SongBsAb: A Dual Prevention Approach against Singing Voice Conversion based Illegal Song Covers

Guangke Chen*, Yedi Zhang†, Fu Song(✉)‡,§, Ting Wang¶, Xiaoning Du‖ and Yang Liu**
*Pengcheng Laboratory †National University of Singapore ‡Key Laboratory of System Software (Chinese Academy of Sciences) and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
§ Nanjing Institute of Software Technology ¶Stony Brook University ‖Monash University **Nanyang Technological University
chengk@pcl.ac.cn yd.zhang@nus.edu.sg songfu@ios.ac.cn
inbox.ting@gmail.com xiaoning.du@monash.edu yangliu@ntu.edu.sg

***Abstract*—Singing voice conversion (SVC) automates song covers by converting a source singing voice from a source singer into a new singing voice with the same lyrics and melody as the source, but sounds like being covered by the target singer of some given target singing voices. However, it raises serious concerns about copyright and civil right infringements. We propose SongBsAb[1], the first proactive approach to tackle SVC-based illegal song covers. SongBsAb adds perturbations to singing voices before releasing them, so that when they are used, the process of SVC will be interfered, leading to *unexpected* singing voices. Perturbations are carefully crafted to (1) provide a dual prevention, i.e., preventing the singing voice from being used as the source and target singing voice in SVC, by proposing a gender-transformation loss and a high/low hierarchy multi-target loss, respectively; and (2) be harmless, i.e., no side-effect on the enjoyment of protected songs, by refining a psychoacoustic model-based loss with the backing track as an additional masker, a unique accompanying element for singing voices compared to ordinary speech voices. We also adopt a frame-level interaction reduction-based loss and encoder ensemble to enhance the transferability of SongBsAb to unknown SVC models. We demonstrate the prevention effectiveness, harmlessness, and robustness of SongBsAb on five diverse and promising SVC models, using both English and Chinese datasets, and both objective and human study-based subjective metrics. Our work fosters an emerging research direction for mitigating illegal automated song covers.**

I. Introduction

The advent of generative AI has revolutionized the realm of AI-generated art, including AI-generated song covers based on singing voice conversion (SVC) [3]. Unlike human-based song covers, SVC empowers individuals without exceptional singing and vocal imitation abilities to create song covers. Consequently, the internet has seen a surge in SVC-covered singing voices and songs. One of the most notable examples is "AI Sun Yanzi", a virtual singer that imitates the singing voice of the famous Mandopop female singer Stefanie Sun (Chinese name Yanzi Sun) and has covered over 1,000 songs from other singers, far more than the total number of songs by Stefanie in her past 23-year career. The most popular cover has garnered millions of views and thousands of shares on Bilibili, China's largest user-generated video streaming site [4], [5]. Another cover is the song "Heart on My Sleeve", which imitates the singing voices of the singers Drake and The Weeknd. It has garnered over 15 million views on TikTok in just two days, and was submitted for a Grammy Award consideration [6].

However, it raises serious concerns about copyright and civil right infringements [4], [7], [8] (cf. § II-B for details), because a song is an intellectual property composed of key elements such as lyrics, melody, and the singer's rendition. Recently, the "Elvis Act" was signed into state law for the first time to protect against exploitative use of generative AI [9], and an open letter issued by the Artist Rights Alliance and signed by more than 200 artists (e.g., Billie Eilish, Katy Perry) calls for responsible AI music practices [10].

Thus, it is increasingly crucial for the music industry and society at large to safeguard the interests and rights of song owners and singers facing potential infringements whenever songs are used as source or target singing voices in SVC. One may detect SVC-covered singing voices after infringements have already been committed, but, this passive solution becomes inefficient and cumbersome with the surge of SVC-covered singing voices and songs due to its low entry barriers. In this work, we propose SongBsAb, the first prevention approach, to effectively tackle SVC-based illegal song covers. SongBsAb is a proactive, dual prevention solution that can fundamentally prevent infringements from happening by adding a subtle perturbation to a singing voice. The song owners (defenders) can employ SongBsAb on singing voices prior to their release. When protected singing voices are used, the process of SVC will be interfered, producing unexpected singing voices to the SVC users. The design of SongBsAb faces and solves the following technical challenges, especially compared to (ordinary) speech voices and their conversion.

Challenge-1: More Involved Rights to Protect. The protec-

[1]BsAb stems from "**Bs**pecific **A**nti**b**ody" which has two different antibodies (SongBsAb's prevention of a song being used as both the target and source songs by identity and lyric disruptions), respectively neutralizing two different types of antigen (SVC's infringement of civil rights and copyrights). Code and audio are available at [1] and the full version of paper refers to [2].

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.240747
www.ndss-symposium.org

tion requirements for songs are complex and distinct, involving more intricate rights than those of speech voices or images. Indeed, songs can be used as either source or target singing voices in SVC, unknown to song owners in advance, thus requiring protection of various rights (e.g., singer's civil rights, and copyrights of lyrics and melodies; cf. § II-B). In contrast, there are no copyright issues regarding melody or textual content for speech voices and images. Therefore, prior works on speech voices [11], [12], [13], [14], [15], which only protect speaker identity, or on images [16], [17], [18], which only protect the identity in faces or artists' painting styles, cannot be ported for the protection of songs. To tackle this challenge, SongBsAb is designed to provide a dual prevention by adding subtle perturbations to singing voices to prevent them from being used as source/target singing voices in SVC. By doing so, SVC-covered singing voices (songs) neither preserve the original lyrics (lyric disruption) nor imitate the singer (identity disruption), thus directly protecting the copyrights of lyrics and the civil rights of the singer. The copyrights of melodies and copyrights to reproduce and distribute songs are indirectly protected as SVC users are discouraged to release unexpected SVC-covered singing voices and gradually abandon SVC due to its weird behavior. We remark that SongBsAb is also effective in disrupting lyrics (resp. identity) only. Inspired by adversarial attacks [19], SongBsAb formulates the perturbations searching as an optimization problem with novel designated loss functions, including a gender-transformation loss and a high/low hierarchy multi-target loss to maximize identity disruption and lyric disruption, respectively.

Challenge-2: Higher Quality Requirements. In contrast to speech voices that are primarily used for conversation, songs are music arts for appreciation and entertainment, and are highly expected to meet high-quality standards [20], [21], [22]. Thus, the prevention should be harmless for the song (including melody, lyrics, and singing style). To tackle this challenge, we harness the simultaneous masking [23] which entails that a faint yet audible sound (the maskee) becomes inaudible when another louder audible sound (the masker) is concurrently occurring [24]. In the real world, a singing voice is typically accompanied with a backing track in the song. We treat both a singing voice and its backing track as maskers and a perturbation as maskee, and use a loss to control the magnitude of the perturbation. It refines the prior simultaneous masking-based loss [25], [26] that only uses the speech voice as the masker, thus significantly improving the harmlessness, as the perturbation will be inaudible as long as it is weaker than any of two maskers.

Challenge-3: More Challenging for Transferability. In practice, SVC models may use distinct encoders from SongBsAb. Hence, the prevention should generalize and transfer to unknown SVC models. Adversarial voices inherently exhibit low transferability [27], and due to the dual prevention, SongBsAb involves more possibly distinct encoders than prior works in the speech domain [11], [12], [13]. Therefore, it is more challenging for SongBsAb to achieve high transferability of both identity and lyric disruptions. To tackle this challenge, we adopt a frame-level interaction reduction-based (FL-IR) loss [28] and encoder ensemble [29], [30], [11]. They improve the transferability from two different perspectives, thus are complementary, i.e., their combination further boost the transferability, making SongBsAb more practical and useful.

We conduct an extensive evaluation to demonstrate the efficacy of SongBsAb. We first evaluate the prevention effectiveness on 5 diverse and promising SVC models using both English and Chinese datasets via 5 objective metrics. SongBsAb can reduce the identity similarity between SVC-covered singing voices and the target singer and enlarge the lyric word error rate, together reducing the singing voice conversion success rate by over 97%. It significantly outperforms two recent promising methods [12], [11] that were designed for preventing ordinary speech voice conversion in terms of both prevention effectiveness and harmlessness. The subjective human study with 3 tasks also confirms the prevention effectiveness and utility of SongBsAb on the enjoyment of protected songs.

We then evaluate transferability on 8 distinct identity encoders and 5 distinct lyric encoders. SongBsAb shows a strong ability to transfer to unknown SVC models, while also surpassing previous works [12], [11].

We finally demonstrate the robustness of SongBsAb in over-the-air scenario and against adaptive SVC users who completely know and aim to bypass SongBsAb by pre-processing protected singing voices via existing voice transformations and tailored optimizations, or by fine-tuning SVC models.

In summary, the main contribution of this work includes:

- We present SongBsAb, the first proactive solution to prevent right infringements caused by SVC-based illegal song covers. It features a dual prevention, capable of causing both the identity disruption and lyric disruption in SVC-covered singing voices, for which we devise a gender-transformation loss and a high/low hierarchy multi-target loss, respectively.
- We propose to utilize backing tracks, a unique accompanying element with singing voices in songs compared to speech voices, as maskers to further improve harmlessness. Our simultaneous masking-based loss effectively enhances the quality of protected songs and thus the utility of SongBsAb.
- While SongBsAb exhibits transferability, we further utilize FL-IR loss and encoder ensemble to enhance transferability for causing both the identity disruption and lyric disruption on unknown SVC models in a complementary way.
- Our work makes the first significant step towards coping with illegal automated song covers. We release our code and audio samples, and discuss possible future works to foster exploration in this emerging research direction.

For convenience, we summarize the abbreviations in TABLE I.

II. Background & Related Work

A. Singing Voice Conversion (SVC)

A song consists of a singing voice and a backing track, stored in separate channels. Singing voice conversion transforms a song's vocal rendition from one singer to another's

style and timbre while preserving the original lyrics and melody [3]. The backing track is removed during conversion and is not part of the process. Mainstream SVC systems use an encoder-decoder architecture [3], as shown in Fig. 1. There are three common encoders: the identity encoder extracts the identity feature from a few target singing voices representing the target singer's singing style and voiceprint, while the pitch encoder and lyric encoder extract pitch and lyric features from the source singing voice of the source singer, characterizing the melody and lyrics. The decoder then fuses these features to produce a singing voice that resembles the target singer covering the source singing voice.

Explicit & Implicit SVC. Since both target and source singing voices contain identity and lyric information, SVC relies on information disentanglement, achieved through either explicit or implicit methods [3]. Explicit methods use pre-trained encoders with disentanglement capabilities and the decoder is then trained with these frozen encoders. Lyric encoding is typically handled by speaker-independent models such as Whisper [31] or Hubert [32], while identity encoding uses content-independent models such as GE2E [33]. In contrast, implicit methods adopt encoders that originally lacked of disentanglement capabilities and employ specialized strategies for disentanglement. For example, NeuCoSVC [34] uses the same encoder for both voices and a KNN-based matching module to retain the source lyrics while shifting identity to target singers. StarGANv2 [35] employs adversarial training to supervise the decoder to capture only the target's identity and source's lyrics. Both methods use signal-processing based (e.g., WORLD [36]) or neural networks-based (e.g., Crepe [37]) pitch encoders, and use generative models such as GANs [38] or diffusion models [39] as decoders due to their strong generative capacity.

Few-shot & non-few-shot. Few-shot trains the encoders and decoder without the target singers used during inference. In contrast, non-few-shot predefines target singers during training, and to align with an unseen target singer in inference, models must be trained from scratch or fine-tuned, requiring more computational resources and a large number of singing voices from the target singer to avoid overfitting [40], [41], [34]. For both, it is common to use a few samples from the target singer during inference and feed the aggregated identity features to the decoder to make outputs sound more like the target singer.

Key differences between SVC and ordinary voice conversion include (1) challenging task [3]: singing voices differ fundamentally [42], [43], [44] and vary more in phoneme duration, pitch, expression, singing style, and speaker characteristics [45], [46], [47], [48], requiring more proper information disentanglement [3]; (2) severe rights infringements (cf. § I): singing voices involve more complex copyright concerns [46]; (3) architecture: SVC uses specialized pitch encoders; and (4) inputs: SVC source voices should be professionally sung [46].

TABLE I: Main Abbreviations.

Abbr.	Full Form	Meaning
SVC	singing voice conversion	N/A
$\mathcal{I}$	target singing voice	input of SVC providing identity information
$\mathcal{L}$	source singing voice	input of SVC providing lyric and melody
$\tilde{\mathcal{I}}$	protected target singing voice	protected version of $\mathcal{I}$
$\tilde{\mathcal{L}}$	protected source singing voice	protected version of $\mathcal{L}$
-	target singer	the singer of $\mathcal{I}$
-	source singer	the singer of $\mathcal{L}$
$y/\tilde{y}$	undefended/defended output singing voice	output of SVC without/with SongBsAb
-	destination singer	the singer of $\tilde{y}$
FL-IR loss	frame-level interaction reduction-based loss	a loss for enhancing the transferability of SongBsAb

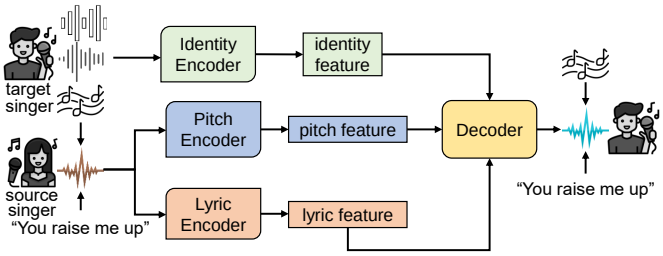


Fig. 1: Mainstream Singing Voice Conversion Systems.

B. The Rights Infringed by SVC

A song is an intellectual property created by multiple contributors, including the lyricist, composer, singer, and record company. The lyricist and composer write the lyrics and melody, typically transferring their copyrights to the record company while retaining authorship and sharing royalties. The record company hires singers to perform the song and becomes the owner by acquiring performance rights through copyright transfer. In rare cases, such as with online singers, the lyricist, composer, and singer are the same individual. SVC may harm the following rights and interests of song owners and singers when their songs are used as source or target singing voices.

Copyrights to reproduce and distribute songs. Copyright laws protect song owners' exclusive rights to reproduce and distribute their songs, as outlined in Article 9 of China's copyright law [49], §106 of Title 17 of the U.S. Code [50], and Section 6 of U.K. copyright law [51]. During singing voice conversion, both source and target singing voices are transferred from their original publication platforms (e.g., music platforms) to a computational platform performing SVC. Without copyright licenses, this process infringes on the rights of song owners for both the source and target singing voices.

Copyrights to perform and display lyrics and melodies. Copyright laws protect song owners' exclusive rights to display and perform their melodies and lyrics [49], [50], [51]. They must be informed and compensated for any usage of their works. Many singers have faced charges for covering songs without permission, and SVC-based automated covers similarly threaten the rights of song owners regarding the lyrics

and melodies of the source singing voices. Additionally, song owners often seek to ensure exclusive performances by singers to maintain their reputation and profits, a goal undermined by SVC. Finally, releasing SVC-covered songs without crediting the lyricists or composers violates their authorship rights.

Civil rights of singers over voices and reputation. First, singers have civil rights over their voices (e.g., Article 1023 of the Civil Code of China [52] and ELVIS Act of Tennessee U.S.A [9]), akin to rights over their likeness, which prohibits the production, use, or publication of their voices without permission. SVC violates this regulation by producing singing voices that sound like the target singer. Second, malicious SVC users may exploit sensitive source singing voices, such as those with political bias, discrimination, or violence, leading to reputational damage and infringement under civil codes [52] or defamation/privacy laws [53], [54], [55]. Third, exceptional vocal skills and unique performance styles are vital for singers' livelihood and careers. SVC-enabled AI singers, with lower entry barriers and costs, could replace traditional singers, potentially breaching unfair competition laws [56], [57]. Finally, within contracts, a singer's voice and public image are tools for the song owner's profit, so imitating a singer's voice or degrading his/her reputation could harm the owner's revenue.

C. Adversarial Examples

Adversarial examples for good. Adversarial examples are deliberately crafted inputs to deceive models and have been widely studied [19], [58], [59], [25], [26], [60], [61], [30], [62], [63], [64]. They also have been utilized for beneficial applications (cf. TABLE IX in Appendix A for a summary).

Error-minimizing noise is applied to personal data so that models trained on them are tricked into believing there is "nothing" to learn [65]. Such noises are improved later to make them robust against adversarial training [66]. Glaze [17] and MIST [16] added perturbations to artists' artworks such that text-to-image models fine-tuned on these artworks fail to mimic the painting styles of the protected artists. UnGANable [18] perturbed face images of a target user so that the face images reconstructed from the face manipulator do not contain the user's identity. V-cloak [14] and VoiceCloak [15] added perturbations to human voices to hide speakers' identities from speaker recognition models, thus achieving voice anonymity. Glaze, MIST, and UnGANable target AI-generated images, V-cloak and VoiceCloak target human-generated speech voices, while our work targets AI-generated singing voices.

AttackVC [12], VSMask [13], and AntiFake [11] are the closest works to ours, all of which target the voice modality and generative models. They add perturbations to ordinary speech voices of a target speaker to make speech voice conversion or synthesis to generate voices not recognized as the target speaker by both speaker recognition models and human perception. Our work focuses on singing voice conversion, a more challenging task than speech voice conversion [3]. SongBsAb differs from them in the following aspects: (1) They only affected the identity of crafted voices and thus cannot prevent singing voices from being used as source singing voices in SVC, while SongBsAb prevents the singing voice from being used as the source or target singing voice (dual prevention). This enables broader applications, protecting not only the civil rights of voices and performing rights of singers but also the copyright of lyrics. Even for protecting singers only, experiments show that SongBsAb significantly outperforms the publicly available AttackVC and AntiFake (cf. § V-B). (2) To decide the destination speaker for better identity disruption, AttackVC and VSMask randomly selected an opposite-gender speaker, and AntiFake utilized the Analytic Hierarchy Process (AHP) to balance computational embedding deviation and human judgment. They both represent a singer with a single voice embedding. Instead, we propose a gender-transformation loss to optimize towards a destination speaker that has the least objective identity similarity with the target singer among a pool of opposite-gender singers and use the centroid of multiple voice embeddings to represent a singer, resulting in the best identity disruption (cf. Appendix H of [2]). (3) To improve harmlessness, AttackVC and VSMask enforced an L_∞ norm-based constraint that may not correlate with human auditory perception [27]. AntiFake used different gain functions for the perturbation strength in different frequency bands and maximized the signal-to-noise ratio. Instead, we utilize the psychoacoustics model [24] to hide perturbations under the auditory perception threshold of humans. Notably, motivated by the fact that a singing voice is commonly accompanied by a backing track in a song, we propose using backing tracks as additional markers, which improves perturbations hiding capacities (cf. TABLE IV). Backing tracks are unique elements of singing voices and have never been explored in the literature to strengthen harmlessness. (4) AttackVC and VSMask evaluated transferability on unknown models, while AntiFake enhanced transferability via encoder ensemble. Besides the encoder ensemble, we propose a frame-level interaction reduction-based (FL-IR) loss to enhance transferability further. The rationales behind the two methods are different (cf. § IV-E), so their combination yields the best results (cf. § V-C), and SongBsAb exhibits superior transferability.

Interaction vs. transferability. Adversarial examples crafted on one surrogate model often can transfer to other target models. However, the transferability may be limited especially when there is a large gap between the surrogate and target models [60], [30]. Wang et al. [28] interprets the transferability from the perspective of interaction I inside perturbations. The interaction between two perturbation units i and j, denoted by I_{ij}, is the change of the importance of the unit i after perturbing unit j. The average interaction over all pairs of perturbation units is defined as:

$$\frac{\mathbb{E}_i(v(\Omega) + v(\emptyset) - v(\Omega \setminus \{i\}) - v(\{i\}))}{n-1}$$

where v is a utility function measuring the importance of perturbation units for deceiving models, n is the number of perturbation units, and Ω, $\emptyset$, $\Omega \setminus \{i\}$, and $\{i\}$ denote the cases of all units being perturbed, no unit being perturbed (i.e., normal example), all units excluding the unit i being

perturbed, and only the unit i being perturbed, respectively. It was shown that interaction is negatively correlated with transferability [28]: a large interaction indicates that the perturbation units need to work closely to jointly fool the surrogate model, thus leading to low transferability, as a large interaction is more likely to be broken on target models.

D. Simultaneous Masking

Simultaneous masking refers to the phenomenon that one faint but audible sound (the maskee) becomes inaudible in the presence of another simultaneously occurring louder audible sound (the masker) [23], [24]. The masker introduces a curve of masking threshold which specifies the minimal sound pressure level of a tone to be human perceptible with respect to the tone frequency. In other words, any signal below this curve is inaudible to human. The masking threshold of a masker can be approximated using the psychoacoustic model [24].

III. Overview of SongBsAb

A. Objective and Design

Our goal is to protect the rights of songs by mitigating SVC-based song cover (*prevention*) no matter songs are used as the source and/or target singing voices in SVC, but without impacting the release, spread and enjoyment of songs (*harmlessness*). These two objectives are achieved by SongBsAb.

Fig. 2 depicts the overview of SongBsAb, where the left part shows the workflow of SVC without SongBsAb while the right part shows that song owners create protected counterparts by adding perturbations with SongBsAb. When protected ones are used as the source and/or target singing voices in SVC, the conversion fails to produce the expected one, achieving the prevention objective, while the perturbations are inaudible by audiences, achieving the harmlessness objective.

It is unknown to song owners in advance if a singing voice will be used as the source or target singing voice in SVC, thus SongBsAb is designed to feature a dual prevention by causing the following two disruptions in SVC-covered singing voices.

- **Identity disruption**. To prevent a singing voice from being used as the target singing voice in SVC, SongBsAb crafts the perturbation on the identity encoder so that the SVC-covered singing voice sounds unlike being covered by the target singer, protecting both the performing and civil rights of the target singer.
- **Lyric disruption**. To prevent a singing voice from being used as the source singing voice in SVC, SongBsAb crafts the perturbation on the lyric encoder so that the SVC-covered singing voice contains unclear and even distinct lyrics from the expected one, protecting the copyrights of the lyrics in the source singing voice.

SongBsAb directly protects the civil rights of singers and the copyrights of lyrics in a straightforward manner, while the copyrights of melodies and the copyrights to reproduce and distribute songs are indirectly protected by SongBsAb, as SongBsAb worsens the performance of singing voice conversion and thus discourages the release, distribution and spread of SVC-covered songs, and the usage of SVC. We will discuss possible solutions to directly protect more rights in § VI.

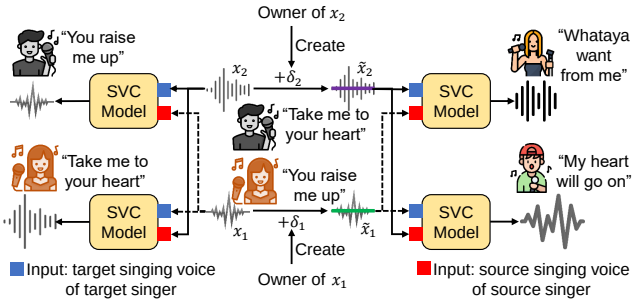


Fig. 2: Overview of SongBsAb. Song owners apply SongBsAb to singing voices (x_1 and x_2) and obtain the protected counterparts ($\tilde{x}_1$ and $\tilde{x}_2$) to prevent them from being used as source or target singing voices (dual prevention), by disrupting lyrics and singer identity in SVC-covered singing voices, respectively.

We note that one may only want to directly protect the civil rights of singers (resp. copyrights of lyrics), for which it suffices to prevent the singing voice from being used as target (resp. source) singing voice in SVC. Thus, SongBsAb is designed to be configurable to provide a sole or dual prevention, by causing one of two disruptions or both.

B. Threat Model

We discuss the threat model of the adversary and defender, where the adversary can be neutral or malicious SVC users.

Adversary's purpose. Neutral users use SVC for entertainment purposes, e.g., fans of a singer hope the singer covers some songs, and music enthusiasts who greatly admire the lyrics and melody of a song wish for it to be covered and spread widely. Malicious users gain improper benefits such as financial gain via SVC. For example, a company might use SVC to release records sung by a target singer, competing with the original record company. They may also use SVC to create singing voices with sensitive contents, for product promotion, advocacy, and so on. Both neutral and malicious users can cause right infringements, regardless of their purposes.

Adversary's capacity. (1) *Singing voices:* We assume that the adversary can collect a few target singing voices and a source singing voice, e.g., downloading or recording songs available on music platforms and then easily extracting singing voices from the songs. (2) *SVC models:* We assume that the adversary has access to a *few-shot* SVC model, which requires much fewer resources (computation and the target singer's singing voices) than a non-few-shot model (cf. § II-A). This makes it accessible to a broader range of adversaries, producing more illegally covered songs. Preventing such adversaries is thus more urgent and expands the application of SongBsAb.

Adversary's knowledge. The adversary may be unaware of prevention or has complete knowledge of SongBsAb (cf. § V-F) under which the adversary may adopt adaptive strategies to bypass the prevention.

Defender. (1) *Subject:* The song owners are the defenders. The composer, lyricist, and singer can be the defenders as well,

e.g., when they are the same person such as online singers. (2) *Purpose:* By applying SongBsAb to their original clean singing voices, defenders can prevent their songs from being used as both target and source songs by disrupting the identity and/or lyrics of SVC-covered singing voices, protecting the civil rights of singers and/or the copyrights of songs. (3) *Knowledge of SVC models:* We first assume that the defender knows the identity and the lyric encoders of the SVC model adopted by the adversary. Later, we will relax this assumption in § V-C. (4) *Knowledge of target singers:* When applying SongBsAb to prevent a song from being used as a target song (target singer) by SVC, the target singer is the singer of the song, thus known to the defender.

C. Practicality of SongBsAb

Platform difference. SongBsAb is applied by song owners before song release so that the same protected songs can be distributed across various platforms. Despite variations in storage or transmission methods (e.g., compression), SongBsAb is robust against transformations like compression (cf. § V-F1).

Unprotected songs. While adversaries may utilize songs covered by individuals (including the adversaries themselves) with excellent vocal skills as source singing voices, they still need songs sung by target singers as target singing voices to replicate their singing styles. Thus, at least the target singing voices are controlled by defenders and protected by SongBsAb, which causes at least identity disruption in SVC-covered singing voices. We consider this case in Appendix G of [2]. In addition, though adversaries may have copies of some songs released prior to SongBsAb and thus they are not protected, SongBsAb remains effective for identity disruption even when the ratio of the protected target singing voices is small (cf. Appendix D of [2]). To enhance the practicality of SongBsAb in real-world usage and better protect copyrights and civil rights, non-technical strategies also can be adopted (cf. § VI for more discussions).

IV. Methodology of SongBsAb

Fig. 3 presents overview of our methodology of SongBsAb. We first formulate the optimization problem of crafting protected songs and then detail the loss functions designed for identity disruption, lyric disruption, utility, and transferability.

A. Problem Formulation

Given a singing voice $x^0 \in \mathbb{R}^{1\times D}$ represented by a waveform with length D, the identity encoder Θ, and the lyric encoder Φ, we attempt to craft a (protected) singing voice x to disrupt the identity and lyrics of SVC-covered singing voices while preserving the quality of songs. Formally, we need to solve the following optimization problem:

$$\min_{x} \left(\begin{array}{c} f_\Theta(x,x^0) + f_\Phi(x) + \lambda_u f_u(x,x^0) \\ +\lambda_\Theta^{te} f_\Theta^{te}(x,x^0) + \lambda_\Phi^{te} f_\Phi^{te}(x,x^0) \end{array} \right)$$
$$\text{subject to } x \in [-1,1]$$

where f_Θ and f_Φ are prevention losses for the identity and lyric disruptions; f_u is the utility loss for harmlessness (i.e., song quality); f_Θ^{te} and f_Φ^{te} are the transferability enhancement

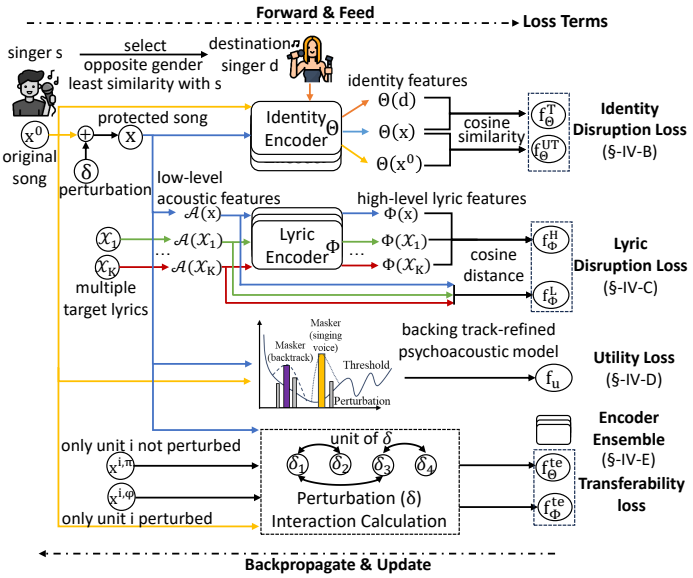


Fig. 3: Overview of the methodology of SongBsAb

loss for the identity and lyric disruptions. The positive factors λ_u, λ_Θ^{te}, and λ_Φ^{te} are used to control the impact of these losses on the perturbation $(x - x^0)$. Following [14], [59], [64], [30], [60], [62], we normalize singing voices by dividing their magnitudes by the maximum value of the bit-width, to avoid causing potential overflow during optimization, so valid singing voices subject to $[-1, 1]$.

B. Identity Disruption: Gender-Transformation Loss

Untargeted loss $f_\Theta^{\mathtt{UT}}$. We attempt to cause SVC-covered singing voices to *sound unlike the singer of x^0* (i.e., the target singer when the protected singing voice x is used as the target singing voice by SVC). Since the identity information for SVC is provided by identity features, we achieve this purpose by ensuring that *the identity feature $\Theta(x)$ of the protected singing voice x deviates from the original identity feature $\Theta(x^0)$*. Formally, we minimize the following loss that quantifies the similarity between identity features $\Theta(x)$ and $\Theta(x^0)$:

$$f_\Theta^{\mathtt{UT}}(x,x^0) = \mathtt{Sim}(\Theta(x),\Theta(x^0))$$

where $\mathtt{UT}$ denotes untargeted and $\mathtt{Sim}(\cdot)$ is the similarity function. In this work, we use the cosine similarity [67] due to its bounded nature which results in bounded losses and therefore more stable optimization [68], [69], [70].

Targeted loss $f_\Theta^{\mathtt{T}}$. Humans can better perceive the vocal difference between opposite-gender singers than between same-gender singers. Hence, to enhance identity disruption, we cause SVC-covered singing voices to *sound like being covered by a singer with the opposite gender* (called *destination* singer) from the original singer (i.e., the singer of x^0).

We design the following process to choose the destination singer that *has the opposite gender and is the most objectively identity-dissimilar from the original singer*. Firstly, we collect a set of auxiliary singers with the *opposite* gender, each of which has a set of singing voices $\mathcal{V}_i$. Then we

represent each auxiliary singer i by the centroid identity feature $\Theta^c_{a,i} = \frac{1}{|\mathcal{V}_i|}\sum_{v\in\mathcal{V}_i}\Theta(v)$ and the original singer by $\Theta^c = \frac{1}{N}\sum_{i=1}^{N}\Theta(x^0_i)$ where N is the number of singing voices. Finally, the destination singer is chosen as the auxiliary singer whose centroid identity feature $\Theta^c_{a,i}$ is the farthest from Θ^c, i.e., $k = \mathrm{argmin}_i\, \texttt{Sim}(\Theta^c_{a,i}, \Theta^c)$. Let $\Theta^c_{\texttt{des}}$ denote $\Theta^c_{a,k}$. This process is characterized by (1) comprising both subjective perception (opposite-gender) and objective perception (least identity-similar) and (2) precise singer representation by the centroid embedding rather than a voice embedding. The defender can select auxiliary singers from open-source datasets without infringing their civil rights.

With the destination singer, we ensure that *the identity feature $\Theta(x)$ of the protected singing voice x approaches that of the destination singer* by minimizing the following loss:

$$f^{\texttt{T}}_{\Theta}(x) = -\texttt{Sim}(\Theta(x), \Theta^c_{\texttt{des}})$$

where T denotes targeted.

Final loss. Putting $f^{\texttt{UT}}_{\Theta}$ and $f^{\texttt{T}}_{\Theta}$ together, our final loss for identity disruption, called gender-transformation loss, is defined as:

$$f_{\Theta}(x, x^0) = f^{\texttt{UT}}_{\Theta}(x, x^0) + \lambda_{\Theta} f^{\texttt{T}}_{\Theta}(x)$$

where $\lambda_{\Theta} > 0$ is the loss balancing factor.

We emphasize that the two loss terms, selecting the least identity-similar and opposite-gender destination singer, and representing destination singers by centroid embeddings, all contribute to identity disruption (cf. Appendix H of [2]).

C. Lyric Disruption: High/Low Hierarchy Multi-Target Loss

High hierarchy loss. Since lyric features provide the lyric information for SVC, to achieve lyric disruption, we ensure that the lyric feature $\Phi(x)$ of the protected singing voice x differs from the original lyric feature $\Phi(x^0)$. Prior work [71] on adversarial attacks against speech-to-text tasks has shown that targeted attacks are more transferable than untargeted attacks regarding mistranscription. Inspired by this, we *choose a singing voice χ with different lyrics from x^0 and pull together the lyric feature $\Phi(x)$ and the lyric feature $\Phi(\chi)$ of χ* by minimizing the following designated loss:

$$f^{H}_{\Phi}(x) = \texttt{Dist}(\Phi(x), \Phi(\chi))$$

where $\texttt{Dist}(\cdot)$ is the distance function, initialized by the cosine distance (i.e., 1 minus cosine similarity) in this work due to the same reason as in § IV-B.

Low hierarchy loss. The loss $f^{H}_{\Phi}(x)$ only minimizes the distance between high-level lyric features $\Phi(x)$ and $\Phi(\chi)$. However, mainstream SVC models also rely on low-level acoustic features [72], including handcrafted ones (e.g., filter Bank [73]) and representations produced by shallow hidden layers (e.g., Hubert-based features [32]), which are extracted from voice waveform and used to derive lyric features. Inspired by this, we hypothesize that *aligning the low-level acoustic features $\mathcal{A}(x)$ with $\mathcal{A}(\chi)$ can improve the alignment of high-level lyric features*, thus enhancing the lyric disruption. Therefore, we define the following low hierarchy lyric disruption loss:

$$f^{L}_{\Phi}(x) = \texttt{Dist}(\mathcal{A}(x), \mathcal{A}(\chi)).$$

Multiple targets. Both $f^{H}_{\Phi}(x)$ and $f^{L}_{\Phi}(x)$ only utilize a single singing voice χ to provide target lyrics. *Due to the phoneme difference among different singing voices, given the protected singing voices x, the difficulty of optimizing $f^{H}_{\Phi}(x)$ and $f^{L}_{\Phi}(x)$ may vary with χ.* To tackle this issue, we propose to *enhance the effect of lyric disruption with multiple target lyrics* by adapting $f^{H}_{\Phi}(x)$ and $f^{L}_{\Phi}(x)$ as follows:

$$f^{H}_{\Phi}(x) = \frac{1}{K}\sum_{k=1}^{K}\texttt{Dist}(\Phi(x), \Phi(\chi_k))$$
$$f^{L}_{\Phi}(x) = \frac{1}{K}\sum_{k=1}^{K}\texttt{Dist}(\mathcal{A}(x), \mathcal{A}(\chi_k))$$

where $\chi_1, \cdots, \chi_K$ are singing voices with distinct lyrics.

Final loss. Our final loss for lyric disruption, called high/low hierarchy multi-target loss, is formulated as follows:

$$f_{\Phi}(x) = \lambda^{H}_{\Phi} f^{H}_{\Phi}(x) + \lambda^{L}_{\Phi} f^{L}_{\Phi}(x)$$

where λ^{H}_{Φ} and λ^{L}_{Φ} are positive balancing factors.

D. Utility: Backing Track-Refined Simultaneous Masking Loss

Basic loss. Since the original singing voice x^0 and the perturbation simultaneously occur when the singing voice x is played, the perturbation can be hidden with simultaneous masking. Specifically, we *treat x^0 as the masker and make the perturbation (maskee) inaudible by forcing it to fall under the masking threshold of the masker.* Let $\theta_a \in \mathbb{R}^{T\times F}$ denote the masking threshold of the audio a where T is the number of frames (audio's short segments) and F is the number of frequencies. Let $p_a \in \mathbb{R}^{T\times F}$ denote the log-magnitude power spectral density of the audio a. Formally, we minimize the following utility loss based on simultaneous masking:

$$f_u(x, x^0) = \frac{1}{T\cdot F}\sum_{t=1}^{T}\sum_{k=1}^{F}\max\{0, p_{x-x^0}(t,k) - \theta_{x^0}(t,k)\}.$$

Refined loss. *A singing voice is typically accompanied by a backing track $\mathcal{M}$ in a different channel of a song.* Thus, we propose to *utilize the backing track as an additional masker* to improve harmlessness: *the perturbation will not be audible as long as it is under one of the masking thresholds* of the singing voice and the backing track. The backing track-refined simultaneous masking utility loss is defined as:

$$f_u(x, x^0) = \frac{1}{T\cdot F}\sum_{t=1}^{T}\sum_{k=1}^{F}\max\{0, p_{x-x^0}(t,k) - \theta_{x^0,\mathcal{M}}(t,k)\}$$

where $\theta_{x^0,\mathcal{M}}(t,k) = \max\{\theta_{x^0}(t,k), \theta_{\mathcal{M}}(t,k)\}$ is the joint masking threshold of the two maskers. Intuitively, minimizing the loss f_u minimizes the density of the perturbation for each frame and frequency until it is no greater than one of the masking thresholds of the singing voice and the backing track.

Remark that the refined loss f_u is tailored for songs since it utilizes the unique backing tracks of songs that do not exist for ordinary speech voices. It adopts a simplified joint psychoacoustic model for the singing voice and backing track, as the modeling of simultaneous masking of multiple-channel

signals is a very complex task [74]. Despite being simplified, it effectively improves harmlessness in our experiments (cf. Appendix J of [2]). More precise modeling is left as future work.

E. Transferability Enhancement

Frame-level interaction reduction-based (FL-IR) loss f_Θ^{te}&f_Φ^{te}. Inspired by the *negative correlation between transferability and interaction inside perturbations* (cf. § II-C), we define f_Θ^{te} and f_Φ^{te} to enhance transferability by reducing interaction:

$$f_\Theta^{te}(x,x^0) = \mathbb{E}_i(f_\Theta^{\mathtt{UT}}(x,x^0) + 1 - f_\Theta^{\mathtt{UT}}(x^{i,\pi},x^0) - f_\Theta^{\mathtt{UT}}(x^{i,\varphi},x^0))$$
$$f_\Phi^{te}(x,x^0) = \mathbb{E}_i(f_\Phi^{H}(x) + f_\Phi^{H}(x^0) - f_\Phi^{H}(x^{i,\pi}) - f_\Phi^{H}(x^{i,\varphi}))$$

where 1 is $f_\Theta^{\mathtt{UT}}(x^0,x^0)$ in defining $f_\Theta^{te}(x,x^0)$, $x^{i,\pi}$ is identical to x except that its i-th unit is not perturbed and $x^{i,\varphi}$ is identical to x^0 except that its i-th unit is perturbed as x.

The computation of the FL-IR loss $f_\Theta^{te}(x)$ and $f_\Phi^{te}(x)$ involves iterating over all the sample points within a singing voice, which however, may contain numerous sample points due to the high sampling rate (e.g., 48KHz) [30], leading to costly and even intractable computational overhead. Observing that singing voices are split into multiple short fragments (called frames) before being fed to SVC models, we address this challenge by calculating the losses at the frame level. Specifically, given the frame length w_l and the frame shift w_s, we first decide the boundaries of each frame. The boundaries of the i-th frame are $i \times w_s$ and $i \times w_s + w_l$. We treat each frame as a whole, that is, all points within a frame are simultaneously perturbed or not perturbed. Then we compute the FL-IR loss by iterating over the frames instead of all the sample points, where $x^{i,\pi}$ becomes identical to x except that all sample points within its i-th frame are not perturbed and $x^{i,\varphi}$ becomes identical to x^0 except that all sample points within its i-th frame are perturbed as x. We also approximate the expectation $\mathbb{E}$ by R times random sampling to further reduce the overhead [28].

Encoder ensemble. It is known that model ensemble can enhance transferability [30], [11], [29]. Thus, we collect various identity/lyric encoders on which average losses $f_\Theta(x)$, $f_\Phi(x)$, $f_\Theta^{te}(x)$ and $f_\Phi^{te}(x)$ are computed. *It improves transferability from a different perspective than the FL-IR loss.* The FL-IR loss reduces the interaction among perturbation units which has a negative correlation with transferability, while encoder ensemble enforces that the features of the singing voice x deviate enough from that of the original one x_0. Our results confirm that both methods can improve transferability, and their combination yields the best transferability (cf. § V-C).

F. Final Approach

Finally, we solve the following optimization problem:

$$\min_x \left(\begin{array}{c} f_\Theta^{\mathtt{UT}}(x,x^0) + \lambda_\Theta f_\Theta^{\mathtt{T}}(x) + \lambda_\Phi^H f_\Phi^H(x) + \lambda_\Phi^L f_\Phi^L(x) \\ +\lambda_u f_u(x,x^0) + \lambda_\Theta^{te} f_\Theta^{te}(x,x^0) + \lambda_\Phi^{te} f_\Phi^{te}(x,x^0) \end{array} \right)$$
$$\text{subject to } x \in [-1,1].$$

Instead of manually setting the balance factors λ_Θ, λ_Φ^H, λ_Φ^L, λ_u, λ_Θ^{te}, and λ_Φ^{te}, we utilize automatic and dynamic loss balance by loss normalization [30], due to its advantage of nearly equally weighing different loss functions with different ranges and scales. Specifically, at each iteration of crafting the singing voice x, we normalize each loss f_k by its mean μ_k and variance σ_k, i.e., $f_k' = \frac{1}{\sqrt{\sigma_k}}(f_k - \mu_k)$. Both μ_k and σ_k are loss-specific and iteratively updated via $\mu_k = \mu_k + \frac{1}{n}(f_k - \mu_k)$ and $\sigma_k = \sigma_k + \frac{1}{n}((f_k - \mu_k)^2 - \sigma_k)$, where n is the current iteration. Finally, the total loss function is defined as the sum of the normalized losses.

Algorithm 1: SongBsAb

Input: original singing voice x^0; number of steps N; learning rate α; identity encoder Θ; lyric encoder Φ; `protect_target`; `protect_source`; `transfer_identity`; `transfer_lyric`
Output: singing voice x

1 Adam $\leftarrow$ initialize Adam optimizer with α;
2 $K \leftarrow 7$; $F \leftarrow [f_\Theta^{\mathtt{UT}}, f_\Theta^{\mathtt{T}}, f_\Phi^H, f_\Phi^L, f_u, f_\Theta^{te}, f_\Phi^{te}]$;
3 **for** k *from* 1 *to* K **do** $\mu_k \leftarrow 0$; $\sigma_k \leftarrow 1$;
4 **for** n *from* 1 *to* N **do**
5 $f_{\mathtt{total}} \leftarrow 0$;
6 **for** k *from* 1 *to* K **do**
7 $f \leftarrow F_k$;
8 **if** $f \in \{f_\Theta^{\mathtt{UT}}, f_\Theta^{\mathtt{T}}\} \wedge$ `protect_target` $=$ `False` **then continue**;
9 **if** $f \in \{f_\Phi^H, f_\Phi^L\} \wedge$ `protect_source` $=$ `False` **then continue**;
10 **if** $f = f_\Theta^{te} \wedge$ `transfer_identity` $=$ `False` **then continue**;
11 **if** $f = f_\Phi^{te} \wedge$ `transfer_lyric` $=$ `False` **then continue** ;
12 $f_k \leftarrow f(x^{n-1}, x^0)$; $\mu_k \leftarrow \mu_k + \frac{f_k - \mu_k}{n}$;
13 $\sigma_k \leftarrow \sigma_k + \frac{1}{n}((f_k - \mu_k)^2 - \sigma_k)$; $f_k \leftarrow \frac{f_k - \mu_k}{\sqrt{\sigma_k}}$;
14 $f_{\mathtt{total}} \leftarrow f_{\mathtt{total}} + f_k$;
15 $x^n \leftarrow \mathrm{Adam}(x^{n-1}, \nabla_{x^{n-1}} f_{\mathtt{total}})$;
16 $x^n \leftarrow \max\{\min\{x^n, 1\}, -1\}$;
17 **return** x^N

We minimize the loss by N-iteration gradient descent using the Adam optimizer (with learning rate α). The overall algorithm is shown in Alg. 1. In each iteration (Lines 4–16), we iteratively (Lines 6– 14) compute each f_k of 7 losses and normalize it using its mean μ_k and variance σ_k (Lines 12–13). Remark that if encoder ensemble is enabled, f_k is the average over all encoders. We then compute the total loss $f_{\mathtt{total}}$ by summing the 7 normalized losses (Line 14), update the singing voice using the Adam optimizer and the gradient w.r.t. the total loss (Line 15), and clip it to be a valid singing voice (Line 16). To be flexible, we provide the following flags: `protect_target`, `protect_source`, `transfer_identity`, and `transfer_lyric`. If the defender does not prevent singing voices from being used as target (resp. source) singing voices, `protect_target` (resp. `protect_source`) can be `False`. Similarly, if the defender has access to the identity (resp. lyric) encoder of the SVC, `transfer_identity` (resp. `transfer_lyric`) can be `False`. When a flag is `False`, SongBsAb will ignore the respective loss (Lines 8-11). We also provide convergence analysis in Appendix B of [2].

Remark that except for the refined utility loss that utilizes

TABLE II: Details of singing voice conversion (SVC) models.

Dis.‡	Model	Few-Shot?	Identity Encoder	Lyric Encoder	Pitch Encoder	Decoder♯ (G / $G+V$)	Sample Rate
Explict	**Lora -SVC**	✓	LSTM [33]	Whisper -Medium [31]	WORLD [36]	BigVGAN♯ [86]	16kHz
	Vits -SVC	✓	LSTM [33]	Whisper -Large [31] & Hubert [32]	Crepe [37]	BigVGAN♯ [86]	32kHz
	Grad -SVC	✓	LSTM [33]	Hubert [32]	Praat [87]	Diffusion [88] +BigVGAN♯ [86]	32kHz
Implicit	**NeuCo -SVC**	✓	WavLM -Large [89]	WavLM -Large [89]	pYIN [90] & REAPER [91]	FiLM UNet [92]	24kHz
	StarGANv2 -SVC	✗	Style Encoder [35]	VGG -BLSTM [93]	JDCNet [94]	StarGANv2 [95] +ParallelWaveGAN [96]	24kHz

(1) ‡: Dis. is short for information disentanglement. (2) ♮: Decoder may directly produce waveforms with a generator (G) or first use G to produce acoustic features and utilize vocoders V to synthesize waveforms ($G+V$). (3) ♯: Their specific architectures and parameters are different.

unique elements of songs, the others could generalize to other domains. In particular, the loss f_Θ for identity disruption, and the FL-IR loss and encoder ensemble for complementarily enhancing transferability could be exploited to prevent ordinary speech voice conversion/synthesis to protect speaker identity. The high/low hierarchy multi-target loss could be used to enhance the generation of adversarial speech examples against speech-to-text models for malicious purposes. Also, the FL-IR loss could be utilized by attackers to strengthen transfer-based adversarial attacks against speech processing systems [30]. We leave these as interesting future works.

V. Evaluation

A. Experimental Setup

Models. We adopt four recent promising SVC models with few-shot conversion capability: Lora-SVC [40], Vits-SVC [75], Grad-SVC [41], and NeuCo-SVC [34]. We also consider one non-few-shot SVC model StarGANv2-SVC [35]. As shown in TABLE II, they are diverse in information disentanglement method, identity, lyric, and pitch encoders, the decoder, and the sampling rate of singing voices. Since we target few-shot SVC models (cf. § III-B), this section only considers 4 few-shot models while StarGANv2-SVC is evaluated in Appendix B.

To evaluate the transferability of SongBsAb, we consider another 8 distinct identity encoders: X-vectors (XV) [76], ECAPA-TDNN (ECAPA) [77], ResNet18 for identification (Res18-I) [78], [79], ResNet34 for identification (Res34-I) [80], [79], ResNet34 for verification (Res34-V) [80], [79], AutoSpeech (Auto) [81], ResNetSE34V2 (Res-SE) [82], and VGGVox-40 (VGG) [83]; and another 5 distinct lyric encoders: Whisper-Tiny [31], Whisper-Base [31], Whisper-Small [31], Wav2vec2 [84], and Decoar2 [85]. These 13 encoders are different from those used in all SVC models.

Datasets. We use two datasets: OpenSinger [46] and NUS-48E [47], whose attributes are shown in TABLE III.

We select target singers, target and source singing voices as follows. Let m denote the number of singers in a dataset. Firstly, we regard each singer as target singer and randomly select t singing voices sung by the singer as target singing voices and randomly select s singing voices from other singers as the source singing voices, leading to $p = m \times s$ pairs of target singer and source singing voice. Then, we run the SVC model and choose 2,000 pairs out of p pairs with top identity similarity, following the practice of previous works [12], [11]. The rationale behind this selection is that the SVC model performs better on these selected pairs, thus they are more necessary to be protected than the others. For OpenSinger, m=76, t=10, s=100, p=7,600. For NUS-48E with a smaller volume, m=12, t=4, s=200, p=2,400. The number of pairs is large enough to cover all singers and songs for both datasets.

TABLE III: The attributes of datasets.

Data set	Language	#Accent	Voice Type	Tempo (bpm§)	Pitch	#Singers†	#Songs♮	#Pairs
Open Singer	Chinese	NA‡	NA	NA	280.4 ± 94.6	76 (48F, 28M)	363	2,000
NUS -48E	English	7	5♭	68 ∼ 150	NA	12 (6F, 6M)	20	2,000

(1) NA‡ means that the respective metadata is not available. (2) bpm§ is short for beats per minutes. (3) #Singers† and #Songs♮ denote the number of singers and songs of the dataset and our selected pairs of target singers and source singing voices cover all singers and songs. (4) "(xF, yM)" denotes x female and y male singers. (5) ♭: Soprano, Alto, Tenor, Baritone, and Bass.

Since both datasets do not contain any backing tracks, for each singing voice, we randomly crop the backing track "Amazing Grace" to match the length of each singing voice. For the loss f_Θ (cf. § IV-B), all the singers with the opposite gender of the target singer are used as auxiliary singers. For the loss f_Φ (cf. § IV-C), $\chi_{k\in[1,K]}$ are selected from all singing voices with different lyrics from the source singing voice, and K is set to 10 after investigation.

Metrics. Besides human study in § V-E as subjective evaluation metrics, we use the following objective metrics. (1) *Identity similarity (IS)*: cosine similarity between the centroid identity feature of the target singer and the identity feature of the SVC-covered output. It measures how well SVC models imitate the timbre of the target singer. We extract identity features with the Resnet18 for verification (Res18-V) model [78], [79] differing from the other encoders.

(2) *Lyric word error rate (WER)* measures the lyric differences between the SVC-covered output and the source singing voice, i.e., the error that SVC models commit in retaining lyrics. $\texttt{WER} = \frac{D+I+S}{N}$ where N is the number of words in the source singing voice, and D, I, and S are the numbers of deletions, insertions, and substitutions, respectively. We use the speech-to-text model Conformer [97] to recognize lyrics.

(3) *Success reduction rate (SRR)* is the reduction of SVC success rate after applying prevention, including SRR for target identity imitation (SRR-I), SRR for source lyric preservation (SRR-L), and overall SRR (SRR-T):

$$\text{SRR-I} = \frac{\sum_{i=1}^{Q}\mathbb{I}(\text{IS}(y_i)\geq\xi_I)-\sum_{i=1}^{Q}\mathbb{I}(\text{IS}(\hat{y}_i)\geq\xi_I)}{Q}$$

$$\text{SRR-L} = \frac{\sum_{i=1}^{Q}\mathbb{I}(\text{WER}(y_i)\leq\xi_L)-\sum_{i=1}^{Q}\mathbb{I}(\text{WER}(\hat{y}_i)\leq\xi_L)}{Q}$$

$$\text{SRR-T} = \frac{\left(\begin{array}{c}\sum_{i=1}^{Q}\mathbb{I}(\text{IS}(y_i)\geq\xi_I \bigwedge \text{WER}(y_i)\leq\xi_L) \\ - \\ \sum_{i=1}^{Q}\mathbb{I}(\text{IS}(\hat{y}_i)\geq\xi_I \bigwedge \text{WER}(\hat{y}_i)\leq\xi_L)\end{array}\right)}{Q}$$

where y_i and $\hat{y}_i$ for $i = 1, \cdots, Q$ are the undefended and defended SVC-covered outputs, respectively; ξ_I and ξ_L are the thresholds for deciding the success of SVC w.r.t. identity imitation and lyric preservation, respectively; and $\mathbb{I}$ is the

indicator function. We set $\xi_I = 0.41$ (the same as [30]) and ξ_L to the average WER of undefended SVC-covered outputs.

The higher (resp. lower) WER, SRR-I, SRR-L, and SRR-T (resp. IS) are, the more effective a prevention method is.

(4) *Signal-to-noise ratio (SNR) [62]*, $SNR = 10\log_{10}\frac{P_x}{P_\delta}$, is widely used to measure the imperceptibility of voice perturbations, where P_x and P_δ are the power of the original singing voice x and the perturbation δ, respectively.

(5) *Perceptual evaluation of speech quality (PESQ) [98]* is an objective perceptual metric that simulates the human auditory system [99], ranging from -0.5 to 4.5.

To compute SNR and PESQ for a stereo song where one channel is the singing voice and the other is the backing track, we merge the song into a mono audio using the "pydub" package [100]. Higher SNR and PESQ indicate better imperceptibility and thus better harmlessness of prevention.

Baselines. SongBsAb is the first for preventing SVC, so we select from TABLE IX the closest baselines (targeting voice modality and generative models) for comparison, AttackVC and AntiFake (using its best target-based scheme [11]), which are designed to prevent ordinary voice conversion or synthesis. VSMask is not considered since it is not publicly available.

Experimental design. We first evaluate the dual prevention effectiveness of SongBsAb (i.e., disrupting both identity and lyrics in SVC-covered singing voices), assuming the defender is aware of the identity and lyrics encoders of adversaries. Next, we relax this assumption by evaluating the transferability of SongBsAb to unknown SVC models and analyze the efficiency of SongBsAb. Finally, we subjectively evaluate SongBsAb via human study and evaluate its robustness in over-the-air scenario and against adaptive adversaries.

B. Dual Prevention of SongBsAb

Setting. To evaluate the dual prevention of SongBsAb, we set `protect_target` and `protect_source` to `True`, the initial learning rate α=0.001 for the Adam optimizer, and the number of iterations N=1,000. The results are shown in TABLE IV.

Results of identity disruption. With each of SongBsAb, AntiFake, and AttackVC, the identity similarity of the defended SVC-covered outputs $\tilde{y}$ is lower compared to the undefended SVC-covered outputs y, indicating that all of them disrupt the identity of SVC-covered outputs away from target singers. However, SongBsAb achieves much lower identity similarity and much higher SRR-I than baselines, regardless of datasets and SVC models, demonstrating that *SongBsAb is significantly more effective than baselines for the prevention of SVC regarding identity disruption.* This is probably because (1) AttackVC perturbs acoustic features and uses the Griffin-Lim algorithm [101] to reconstruct voices, which is a lossy procedure that may interrupt the perturbation [62]. It is evidenced by the much higher SRR-I of AttackVC-W, a modified version that directly perturbs voices. (2) AttackVC randomly chooses the destination speaker from some opposite-gender speakers, while SongBsAb selects the opposite-gender singer having the least identity similarity with the target singer. (3) AntiFake only penalizes the distance of embeddings between the protected voice and the destination speaker while SongBsAb additionally penalizes the similarity of embeddings between the protected and original voices (i.e., $f_\Theta^{\mathtt{UT}}$, cf. § IV-B). (4) Both AttackVC and AntiFake represent the destination speaker by a voice embedding, while SongBsAb uses the centroid of multiple voice embeddings. The ablation study reported in Appendix H of [2] justifies the reasons (2)–(4).

TABLE IV: Comparison of prevention effectiveness and harmlessness between SongBsAb and baselines.

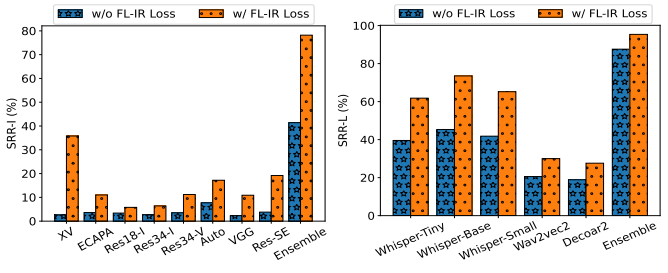
Dataset	SVC Model	Approach	Prevention Effectiveness							Harmlessness			
			Identity Similarity ↓		Lyric WER (%) ↑		SRR (%) ↑			SNR (dB) ↑		PESQ ↑	
			y	$\tilde{y}$	y	$\tilde{y}$	SRR-I	SRR-L	SRR-T	$\tilde{\mathcal{I}}$	$\tilde{\mathcal{L}}$	$\tilde{\mathcal{I}}$	$\tilde{\mathcal{L}}$
Open Singer	Lora -SVC	AntiFake	0.54	0.15	13.9	13.2	65.5	9.8	68.0	24.6	26.6	3.1	3.3
		AttackVC		0.55		13.9	0.1	9.3	9.4	-5.0	-4.5	2.0	2.3
		AttackVC-W		0.24		13.1	41.4	8.8	44.5	10.1	12.4	1.3	1.4
		SongBsAb		**0.05**		**76.1**	**88.1**	**92.2**	**99.3**	**26.5**	**30.6**	**3.9**	**4.2**
	Vits -SVC	AntiFake	0.51	0.15	14.9	15.2	71.1	11.2	73.6	24.5	25.5	3.0	3.1
		AttackVC-W		0.26		14.7	40.1	9.5	44.8	10.2	11.1	1.4	1.7
		SongBsAb		**0.09**		**90.4**	**82.6**	**92.7**	**99.1**	**26.3**	**27.5**	**3.9**	**4.0**
	Grad -SVC	AntiFake	0.48	0.17	32.1	31.4	77.8	8.9	78.9	24.7	24.8	3.2	3.1
		AttackVC-W		0.23		30.9	59.2	8.2	62.1	10.0	10.3	1.5	1.7
		SongBsAb		**0.11**		**103.6**	**85.2**	**94.6**	**99.4**	**26.6**	**27.7**	**4.1**	**4.0**
	NeuCo -SVC	AntiFake	0.65	0.33	18.1	20.8	70.7	14.1	72.2	18.8	19.2	2.2	2.4
		AttackVC-W		0.28		20.1	78.5	13.0	79.3	10.7	11.2	1.4	1.6
		SongBsAb		**0.22**		**86.5**	**88.8**	**90.4**	**98.9**	**27.6**	**28.3**	**3.8**	**4.1**
NUS -48E	Lora -SVC	AntiFake	0.47	0.22	23.3	22.0	70.8	5.7	72.4	26.5	26.5	3.2	3.0
		AttackVC-W		0.25		23.1	79.2	7.3	80.4	6.2	6.1	1.3	1.2
		SongBsAb		**0.12**		**79.9**	**87.7**	**93.6**	**98.7**	**32.4**	**28.3**	**4.4**	**4.3**
	Vits -SVC	AntiFake	0.48	0.19	18.4	19.4	75.8	10.3	78.3	26.3	24.2	3.3	2.0
		AttackVC-W		0.25		19.3	69.1	10.8	72.6	6.2	5.9	1.5	1.3
		SongBsAb		**0.12**		**78.4**	**80.1**	**92.5**	**98.0**	**32.1**	**25.7**	**4.4**	**4.2**
	Grad -SVC	AntiFake	0.45	0.24	41.1	41.2	74.8	7.8	78.6	26.0	23.3	3.4	3.2
		AttackVC-W		0.24		43.6	69.7	10.9	73.9	6.0	5.8	1.5	1.5
		SongBsAb		**0.16**		**94.5**	**83.4**	**91.7**	**97.3**	**31.8**	**26.4**	**4.3**	**4.2**
	NeuCo -SVC	AntiFake	0.59	0.24	22.6	22.7	79.0	10.4	79.1	14.9	16.9	1.9	2.2
		AttackVC-W		0.22		21.7	81.9	8.2	82.4	6.8	7.3	1.3	1.7
		SongBsAb		**0.16**		**76.6**	**86.9**	**94.4**	**98.9**	**26.4**	**27.3**	**3.7**	**4.3**

(1) Acronyms refer to TABLE I and § V-A-"Metrics". (2) For each combination of datasets and models, the best results among all prevention approaches are highlighted in **bold**. (3) ↑: the higher, the more effective or harmless the approach is. (4) ↓: the lower, the more effective the approach is. (5) AttackVC is only considered on OpenSinger and Lora-SVC since it is much less effective than its variants AttackVC-W.

Results of lyric disruption. With SongBsAb, the lyric WER of $\tilde{y}$ is 53%-75% higher than that of y. On each combination of SVC models and datasets, SongBsAb achieves more than 90% overall SVC success reduction rate (SRR-T), higher than both SRR-I and SRR-L. In comparison, the SRR-T of AttackVC and AntiFake is nearly identical to SRR-I, and the WER of $\tilde{y}$ is very close to that of y. These demonstrate that *both baselines cannot disrupt lyrics, while SongBsAb is effective in the dual prevention of SVC by both identity and lyric disruptions.*

Results of harmlessness. The SNR and PESQ of protected singing voices crafted by SongBsAb exceed 25 dB and 3.7, respectively, higher than that of protected singing voices crafted by the two baselines, especially for PESQ, indicating that *SongBsAb outperforms both baselines regarding harmlessness and perturbations' side-effect on songs.* This is attributed to our refined simultaneous masking loss utilizing backing tracks as additional maskers (cf. Appendix J of [2] for justification).

Ablation study. We conduct ablation study to evaluate: (E1) the impact of the ratio of protected target singing voices; (E2) the effectiveness of SongBsAb w.r.t. singer genders and song genres; (E3) the single prevention of SongBsAb for disrupting lyric or identity but not both; (E4) the effectiveness of the gender-transformation loss and high/low hierarchy multi-target loss; and (E5) the effectiveness of the refined utility loss. The results show that: (R1) SongBsAb is not impacted by the ratio for disrupting lyric, and is still effective for disrupting identity even only a small fraction of target singing voices are protected



(a) Identity disruption (b) Lyric disruption

Fig. 4: Transferability of SongBsAb.

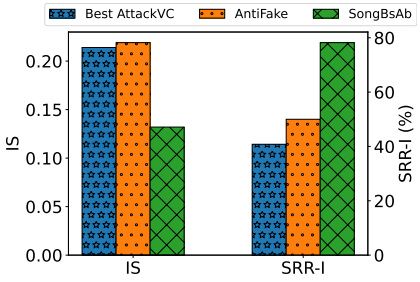


Fig. 5: Comparison of transferability for identity disruption in terms of identity similarity. AttackVC uses a single encoder, and Best AttackVC means the best result among all encoders.

and becomes more effective when increasing the ratio; (R2) SongBsAb exhibits universality across different singer genders and song genres; (R3) SongBsAb is still effective for single prevention of disrupting identity or lyric; (R4) our gender-transformation loss outperforms the loss without or with only the loss term f_{Θ}^{UT}, the loss randomly selecting a destination singer with the opposite gender, and the loss representing the destination singer by a voice embedding. Our high/low hierarchy multi-target loss achieves better lyric disruption than the low hierarchy loss, high hierarchy loss, and the loss without multiple targets; and (R5) the refined utility loss with backing tracks as additional maskers achieves better harmlessness than the basic utility loss solely using the singing voice as the masker. More details refer to Appendix D-J of [2].

We will mainly consider the OpenSinger dataset and Lora-SVC, as they generally achieve the best SVC in TABLE IV.

C. Transferability of SongBsAb

We evaluate the transferability for disrupting identity and lyric separately to avoid interference. For the FL-IR loss, we set $R = 32$, $w_l = w_s = \frac{L}{200}$ where L is the number of sample points of a singing voice. The impact of w_l and w_s is evaluated in Appendix K of [2]. Remark that we also evaluate the transferability via human study in § V-E.

Transferability for identity disruption. Each of the 8 identity encoders (cf. § V-A) is used to craft protected target singing voices. The results are shown in Fig. 4a. SRR-I is largely improved after applying either our FL-IR loss (regardless of the identity encoder) or encoder ensemble. Applying both the FL-IR loss and encoder ensemble yields the best transferability, *confirming the effectiveness and complementarity of the FL-IR loss and encoder ensemble in boosting transferability for identity disruption.* According to Fig. 5, SongBsAb achieves lower identity similarity (IS) and higher SRR-I than both AttackVC and AntiFake, *indicating the superiority of SongBsAb over baselines for transferring to unknown identity encoders.* This is because AttackVC and AntiFake are not incorporated with the encoder ensemble or the FL-IR loss.

Transferability for lyric disruption. Each of the 5 lyric encoders (cf. § V-A) is used to craft protected source singing voices. The results are shown in Fig. 4b. We can observe that *SongBsAb has inherent transferability for lyric disruption, and the FL-IR loss and encoder ensemble can further enhance the transferability for lyric disruption.* More results of transferability with different SVC models and metrics are reported in Appendix L of [2], from which we draw the same conclusions.

TABLE V: Pairs of singing voices for human study task 1.

Pair Name	No.	Description
Normal	9	A pair consists of two original singing voices from a target singer and a source singer, respectively.
Undefended Output	9	A pair is built by replacing the source singer's voice in a Normal pair with its SVC-covered output using the identity of corresponding target singer, where SongBsAb is disabled.
Defended Output	5	The SVC-covered output in each of 5 randomly selected Undefended Output pairs is replaced with another SVC-covered output, where SongBsAb is enabled and uses the same identity encoder as the SVC model.
Defended Output (Transfer)	5	They are built the same as Defended Output pairs except that SongBsAb uses different identity encoders from the SVC model.
Prot	4	The target singing voices protected by SongBsAb for building 5 Defended Output pairs (the duplicated one is removed), with their original counterparts.
Prot (Transfer)	5	The target singing voices protected by SongBsAb for building 5 Defended Output (Transfer) pairs, with their original counterparts.
Special	3	Each pair consists of two original singing voices from two singers with opposite genders. If a participant fails to choose *different* for any of them, we exclude all his/her submissions.

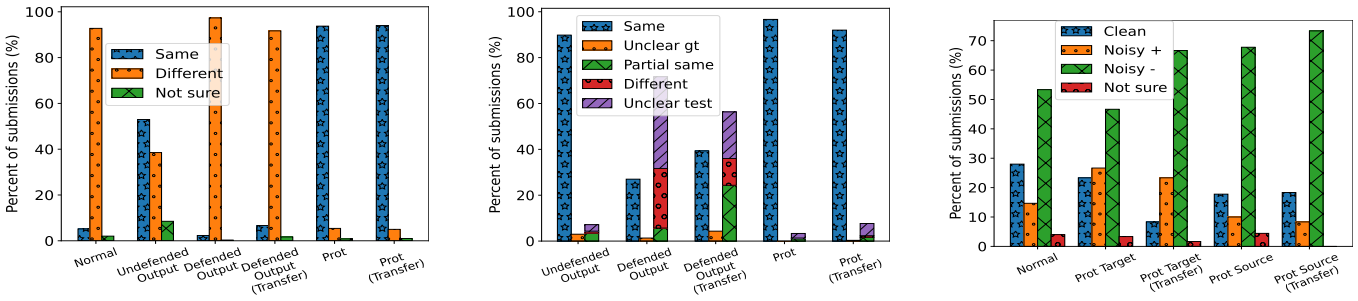
D. Run Time and Efficiency Analysis

The optimization process of SongBsAb took an average of 287 seconds (0.287 seconds per iteration $\times$ 1000 iterations) using an NVIDIA RTX 2080Ti GPU, comparable to the baselines. Given the relatively short runtime and that SongBsAb is applied offline before song release, thus holding a minimal real-time requirement and a high tolerance for runtime, we regard SongBsAb as a computationally efficient toolkit.

E. Human Study

To further confirm the effectiveness and harmlessness of SongBsAb in practice, we conduct a human study as subjective evaluation metrics. The human study was approved by the Institutional Review Board (IRB) of our institutes. We design the following 3 tasks for human study in the form of questionnaires on Credamo [102], an online opinion research questionnaire completion platform. Here we report the results on the Chinese dataset OpenSinger while the results on the English dataset NUS-48E are similar (cf. Appendix M in [2]).

Low-quality answers filtering. We set up special questions as concentration tests. Each task contains 3 different questions inserted at random positions, and each question is designed to



(a) Identify Singer (b) Identify Lyric (c) Clean or Noisy

Fig. 6: Results of human study. "Noise +" and "Noise -" denote the answers "*noisy w/ influence*" and "*noisy w/o influence*".

TABLE VI: Pairs of singing voices for human study task 2.

Pair Name	No.	Description
Undefended Output	10	A ground-truth voice is an original source singing voice and its test voice is an SVC-covered output using the lyrics of the ground-truth voice, where SongBsAb is disabled.
Defended Output	5	The SVC-covered output in each of 5 randomly selected Undefended Output pairs is replaced by another SVC-covered output, where SongBsAb is enabled and uses the same lyric encoder as the SVC model.
Defended Output (Transfer)	5	They are built the same as Defended Output pairs except that SongBsAb uses different lyric encoders from the SVC model.
Prot	5	The source singing voices protected by SongBsAb for building 5 Defended Output pairs, with their original counterparts as ground-truth voices.
Prot (Transfer)	5	The source singing voices protected by SongBsAb for building 5 Defended Output (Transfer) pairs, with their original counterparts as ground-truth voices.
Special	3	Each pair consists of two original singing voices in Chinese and English, respectively. If a participant fails to choose *different* for any of them, we exclude all his/her submissions.

be trivial and tailored for each task. Details refer to TABLE V, TABLE VI, and Task-3.

Participants. We recruited 120 participants (after filtering) for each task, and each participant can only participate in one task, resulting in 360 participants. We restricted participants to be within China. Overall, the participants come from 27 provinces and 112 cities in China, offering a reasonable representation.

Spent Time. Participants have adequate time to review each sample and complete the whole task without any time restriction. Statistically, they spent 20 ± 9, 17 ± 9, and 9 ± 6 minutes for Task 1, Task 2, and Task 3, respectively. In contrast, filtered participants by special questions spent 8 ± 5, 7 ± 3, and 6 ± 3 minutes, indicating a positive correlation between spent time and answer quality.

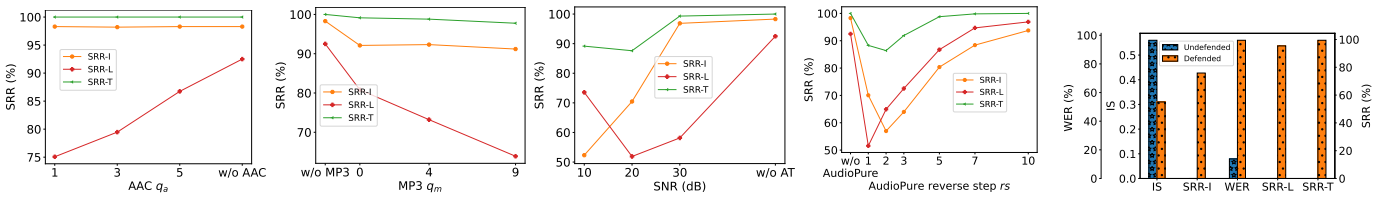
Task 1: identify singer. To evaluate SongBsAb's effectiveness in disrupting identity, participants are asked to tell whether each pair of singing voices is sung by the same singer, with options: *same*, *different*, or *not sure*. We randomly created 37 pairs; see TABLE V for details.

The results are shown in Fig. 6a. Over 92% of participants choose *different* for Normal pairs, confirming the quality of submissions. By contrast, most participants choose *same* for Undefended Output pairs, demonstrating the identity conversion capacity of SVC models. Remark that it is reasonable that a few participants choose *different* for Undefended Output pairs, as when humans consecutively listen to the undefended SVC-covered and original target singing voices, they are conservative in considering them as sung by the same singer, consistent with previous human studies [59], [60]. Remarkably, 97% and 91% of participants choose *different* for Defended Output and Defended Output (Transfer) pairs, 58% and 52% higher than that of the undefended counterparts, respectively. It indicates that SongBsAb is very effective for disrupting the target singer's identity in SVC-covered singing voices, even using different identity encoders. More than 93% of participants choose *same* for Prot and Prot (Transfer) pairs, confirming the harmlessness of SongBsAb on preserving the singer's identity in the protected singing voices.

Task 2: identify lyric. To evaluate the effectiveness of SongBsAb for disrupting lyric, participants are asked to tell if a ground-truth voice and a test voice contain the same lyrics. We provide 5 options: *same*, *partially same*, *different*, *unclear ground-truth*, and *unclear test*, where the first three options denote intelligible lyrics, and "unclear" in other options means that lyrics are too vague to be recognized, i.e., unintelligible. We randomly build 30 pairs; see TABLE VI for details.

The results are shown in Fig. 6b. Nearly 90% of participants choose *same* for Undefended Output pairs, confirming the capacity of SVC models for preserving the lyrics. By contrast, more than 71% and 56% of participants believe that the SVC-covered singing voices in Defended Output and Defended Output (Transfer) pairs contain either unclear or (partially) different lyrics from the ground-truth ones, much higher than that of the undefended counterparts (7%). It confirms the effectiveness of SongBsAb for disrupting lyrics in SVC-covered singing voices. Moreover, over 92% of participants choose *same* for Prot and Prot (Transfer) pairs, indicating the harmlessness of SongBsAb on preserving the lyrics in the protected singing voices.

Task 3: clean or noisy. The above two tasks have confirmed the harmlessness of SongBsAb on preserving the identity and lyrics in protected singing voices, respectively. This task performs a much stricter study by asking participants if a given song contains any background noise and if so, how the noise influences their enjoyment of the song, provided with 4 options: *clean*, *noisy w/ influence*, *noisy w/o influence*, and *not sure*. We randomly select 5 normal songs, 5 protected source songs and 5 protected target songs. Among 5 target songs, 3 (resp. 2) songs are crafted using the same (resp.



(a) AAC (b) MP3 (c) AT (d) AudioPure (e) Optimization-based

Fig. 7: Robustness of SongBsAb against transformation-based and optimization-based adaptive adversaries

different) identity encoders than the SVC model, denoted by "Prot Target" (resp. "Prot Target (Transfer)"). Similarly, the 5 source songs consist of 3 "Prot Source" songs and 2 "Prot Source (Transfer)" songs. We remark that these songs contain the backing tracks since in practice, singing voices are usually accompanied by backing tracks. We additionally insert 3 silent audios with zero magnitude as the concentration test. If a participant didn't choose *clean* or *not sure* for any of silent audios, we exclude all his/her submissions.

The results are shown in Fig. 6c. Although the number of *clean* answers of four types of adversarial songs decreases compared to the normal songs, a large majority of them do not influence the perception and enjoyment. This demonstrates that SongBsAb can maintain the song quality and enjoyment of protected songs in practice.

F. Robustness of SongBsAb

1) Robustness against Adaptive Adversaries: Here we consider adaptive adversaries who know and attempt to bypass SongBsAb. We design three types of adversaries, i.e., transformation-based and optimization-based adversaries by modifying singing voices, and fine-tuning-based adversaries by modifying SVC models.

Transformation-based adversaries. The adversary preprocesses protected singing voices via some transformations before SVC. We consider three typical methods in the audio domain: AAC compression (AAC) [62], MP3 compression (MP3) [62], and Audio Turbulence (AT) [61], and one recent advanced method AudioPure [103]. AAC and MP3 perform different speech compression schemes controlled by the compression quality parameters q_a and q_m, respectively. AT adds white Gaussian noise to each voice within a pre-defined SNR limit (cf. § V-A). We set q_a of AAC as 1, 3 and 5, q_m of MP3 as 0, 4 and 9, and the SNR of AT as 10, 20 and 30 dB, following the setting in [62]. AudioPure first adds noise to the input voice and then runs a reverse process with rs reverse steps to recover the purified voice from the noisy one. We set rs to 1, 2, 3, 5, 7, and 10, the same as [103].

The results are shown in Fig. 7a–Fig. 7d, where the SVC success reduction rate (SRR) decreases compared to no transformation, indicating the reduction of the prevention effect of SongBsAb. However, regardless of the preprocessing methods and their specific parameters, the SRR-I, SRR-L, and SRR-T are larger than 52%, 51%, and 87%, respectively. These demonstrate the robustness of SongBsAb against transformation-based adversaries. There are two main reasons. Firstly, they transform singing voices without any guidance, so although interfering with perturbations, they may just push the target and source singing voices towards another different singer and lyrics rather than the target singer and source lyrics. Thus, the SVC success rate is still low. Secondly, transformations have side-effects on the quality of singing voices [30], [62], [104] by injecting noises (AT and AudioPure) or lossily compressing (AAC and MP3). SVC models accepting low-quality inputs tend to produce poor-quality outputs with low target identity similarity and high source lyric WER.

Optimization-based adversaries. The adversary tries to restore the features of the expected lyrics and the target singer's identity from protected singing voices by applying SongBsAb in a "reverse" direction. The key challenge is to determine the "reverse" direction. For lyrics, the adversary can use a Text-to-Speech tool (we use the iFlytek TTS [105]) to craft a voice with the expected lyrics, which replaces the voice χ in f_Φ (cf. § IV-C) as the "reverse" direction. However, since adversaries cannot acquire the target singer's clear singing voices (otherwise they can be directly used for SVC), the "reverse" direction for the target singer is unknown. We instead increase adversaries' capacity by assuming that they can probe queries to a speaker recognition system enrolled by the target singer and use recognition scores to guide the optimization. Specifically, we use the speaker recognition system XV (cf. § V-A) and natural evolution strategy (NES) to estimate the gradient of the loss f^{T}_{Θ}. The same as [59], [30], the parameter `samples_per_draw` and the number of iterations of NES are set to 50 and 1,000, respectively, thus the number of total queries is $50 \times 1,000 = 50,000$.

The results are shown in Fig. 7e, where IS denotes the (cosine) identity similarity. SongBsAb can still reduce the overall singing voice conversion success rate by 99% (SRR-T is 99%), indicating that the optimization-based adversary fails to circumvent SongBsAb on almost all the protected singing voices. The reason is two-fold. Firstly, while adversaries may obtain a relatively precise reverse direction for expected lyrics, the estimated gradient by NES is less informative than the exact gradient, preventing an accurate reverse direction for target identity. Secondly, although the optimization-based adversaries outperform the transformation-based ones by using guidelines, they still suffer from degrading input quality since perturbations are applied twice to singing voices.

Fine-tuning-based adversaries. This adversary fine-tunes the identity and lyric encoders, aimed to produce identity and lyric

TABLE VII: Robustness of SongBsAb against fine-tuning adversaries in terms of SRR

E \ D	w/o FT	FT
w/o FT	100%	
FT (f_1)	83.7%	76.6%
FT (f_1+f_2)	91.2%	68.9%

E: Encoders; D: Decoders
FT: Fine-tuning

TABLE VIII: Over-the-air robustness of SongBsAb in terms of SRR

L \ M	iPhone	OPPO
JBL	90%	90%
Xiaodu	92%	91%
iPad	89%	85%

L: Loudspeakers
M: Microphones

features for the protected singing voices that are close to that of original singing voices. We design two different fine-tuning approaches: $\arg\min_{\vartheta} f_1$ and $\arg\min_{\vartheta} f_1 + f_2$, where $\vartheta \in \{\Theta, \Phi\}$ is the encoder, $f_1 = \texttt{Dist}(\vartheta(x^0), \vartheta(x))$, and f_2 is the encoder's original training loss, which intends to preserve the functionality of recognizing singers and lyrics of the identity and lyric encoders, respectively. With the fine-tuned identity and lyric encoders, the adversary may or may not fine-tune the decoder to align with encoders' modification, leading to four types of adversaries. For both encoders and the decoder, we adopt their respective official training settings, since they have been optimized and tailored towards more effective training.

The results are shown in TABLE VII. Unsurprisingly, regardless of the loss for fine-tuning encoders, additional fine-tuning of the decoder further reduces the effectiveness of SongBsAb compared to only fine-tuning encoders. This is because the fine-tuned decoder aligns with the modified feature space of fine-tuned encoders. When the decoder is not fine-tuned, $f_1 + f_2$ is less effective in bypassing SongBsAb than f_1. The reason is that f_2 introduces larger modifications to the feature space for preserving the functionality destroyed by f_1 which forcefully pulls together two distinct features, while the un-fine-tuned decoder is trained to cooperate well with the unmodified feature space. When the decoder is fine-tuned, $f_1 + f_2$ exhibits a larger impact on SongBsAb than f_1, due to the more functional feature space achieved by $f_1 + f_2$, based on which the decoder is more effective for SVC.

Under the strongest adversary, SongBsAb still achieves over 68% SRR, indicating that fine-tuning adversaries cannot render SongBsAb to be ineffective. This is due to the continuous and large input and output spaces of generative SVC models, so even less effective perturbation in inputs still directly affects the covered songs. It is consistent with the finding that adversarial training has an upper bound of defeating adversarial examples [106].

2) Over-the-air robustness of SongBsAb: The adversary may obtain singing voices by recording them using microphones, during which perturbations used for prevention may be disrupted [60], [30]. We evaluate the robustness of SongBsAb by playing singing voices via 3 loudspeakers (JBL clip3 portable loudspeaker, Xiaodu smart speaker, and iPad Pro 10.5) and recording the air channel-transmitted singing voices using 2 microphones (iOS iPhone 15 Plus and Android OPPO), leading to 6 diverse combinations of hardware settings. We randomly select 100 pairs of target singers and protected source singing voices with 3 protected target singing voices per target singer. To ensure the quality of recorded singing voices for SVC [59], [107], [108], [109], [110], we conduct experiments in a relatively quiet room with air-conditioner noise, and set the distance between microphones and loudspeakers to 2 meters. The results are shown in TABLE VIII. Though the effectiveness of SongBsAb varies slightly with loudspeakers and microphones, SongBsAb achieves at least 85% SRR regardless of devices, confirming the over-the-air robustness of SongBsAb. Over-the-air transmissions introduce hardware distortion, ambient noise, and reverberation to perturbations [60], [110], which can be regarded as transformations, so SongBsAb's over-the-air robustness shares the reason with robustness against transformation-based adversaries.

VI. Discussion

In this section, we discuss the limitations of SongBsAb, insights, and potential future works motivated by this work.

Copyrights of melodies. SongBsAb directly protects the civil rights of target singers and the copyright of lyrics, but only indirectly safeguards melody copyright by discouraging the sharing of SVC-covered songs. To directly disrupt the melody, we crafted perturbations by maximizing the root mean square error of pitch features between the protected and original singing voices. However, significant deviation in pitch features leads to noticeable changes in melody, likely due to the tight mapping between singing voices and low-dimension pitch features. Future work should study extracting pitch features and the correlation with singing voices to enable pitch disruption in SVC while preserving melody in the protected songs.

Real-world usage of SongBsAb. While we have confirmed the effectiveness of SongBsAb to protect both Chinese and English songs with different singer genders (cf. Appendix E of [2]), accents, and voice types (cf. TABLE III), and different song genres (cf. Appendix F of [2]), tempos, and pitches (cf. TABLE III), songs can be in other languages and diverse in other aspects, e.g., degree of pitch fluctuations and instrument types and loudness. We do not consider these factors due to the unavailability of suitable datasets, which should be examined in future to better understand the applicability of SongBsAb in the real world.

Arm race between adaptive adversaries and defenders. The defender, aware of transformations in § V-F1, can apply them to intermediate singing voices at each iteration such that the protected singing voices gain sufficient robustness against the transformations [111], [112]. However, this cannot protect previously released songs that have been kept by adversaries [113]. Adversaries might also try to bypass SongBsAb with a binary detector to identify and discard protected singing voices. However, this detector will inevitably produce false negatives, allowing SongBsAb to succeed, even with a small proportion of protected voices (cf. Appendix D of [2]). Additionally, the defender can counter this detection by incorporating the detector's outputs into the loss to deceive both the encoders and the detector [114], [115], [116].

Harmlessness of perturbations. We propose using the backing track as an additional masker, enhancing the hiding capacity of perturbations compared to previous methods that relied solely on the voice as the masker. This indicates that harmlessness can be improved by utilizing unique elements of singing voices versus ordinary speech. We believe this insight is valuable for future research. For example, different backing tracks may have varying masking capacities, so future studies could explore this correlation to identify or design backing tracks that more effectively conceal perturbations. While SongBsAb restricts perturbations to stay below the hearing threshold for *all* frequencies, an alternative is to position perturbations outside the human hearing range (20-20 kHz [117]) using methods like the ultrasound transformation model [118] or imposing larger penalties on audible frequency bands [11]. This approach eliminates the constraint on perturbation magnitude [118], but it may also reduce the frequency information available for optimizing prevention losses. We leave this as interesting future work.

Other song cover techniques. Other techniques for automated song covers include singing voice synthesis (SVS) [119], which uses musical scores with lyrics and the target singer's voices to generate a performance as if sung by the target. The main difference between SVC and SVS is how melody and lyric information are provided. While our identity disruption methods can be adapted for SVS, future work should focus on lyric and melody disruption for SVS, potentially using adversarial examples from natural language processing [120].

Non-technical efforts for rights protection. Complex music copyrights and civil rights protection also require collaborative non-technical efforts. Firstly, applying and enforcing legal frameworks is challenging due to outdated definitions of infringement in the generative AI era and difficulties in cross-jurisdictional protection. Regulatory bodies must address these issues by collaborating with industry stakeholders to update requirements and promote international treaties. To enhance the effectiveness of SongBsAb, song owners should minimize unprotected songs. For individual-covered songs, they need to conduct regular inspections and monitoring on various platforms to combat unlicensed covers. For songs released before SongBsAb, they should "patch" them on all controlled platforms and fight piracy and the secondary distribution that leads to "unpatchable" songs.

Impact on authorized SVC. SongBsAb does not hinder authorized SVC, as song owners can maintain both unaltered and perturbed versions of a song, providing the unaltered version to authorized SVC entities. To prevent leaks of unaltered songs that could undermine SongBsAb, song owners can embed entity-specific watermarks into songs for traceability, allowing them to identify the source of leakage and seek compensation.

VII. Conclusion

We have proposed SongBsAb, the first proactive approach that can be utlized by song owners to mitigate SVC-based illegal song covers for protecting their copyright and singers' civil rights. SongBsAb features a dual prevention, preventing singing voices from being used as the source and target singing voices in SVC, by perturbing singing voices prior to their release with a gender-transformation loss and a high/low hierarchy multi-target loss; preserves the quality of singing voices with a refined simultaneous masking loss; exhibits strong transferability to unknown SVC models with a transferability enhancement loss and encoder ensemble; and possesses robustness in over-the-air scenario and against adaptive adversaries. We make the first significant step towards coping with illegal automated song covers. Our open-source code, audio samples, and discussions on future works can foster researchers in exploring this direction further.

Acknowledgments

We thank our anonymous discussion lead and the reviewers for their constructive feedback and suggestions.

This research is partially supported by Joint Funds of the National Natural Science Foundation of China (Grant No. U22A2036), National Natural Science Foundation of China (No. 62102202), CAS Project for Young Scientists in Basic Research (Grant No. YSBR-040), ISCAS New Cultivation Project (Grant No. ISCAS-PYFX-202201), and ISCAS Basic Research (Grant No. ISCAS-JCZD-202302).

It is also supported by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008). It is also supported by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

References

[1] "Official website of SongBsAb," https://sites.google.com/view/songbsab, 2024.

[2] G. Chen, Y. Zhang, F. Song, T. Wang, X. Du, and Y. Liu, "SongBsAb: A Dual Prevention Approach against Singing Voice Conversion based Illegal Song Covers," *CoRR*, vol. abs/2401.17133, 2024. [Online]. Available: https://arxiv.org/abs/2401.17133

[3] W. Huang, L. P. Violeta, S. Liu, J. Shi, Y. Yasuda, and T. Toda, "The singing voice conversion challenge 2023," *CoRR*, vol. abs/2306.14422, 2023.

[4] Rita Liao, "China has its DrakeGPT moment as AI singer goes viral," https://techcrunch.com/2023/05/10/china-ai-singer-stefanie-sun, 2023.

[5] "Bilibili: China's largest user-generated video streaming site," https://www.bilibili.com, 2010.

[6] "Heart on My Sleeve: AI-generated song mimicking Drake and The Weeknd submitted for Grammy consideration," https://www.independent.co.uk/arts-entertainment/music/news/drake-and-weeknd-ai-song-heart-on-my-sleeve-b2406902.html, 2023.

[7] Global Times, "Beware: AI-generated singing voices pleasing to the ear," https://www.globaltimes.cn/page/202305/1290425.shtml, 2023.

[8] M. E. Salazar, M. Dean, N. Moran, J. Morelle, and R. Wittman, "No AI FRAUD Act," https://www.congress.gov/bill/118th-congress/house-bill/6943/text/ih, 2024.

[9] Ashley King, "The ELVIS Act Has Officially Been Signed Into Law — First State-Level AI Legislation In the US," https://www.digitalmusicnews.com/2024/03/21/elvis-act-signed-tennessee, 2024.

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

[11] Z. Yu, S. Zhai, and N. Zhang, "Antifake: Using adversarial audio to prevent unauthorized speech synthesis," in *CCS*, 2023.

[12] C. Huang, Y. Y. Lin, H. Lee, and L. Lee, "Defending your voice: Adversarial attack on voice conversion," in *SLT*, 2021.

[13] Y. Wang, H. Guo, G. Wang, B. Chen, and Q. Yan, "Vsmask: Defending against voice synthesis attack via real-time predictive perturbation," in *WiSec*, 2023.

[14] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-cloak: Intelligibility-, naturalness- & timbre-preserving real-time voice anonymization," in *USENIX Security*, 2023.

[15] M. Chen, L. Lu, J. Wang, J. Yu, Y. Chen, Z. Wang, Z. Ba, F. Lin, and K. Ren, "Voicecloak: Adversarial example enabled voice de-identification with balanced privacy and utility," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2023.

[16] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples," in *ICML*, 2023.

[17] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," in *USENIX Security*, 2023.

[18] Z. Li, N. Yu, A. Salem, M. Backes, M. Fritz, and Y. Zhang, "Unganable: Defending against gan-based face manipulation," in *USENIX Security*, 2023.

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[20] E. C. Carterette, "The science of the singing voice," 1989.

[21] I. R. Titze and D. W. Martin, "Principles of voice production," 1998.

[22] N. Cook, *Music: A very short introduction*. Oxford University Press, 2021.

[23] M. Redon, "Auditory Masking: Using Sound to Control Sound," https://www.ansys.com/blog/what-is-auditory-masking, 2023.

[24] Y. Lin, W. H. Abdulla *et al.*, "Audio watermark," *Springer, Cham.*, vol. 3, no. 319, p. 07974, 2015.

[25] Y. Qin, N. Carlini, G. W. Cottrell, I. J. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *ICML*, 2019.

[26] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *NDSS*, 2019.

[27] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *S&P*, 2021.

[28] X. Wang, J. Ren, S. Lin, X. Zhu, Y. Wang, and Q. Zhang, "A unified approach to interpreting and boosting adversarial transferability," in *ICLR*, 2021.

[29] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[30] G. Chen, Y. Zhang, Z. Zhao, and F. Song, "QFA2SR: query-free adversarial transfer attacks to speaker recognition systems," in *USENIX Security*, 2023.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[32] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2021.

[33] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[34] B. Sha, X. Li, Z. Wu, Y. Shan, and H. Meng, "Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion," in *ICASSP*, 2024.

[35] Y. A. Li, A. Zare, and N. Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Interspeech*, 2021.

[36] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, 2016.

[37] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *ICASSP*, 2018.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[39] F. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[40] "Singing Voice Conversion based on Whisper & neural source-filter BigVGAN," https://github.com/PlayVoice/lora-svc, 2022.

[41] "Grad-SVC based on Grad-TTS from HUAWEI Noah's Ark Lab," https://github.com/PlayVoice/Grad-SVC, 2023.

[42] H. Tamaru, S. Takamichi, N. Tanji, and H. Saruwatari, "Jvs-music: Japanese multispeaker singing-voice corpus," *CoRR*, vol. abs/2001.07044, 2020.

[43] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *INTERSPEECH-Eurospeech*, 2005.

[44] E. C. Carterette, "The science of the singing voice," 1989.

[45] W. Zhou, F. Zhang, Y. Liu, W. Guan, Y. Zhao, and H. Qu, "Zero-shot sing voice conversion: built upon clustering-based phoneme representations," *CoRR*, vol. abs/2409.08039, 2024.

[46] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with A large-scale corpus," in *MM '21: ACM Multimedia Conference 20 - 24, 2021*, 2021.

[47] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013.

[48] X. Li, S. Liu, and Y. Shan, "A hierarchical speaker representation framework for one-shot singing voice conversion," in *Interspeech*, 2022.

[49] "Copyright Law of the People's Republic of China," https://www.gov.cn/guoqing/2021-10/29/content_5647633.htm, 2021.

[50] "Copyright Law United States and Related Laws Contained in Title 17 of the United States Code," https://www.copyright.gov/title17/title17.pdf, 2022.

[51] "UK Copyright Law," https://copyrightservice.co.uk/_f/5716/9839/4538/edupack.pdf, 2022.

[52] "Civil Code of the People's Republic of China," https://www.gov.cn/xinwen/2020-06/01/content_5516649.htm, 2020.

[53] "Defamation and Privacy Law in England & Wales," https://www.carter-ruck.com/law-guides/defamation-and-privacy-law-in-england-wales, 2013.

[54] "United States Defamation Laws," https://constitution.congress.gov/browse/essay/amdt1-7-5-7/ALDE_00013808, 1791.

[55] "United States Privacy Laws," http://www.rbs2.com/privacy.htm, 1997.

[56] "Unfair Competition Law, California Business and Professions Code sections 17200–17209 ("UCL")," https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=BPC&sectionNum=17200., 1993.

[57] "Anti-Unfair Competition Law of the People's Republic of China," https://www.gov.cn/xinwen/2017-11/05/content_5237325.htm, 2022.

[58] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *S&P*, 2017.

[59] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real Bob? adversarial attacks on speaker recognition systems," in *S&P*, 2021.

[60] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[61] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *USENIX Security*, 2018.

[62] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, F. Wang, and J. Wang, "Towards understanding and mitigating audio adversarial examples for speaker recognition," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[63] M. Chen, L. Lu, Z. Ba, and K. Ren, "Phoneytalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition," in *INFOCOM*, 2022.

[64] J. Deng, Y. Chen, and W. Xu, "Fencesitter: Black-box, content-agnostic, and synchronization-free enrollment-phase attacks on speaker recognition systems," in *CCS*, 2022.

[65] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *ICLR*, 2021.

[66] S. Fu, F. He, Y. Liu, L. Shen, and D. Tao, "Robust unlearnable examples: Protecting data privacy against adversarial learning," in *ICLR*, 2022.

[67] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Odyssey*, 2010.

[68] Fiveable Inc. (2024) Bounded loss functions. [Online]. Available: https://library.fiveable.me/key-terms/theoretical-statistics/bounded-loss-functions

[69] A. Akbari, M. Awais, M. Bashar, and J. Kittler, "How does loss function affect generalization performance of deep learning? application to human age estimation," in *ICML*, 2021.

[70] M. Akhtar, M. Tanveer, and M. Arshad, "Hawkeye: Advancing robust regression with bounded, smooth, and insensitive loss function," *CoRR*, vol. abs/2401.16785, 2024.

[71] Y. Ge, L. Zhao, Q. Wang, Y. Duan, and M. Du, "Advddos: Zero-query adversarial attacks against commercial speech recognition systems," *IEEE Trans. Inf. Forensics Secur.*, 2023.

[72] H. yi Lee. (2020) The most popular acoustic features. [Online]. Available: http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/ASR%20(v12).pdf

[73] H. F. Pardede, V. Zilvan, D. Krisnandi, A. Heryana, and R. B. S. Kusumo, "Generalized filter-bank features for robust speech recognition against reverberation," in *Proceedings of the 6th International Conference on Computer, Control, Informatics and its Applications*, 2019, pp. 19–24.

[74] J. R. Stuart, "The psychoacoustics of multichannel audio," in *Audio Engineering Society Conference: UK 11th Conference: Audio for New Media (ANM)*, 1996.

[75] "Variational Inference with adversarial learning for end-to-end Singing Voice Conversion based on VITS," https://github.com/PlayVoice/whisper-vits-svc, 2022.

[76] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.

[77] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020.

[78] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker recognition: Modular or monolithic?" in *Interspeech*, 2019.

[79] "AutoSpeech: Neural Architecture Search for Speaker Recognition," https://github.com/VITA-Group/AutoSpeech, 2020.

[80] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[81] S. Ding, T. Chen, X. Gong, W. Zha, and Z. Wang, "Autospeech: Neural architecture search for speaker recognition," in *Interspeech*, 2020.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[83] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017, pp. 2616–2620.

[84] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[85] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *CoRR*, vol. abs/2012.06659, 2020.

[86] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," in *ICLR*, 2023.

[87] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[88] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, M. Meila and T. Zhang, Eds., 2021.

[89] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. Signal Process.*, 2022.

[90] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *ICASSP*, 2014.

[91] (2019) REAPER: Robust Epoch And Pitch EstimatoR. [Online]. Available: https://github.com/google/REAPER

[92] X. Li, S. Liu, and Y. Shan, "A hierarchical speaker representation framework for one-shot singing voice conversion," in *Interspeech*, 2022.

[93] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, 2017, pp. 4835–4839.

[94] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, 2019.

[95] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[96] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020.

[97] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020.

[98] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[99] Y. Xiang, G. Hua, and B. Yan, *Digital audio watermarking: fundamentals, techniques and challenges*. Springer, 2017.

[100] "Pydub lets you do stuff to audio in a way that isn't stupid," https://github.com/jiaaro/pydub, 2011.

[101] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, 1983, pp. 804–807.

[102] "The Credamo platform," https://www.credamo.world, 2017.

[103] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," in *ICLR*, 2023.

[104] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *ASIACCS*, 2020.

[105] iFlytek, "The text-to-speech API provided by iFlytek." https://global.xfyun.cn/products/text-to-speech.

[106] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[107] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *CCS*, 2020, pp. 1121–1134.

[108] Z. Qin, X. Zhang, and S. Li, "A robust adversarial attack against speech recognition with uap," *High-Confidence Computing*, vol. 3, no. 1, p. 100098, 2023.

[109] L. Schönherr, T. Eisenhofer, S. Zeiler, T. Holz, and D. Kolossa, "Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems," in *ACSAC*, 2020, pp. 843–855.

[110] T. Chen, L. Shangguan, Z. Li, and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *NDSS*, 2020.

[111] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 274–283.

[112] R. S. Zimmermann, W. Brendel, F. Tramèr, and N. Carlini, "Increasing confidence in adversarial robustness evaluations," in *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, 2022, pp. 13 174–13 189.

[113] R. Hönig, J. Rando, N. Carlini, and F. Tramèr, "Adversarial perturbations cannot reliably protect artists from generative ai," *CoRR*, vol. abs/2406.12027, 2024.

[114] N. Carlini and D. A. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[115] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *CoRR*, vol. abs/1902.06705, 2019.

[116] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *NeurIPS*, 2020, pp. 1633–1645.

[117] C. Sujatha, *Fundamentals of Acoustics*, 2023, pp. 161–217.

[118] X. Li, C. Yan, X. Lu, Z. Zeng, X. Ji, and W. Xu, "Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time," in *NDSS*, 2024.
[119] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *AAAI*, 2022.
[120] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible NLP attacks," in *S&P*, 2022.

APPENDIX

A. Summary of Related Works that Exploiting Adversarial Examples for Good

TABLE IX summarizes related works that exploit adversarial examples for beneficial purposes, comparing prior works with our proposed SongBsAb. More detailed discussion refers to § II-C. Remarkably, we use the term "harmlessness" instead of "imperceptibility" in this work when exploiting adversarial examples for good. This is because although protected samples (e.g., voices, images) are imperceptible to the adversary when they have not been used by the adversary in conversion/synthesis/training, the adversary will largely become aware of the protection/prevention when the unexpected output is produced by conversion/synthesis/training.

B. Effectiveness of SongBsAb on the Non-Few-Shot SVC Model StarGANv2

Recall that we focus on few-shot SVC models, which require much fewer computational resources and target singers' voices than non-few-shot models (cf. § III-B), allowing for direct use by individuals without any training process and making them accessible to a broader range of adversaries, thus expanding the potential applications of SongBsAb. We have confirmed the effectiveness of SongBsAb on four different few-shot SVC models. Here, we evaluate SongBsAb against adversaries who have sufficient singing voices of target singers and computational resources such that they can train from scratch or fine-tune SVC models with these singing voices.

Specifically, we consider the promising non-few-shot model StarGANv2 [35] (details refer to TABLE II) and use its style encoder inference mode (in contrast to the mapping network mode). Although the released pre-trained model is trained only using ordinary voices, it exhibits emerging capacities including generalizing to singing conversion [35]. Therefore, we fine-tune the model respectively for each of the 12 singers in the English dataset NUS-48E instead of training from scratch. Note that StarGANv2 does not support Chinese. Other experimental settings are the same as in § V-A.

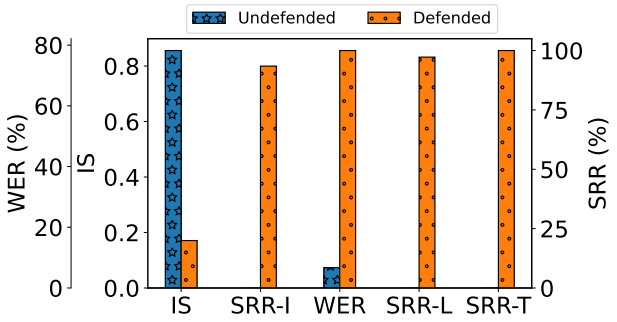


Fig. 8: The prevention effectiveness of SongBsAb against the non-few-shot model StarGANv2.

The results are shown in Fig. 8. The identity similarity (resp. lyric WER) of SVC-covered singing voices is much lower (resp. higher) than that of undefended output singing voices. Overall, SongBsAb is able to achieve nearly 100% SRR-I, SRR-L, and SRR-T. These results demonstrate the effectiveness of SongBsAb against this non-few-shot model. We leave the evaluation of the prevention effectiveness of SongBsAb on more non-few-shot SVC models as future work.

TABLE IX: Comparison between SongBsAb and related works

	Target Model	Purpose	Harmlessness	Transfer ↑	Application
Unlearnable [65]	image recognition (image$^\sharp$ & D$^\natural$)	making data unlearnable	L_∞ norm	✗	preventing unauthorized data exploitation for training
Robust Unlearnable [66]					
§**Glaze** [17]	text-to-image (image & G‡)	style disruption			protecting copyrights of artworks
§**MIST** [16]					
UnGANable [18]	GAN-based face manipulator (image & G)	identity disruption			preventing abuse of biometric data
V-cloak [14]	speaker recognition (voice† & D)		psychoacoustics model		
VoiceCloak [15]			L_∞ norm		
***AttackVC** [12]	speech voice conversion/synthesis (voice & G)				
***VSMask** [13]					
***AntiFake** [11]			frequency penalty & SNR	encoder ensemble	
Our work (**SongBsAb**)	singing voice conversion (voice & G)	identity disruption and lyric disruption	psychoacoustics model (with backing tracks)	FL-IR loss and encoder ensemble	protecting rights of singers and lyrics (direct), and melodies (indirect)

(1) "Transfer ↑": transferability enhancement; image$^\sharp$/voice†: image/voice modality; D$^\natural$/G‡: discriminative/generative models. (2) §/$*$: achieving purposes via artist/speaker style transfer, which is analogous to each other, so their prevention techniques are generally the same, involving pulling artist or speaker style features towards targets. (3) We also experimentally compare SongBsAb with the closet works AttackVC and AntiFake, all of which are of voice modality and target generative models (VSMask is not considered since it is unavailable).