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

Guangke Chen[1], Yedi Zhang[2], Fu Song[3]
Ting Wang[4], Xiaoning Du[5], Yang Liu[6]

1 鹏城实验室 Pengcheng Laboratory
2 NUS National University of Singapore
3 ISCAS 中国科学院软件研究所 Institute of Software Chinese Academy of Sciences
4 Stony Brook University
5 MONASH University
6 NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

AI Generative Music

AI-based New Music Generation

MusicGPT

★ AI-based Automated Song Cover

covers.ai

Jammable

Fineshare

Singify

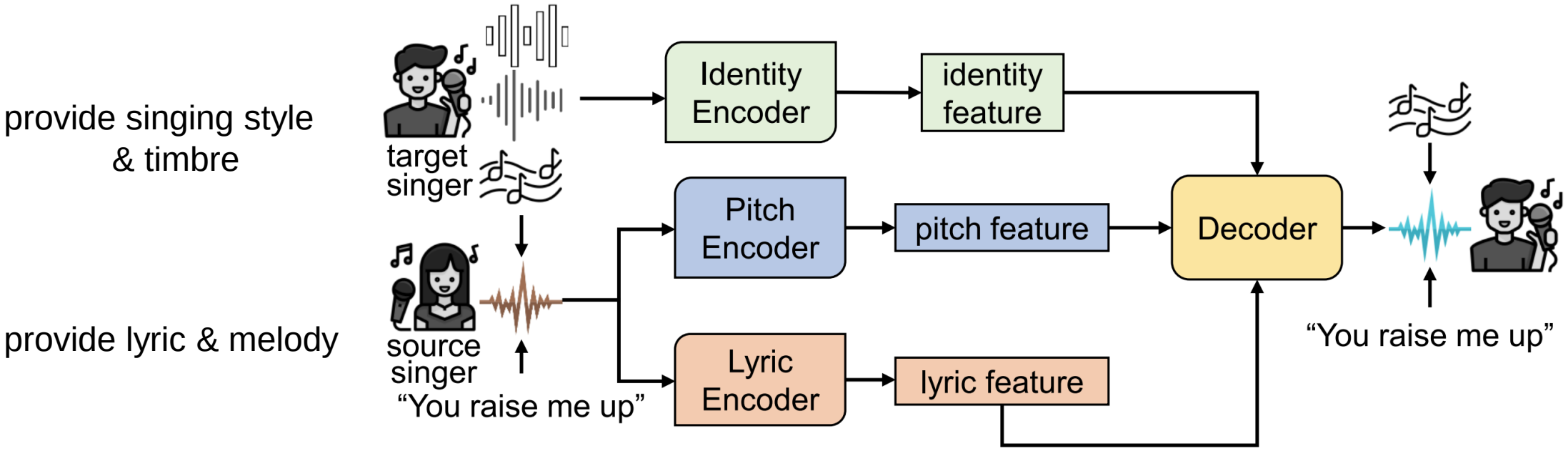
Media.io

TopMediAi

Musicfy

AI-based Automated Song Cover by Singing Voice Conversion (SVC)

- transforms a song's vocal rendition from one singer to another's singing style and timbre while preserving the original lyrics and melody



SVC-based song covers

Singing Voice Conversion (SVC): Challenging Music Industry

■ low entry barriers →
wide spread of AI-based song cover

Popular “AI Sun Yanzi” in China

An artificial intelligence-generated song, which mimics the voices of Drake and The Weeknd with terrifying accuracy, has been submitted for Grammy consideration.

■ **The Impact on Music Industry:**

- Infringement of singers' civil rights over voices & reputation
- Infringement of record companies' rights to release & distribute songs
- Infringement of the copyright of lyrics and melodies
- Erosion of singers' skill competitiveness (rely on for livelihood)
- Unfair competition faced by record companies

What should we do?

- **Reactive Detection:**

reactive detection

Method	SingFake-T02	
	Mixture	Vocals
AASIST	58.12	37.91
Spectrogram+ResNet	51.87	37.65
LFCC+ResNet	45.12	54.88
Wav2Vec2+AASIST	56.75	57.26

low accuracy

Singing Voice Deepfake Detection Challenge [1]

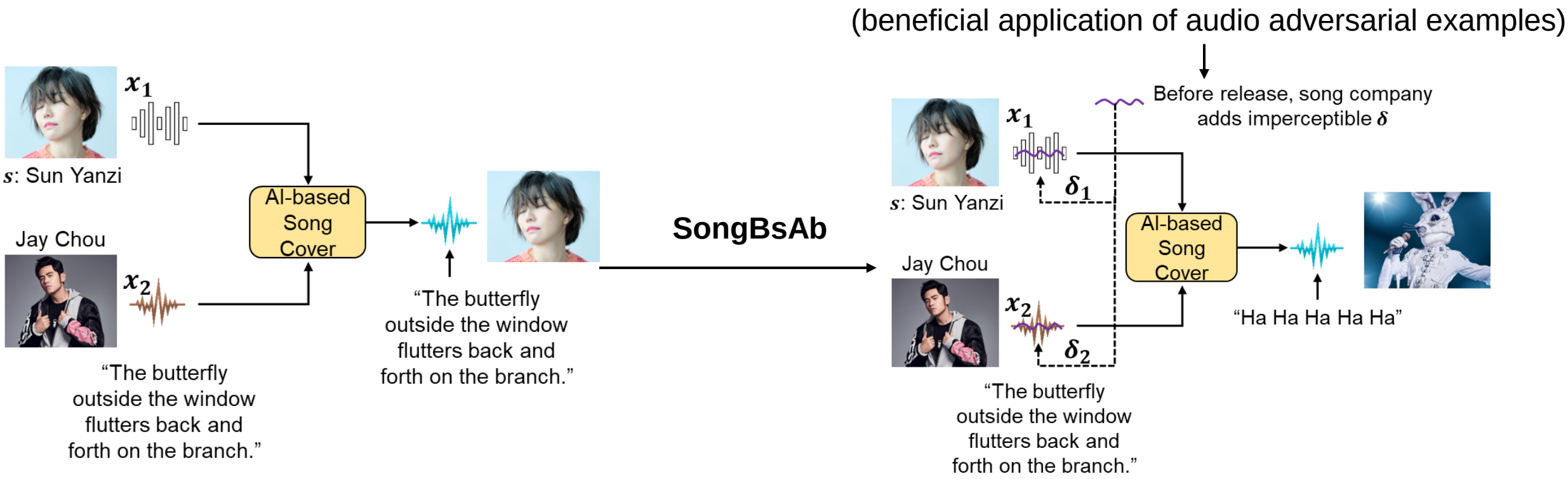
Reactive detection

- Infringement already committed
- High quality, hard to detect
- Large number, inefficient to detect

[1] 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.

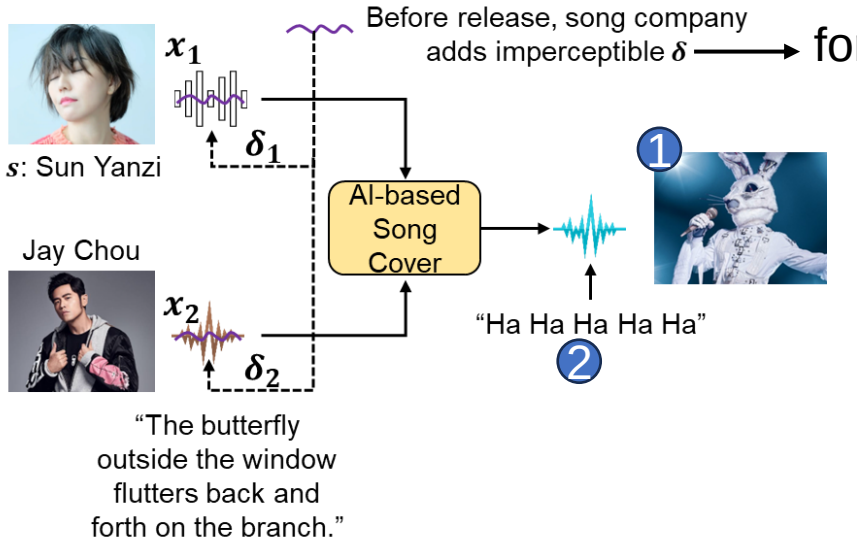
What should we do?

- **Proactive Prevention:**



SongBsAb: Methodology

- **Challenge:** do not know in advance if songs will be used for target or source songs
- **Solution:** Dual Prevention



- ❶**Identity disruption**

Gender-Transformation Loss

$$f_{\text{ID}} = \texttt{Distance}(\Theta(x+\delta), \Theta(s'))$$

Θ singer timbre extractor
s' opposite-gender singers

- ❷**Lyric disruption**

High/Low Hierarchy Loss

$$f_{\text{LD}} = \texttt{Distance}(\Phi_{\text{H}}(x+\delta), \Phi_{\text{H}}(x')) + \texttt{Distance}(\Phi_{\text{L}}(x+\delta), \Phi_{\text{L}}(x'))$$

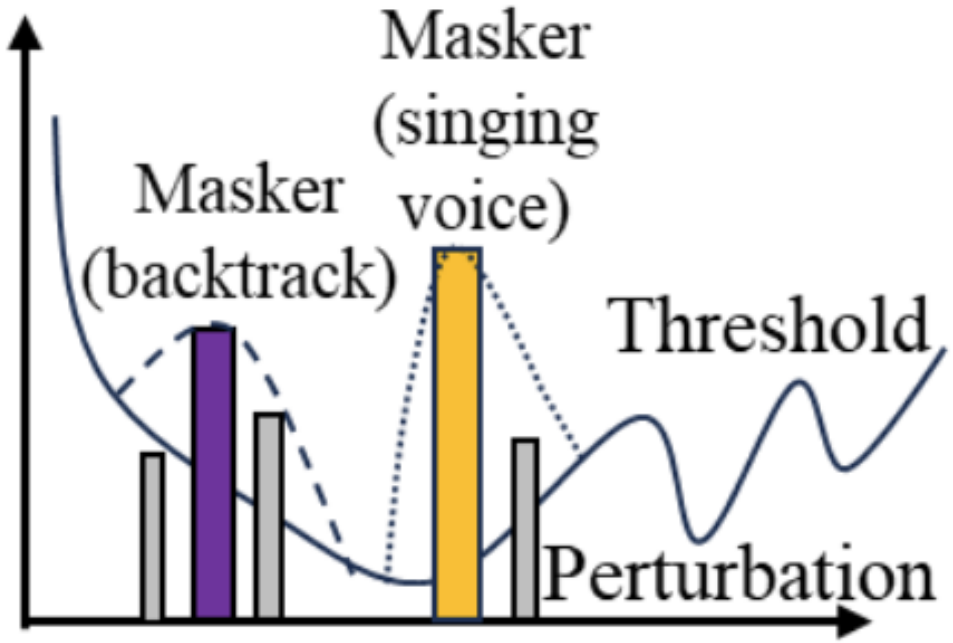
Φ_{H} high-level lyric features
Φ_{L} low-level lyric features
x' A song with different lyrics

$$\arg\min_{\delta} f_{\text{ID}} + f_{\text{LD}}$$

SongBsAb: Methodology

- **Challenge:** high quality requirements of songs

- **Solution:** backing track-refined simultaneous masking



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

SongBsAb: Methodology

- **Challenge:** Transferability to unknown SVC models exploited by adversaries
- **Solution:** encoder ensemble & frame-level interaction reduction loss

- **encoder ensemble:** craft perturbation on multiple local white-box SVC models
- **frame-level interaction reduction loss:**

perturbation interaction [2]: $\frac{\mathbb{E}_i(v(\Omega) + v(\emptyset) - v(\Omega \setminus \{i\}) - v(\{i\}))}{n-1}$

no perturbation

Ω: set of perturbation units

$\emptyset$: no perturbation

$\Omega/\{i\}$: only unit i not perturbed

$\{i\}$: only unit i perturbed

interaction is negatively correlated with transferability ⟶ minimize the interaction loss at frame-level

[2] 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.

SongBsAb: Evaluation

- **Few-shot SVC model:** Lora-SVC, Vits-SVC, Grad-SVC, NeuCo-SVC
- **Dataset:** OpenSinger (Chinese), NUS-48E (English)
- **Metric:** Identity Similarity with target singer; lyric Word Error Rate (WER)
- **Baseline:** Attack-VC [3], AntiFake [4]

[3] C. Huang, et al, “Defending your voice: Adversarial attack on voice conversion,” in SLT, 2021.
[4] Z. Yu, et al, “Antifake: Using adversarial audio to prevent unauthorized speech synthesis,” in CCS, 2023.

y Covered songs w/o prevention

$\tilde{y}$ Covered songs w/ prevention

Dataset	SVC Model	Approach	Prevention Effectiveness			
			Identity Similarity ↓		Lyric WER (%) ↑	
			y	$\tilde{y}$	y	$\tilde{y}$
Open Singer	Lora -SVC	AntiFake	0.54	0.15	13.9	13.2
		AttackVC		0.55		13.9
		AttackVC-W		0.24		13.1
		SongBsAb		**0.05**		**76.1**
	Vits -SVC	AntiFake	0.51	0.15	14.9	15.2
		AttackVC-W		0.26		14.7
		SongBsAb		**0.09**		**90.4**
	Grad -SVC	AntiFake	0.48	0.17	32.1	31.4
		AttackVC-W		0.23		30.9
		SongBsAb		**0.11**		**103.6**
	NeuCo -SVC	AntiFake	0.65	0.33	18.1	20.8
		AttackVC-W		0.28		20.1
		SongBsAb		**0.22**		**86.5**

Dataset	SVC Model	Approach	Prevention Effectiveness			
			Identity Similarity ↓		Lyric WER (%) ↑	
			y	$\tilde{y}$	y	$\tilde{y}$
NUS -48E	Lora -SVC	AntiFake	0.47	0.22	23.3	22.0
		AttackVC-W		0.25		23.1
		SongBsAb		**0.12**		**79.9**
	Vits -SVC	AntiFake	0.48	0.19	18.4	19.4
		AttackVC-W		0.25		19.3
		SongBsAb		**0.12**		**78.4**
	Grad -SVC	AntiFake	0.45	0.24	41.1	41.2
		AttackVC-W		0.24		43.6
		SongBsAb		**0.16**		**94.5**
	NeuCo -SVC	AntiFake	0.59	0.24	22.6	22.7
		AttackVC-W		0.22		21.7
		SongBsAb		**0.16**		**76.6**

SongBsAb: Audio Demo

Input: target singer (style & timbre)

Output w/o SongBsAb

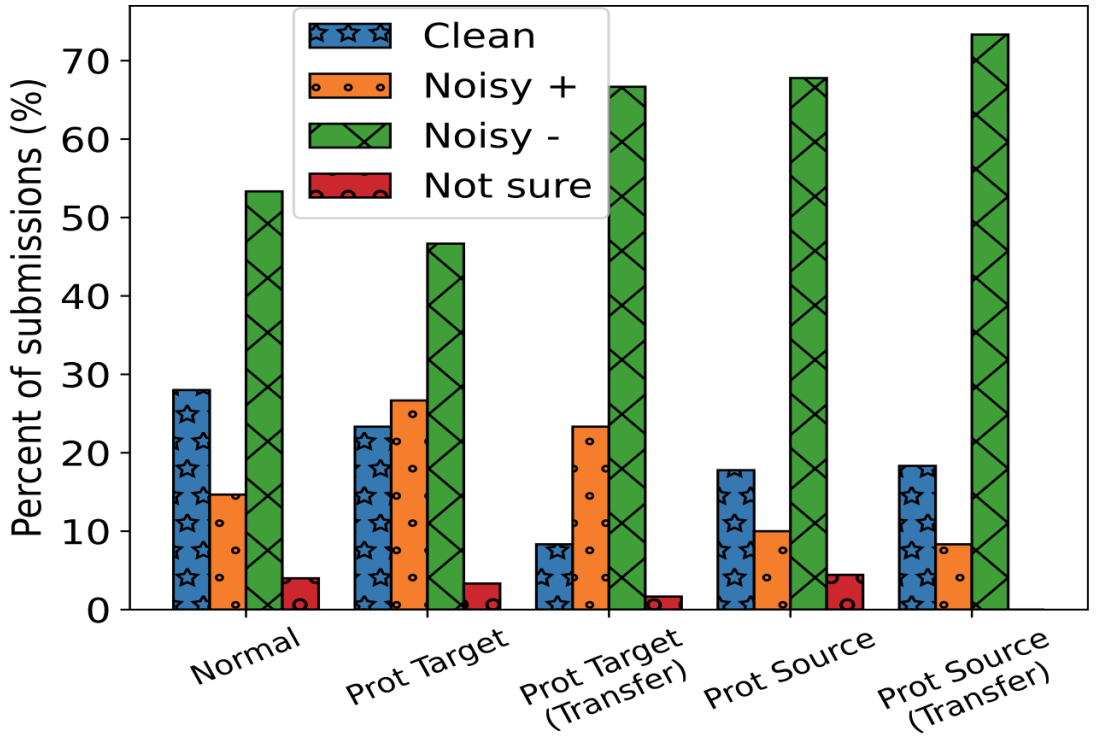
Input: source singer (lyric & melody)

Output w/ SongBsAb

SongBsAb: Evaluation

- Impact on song quality and enjoyment experience

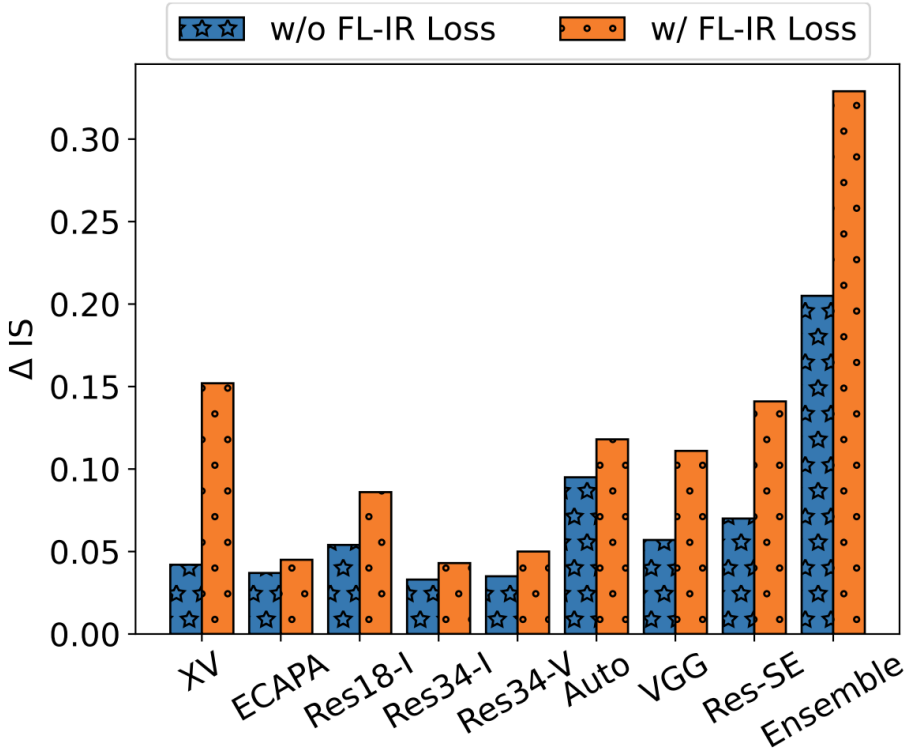
human study: if a given song contains any background noise and if so, how the noise influences their enjoyment of the song



“Noise +” and “Noise -” denote the answers “noisy w/ influence” and “noisy w/o influence”

SongBsAb: Evaluation

- Transferability

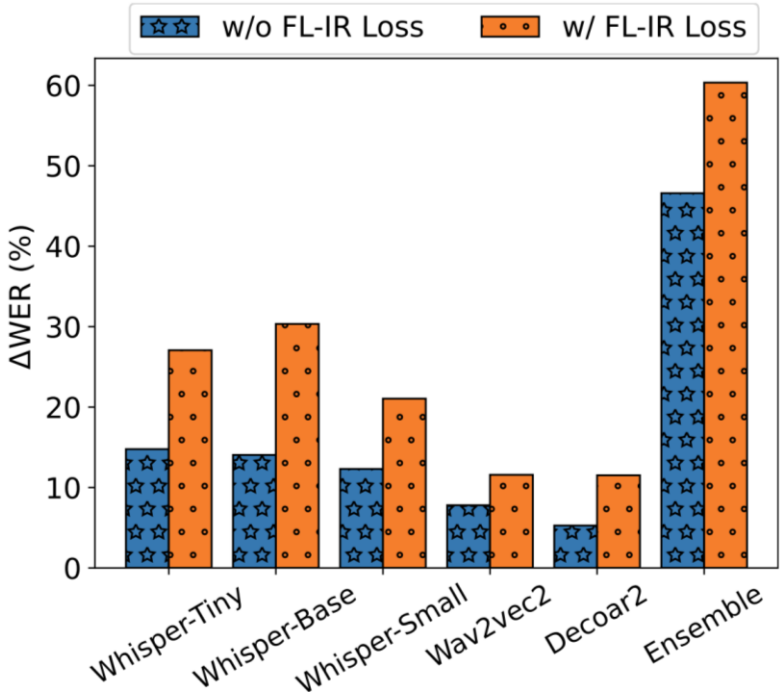


Identity Disruption

ΔIS: decrease of identity similarity

Ensemble: encoder ensemble

FL-IR: frame-level interaction reduction



Lyric Disruption

ΔWER: increase of lyric Word Error Rate

Ensemble: encoder ensemble

FL-IR: frame-level interaction reduction

Take away

- The first proactive prevention against singing voice conversion-based illegal song covers
- Dual prevention: identity & lyric disruption
- Backing track-refined simultaneous masking to preserve song quality
- Encoder ensemble & frame-level interaction reduction loss to enhance transferability
- Application: copyright & civil rights protection by record companies & singers

Website: https://sites.google.com/view/songbsab

Any Question?
Thanks!