My Music My Choice: Adversarial Protection Against Vocal Cloning in Songs

İlke Demir
Cauth AI
ilke@cauth.ai

Gerald Pena Vargas
Binghamton University
gpenava1@binghamton.edu

Alicia Unterreiner
Binghamton University
aunterr1@binghamton.edu

David Ponce
Binghamton University
dponce1@binghamton.edu

Umur Aybars Çiftçi
Binghamton University
uciftci@binghamton.edu

Abstract

Vocal cloning poses unprecedented threats to the music industry, enabling unauthorized reproduction of artists' vocal characteristics in songs. We introduce My Music My Choice (MMMC), a lightweight adversarial protection framework designed to proactively defend against vocal cloning in musical contexts. MMMC generates imperceptibly modified audio that preserves original vocals while significantly degrading the output of generative AI. MMMC achieves high-quality protected vocals with 0.944 STOI, reducing cloned output quality to 0.558 STOI. Our approach demonstrates effective transfer across different musical styles and maintains protection when vocals are reconstructed into full songs, providing a practical defense mechanism for artists and music creators.

1 Introduction

The music industry faces an existential threat from sophisticated generative AI methods such as retrieval-based voice conversion (RVC) and singing voice conversion (SVC) technologies that can clone any artist's voice from minimal audio samples. Recent advances achieved near-human quality in converting singing voices while preserving musical content and pitch information. Unlike speech-based voice cloning, singing voice conversion must handle complex musical elements including pitch contours, vibrato, breath control, and harmonic interactions with instrumental backing tracks.

The proliferation of open-source VC tools has democratized vocal cloning capabilities, enabling unauthorized creation of songs (46), fake collaborations (50), and musical deepfakes (41) that can damage artists', publishers', and distributors' reputations and economic interests. While detection methods show promise for identifying synthetic singing voices, retroactive detection may be too late in the life-cycle of a voice clip reliably fooling both humans and machines (47).

Instead, we explore proactive actions by protecting against replication. Previously, adversarial prevention techniques are modeled as attacks on speaker verification (22), speech-based interfaces (1), or speaker embeddings (52), but not on the generative output quality to completely disable such applications. Current adversarial protection methods designed for speech (52; 32) fail to address the unique challenges of songs. We present My Music My Choice (MMMC), the first adversarial protection framework specifically engineered for defending voices in songs against conversion attacks, empowering artists to have control over their vocals by preventing generative AI replication. MMMC learns to generate imperceptible adversarial samples while maximizing the degradation of VC quality.

We evaluate our approach attacking two VCs and evaluating both the protected version and the cloned output by intelligibility, noise, opinion, and perceptual scores; in vocal and song forms. MMMC

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

outputs stellar quality songs with 0.912 Short-Time Objective Intelligibility measure (STOI) and 3.882 Mean Opinion Score (MOS), while also deteriorating VC outputs with 0.420 STOI and 2.038 MOS. We perform ablation studies to interpret loss terms and parameters. In a world where celebrities' explicit requests are denied (18), we hope MMMC will pioneer and democratize this line of defense.

2 Related Work

Voice Cloning: VC is widely used in Singing Voice Synthesis (SVS), Singing Voice Conversion (SVC) and Text-to-Speech (TTS) systems. Early works use parametric statistical models and spectrum matching (20; 21) which later paved the way for complex approaches (42; 23). SVC aims to transform the vocal characteristics of source and target singers while preserving musical content, pitch, and timing, by combining VITS with content encoders and pitch extraction (44), learning speaker information by retrieval (39), introducing diffusion models for control (27), using pitch- and singer-invariant features (9), and leveraging differentiable signal processing (53). The SVC Challenge (15) established standardized evaluation protocols and revealed current state-of-the-art capabilities.

Protection and Detection: Adversarial attacks are developed to exploit vulnerabilities in automatic speaker verification (ASV) and speech recognition (ASR) systems (19; 2). Unlike audio deepfakes, they aim to deceive by injecting imperceptible perturbations, to imitate a target speaker and gain unauthorized access for ASV (28; 51) and to manipulate audio for misinformation for ASR (1; 34; 29). SongBsAb (3) combines adversarial perturbations with psycho-acoustic modeling for musical contexts whereas SingGraph (5) combines music understanding models with linguistic analysis through graph modeling. However, detection accuracy significantly degrades in the presence of background music. Speaker verification (11), anti-spoofing (17), liveness detection (55), and deepfake detection (31) algorithms have also been proposed as retroactive techniques against unauthorized voice cloning.

Adversarial Audio Defense: As generative models empower large-scale replication of unauthorized content;adversarial protection approaches are proposed for faces (25; 7), bodies (8), artistic styles (40), or any visual content (36) against generative AI misuse. In audio domain, adversarial methods are used as defenses, to mask the identity of the speaker (6; 33; 32; 10); to imperceptibly push the speaker embedding (52; 45; 14); to watermark samples (37; 26); and to clean (48), mark (16), and block (49) possibly malicious audio. Attacking only the speaker embedding, in turn, contributes to the same problem, only with someone else's identity. Physical-world attacks, ensemble attack strategies, and AI-induced distortions pose significant challenges to current protection methods (54). To effectively provide a proactive solution,output should be nullified, similar to the motivation of (36) in visual domain. Successively, both adversarial attacks and audio deepfakes can be prevented.

3 My Music My Choice

VC approaches contain common modules for source separation, vocoder, synthesizer, pitch extraction, or encoder. While attacking these separately can degrade performance, some may cancel adversarial changes or there may be alternatives for a single module. For such, we design MMMC to attack VC systems in an end-to-end manner.Given an attack model $X(S, v) = v'$ with speaker embedding S and voice input v, the output v' is assumed to have good quality and characteristics of speaker S (Fig. 1, top left). We model MMMC as a protector model P to corrupt these expectations as,

$$P(v, X) = v^+ \text{ such that } v^+ \sim v \text{ and } X(\_, v^+) \neq X(\_, v) \tag{1}$$

Fig. 1 depicts this threat model, where v, v', v^+, v^- represent input (black), converted input (black with speaker), protected (green), and broken (red) voices. MMMC is designed as a plug-and-play black-box attack on any VC, where the attack model is kept frozen with no access to gradients.

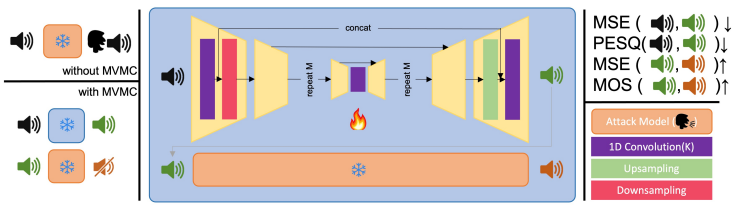

Figure 1: **Overview.** (Left) Attack model imitates a voice (black). With MMMC (green), it produces distorted audio (red). MMMC model (mid) and loss terms (right) are depicted to relate with vocal I/O.

3.1 Architecture

MMMC employs a U-Net architecture (38) operating directly on waveforms ω, composing vocals $v = \{\omega_0, \ldots, \omega_{|v|/|\omega|}\}$ to handle complex spectro-temporal patterns in singing voices. The architecture consists of down/upsampling blocks with skip connections, enabling hierarchical learning of fine-grained vocal details. As opposed to complicated architectures operating on lossy representations such as Mel filters or spectrogram images, a simple yet efficient model provides enough flexibility. MMMC's architecture (Fig. 1 (mid)) consists of $M = 12$ downsampling blocks that extract higher level features at coarser time scales. Each contains 1D convs with a kernel size $K = 15$, operating in half time scale of the previous block's input. Then, these features are combined with the locally computed high resolution features in $M = 12$ upsampling blocks operating in double time scale.

3.2 Training Objective

To abide by the design in Eqn. 1, we combine several loss terms in one multi-objective loss as $L = \alpha_r L_r + \alpha_p L_p + \alpha_d L_d + \alpha_o L_o$ where α_* weights balance loss terms. We minimize this objective, thus, all loss terms are designed to be decreasing. We denote modified versions with superscripts for all, such as $P(v) = v^+$, $X(\_, v) = v'$, $X(\_, v^+) = v^-$.

Reconstruction Loss: To preserve voice quality during protection, we define a reconstruction loss L_r as the MSE between input (ω_*) and protected (ω_*^+) wave forms.

$$L_r = \frac{1}{|v|} \sum_{i=0}^{|v|} ||\omega_i - \omega_i^+||^2 \text{ where } v = \{\omega_0, \ldots, \omega_{|v|}\} \tag{2}$$

Perceptual Loss: L_p keeps v^+ perceptually unchanged from v while flexing structural deviation from L_r. We utilize PESQ (12) as $L_p = 0.2(0.5 + PESQ(v, v^+))$, normalized to $[0, 1]$.

Distortion Loss: L_d maximizes degradation of v^- through inverted MSE, measuring dissimilarity between v^+ and v^-. Unlike L_r, we formulate this term to inversely contribute to overall loss.

$$D = \frac{1}{|v^+|} \sum_{i=0}^{|v^+|} ||\omega_i^+ - \omega_i^-||^2 \text{ and } L_d = \frac{1}{\log(1 + D + \epsilon)} \tag{3}$$

Opinion Loss: L_o targets degrading semantic properties of v^-. To accomplish this ill-defined task, we minimize Mean Opinion Score (MOS) as $L_o = (1 - 0.2(MOS(v^+, v^-)))$, normalized to $[1, 0]$, predicted by (30). While L_d pushes for structural dissimilarity, L_o pushes for unintelligible outputs.

3.3 Training & Testing

We use MUSDB18 (35), providing 150 music tracks (10+ hours) across diverse genres. This dataset allows us to test the performance on actual songs, its effects on both the singer performance and listening quality, edge cases where regular speech does not cover, and on both isolated vocal protection and full song reconstruction scenarios. We use So-VITS-SVC (44) and RVC (39) as our primary attack models due to their SOTA performance and adoption. We train for 20 epochs using Adam, 80/20 train/test split, with $\alpha_r = 0.29$, $\alpha_d = 0.02$, $\alpha_p = 0.29$, and $\alpha_o = 0.65$, optimized for musical vocal characteristics. One epoch takes approximately 15 minutes. We use three setups to train, 1 2080TI, 6 A100s, and 8 V100s; with corresponding batch sizes and LRs of (16, 0.001) for RVC, (32, 0.01) for So-VITS, and (256, 0.01) for RVC. GPU memory is mostly spent on MOS queries.

4 Results & Experiments

We quantify MMMC's success in input preservation ($v \sim v^+$) and output degradation ($v' \neq v^-$) by four metrics: STOI (43) for intelligibility, SI-SDR (24) for distortion, MOS (30) for subjective opinion, and PESQ (12) for perceptual quality. MMMC achieves higher scores on the left, and lower on the right in Tab. 1: high intelligibility (0.912/1.0 STOI), high perceptual quality (3.874/4.5 PESQ), varying energy (SI-SDR), and high subjective quality (3.882/5 MOS) for v^+; whereas low intelligibility (0.420/1.0 STOI), very low perceptual quality (1.079/4.5 PESQ), high distortion (large

negative SI-SDR) and low subjective quality (2.038/5 MOS) for v^-. For reference, 3+ MOS represents relatively good quality (MOS for TIMIT (13) is measured at 3.45±0.52). Sample spectrograms of low, random, and high scoring quadruplets of (v, v^+, v', v^-) are demonstrated in Fig. 2.

Table 1: **MMMC Evaluation.** Vocal tracks (first row) and full songs (second row) are successfully protected with high fidelity on v^+ (left half) and high corruption on v^- (right half).

	v vs. v^+ (↑)				v' vs. v^- (↓)			
	STOI	PESQ	SI-SDR	MOS	STOI	PESQ	SI-SDR	MOS
Vocals	0.912	3.874	7.347	3.882	0.420	1.079	-26.792	2.038
Reconstructed	0.944	3.765	7.380	3.755	0.558	1.189	-9.631	3.932

We validate MMMC's protection on both isolated tracks and full songs, by placing protected vocals back. Vocals are successfully preserved in v^+s and distorted in v^-s (first row). For full songs, protection effectiveness persists, however the presence of instruments partially masks vocal artifacts, resulting in higher subjective quality as expected.
MMMC shows robustness across diverse styles within MUSDB18, including pop, rock, and world music genres; adapting to different techniques, from whispered vocals to powerful belting, without style-specific optimization.

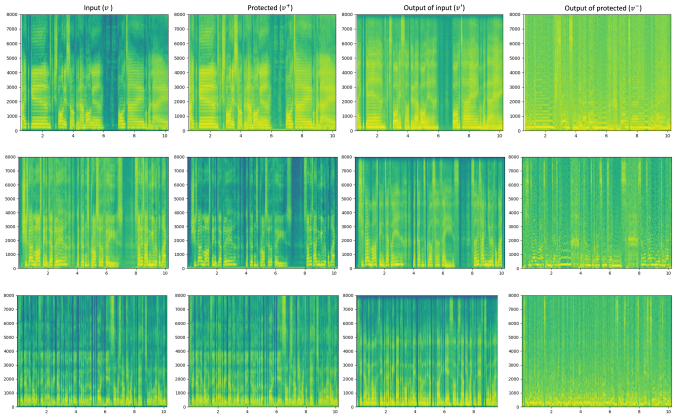

Figure 2: **Spectrogram Samples** of low (top), random (mid), and high (bot) scoring quadruplets of (v, v^+, v', v^-) values. Similarity of first two and dissimilarity of last two visually verify results.

In Tab. 2, we modify the multi-objective training loss of MMMC on a speech dataset to include only the marked terms.Without L_p, perceptual and subjective quality of v^+ drops (row 1). Without L_r, there is significant noise introduced as we lose the signal (row 2). L_o being a semantic corruption factor over L_d, it helps reduce the intelligibility of v^- (row 3). Without L_p, all v^+ metrics drop as the low weight on L_r limits reconstruction to preserve the voice as much as L_p does perceptually.

Table 2: **Loss Contributions**. We systematically omit loss terms to evaluate and reason their impact.

				v vs. v^+ (↑)				v' vs. v^- (↓)			
L_r	L_d	L_p	L_o	STOI	PESQ	SI-SDR	MOS	STOI	PESQ	SI-SDR	MOS
✓	✓			0.975	2.152	15.5091	4.000	0.598	1.050	-21.991	1.788
	✓	✓		0.998	4.568	-26.429	4.439	0.601	1.050	-22.200	1.809
✓	✓	✓		0.983	3.827	0.132	4.399	0.645	1.068	-30.295	1.626
✓	✓		✓	0.774	1.186	-22.647	4.346	0.543	1.045	-28.226	1.798
✓	✓	✓	✓	0.989	4.151	4.782	4.398	0.598	1.049	-28.007	1.783

5 Conclusion

In this generative landscape where every song is under constant threat of being stolen, we say "My Music My Choice" to let everyone protect their audio. We design MMMC as a black-box adversarial attack on voice cloning, balancing vocal quality with protection effectiveness.We quantitatively evaluate imperceptibly changed protected songs and disturbingly changed attack model outputs by four metrics on separated vocals and full songs in diverse genres. In future, we aim to evaluate and support generalization to other VC methods, quantify robustness, survey human evaluations, analyze limitations, and design attacks on other common tracks.We would also like our adversarial generation algorithm utilized for more elaborate uses, such as ownership verification by audio steganography (4) or multi-modal content protection standards.

References

[1] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., Zhou, W.: Hidden voice commands. In: 25th USENIX security symposium (USENIX security 16). pp. 513–530 (2016)

[2] Carlini, N., Wagner, D.: Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE security and privacy workshops (SPW). pp. 1–7. IEEE (2018)

[3] Chen, G., Zhang, Y.: Songbsab: A dual prevention approach against singing voice conversion based illegal song covers. In: 32nd Annual Network and Distributed System Security Symposium (2025)

[4] Chen, L., Wang, R., Dong, L., Yan, D.: Imperceptible adversarial audio steganography based on psychoacoustic model. Multimedia Tools and Applications **82**(17), 26451–26463 (2023)

[5] Chen, X., Wu, H., Jang, J.S.R., Lee, H.y.: Singing voice graph modeling for singfake detection. arXiv preprint arXiv:2406.03111 (2024)

[6] Chiquier, M., Mao, C., Vondrick, C.: Real-time neural voice camouflage. arXiv preprint arXiv:2112.07076 (2021)

[7] Ciftci, U.A., Yuksek, G., Demir, I.: My face my choice: Privacy enhancing deepfakes for social media anonymization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1369–1379 (2023)

[8] Ciftci, U.A., Tanriverdi, A.K., Demir, I.: My body my choice: Human-centric full-body anonymization. arXiv preprint arXiv:2406.09553 (2024)

[9] Deng, C., Yu, C., Lu, H., Weng, C., Yu, D.: Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7749–7753. IEEE (2020)

[10] Deng, J., Teng, F., Chen, Y., Chen, X., Wang, Z., Xu, W.: {V-Cloak}: Intelligibility-, naturalness- & {Timbre-Preserving}{Real-Time} voice anonymization. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5181–5198 (2023)

[11] Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In: Interspeech 2020. pp. 3830–3834 (2020)

[12] Dong, X., Williamson, D.S.: An attention enhanced multi-task model for objective speech assessment in real-world environments. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 911–915. IEEE (2020)

[13] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. NASA STI/Recon technical report n **93**, 27403 (1993)

[14] Huang, C.y., Lin, Y.Y., Lee, H.y., Lee, L.s.: Defending your voice: Adversarial attack on voice conversion. In: 2021 IEEE Spoken Language Technology Workshop (SLT). pp. 552–559. IEEE (2021)

[15] Huang, W.C., Violeta, L.P., Liu, S., Shi, J., Toda, T.: The singing voice conversion challenge 2023. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1–8. IEEE (2023)

[16] Hussain, S., Neekhara, P., Dubnov, S., McAuley, J., Koushanfar, F.: {WaveGuard}: Understanding and mitigating audio adversarial examples. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 2273–2290 (2021)

[17] Kamble, M.R., Sailor, H.B., Patil, H.A., Li, H.: Advances in anti-spoofing: from the perspective of asvspoof challenges. APSIPA Transactions on Signal and Information Processing **9**, e2 (2020)

[18] Kastrenakes, J.: Scarlett johansson told openai not to use her voice — and she's not happy they might have anyway. https://www.theverge.com/2024/5/20/24161253/scarlett-johansson-openai-altman-legal-action, accessed: 2024-05-21

[19] Khan, A., Malik, K.M., Ryan, J., Saravanan, M.: Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. Artificial Intelligence Review **56**(Suppl 1), 513–566 (2023)

[20] Kobayashi, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S.: Statistical singing voice conversion with direct waveform modification based on the spectrum differential. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)

[21] Kobayashi, K., Toda, T., Neubig, G., Sakti, S., Nakamura, S.: Statistical singing voice conversion based on direct waveform modification with global variance. In: Sixteenth Annual Conference of the International Speech Communication Association. Citeseer (2015)

[22] Kreuk, F., Adi, Y., Cisse, M., Keshet, J.: Fooling end-to-end speaker verification with adversarial examples. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1962–1966. IEEE (2018)

[23] Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al.: Voicebox: Text-guided multilingual universal speech generation at scale. Advances in neural information processing systems **36**, 14005–14034 (2023)

[24] Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: Sdr–half-baked or well done? In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 626–630. IEEE (2019)

[25] Li, Z., Yu, N., Salem, A., Backes, M., Fritz, M., Zhang, Y.: {UnGANable}: Defending against {GAN-based} face manipulation. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 7213–7230 (2023)

[26] Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W., Yu, N.: Detecting voice cloning attacks via timbre watermarking. arXiv preprint arXiv:2312.03410 (2023)

[27] Liu, S., Cao, Y., Su, D., Meng, H.: Diffsvc: A diffusion probabilistic model for singing voice conversion. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 741–748. IEEE (2021)

[28] Liu, S., Wu, H., Lee, H.y., Meng, H.: Adversarial attacks on spoofing countermeasures of automatic speaker verification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 312–319. IEEE (2019)

[29] Lu, Z., Han, W., Zhang, Y., Cao, L.: Exploring targeted universal adversarial perturbations to end-to-end asr models. arXiv preprint arXiv:2104.02757 (2021)

[30] Manocha, P., Kumar, A.: Speech quality assessment through mos using non-matching references. arXiv preprint arXiv:2206.12285 (2022)

[31] Müller, N.M., Czempin, P., Dieckmann, F., Froghyar, A., Böttinger, K.: Does audio deepfake detection generalize? arXiv preprint arXiv:2203.16263 (2022)

[32] O'Reilly, P., Bugler, A., Bhandari, K., Morrison, M., Pardo, B.: Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. Advances in Neural Information Processing Systems **35**, 30058–30070 (2022)

[33] Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.Y.: Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems. pp. 82–94 (2018)

[34] Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C.: Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: International conference on machine learning. pp. 5231–5240. PMLR (2019)

[35] Rafii, Z., Liutkus, A., Stöter, F.R., Mimilakis, S.I., Bittner, R.: The MUSDB18 corpus for music separation (Dec 2017). https://doi.org/10.5281/zenodo.1117372, https://doi.org/10.5281/zenodo.1117372

[36] Rhodes, A., Bhagat, R., Çiftçi, U.A., Demir, I.: My art my choice: Adversarial protection against unruly ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8389–8394 (2024)

[37] Roman, R.S., Fernandez, P., Défossez, A., Furon, T., Tran, T., Elsahar, H.: Proactive detection of voice cloning with localized watermarking. arXiv preprint arXiv:2401.17264 (2024)

[38] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

[39] RVC-Project: Retrieval-based-voice-conversion-webui. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

[40] Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 2187–2204 (2023)

[41] Sherman, N.: World's biggest music labels sue over ai copyright. https://www.bbc.com/news/articles/ckrrr8yelzvo, accessed: 2025-08-21

[42] Sisman, B., Vijayan, K., Dong, M., Li, H.: Singan: Singing voice conversion with generative adversarial networks. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 112–118. IEEE (2019)

[43] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE international conference on acoustics, speech and signal processing. pp. 4214–4217. IEEE (2010)

[44] svc-develop team: Softvc vits singing voice conversion. https://github.com/svc-develop-team/so-vits-svc

[45] Wang, Y., Guo, H., Wang, G., Chen, B., Yan, Q.: Vsmask: Defending against voice synthesis attack via real-time predictive perturbation. arXiv preprint arXiv:2305.05736 (2023)

[46] Weatherbed, J.: Record labels claim ai generator suno illegally ripped their songs from youtube. https://www.theverge.com/news/782448/riaa-suno-ai-lawsuit-update-stream-ripping-youtube, accessed: 2025-09-22

[47] Wenger, E., Bronckers, M., Cianfarani, C., Cryan, J., Sha, A., Zheng, H., Zhao, B.Y.: " hello, it's me": Deep learning-based speech synthesis attacks in the real world. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. pp. 235–251 (2021)

[48] Wu, S., Wang, J., Ping, W., Nie, W., Xiao, C.: Defending against adversarial audio via diffusion model. arXiv preprint arXiv:2303.01507 (2023)

[49] Xu, X., Fu, C., Du, X., Ratazzi, E.P.: Voiceguard: An effective and practical approach for detecting and blocking unauthorized voice commands to smart speakers. In: 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). pp. 582–596. IEEE (2023)

[50] Yang, A., Dasrath, D.: Tupac's estate threatens to sue drake over diss track using what appears to be late rapper's ai-generated voice. https://www.nbcnews.com/pop-culture/pop-culture-news/tupac-shakur-estate-threatens-to-sue-drake-ai-use-dis-track-rcna149242, accessed: 2025-08-21

[51] Yu, Z., Chang, Y., Zhang, N., Xiao, C.: {SMACK}: Semantically meaningful adversarial audio attack. In: 32nd USENIX security symposium (USENIX security 23). pp. 3799–3816 (2023)

[52] Yu, Z., Zhai, S., Zhang, N.: Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. pp. 460–474 (2023)

[53] yxlllc: Ddsp-svc: Real-time end-to-end singing voice conversion system based on ddsp (differentiable digital signal processing) (2023), `https://github.com/yxlllc/DDSP-SVC`, gitHub repository

[54] Zhang, B., Cui, H., Nguyen, V., Whitty, M.: Audio deepfake detection: What has been achieved and what lies ahead. Sensors (Basel, Switzerland) **25**(7), 1989 (2025)

[55] Zhang, L., Tan, S., Wang, Z., Ren, Y., Wang, Z., Yang, J.: Viblive: A continuous liveness detection for secure voice user interface in iot environment. In: Annual Computer Security Applications Conference. pp. 884–896 (2020)