibility between the perturbation and the musical accompaniment, and (2) a psychoacoustic masking term that suppresses the audibility of the perturbation in the time-frequency domain.

4.1 Separation Process Disruption

The separation process disruption module is designed to compromise the source separation stage foundational to modern singing voice conversion frameworks. First, we perturb the model's learned masks to induce the system to misattribute vocal regions to the backing track and vice versa. Second, we obfuscate the semantic features of the underlying tracks, effectively preventing the model from successfully disentangling the vocal track from the backing track. This disruption ensures that any subsequent voice conversion is performed on a corrupted or poorly isolated source.

Mask-Misguiding Loss. Modern source separation models typically learn an intermediate latent mask (or prediction matrix) to determine whether specific time-frequency regions belong to the vocal or the backing track [34]. Although architectures vary, the functional role of these masks is largely consistent across source separation frameworks. Leveraging this commonality, we design a mask-misguiding loss to systematically perturb the mask prediction process. Let $\mathcal{M}$ denote the mask-generation module of a source separation model. We define:

$$\begin{aligned} \mathbf{M}_0 &= \mathcal{M}(x), \\ \mathbf{M}_{\text{adv}} &= \mathcal{M}(x_{\text{adv}}), \\ \mathbf{M}_r &= \mathcal{M}(x_r), \end{aligned} \tag{2}$$

where x is the original sample, x_{adv} is the protected sample, and x_r is a reference sample. In our implementation, x_r is drawn from a proxy singer that is acoustically distinct from the protected singer. Following prior work [14, 33], we select the candidate with the largest timbre distance from the protected singer, as a more distant reference provides a clearer displacement direction in the embedding space and yields more stable optimization than a proximate one. Our objective is to drive the predicted mask of the protected sample, $\mathbf{M}_{\text{adv}}$, away from the original mask $\mathbf{M}_0$ and toward the reference mask $\mathbf{M}_r$, thereby inducing systematic separation errors.

The mask-misguiding loss is defined as

$$\mathcal{L}_{\text{MM}} = \mathcal{L}_{\text{MSE}}(\mathbf{M}_{\text{adv}}, \mathbf{M}_r) - \lambda_{mm}\mathcal{L}_{\text{MSE}}(\mathbf{M}_{\text{adv}}, \mathbf{M}_0), \tag{3}$$

where $\mathcal{L}_{\text{MSE}}$ denotes the mean squared error and λ_{mm} is a hyperparameter. By driving the predicted mask toward a target reference while diverging from the ground truth, this loss sabotages the model's ability to delineate vocal and backing regions. Because such masking mechanisms are a foundational abstraction in source separation, this strategy achieves high cross-model transferability.

Table 2: Properties of the ensemble encoders.

Encoder	Architecture	Emb. Size	Dataset
ResNet34 [2]	ResNet	256	VoxCeleb [44]
HuBERT-Soft [58]	BERT	256	LibriSpeech [45]
AdaIN [21]	VAE	128	VCTK [62]
GE2E [59]	LSTM	256	LibriTTS + VCTK + VoxCeleb + CommonVoice [10]

Emb. Size = Embedding Size

Cross-Track Loss. While the Mask-Misguiding Loss perturbs the internal decision mechanism of source separation models, it does not explicitly constrain the structure of the separated outputs. In practice, most music source separation systems decompose an input mixture into two primary tracks: a vocal track and a backing track. A natural yet underexplored confusion strategy is therefore to directly disrupt the semantic distinction between these two outputs. To this end, we introduce the Cross-Track Loss (CT Loss). The core intuition is to encourage the separated vocal track to resemble the backing track, and vice versa. Such bidirectional confusion blurs the boundary between vocal and accompaniment components, causing residual accompaniment signals to persist in the vocal track.

Formally, given a clean music sample x and its adversarial version x_{adv}, we denote the vocal and accompaniment outputs of the source separation model as:

$$\begin{aligned} V_{\text{org}} &= S_{\text{vocal}}(x), & A_{\text{org}} &= S_{\text{back}}(x), \\ V_{\text{adv}} &= S_{\text{vocal}}(x_{\text{adv}}), & A_{\text{adv}} &= S_{\text{back}}(x_{\text{adv}}). \end{aligned} \tag{4}$$

where $S_{vocal}(\cdot)$ and $S_{back}(\cdot)$ denote the vocal and accompaniment separation heads, respectively. The Cross-Track Confusion Loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{CT}} = \mathcal{L}_{\text{MSE}}(V_{\text{adv}}, A_{\text{org}}) - \lambda_{ct}\mathcal{L}_{\text{MSE}}(V_{\text{adv}}, V_{\text{org}}) \\ + \mathcal{L}_{\text{MSE}}(A_{\text{adv}}, V_{\text{org}}) - \lambda_{ct}\mathcal{L}_{\text{MSE}}(A_{\text{adv}}, A_{\text{org}}) \end{aligned} \tag{5}$$

We combine them to form the overall separation process loss as:

$$\mathcal{L}_{\text{SEP}} = \mathcal{L}_{\text{MM}} + \lambda_{sep}\mathcal{L}_{\text{CT}}, \tag{6}$$

where λ_{sep} is a weight coefficient that balances the contributions of mask-misguiding loss and cross-track loss.

4.2 Identity Embedding Obfuscation

Following source separation, singing voice conversion frameworks perform identity encoding, utilizing pre-trained encoders to extract singer-specific embeddings for synthesis. To safeguard the singer's identity, we design an identity embedding obfuscation module to explicitly perturb these latent representations through a combination of untargeted and targeted obfuscation objectives.

In real-world scenarios, attackers may deploy a wide array of identity encoders with varying architectures, which

often limits the transferability of perturbations optimized for a single model. To mitigate this, we adopt an ensemble-based optimization strategy [14, 61, 64] that jointly disrupts the embeddings produced by a diverse set of representative encoders. Let $\mathcal{E}$ denote a set of heterogeneous encoders, encompassing LSTM, ResNet, BERT, and VAE-based architectures. As summarized in Table 2, our ensemble covers a broad spectrum of network topologies and training corpora, including an LSTM-based GE2E encoder, a ResNet-based verification model, a self-supervised HuBERT representation model, and a VAE-based encoder utilizing Adaptive Instance Normalization (AdaIN).

Untargeted Identity Obfuscation Loss. The untargeted objective is designed to prevent the converted singing voice from resembling the protected singer. This is achieved by maximizing the distance between the adversarial and original embeddings across the entire ensemble $\mathcal{E}$. The untargeted identity loss is formulated as:

$$\mathcal{L}_{\text{ID}}^{\text{untar}} = -\sum_{E\in\mathcal{E}} \mathcal{L}_{\text{MSE}}(E(x_{\text{adv}}), E(x)). \tag{7}$$

Targeted Identity Obfuscation Loss. While untargeted disruption effectively degrades identity similarity, it can result in stochastic or ill-defined latent representations that hinder optimization stability. To provide more structured guidance and enhance cross-model transferability, we introduce a targeted identity obfuscation loss. Specifically, we select a reference vocal signal x_r from a proxy singer who is acoustically distinct from the protected singer. We then constrain the protected sample to converge toward this target identity within the embedding space. The targeted identity loss is formulated as:

$$\mathcal{L}_{\text{ID}}^{\text{tar}} = \sum_{E\in\mathcal{E}} \mathcal{L}_{\text{MSE}}(E(x_{\text{adv}}), E(x_r)). \tag{8}$$

The overall identity embedding loss is obtained by combining the untargeted and targeted objectives:

$$\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{ID}}^{\text{tar}} + \lambda_{id}\mathcal{L}_{\text{ID}}^{\text{untar}}, \tag{9}$$

where λ_{id} controls the balance between identity deviation and targeted guidance.

4.3 Perceptual Quality Optimization

To remain practical for music distribution, the protected sample must maintain high perceptual fidelity. Conventional signal-level constraints (e.g., L_p norms) often fail to account for the structured nature of human auditory perception in complex musical mixtures. To address this, ORPHEUS optimizes imperceptibility through two complementary strategies: harmonic alignment and psychoacoustic masking, while strictly constraining the perturbation amplitude δ within the range of $[-0.015, 0.015]$. The harmonic alignment objective enforces spectral consistency between the perturbation and the backing track, and the psychoacoustic masking leverages the human auditory system's limitations to render the adversarial noise inaudible.

Harmonic alignment Loss. We introduce a harmonic alignment Loss to enhance the perceptual masking of adversarial perturbations by enforcing spectral consistency with the backing track. Intuitively, perturbations that adhere to the tonal structure of the background music are more effectively masked by the musical context. To implement this, we project the audio signal into a chroma-based tonal representation [23, 30]. For a given chroma vector $\mathbf{c}(x) \in \mathbb{R}^{12}$, which represents the energy distribution across the 12 semitone pitch classes, we normalize the vector and project it into a tonal space using a predefined transformation matrix $\mathbf{T}$ as:

$$z(x) = T \cdot \frac{c(x)}{\|c(x)\|_1}. \tag{10}$$

The harmonic alignment loss is then defined as the Euclidean distance between the tonal projections of the original audio and the adversarial perturbation δ:

$$\mathcal{L}_{\text{HAR}} = \|z(x) - z(\delta)\|_2. \tag{11}$$

By minimizing this distance, we ensure that the perturbation's energy is concentrated within the existing harmonic framework of the music, rendering it less detectable to the human ear.

Psychoacoustic loss. To further ensure imperceptibility, we incorporate a psychoacoustic masking loss grounded in auditory masking theory [3, 14, 33]. For the original signal x, we compute a time-frequency masking threshold $\Theta(x)$ using a Bark-scale psychoacoustic model, which defines the maximum energy of a signal that remains inaudible in the presence of a masker. The psychoacoustic loss penalizes the spectral discrepancy between the original and adversarial signals relative to this threshold:

$$\mathcal{L}_{\text{PSY}} = \mathbb{E}\left[\left(\frac{\text{STFT}(x) - \text{STFT}(x_{\text{adv}})}{\Theta(x)}\right)^2\right], \tag{12}$$

where $\text{STFT}(\cdot)$ denotes the Short-Time Fourier Transform, and the expectation $\mathbb{E}[\cdot]$ is taken over time–frequency bins.. By normalizing the error by $\Theta(x)$, the optimization prioritizes the suppression of perturbations in frequency regions where the human ear is most sensitive (i.e., where the masking threshold is low).

The overall perceptual quality loss is formulated as:

$$\mathcal{L}_{\text{PQ}} = \mathcal{L}_{\text{HAR}} + \lambda_{\text{pq}}\mathcal{L}_{\text{PSY}} \tag{13}$$

4.4 Method Summary

The overall objective in Eq. (1) is instantiated as a weighted combination of the three loss terms:

$$\mathcal{L} = \lambda_{\text{SEP}}\mathcal{L}_{\text{SEP}} + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}} + \lambda_{\text{PQ}}\mathcal{L}_{\text{PQ}}, \tag{14}$$

Algorithm 1 ORPHEUS

Require: Music x, target reference x_r, mask extractor $M(\cdot)$, separation heads $S_{\text{vocal}}(\cdot), S_{\text{acc}}(\cdot)$, identity encoder set $\mathcal{E}$, tonality embedder $z(\cdot)$, masking estimator $\Theta(\cdot)$, budget ε, iterations T, learning rate η, weights $\lambda_{sep}, \lambda_{id}, \lambda_{\text{pq}}, \lambda_{\text{SEP}}, \lambda_{\text{ID}}, \lambda_{\text{PQ}}$

Ensure: Protected music x_{adv}

1: Initialize $p \leftarrow 0$; optimizer $\text{Adam}(p, \eta)$
2: Cache (no grad): $M_0 \leftarrow M(x)$, $M_{\text{r}} \leftarrow M(x_{\text{r}})$, $\Theta_x \leftarrow \Theta(x)$
3: **for** $t = 1$ to T **do**
4: $\quad \delta \leftarrow \varepsilon \tanh(p)$
5: $\quad x_{\text{adv}} \leftarrow x + \delta$
6: $\quad M_{\text{adv}} \leftarrow M(x_{\text{adv}})$
7: $\quad C_1 \leftarrow \mathcal{L}_{\text{MM}}(M_{\text{adv}}, M_{\text{r}}, M_0)$
8: $\quad V_{\text{org}} \leftarrow S_{\text{vocal}}(x)$; $A_{\text{org}} \leftarrow S_{\text{acc}}(x)$
9: $\quad V_{\text{adv}} \leftarrow S_{\text{vocal}}(x_{\text{adv}})$; $A_{\text{adv}} \leftarrow S_{\text{acc}}(x_{\text{adv}})$
10: $\quad C_2 \leftarrow \mathcal{L}_{\text{CT}}(V_{\text{adv}}, A_{\text{adv}}, V_{\text{org}}, A_{\text{org}})$
11: $\quad C_3 \leftarrow C_1 + \lambda_{sep} C_2$
12: $\quad C_4 \leftarrow \mathcal{L}_{\text{ID}}(x_{\text{adv}}, x, x_{\text{r}}, \mathcal{E}; \lambda_{id})$
13: $\quad C_5 \leftarrow \mathcal{L}_{\text{HAR}}(z(x), z(\delta))$
14: $\quad C_6 \leftarrow \mathcal{L}_{\text{PSY}}(x, x_{\text{adv}}, \Theta_x, \delta)$
15: $\quad C_7 \leftarrow C_5 + \lambda_{\text{pq}} C_6$
16: $\quad \mathcal{L} \leftarrow \lambda_{\text{SEP}} C_3 + \lambda_{\text{ID}} C_4 + \lambda_{\text{PQ}} C_7$
17: $\quad p \leftarrow \text{Adam}(p, \nabla_p \mathcal{L})$
18: **end for**
19: **return** x_{adv}

where the three terms govern robustness to source separation, resistance to identity extraction, and preservation of auditory quality, respectively.

Algorithm 1 summarizes the optimization procedure of ORPHEUS. Following prior audio adversarial defense methods [26,33], we parameterize the perturbation as $\delta = \varepsilon \tanh(p)$. The $\tanh(\cdot)$ function maps the unconstrained optimization variable p into $(-1, 1)$, ensuring that the perturbation remains smoothly bounded upon scaling by ε. This reparameterization enables direct optimization of p via Adam while guaranteeing that the resulting perturbation satisfies the prescribed magnitude constraint.

In each iteration, the protected music sample is synthesized by adding the bounded perturbation to the original signal. The optimization then evaluates a multi-objective loss function consisting of three primary components. First, the separation process loss integrates mask-misguiding and cross-track losses to degrade the fidelity and reliability of the isolated vocal track (Lines 6–11). Second, the identity embedding Loss is calculated across an ensemble of heterogeneous encoders, simultaneously driving the vocal track away from the protected singer and toward a distant reference identity (Line 12). Third, the perceptual quality loss enforces harmonic alignment and psychoacoustic masking to ensure the perturbation remains inaudible (Lines 13–15). The final objective is computed as a weighted sum of these terms, and the perturbation parameters are updated via backpropagation. After T iterations, the resulting adversarial music is produced as the protected output sample.

Table 3: The detailed information of singing voice conversion models.

Models	Content Encoder	Identity Encoder	Pitch Encoder	Decoder / Vocoder
HQ-SVC	Decoupled codec jointly content & speaker	Same codec	Explicit pitch & loudness modeling	NSF-HiFiGAN [29]
Neuco-SVC	WavLM-Large [17]	WavLM-Large	pYIN [42] & REAPER [1]	FiLM UNet [35]
Seed-VC	Whisper [48]	Encoder timbre head	External F0 extractor	BigVGAN [32]
KNN-SVC	WavLM features with nearest neighbors	Implicit via kNN candidates	Additive synthesis & harmonic features	HiFi-GAN

5 Evaluations

This section details the experimental setup, covering the target singing voice conversion models, datasets, baselines, hyperparameters, and evaluation metrics. All experiments were conducted on an NVIDIA A800 PCIe GPU.

5.1 Singing Voice Conversion Models

To simulate capable attackers utilizing advanced DeepFake technologies and to evaluate the transferability of ORPHEUS, we target state-of-the-art singing voice conversion models with zero-shot capabilities. Following a screening of various open-source implementations, we selected four representative models: HQ-SVC, NeuCo-SVC, Seed-VC, and KNN-SVC. We provide a more detailed introduction to each model in Table 3.

HQ-SVC. HQ-SVC is a generative model designed for high-quality zero-shot singing conversion. It employs a decoupled audio encoder to jointly extract content and speaker features, integrating conditions such as pitch and energy within a diffusion modeling framework to generate high-resolution acoustic representations. We utilize the official pre-trained model for all experiments.

NeuCo-SVC. NeuCo-SVC is a zero-shot system based on a neural concatenative framework. It utilizes a pre-trained Self-Supervised Learning (SSL) model for content extraction and employs nearest-neighbor matching in the feature space to replace source representations with target features for timbre transfer. Additionally, it integrates neural harmonic generation and waveform synthesis to model pitch and harmonic structures. We adopt the official implementation and provided pre-trained models.

Seed-VC. Seed-VC utilizes a diffusion Transformer architecture for zero-shot speech and singing conversion. It employs pre-trained speech representation models (e.g., Whisper) to extract content features and integrates target timbre information via context conditioning. It further supports pitch (F0) conditioning to ensure melodic consistency. We evaluate this model using the official pre-trained checkpoint.

KNN-SVC. KNN-SVC is a zero-shot method based on nearest-neighbor retrieval. It performs frame-level searches in the self-supervised feature space to replace source features with target representations for timbre transfer. The method incorporates additive synthesis and temporal smoothing to enhance harmonic modeling and continuity. We utilize the public implementation with default configurations.

5.2 Dataset and Baselines

Dataset. To evaluate the robustness of our adversarial perturbations against source separation, we utilized two standard benchmark datasets: MUSDB18-HQ [49] and DSD100 [40]. MUSDB18-HQ, the official dataset for the SiSEC 2018 [55] music separation task, consists of 150 diverse full-length tracks. DSD100, originating from the SiSEC 2016 campaign, comprises 100 complete music works. Both datasets are sampled at 44.1 kHz. Consistent with prior studies, we segmented each track into 10-second clips to mitigate computational overhead.

Baselines. In our experiments, we consider the following four baselines.

- *No-Defence.* The attacker directly performs source separation and singing voice conversion on the raw music without any protection, following the baseline setting in SpeakerMask [39].

- *AntiFake.* A proactive defense that injects adversarial perturbations to disrupt speaker identity in speech synthesis and cloning. We use the official implementation and apply it to music signals for singing voice conversion evaluation.

- *VoiceGuard.* A time-domain adversarial defense with psychoacoustic constraints designed for VC systems. We apply it to music samples and evaluate its transferability to singing voice conversion models.

- *SongBsAb.* A singing-specific protection method that perturbs vocal signals while using a backing track as an additional masker. In the original implementation, however, SongBsAb is evaluated on datasets that contain only isolated vocal tracks, and the backing tracks used for masking are synthetically added rather than taken from the corresponding original backings. To ensure a fair and realistic comparison under authentic music conditions, we adopt the public implementation of SongBsAb but replace the synthetic backing tracks with the original backing tracks from our dataset, so that the vocal and backing remain naturally matched and musically coherent.

5.3 Metrics

Speaker Identity Similarity (SIS) [28]. We evaluate identity protection by measuring the similarity between speaker embeddings of cloned singing voices generated from unprotected and protected music. The cosine similarity between the two embeddings is used as the identity similarity score. We adopt two speaker verification models, CAM++ [60] and ResNetSE [24], and define:

$$\begin{aligned} \text{C-SIS} &= \cos(f_{\text{CAM++}}(y_{\text{raw}}), f_{\text{CAM++}}(y_{\text{adv}})), \\ \text{R-SIS} &= \cos(f_{\text{ResNetSE}}(y_{\text{raw}}), f_{\text{ResNetSE}}(y_{\text{adv}})) \end{aligned} \tag{15}$$

where y_{raw} and y_{adv} denote the cloned singing voices from the original and protected music, respectively, and $f(\cdot)$ denotes the speaker embedding extractor.

Signal-to-Noise Ratio (SNR) [16]. To evaluate the signal-level distortion introduced by adversarial perturbations, we compute the signal-to-noise ratio between the original music signal x and its protected version x_{adv}. A higher SNR indicates that the perturbation has lower energy relative to the original signal, and thus better preserves the audio quality.

Mean Opinion Score (MOS) [31]. We assess perceptual audio quality using both subjective and objective metrics. The subjective Mean Opinion Score (MOS) is obtained from human listeners, who rate the overall audio quality on a scale from 1 to 5, where higher scores indicate better perceived quality.

Naturalness [11]. We further report an objective naturalness score predicted by the SongEval [63] model, which is specifically designed to evaluate the perceptual naturalness of singing voice and music signals. This metric provides a scalable and consistent approximation of human judgments and complements the subjective MOS evaluation.

5.4 Experimental Results

We employ Band-Split-RNN as the surrogate source separation model for perturbation optimization, selected for its state-of-the-art separation performance and ability to estimate precise time–frequency masks. The adversarial perturbations are optimized using the Adam optimizer with a learning rate of $\alpha = 10^{-3}$ for $T = 1{,}000$ iterations, and the perturbation magnitude is constrained by a budget of $\varepsilon = 0.015$. Using this surrogate, we generate perturbations via ORPHEUS and rigorously evaluate their transferability across diverse unseen source separation and singing voice conversion architectures. Table 4 reports the defense performance under the complete black-box attack pipeline; Table 5 quantifies the degradation in effectiveness induced by source separation; and Figure 3 benchmarks the audio quality in terms of SNR and perceptual naturalness.

Baseline Vulnerability. Under the No-Defense baseline, combinations of various separation models (Demucs, Spleeter,

Table 4: Comparison of black-box effectiveness between ORPHEUS and existing works. The best and second-best results are highlighted with bold and underlined, respectively.

Source separation	Singing voice conversion	No-Defence		VoiceGuard		Antifake		Songbsab		ORPHEUS (ours)	
		C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)
Demucs	**HQ-SVC**	0.3822	0.3288	0.3462	0.3214	0.3458	0.3295	0.3154	0.3140	**0.2582**	**0.2578**
	Neuco-SVC	0.4001	0.4865	0.3684	0.4633	0.3804	0.4754	0.3562	0.4108	**0.3111**	**0.3654**
	Seed-VC	0.5753	0.4790	0.4980	0.4534	0.4574	0.4648	0.4189	0.3672	**0.2493**	**0.2957**
	KNN-SVC	0.5121	0.5231	0.4625	0.4902	0.4658	0.5030	0.2778	0.3542	**0.2326**	**0.3382**
	Avg.	0.4674	0.4543	0.4187	0.4320	0.4123	0.4431	0.3420	0.3615	**0.2628**	**0.3142**
Spleeter	**HQ-SVC**	0.3437	0.3694	0.2777	0.2842	0.2842	0.3105	0.2432	0.3019	**0.2117**	**0.2638**
	Neuco-SVC	0.3829	0.4867	0.3563	0.4678	0.3655	0.4757	0.3037	0.3988	**0.2897**	**0.3811**
	Seed-VC	0.5695	0.5563	0.4877	0.4747	0.5075	0.4930	0.4588	0.4509	**0.3017**	**0.3770**
	KNN-SVC	0.4150	0.4920	0.3845	0.4607	0.4085	0.4701	0.3442	0.3987	**0.2258**	**0.3317**
	Avg.	0.4277	0.4761	0.3765	0.4218	0.3914	0.4373	0.3374	0.3875	**0.2572**	**0.3384**
Open-unmix	**HQ-SVC**	0.3688	0.3374	0.3223	0.3293	0.3426	0.3251	0.3060	0.3286	**0.2546**	**0.2896**
	Neuco-SVC	0.3963	0.4810	0.3858	0.4821	0.3855	0.4823	0.3770	0.4408	**0.3511**	**0.3930**
	Seed-VC	0.6091	0.5597	0.5833	0.5556	0.4878	0.5441	0.4253	0.4383	**0.3157**	**0.3535**
	KNN-SVC	0.4657	0.4924	0.4592	0.4880	0.4534	0.4874	0.3983	0.4495	**0.2602**	**0.3338**
	Avg.	0.4599	0.4676	0.4376	0.4637	0.4173	0.4597	0.3766	0.4143	**0.2954**	**0.3424**
Total		0.4517	0.4668	0.4109	0.4392	0.4070	0.4467	0.3520	0.3878	**0.2718**	**0.3317**

Table 5: Impact of Source Separation on SIS.

	VoiceGuard	AntiFake	Songbsab
No-Separation	0.3703	0.3669	0.3058
With-Separation	0.4250 (+0.0547)	0.4268 (+0.0599)	0.3699 (+0.0641)

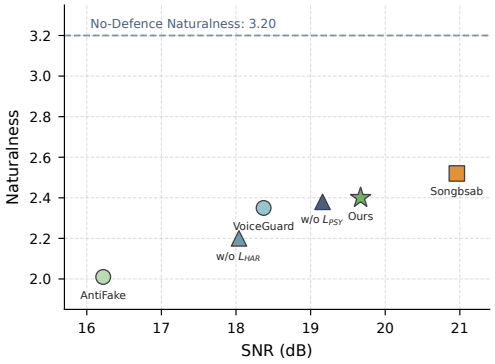


Figure 3: Audio-quality comparison and ablation.

Open-Unmix) and singing voice conversion systems consistently exhibit high identity similarity. The average C-SIS and R-SIS scores hover around 0.45, confirming that attackers can faithfully reconstruct the target singer's timbre even after source separation. This validates the feasibility and severity of the proposed attack pipeline.

Ineffectiveness of Speech-Based Defenses. VoiceGuard and AntiFake fail to significantly reduce identity similarity, yielding results comparable to the No-Defense baseline. This suggests that their adversarial perturbations are largely filtered out during the separation stage, rendering them ineffective against downstream singing voice conversion. For instance, in the Demucs + NeuCo-SVC scenario, the R-SIS remains above 0.46, indicating a failure to disrupt identity modeling. The root cause is that these methods optimize solely for speaker encoding or conversion, targeting the identity embedding space without disrupting the time-frequency structure or vocal-backing discrimination cues relied upon by separation models. Consequently, their perturbations are suppressed during the separation model's projection.

Limitations of Music-Specific Baseline. In contrast, SongBsAb leverages backing masking to achieve lower identity similarity than VoiceGuard and AntiFake, indicating that incorporating musical structure improves robustness. By interfering with the separability of vocal and backing tracks, SongBsAb allows some perturbations to persist in the separated vocal track. However, it still yields high similarity scores against certain singing voice conversion models (e.g., Seed-VC, NeuCo-SVC). This inconsistency suggests that its anti-separation effect is incidental rather than explicitly designed, limiting its stability and cross-model generalization.

Degradation of Adversarial Methods. As shown in Table 5, in the idealized 'No-Separation' scenario, VoiceGuard, AntiFake, and SongBsAb effectively suppress SIS scores to the 0.30–0.37 range. However, introducing source separation triggers a significant rebound in identity similarity, with scores increasing by 0.0547, 0.0599, and 0.0641, respectively. This degradation confirms that source separation acts as an inadvertent purification step, filtering out adversarial perturbations and restoring vocal identity traits in the separated tracks.

Superiority of ORPHEUS. ORPHEUS consistently achieves

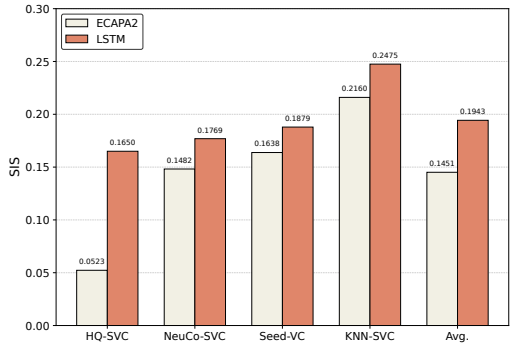


Figure 4: Additional speaker-identity evaluation of ORPHEUS using ECAPA2 and an LSTM-based speaker feature extractor.

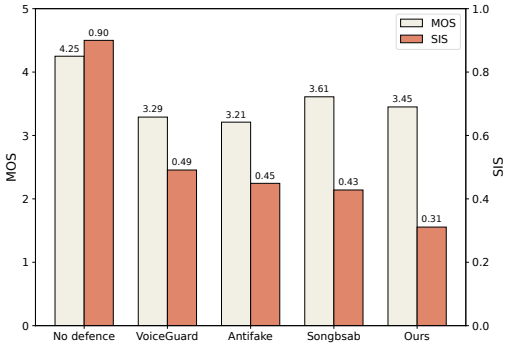


Figure 5: Human study

the lowest C-SIS and R-SIS scores across all combinations of separation and singing voice conversion models, significantly outperforming all baselines. For example, the overall average C-SIS drops from 0.4517 (No-Defense) to 0.2718, and R-SIS from 0.4668 to 0.3317. This demonstrates that ORPHEUS perturbations successfully survive diverse source separation processes to disrupt downstream identity modeling. This black-box transferability stems from two key designs: (1) Joint Optimization: The mask-misguiding and cross-track losses embed perturbations specifically at the vocal-backing discrimination boundary, preventing them from being filtered out; (2) Ensemble Strategy: The multi-encoder identity perturbation introduces offsets across diverse embedding spaces, ensuring generalization to unseen singing voice conversion architectures.

Perceptual Quality Analysis. Since ORPHEUS injects adversarial perturbations into music signals to defend against singing voice cloning, the perceptibility of these perturbations and their impact on audio quality are critical for assessing its practical applicability. As shown in Figure 3, ORPHEUS achieves an SNR of 19.67 dB and a naturalness score of 2.40, both of which are close to the best-performing baseline (SongBsAb) and substantially outperform VoiceGuard and AntiFake.

Evaluation with additional speaker-identity backbones. In the main evaluation, we use CAM++ and ResNetSE as speaker verification models because they are open-source, widely adopted, and reproducible choices for measuring speaker identity similarity. To further validate that the observed protection effect is not specific to these two evaluators, we additionally evaluate ORPHEUS using two alternative speaker-identity backbones: ECAPA2 and the LSTM-based speaker feature extractor adopted in prior singing voice conversion systems such as Grad-SVC [4] and Lora-SVC [6]. Figure 4 reports the average identity similarity of ORPHEUS under the same source separation and singing voice conversion settings. For each singing voice conversion model, we average the results over three source separation models, i.e., Demucs, Open-Unmix, and Spleeter. The results show consistently low similarity scores under both ECAPA2 and the LSTM-based feature extractor, indicating that the protection effect of ORPHEUS generalizes beyond the primary CAM++ and ResNetSE evaluators.

Although SongBsAb attains slightly higher SNR and naturalness, this mainly results from explicitly using the backing track as an additional masker and optimizing primarily for perceptual concealment. In contrast, ORPHEUS simultaneously enforces robustness to source separation and identity disruption, which introduces a modest trade-off in perceptual metrics. Nevertheless, the gap to SongBsAb is small, while ORPHEUS significantly outperforms it in black-box effectiveness and robustness under realistic attack pipelines.

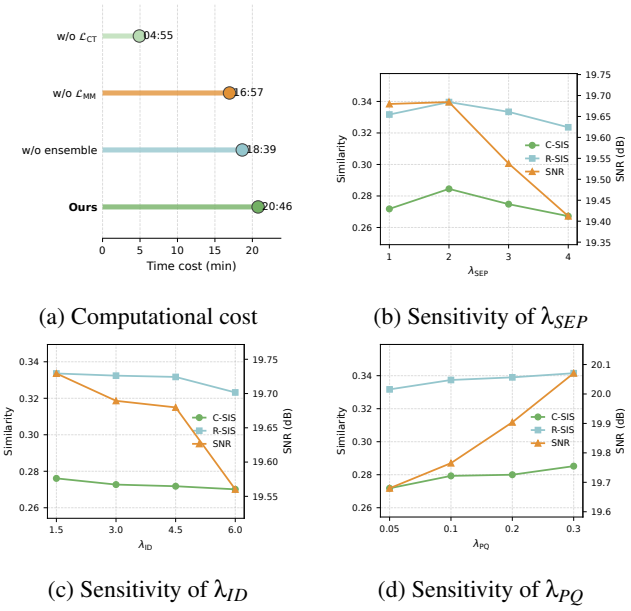
5.5 Human Study

In real-world scenarios, attackers aim to deceive human listeners using singing voice conversion systems, and the ultimate recipients of protected music are also human. Therefore, to further validate the practical effectiveness of ORPHEUS, we investigate human perceptual responses to the protected music, as well as the perceived similarity between synthesized and original singing voices.

Participants and Study Protocol. We recruited 80 participants through anonymous online questionnaires distributed via Wenjuanxing, a widely used online survey platform in China that has been adopted in prior academic studies [22]. The study was conducted in accordance with standard human-subject ethical practices. All participants provided informed consent before participation, all responses were anonymized, and no personally identifiable information was collected. Each participant received a modest compensation of 2.5 RMB upon completion. To ensure a fair comparison, each participant evaluated samples from all defense settings, including No-Defense, VoiceGuard, AntiFake, SongBsAb, and ORPHEUS, with randomized presentation order to reduce bias. All participants were proficient in English. In addition to participants from China, we also recruited non-Chinese participants, in-

Table 6: Ablation study on SIS under various optimization.

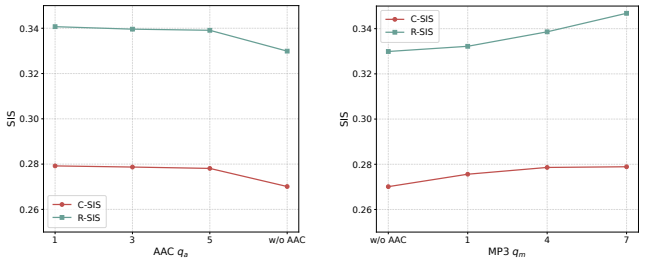
Singing voice conversion	w/o $\mathcal{L}_{CT}$		w/o $\mathcal{L}_{MM}$		w/o ensemble		ORPHEUS (ours)	
	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)	C-SIS(↓)	R-SIS(↓)
HQ-SVC	0.2833	0.2810	0.2962	0.2885	0.2877	0.2803	**0.2582**	**0.2578**
Neuco-SVC	0.3288	0.4121	0.2924	0.4009	0.3609	0.4893	**0.3111**	**0.3654**
Seed-VC	0.3084	0.3682	0.3067	0.3762	0.3023	0.3628	**0.2493**	**0.2957**
KNN-SVC	0.2501	0.3581	0.2590	0.3642	0.2514	0.3536	**0.2326**	**0.3382**
Avg.	0.2926	0.3548	0.2886	0.3565	0.3006	0.3715	**0.2628**	**0.3142**



(a) Computational cost (b) Sensitivity of λ_{SEP}

(c) Sensitivity of λ_{ID} (d) Sensitivity of λ_{PQ}

Figure 6: Ablation results on computational cost and hyperparameter sensitivity.



(a) AAC (b) MP3

Figure 7: Robustness of ORPHEUS against transformation-based attacks.

cluding participants from Italy and Singapore, to improve the diversity of the participant pool and mitigate potential language-related bias. Among the participants from China, the sample covered 10 provinces. The average completion time was approximately 650 seconds, indicating that participants devoted sufficient attention to the listening tasks and that the collected ratings are reliable.

Task 1: Perceptual Speaker Similarity Assessment. This task evaluates the perceptual resemblance between the synthesized vocals and the original singer's timbre. Participants evaluated 20 randomly selected samples and rated their similarity to the target identity (Figure 5). Key observations include: (1) In the No-Defense baseline, participants consistently perceive the singing voice conversion output as originating from the target singer, aligning with prior subjective studies. (2) ORPHEUS reduces the average perceived similarity by 65% relative to the baseline, significantly outperforming existing proactive defenses. These results confirm that ORPHEUS effectively disrupts timbre modeling, thereby preventing unauthorized voice cloning via singing voice conversion systems.

Task 2: Perceptual Quality Assessment. This task quantifies the impact of ORPHEUS on the listening experience. We conducted a Mean Opinion Score (MOS) test, asking participants to rate music samples on a scale of 1 to 5 (where 5 represents the best quality). As illustrated in Figure 5, a score of ≥ 3.0 is defined as the threshold for acceptable quality [64, 65], following prior work. The results indicate that while adversarial perturbations induce a minor degradation compared to clean audio, the quality consistently remains above the acceptability threshold. This demonstrates that ORPHEUS effectively balances robust defense with high perceptual fidelity, indicating a practical tradeoff between robust identity protection and perceptual audio quality.

5.6 Ablation Study

We conduct an ablation study to evaluate: (E1) the effectiveness of the Cross-Track Loss $\mathcal{L}_{CT}$; (E2) the effectiveness of the Mask-Misguiding Loss $\mathcal{L}_{MM}$; (E3) the impact of the encoder ensemble strategy defined in Sec. 4.2; (E4) the overall robustness of ORPHEUS when combining all components; (E5) the contribution of the harmonic alignment loss $\mathcal{L}_{HAR}$ to perceptual quality; and (E6) the contribution of the psychoacoustic loss $\mathcal{L}_{PSY}$ to perceptual quality.

Table 6 reports C-SIS and R-SIS averaged over the three source separation models (Demucs, Spleeter, and Open-

Table 7: Robustness of ORPHEUS against perturbation purification.

	C-SIS	R-SIS
SG	0.276	0.341
DF	0.288	0.366
DO	0.325	0.382

SG: Spectral Gating;
DF: Defence-FakeBob;
DO: Defence-ORPHEUS.

Table 8: Over-the-air robustness of ORPHEUS in terms of SIS.

M \ L	iPhone	HUAWEI
iPad	0.3231	0.3354
Sansui	0.3223	0.3541

L: Loudspeakers;
M: Microphones.

Unmix) for each singing voice conversion model, reflecting transferability under diverse and unseen separation conditions. For the variant "w/o ensemble", the identity similarity is computed as the average over the four encoders in the set $\mathcal{E}$. In addition, Fig. 6a reports the per-sample runtime of the variants in (E1)–(E4), and Fig. 3 reports the audio-quality comparison for the variants in (E5)–(E6).

Our findings can be summarized as follows. (R1) Removing $\mathcal{L}_{\text{CT}}$ consistently increases both C-SIS and R-SIS across all singing voice conversion models, indicating that explicitly confusing the separated vocal and backing tracks is critical for preventing residual identity cues from surviving the separation stage. (R2) Removing $\mathcal{L}_{\text{MM}}$ also leads to noticeable performance degradation, showing that perturbing the prediction masks of separation models is essential for disrupting the vocal–backing decision boundary and achieving robustness across different separation architectures. (R3) Without the encoder ensemble, identity similarity increases substantially, especially in R-SIS, demonstrating that jointly optimizing over multiple heterogeneous identity encoders is crucial for improving transferability to unseen singing voice conversion systems. (R4) Notably, even without the ensemble, ORPHEUS still maintains relatively low SIS compared to the full removal of separation-related losses. This suggests that the robustness of ORPHEUS does not solely rely on identity-space perturbation. Instead, by jointly applying $\mathcal{L}_{\text{CT}}$ and $\mathcal{L}_{\text{MM}}$, backing components are intentionally preserved in the separated vocal track, which corrupts the acoustic structure used by downstream singing voice conversion models and suppresses identity reconstruction, thereby yielding inherent cross-model generalization. (R5) Fig. 6a further reveals a clear computation–effectiveness tradeoff. The full objective requires 20:46 per sample, while removing $\mathcal{L}_{\text{CT}}$, $\mathcal{L}_{\text{MM}}$, and the encoder ensemble reduces the runtime to 04:55, 16:57, and 18:39, respectively. This additional cost comes from jointly optimizing separation disruption, identity obfuscation, and transferability, which yields the strongest robustness. Nevertheless, ORPHEUS is designed as an offline pre-release defense: the perturbation is optimized once before music publication rather than generated in real time. Therefore, the increased runtime is acceptable for the intended deployment scenario, although accelerating the optimization remains an important direction for future work. (R6) The audio-quality ablations in Fig. 3 show that both perceptual objectives are necessary. Removing $\mathcal{L}_{\text{HAR}}$ causes a pronounced degradation in both SNR and naturalness, indicating that psychoacoustic masking alone is insufficient to preserve musical quality. By contrast, removing $\mathcal{L}_{\text{PSY}}$ leads to a smaller but still visible drop relative to the full model, showing that harmonic alignment and psychoacoustic masking play complementary roles: the former improves compatibility with the musical backing, while the latter suppresses locally audible artifacts.

Hyperparameter Study. We further evaluate the sensitivity of ORPHEUS to the three top-level loss weights λ_{SEP}, λ_{ID}, and λ_{PQ}. We vary one coefficient at a time while keeping the others fixed, and report C-SIS, R-SIS, and SNR. As shown in Figs. 6b–6d, ORPHEUS remains stable across a reasonable range of weight values. Changing λ_{SEP} and λ_{ID} leads to only moderate variations in identity similarity and SNR, while increasing λ_{PQ} slightly improves SNR by assigning more weight to perceptual quality. Compared with the undefended samples, which yield much higher identity similarity in the main evaluation, ORPHEUS consistently maintains substantially lower C-SIS and R-SIS across different hyperparameter settings. Overall, these results indicate that ORPHEUS is not highly sensitive to a specific hyperparameter setting and maintains a stable tradeoff among separation robustness, identity protection, and audio quality.

5.7 Robustness of ORPHEUS

To rigorously assess the resilience of ORPHEUS against adaptive adversaries, we evaluate three distinct attack vectors: transformation-based attacks, perturbation purification, and over-the-air recording.

Transformation-based Attacks. Real-world adversaries may employ lossy audio compression as a preprocessing step to sanitize adversarial perturbations and restore input quality. To assess the resilience of ORPHEUS against such transformations, we evaluate its robustness under two ubiquitous codecs: AAC compression (AAC) and MP3 compression (MP3) [16]. These standards utilize different compression algorithms controlled by quality parameters. In our experiments, we vary the AAC quality parameter $q_a \in \{1,3,5\}$ [14] and the MP3 quality parameter $q_m \in \{1,4,7\}$ to simulate varying levels of compression artifacts.

Figure 7 shows the robustness of ORPHEUS under AAC and MP3 compression with different quality settings. Overall, both C-SIS and R-SIS remain stable across a wide range of compression parameters. Under AAC compression, C-SIS only varies from 0.2792 to 0.2781 for quality parameters $q_a \in 1,3,5$ and decreases slightly to 0.2718 without compression, while R-SIS changes marginally from 0.3407 to 0.3391 and 0.3317, respectively. Similarly, for MP3 compression, C-SIS increases moderately from 0.2718 (no compression) to

0.2789 at the highest quality setting, and R-SIS rises from 0.3317 to 0.3468. These variations are minor compared to the similarity scores of undefended samples, indicating that the identity suppression capability of ORPHEUS is largely preserved after lossy encoding. Although higher-quality compression retains more acoustic details and slightly increases identity similarity, the perturbations remain effective against both codecs. Overall, the results demonstrate that ORPHEUS is resilient to transformation-based attacks and maintains strong identity protection under realistic audio compression pipelines.

Perturbation purification. Adversaries may attempt to sanitize protected audio prior to singing voice conversion by applying denoising or purification methods. We therefore evaluate ORPHEUS against three purification settings. The first is Spectral Gating (SG) [52], which serves as a generic denoising baseline. The second is Defence-FakeBob (DF) [20], an open-source adversarial purification method. The third is DO (Defence-ORPHEUS), a defense-aware adaptive purification method derived from the optimization principle of DF. Specifically, DO redefines the purification objective to maximize the similarity between cloned outputs produced before and after purification, as measured by C-SIS and R-SIS, thereby explicitly undermining the protection of ORPHEUS and recovering speaker identity more effectively.

Table 7 reports the results. SG yields C-SIS and R-SIS values of 0.276 and 0.341, indicating that generic denoising alone is insufficient to neutralize the protection effect. DF raises the similarity scores to 0.288 and 0.366. The defense-aware variant DO further increases them to 0.325 and 0.382, demonstrating that a tailored adaptive purification strategy can more effectively erode the defense than the original DF baseline. Nevertheless, even under DO, the recovered similarity remains substantially lower than that of the unprotected baseline in Table 4. These results suggest that the perturbations introduced by ORPHEUS are not merely additive noise amenable to filtering, but are structurally entangled with identity-relevant components of the signal. Consequently, purification can only partially recover speaker identity, and ORPHEUS retains substantial protection even under this stronger adaptive attack.

Robustness against Over-the-Air Recording. Real-world adversaries may capture protected audio via microphones, introducing physical channel distortions [15] such as reverberation, frequency response variation, and ambient noise, which can potentially attenuate or distort adversarial perturbations. To evaluate the resilience of ORPHEUS under such analog-to-digital transformations, we simulate an over-the-air re-recording attack scenario [61].

Specifically, we employ two playback devices, an iPad (2021) and a Sansui loudspeaker, together with two recording terminals, an iPhone 15 Pro Max (iOS) and a HUAWEI P70 Pro (HarmonyOS), to transmit the protected music through the air. For each playback–recording pair, 100 randomly selected music tracks are replayed and re-captured in a quiet indoor environment with a fixed distance of 0.5 meters between the loudspeaker and the microphone.

Table 8 reports the average Speaker Identity Similarity, computed by averaging C-SIS and R-SIS, under different playback and recording device combinations. The consistently low SIS values across all device pairs indicate that the adversarial perturbations introduced by ORPHEUS largely survive the physical channel and remain effective after re-recording, demonstrating strong robustness against over-the-air attacks.

Notably, over-the-air transmission inevitably introduces hardware-induced distortions, environmental noise, and room reverberation, which can be regarded as a class of signal transformations [18]. Therefore, the robustness of ORPHEUS under re-recording shares the same underlying rationale as its robustness against transformation-based attacks: the perturbations are optimized to be resilient to such distortions and thus remain effective after these physical-domain transformations.

6 Discussion

In this section, we discuss the limitations of ORPHEUS under realistic deployment conditions and outline several important directions for future work.

Beyond Singing Voice Conversion. This work focuses on defending against singing voice conversion, where the attacker converts an existing vocal track into the timbre of a target singer. Another important class of automatic song cover techniques is singing voice synthesis (SVS) [36], which generates singing voices directly from musical scores and lyrics conditioned on a target singer's voice. While the proposed identity-disrupting perturbations could potentially be adapted to interfere with the acoustic rendering stage of SVS, protecting against SVS may additionally require disrupting lyric and melody representations at the symbolic or phoneme level. Exploring adversarial protection mechanisms that jointly operate in both acoustic and symbolic domains to counter SVS-based misuse is an important direction for future work.

Real-World Usage of ORPHEUS. The effectiveness of ORPHEUS relies on the assumption that music owners apply the protection before releasing their content. In practice, however, it is difficult to guarantee that all publicly accessible recordings of a singer are consistently protected. Unprotected legacy recordings, live performance captures, or third-party uploads may still be available online and could be exploited by adversaries to obtain clean vocal samples, potentially weakening the overall protection. Moreover, the long-term effectiveness of ORPHEUS may be affected by future advances in audio processing and generative modeling. More powerful source separation or purification techniques, as well as end-to-end singing voice conversion systems with stronger robustness, could partially suppress or bypass the injected perturbations. This highlights an inherent "future-proofing" challenge shared by all adversarial defenses and underscores the need for continuous updates and co-evolution of protection mechanisms

alongside advances in generative audio technologies.

Efficiency and Practical Deployment. The current implementation of ORPHEUS requires approximately 1,000 optimization iterations per audio sample, requiring roughly 20 minutes to process a single track on an NVIDIA A800 PCIe GPU. While this computational cost precludes real-time streaming applications, it remains acceptable for our target scenario: offline music publishing, where tracks undergo one-time protection prior to release. Future work will explore acceleration strategies, such as training feed-forward perturbation generators, leveraging model distillation, or precomputing universal perturbation templates to enable low-latency or real-time protection.

7 Conclusion

In this paper, we present ORPHEUS, the first proactive defense framework explicitly designed for realistic singing voice conversion pipelines involving source separation. Our analysis reveals that prior defenses, which are optimized solely for clean vocals, are rendered ineffective by the separation process. To address this, ORPHEUS jointly disrupts separation mechanisms and identity modeling. By leveraging prediction-matrix and cross-track confusion, it embeds perturbations directly into vocal-backing decision boundaries to ensure persistence post-separation. Additionally, our ensemble-based strategy and psychoacoustic constraints guarantee robust transferability to unseen singing voice conversion systems while maintaining perceptual fidelity. Extensive evaluations confirm that ORPHEUS significantly outperforms state-of-the-art defenses in both robustness and imperceptibility. These findings underscore the critical necessity of domain-aware protection strategies for securing content against real-world singing voice conversion.

Acknowledgements

We sincerely thank the anonymous reviewers and the shepherd for their constructive feedback on our work. This research was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123603), National Natural Science Foundation of China NSFC (No.62472325), and Hubei Provincial Key R&D Program, China (Grant No. 2025BAB021).

Ethical Considerations

This work aims to mitigate the misuse of singing voice conversion systems for unauthorized voice cloning and copyright infringement by providing a proactive, imperceptible protection mechanism for music creators. Nevertheless, research on adversarial perturbations and generative models inherently involves dual-use risks and requires careful ethical reflection.

Human Subjects and Listening Studies. Our evaluation includes subjective listening tests to assess perceptual quality and identity similarity. All participants were adults who provided informed consent prior to participation. The study did not collect personally identifiable information, and all responses were anonymized. The audio materials were drawn from publicly available datasets and synthesized samples; participants were explicitly informed that the content was artificial and used solely for research purposes. No deceptive or harmful content was involved, and the study protocol followed standard ethical guidelines for human-subject audio perception experiments.

Impact on Legitimate Applications. Singing voice conversion and related technologies also have beneficial uses, including assistive technologies, creative production with proper authorization, and artistic experimentation. ORPHEUS is designed to be applied by content owners prior to public release and does not interfere with authorized singing voice conversion scenarios, where unprotected or watermarked originals can be provided to legitimate users. Thus, the method is intended to complement, rather than hinder, lawful and ethical applications of voice conversion.

Societal Benefit. By explicitly modeling realistic attack pipelines that include source separation and by demonstrating robustness against over-the-air recording, our work highlights practical vulnerabilities and provides a technically grounded defense. We believe that advancing such proactive protection mechanisms contributes to safeguarding artists' rights, reducing the impact of deepfake music, and supporting the responsible deployment of generative audio technologies.

Open Science

We commit to the open science principles of the community and provide a permanent, citable link to the full codebase, ensuring long-term accessibility and reproducibility of our results. To facilitate verification and further study, we release the core artifacts of this work. Specifically, we provide a public repository at https://zenodo.org/records/20407256, which contains the complete implementation of ORPHEUS.

References

[1] Reaper: Robust epoch and pitch estimator. https://github.com/google/REAPER, 2015.

[2] Speaker verification using resnetse and ecapa-tdnn. https://github.com/ranchlai/speaker-verification, 2021.

[3] Psychoacoustic loss functions. https://github.com/TUIlmenauAMS/PsychoacousticLoss, 2022.

[4] Grad-svc based on grad-tts from huawei noah's ark lab. https://github.com/PlayVoice/Grad-SVC/tree/20230920-V2-96, 2023.

[5] No ai fraud act. https://www.congress.gov/bill/118th-congress/house-bill/6943/text/ih, 2023.

[6] Singing voice conversion based on whisper neural source-filter bigvgan. https://github.com/PlayVoice/lora-svc, 2023.

[7] 200+ artists urge tech platforms: Stop devaluing music. https://artistrightsnow.medium.com/200-artists-urge-tech-platforms-stop-devaluing-music-559fb109bbac, 2024.

[8] Ai artist xania monet climbs the charts — and signs a multimillion-dollar record deal. https://www.billboard.com/pro/ai-music-artist-xania-monet-multimillion-dollar-record-deal/, 2025.

[9] China ai grand ceremony 2025. https://tv.cctv.com/2025/08/14/VIDAuO5Po88MxjFxndbbFZmC250814.shtml, 2025.

[10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[11] Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 818–824. IEEE, 2024.

[12] Bingsong Bai, Yizhong Geng, Fengping Wang, Cong Wang, Puyuan Guo, Yingming Gao, and Ya Li. Hq-svc: Towards high-quality zero-shot singing voice conversion in low-resource scenarios. *arXiv preprint arXiv:2511.08496*, 2025.

[13] Gongyu Chen, Xiaoyu Zhang, Zhenqiang Weng, Junjie Zheng, Da Shen, Chaofan Ding, Wei-Qiang Zhang, and Zihao Chen. Yingmusic-svc: Real-world robust zero-shot singing voice conversion with flow-grpo and singing-specific inductive biases. *arXiv preprint arXiv:2512.04793*, 2025.

[14] Guangke Chen and Yedi Zhang. Songbsab: A dual prevention approach against singing voice conversion based illegal song covers. In *32nd Annual Network and Distributed System Security Symposium*, 2025.

[15] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, and Yang Liu. As2t: Arbitrary source-to-target adversarial attack on speaker recognition systems. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[16] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui Wang. Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, 20(5):3970–3987, 2022.

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

[18] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[19] Wei Chen, Binzhu Sha, Jing Yang, Zhuo Wang, Fan Fan, and Zhiyong Wu. Singing voice conversion with accompaniment using self-supervised representation-based melody features. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[20] Zesheng Chen, Li-Chi Chang, Chao Chen, Guoping Wang, and Zhuming Bi. Defending against fakebob adversarial attacks in speaker verification systems with noise-adding. *Algorithms*, 15(8):293, 2022.

[21] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

[22] Alessandro Del Ponte, Lianjun Li, Lina Ang, Noah Lim, and Wei Jie Seow. Evaluating sojump. com as a tool for online behavioral research in china. *Journal of Behavioral and Experimental Finance*, 41:100905, 2024.

[23] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research.

[26] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 552–559. IEEE, 2021.

[27] Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. The singing voice conversion challenge 2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[28] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, et al. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. *arXiv preprint arXiv:2307.07218*, 2023.

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

[30] Yuexuan Kong, Gabriel Meseguer-Brocal, Vincent Lostanlen, Mathieu Lagrange, and Romain Hennequin. S-key: Self-supervised learning of major and minor keys from audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[31] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[32] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

[33] Jingyang Li, Dengpan Ye, Long Tang, Chuanxi Chen, and Shengshan Hu. Voice guard: Protecting voice privacy with strong and imperceptible adversarial perturbation in the time domain. In *IJCAI*, pages 4812–4820, 2023.

[34] Kai Li, Guo Chen, Wendi Sang, Yi Luo, Zhuo Chen, Shuai Wang, Shulin He, Zhong-Qiu Wang, Andong Li, Zhiyong Wu, et al. Advances in speech separation: Techniques, challenges, and future trends. *arXiv preprint arXiv:2508.10830*, 2025.

[35] Xu Li, Shansong Liu, and Ying Shan. A hierarchical speaker representation framework for one-shot singing voice conversion. *arXiv preprint arXiv:2206.13762*, 2022.

[36] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028, 2022.

[37] Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

[38] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[39] Zihao Liu, Yan Zhang, and Chenglin Miao. Protecting your voice from speech synthesis attacks. In *Proceedings of the 39th Annual Computer Security Applications Conference*, pages 394–408, 2023.

[40] Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave. The 2016 signal separation evaluation campaign. In Petr Tichavský, Massoud Babaie-Zadeh, Olivier J.J. Michel, and Nadège Thirion-Moreau, editors, *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pages 323–332, Cham, 2017. Springer International Publishing.

[41] Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023.

[42] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 659–663. IEEE, 2014.

[43] Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer, 2015.

[44] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

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

[46] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pages 8599–8608. PMLR, 2021.

[47] Gege Qi, Yuefeng Chen, Yao Zhu, Binyuan Hui, Xiaodan Li, Xiaofeng Mao, Rong Zhang, and Hui Xue. Transaudio: Towards the transferable adversarial audio attack via learning contextualized perturbations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[48] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[49] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18-hq - an uncompressed version of musdb18, https://doi.org/10.5281/zenodo.3338373, August 2019.

[50] Lekshmi Chandrika Reghunath and Rajeev Rajan. Predominant audio source separation in polyphonic music. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):49, 2023.

[51] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

[52] Tim Sainburg and Timothy Q Gentner. Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, 15:811737, 2021.

[53] Binzhu Sha, Xu Li, Zhiyong Wu, Ying Shan, and Helen Meng. Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12577–12581. IEEE, 2024.

[54] Keren Shao, Ke Chen, Matthew Baas, and Shlomo Dubnov. knn-svc: Robust zero-shot singing voice conversion with additive synthesis and concatenation smoothness

optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[55] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305. Springer, 2018.

[56] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667, 2019.

[57] Naoya Takahashi, Shota Inoue, and Yuki Mitsufuji. Adversarial attacks on audio source separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 521–525. IEEE, 2021.

[58] Benjamin Van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6562–6566. IEEE, 2022.

[59] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[60] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023.

[61] Zhaolin Wei, Xiuwen Shi, Dengpan Ye, Yuhan Lin, Zhigang Wang, JiaCheng Deng, Ziyi Liu, and Long Tang. Phonofence: A cross-task defense framework for deepfake via phoneme-level adversarial perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7691–7699, 2025.

[62] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm).*, 2019.

[63] Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

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

[65] Zhisheng Zhang, Derui Wang, Qianyi Yang, Pengyang Huang, Junhan Pu, Yuxin Cao, Kai Ye, Jie Hao, and Yixian Yang. Safespeech: Robust and universal voice protection against malicious speech synthesis. In *34th USENIX Security Symposium (USENIX Security 25)*, Seattle, WA, USA, 2025.