Not All Deepfakes Are Created Equal: Triaging Audio Forgeries for Robust Deepfake Singer Identification

Davide Salvi
Universal Music Group
London, United Kingdom

Hendrik Vincent Koops*
Universal Music Group
London, United Kingdom

Elio Quinton
Universal Music Group
London, United Kingdom

Abstract

The proliferation of highly realistic singing voice deepfakes presents a significant challenge to protecting artist likeness and content authenticity. Automatic singer identification in vocal deepfakes is a promising avenue for artists and rights holders to defend against unauthorized use of their voice, but remains an open research problem. Based on the premise that the most harmful deepfakes are those of the highest quality, we introduce a two-stage pipeline to identify a singer's vocal likeness. It first employs a discriminator model to filter out low-quality forgeries that fail to accurately reproduce vocal likeness. A subsequent model, trained exclusively on authentic recordings, identifies the singer in the remaining high-quality deepfakes and authentic audio. Experiments show that this system consistently outperforms existing baselines on both authentic and synthetic content.

1 Introduction

Recent advances in singing voice cloning technology enable the generation of "deepfakes" that are virtually indistinguishable from authentic recordings, posing a significant threat to content authenticity and artist likeness protection. In the same way that audio fingerprinting protects recordings from unauthorized use [1, 2, 3], we propose a system to protect a singer's vocal likeness. Our goal is to be able to identify a singer's voice in both authentic and deepfake recordings, providing a tool for artists and rights holders to defend against unauthorized uses.

The scientific community has largely pursued two separate countermeasures: deepfake detection, which aims to distinguish synthetic from authentic audio [4, 5, 6, 7], and singer identification, which verifies a vocalist's identity [8, 9]. However, identifying a singer within a deepfake remains a challenge [10]. This paper investigates singer identification across authentic and synthetic signals.

We argue that the potential for harm from a deepfake correlates with its quality: a highly realistic fake is more dangerous than a poor one where the singer is unrecognizable. Based on this observation, we introduce a two-stage pipeline designed for maximum effectiveness against the most threatening deepfakes. First, a discriminator model filters out low-quality forgeries, which do not faithfully reproduce the vocal likeness. Second, a singer identification model trained only on authentic recordings matches the test recording to known vocal likenesses. Our experiments show that our system consistently outperforms existing baselines across both authentic and deepfake content.

2 Method

Motivated by the notion that the potential harm that can be caused to an artist by an unauthorized deepfake is proportional to its perceptual quality, we propose a two-steps pipeline, depicted in

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

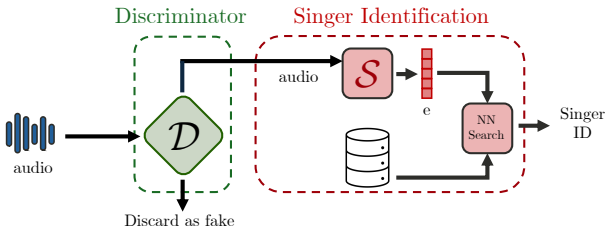


Figure 1: Proposed two-stage pipeline for singer identification. Stage 1 ($\mathcal{D}$) filters low-quality deepfakes. Tracks classified as authentic proceed to Stage 2, where singer identity is determined by nearest neighbor search using cosine distance of extracted embeddings (e).

Figure 1. First, the recording under test is processed by a discriminator $\mathcal{D}$, which objective is to filter out poor quality deepfakes. Because they typically exhibits significant artifacts and do not faithfully replicate the singer's vocal likeness, their potential for harm is comparatively much lower than that of higher quality deepfakes, and they are likely to be easier to detect. Recordings that are deemed either high quality deepfakes or authentic by $\mathcal{D}$ are passed to the second stage, where a singer identification model $\mathcal{S}$ extracts a vocalist likeness embedding to be compared with a database of known vocalist likeness embeddings.

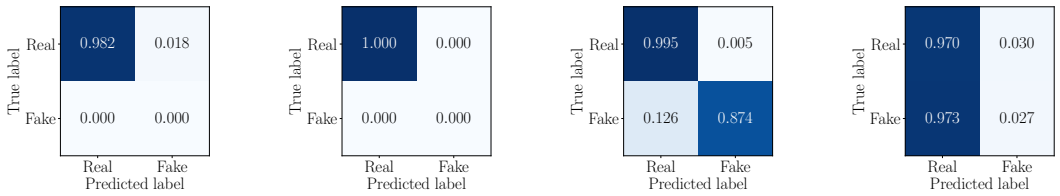
For the discriminator ($\mathcal{D}$) we use a Light Convolutional Neural Network (LCNN) [11, 12], a compact and efficient architecture introduced for spoofing detection. It is trained to predict whether a track is a deepfake or authentic on the CTRSVDD dataset. Because we expect that poor quality deepfakes should be easy to detect because they features salient and unnatural artifacts, we purposefully keep $\mathcal{D}$ light and its training regime simple. Our expectation is that $\mathcal{D}$ would be effective at detecting poor deepfakes but would be fooled by high quality ones.

For the singer identification model ($\mathcal{S}$) we apply the ECAPA-TDNN architecture [13] to singing voice data and train it as a multi-class classifier. We selected this model architecture, shown to perform well in speech applications, on the basis that our task also focuses on human voice. In addition, we consider two baselines. 1- SSL: A Self-Supervised Learning baseline using a Siamese EfficientNet-B0 encoder, specifically designed to learn singer representations from mel-spectrograms [14]. 2- RESNET-TDNN: A hybrid model [15] pre-trained on the VoxCeleb speech dataset [16] to evaluate speech-to-singing speaker identification transferability. $\mathcal{S}$ and SSL are trained only on authentic content. Again, this is an important property since it makes the method easier to operate and scale, and paired authentic and deepfake content is rarely available in practice.

To assess the generalization capability of our proposed pipeline across different data distributions, we use four datasets of music recordings that contain singing voice - see Table 1. These include two datasets with only authentic recordings (PRIVATE and ARTIST20) and two with both authentic and deepfake tracks (CTRSVDD and WILDSVDD). All recordings are resampled to 16 kHz for consistency. **PRIVATE**: A proprietary corpus of 134,826 tracks from 2,000 different singers, containing authentic commercial recordings for training $\mathcal{S}$ and evaluation. **ARTIST20** [17]: Open dataset of authentic recordings, used for testing only. **CTRSVDD** [18]: Authentic and deepfake a cappella vocals from SVDD 2024 Challenge, used for training $\mathcal{D}$ and evaluation. **WILDSVDD** [19]: Designed for real-world deepfake detection, this dataset introduced by Zhang et al. in [19] includes authentic and synthetic tracks sourced from social media. Many tracks listed in the original dataset are no longer available.

		Authenticity		$\mathcal{D}$ Discriminator		$\mathcal{S}$ Singer Identification		
Dataset	Audio Contents	Authentic	Deepfake	Training	Testing	Training	Testing	Reference
PRIVATE	Full mix	✓	×	×	✓	✓	✓	—
ARTIST20	Full mix	✓	×	×	✓	×	✓	Ellis in [17]
CTRSVDD	Vocals only	✓	✓	✓	✓	×	✓	Zang et al. in [18]
WILDSVDD	Full mix	✓	✓	×	✓	×	✓	Zhang et al. in [19]

Table 1: Overview of used datasets. *"Full mix"* refers to a mix of vocals with (background) music. *"Authenticity"* refers to bona fide (*"Authentic"*) or deepfake audio (*"Deepfake"*).



(a) PRIVATE (b) ARTIST20 (c) CTRSVDD (d) WILDSVDD

Figure 2: Confusion matrices of the considered singing voice discriminator $\mathcal{D}$ across the four different datasets. False Positive Rate (FPR) (i.e., authentic tracks misclassified as deepfakes) is exceptionally low across all datasets (see top right corners of the confusion matrices). This means that when $\mathcal{D}$ identifies a track as authentic, it is highly reliable.

We focus on six artists, resulting in 247 tracks evenly split between authentic and fake recordings. We use this dataset only for the evaluation of the $\mathcal{D}$ and $\mathcal{S}$ models.

For all datasets, we create an alternative version where vocals and background music are separated using BS-ROFORMER [20], followed by non-vocal segment removal ($\pm$ 40% per track) via an energy-based Voice Activity Detector (VAD) [21]. This allows us to focus the training on samples that always contain vocals, and to use the separated background music as an augmentation.

The model $\mathcal{D}$ uses mel-spectrogram as input, with 512 FFT bins, 80 mel bins, a hamming window length of 400 samples, and a hop length of 160 samples. It was trained using Binary Cross Entropy loss, a cosine annealing learning rate (10^{-4} to 10^{-7}), and 10^{-4} weight decay. Class imbalance was addressed via random oversampling, ensuring equal numbers of authentic and deepfake samples in each batch. Singer identification ($\mathcal{S}$) models ECAPA-TDNN and SSL were trained on 10-second log-mel spectrogram windows (512 FFT, 400 window, 160 hop, 80 mel bins), batch size of 64, early stopping (patience 10), cosine annealing learning rate (10^{-4} to 10^{-7}), and 10^{-5} weight decay. Data augmentation (35% probability) included random background music, noise (impulsive/stationary) [22], and pitch shifting ($\pm$2 semitones).

To simulate real-world conditions at inference time, we use the datasets in their original condition (i.e. no source separation or VAD). Five 10 s windows are extracted from each song for model inference. For $\mathcal{D}$ we average window predictions for a final song classification. For vocalist identification models, we average the embeddings from the last dense layer of model $\mathcal{S}$, representing vocalist identity. Singer identity estimation employs cosine distance against reference embeddings, and performance is evaluated using standard speaker identification metrics (e.g., ROC, ROC, EER).

3 Experiments & Results

Singer Identification: We first evaluate the performance of the singer identification models across all datasets. Table 2 reveals that ECAPA-TDNN consistently outperforms baselines. RESNET-TDNN's performance is comparable to ECAPA-TDNN's only on CTRSVDD (AUC differences <1%), which we attribute to the accappella content of CTRSVDD being the most similar to speech. Our music-specific training regimen proves advantageous on all other datasets that contain background music.

All models tend to exhibit poorer performance on datasets containing deepfakes. To investigate this further, we analyzed the ECAPA-TDNN's performance on the CTRSVDD dataset, breaking down its effectiveness against each deepfake generation algorithm included in the corpus - see Table 3. Compelling performance is achieved on authentic data and algorithms A01-A05, A12 (AUC > 90%), while it degrades significantly for algorithms A07-A10 and A13 (EER > 30%). Upon manual inspection, we observed that performance seem to correlate with the quality of the deepfakes. The cloned voices generated by algorithms A07-A10 and A13 often do not closely resemble the original singer. This lack of fidelity likely undermines the effectiveness of the singer identification process. It also highlights a challenge in the evaluation deepfake singer identification methods: how to deal with cases where the vocalist is not perceptually recognizable?

		Private dataset		Artist20		CTRSVDD		WILDSVDD		Average	
Model	Ref.	EER (%) ↓	AUC (%) ↑	EER (%) ↓	AUC (%) ↑	EER (%) ↓	AUC (%) ↑	EER (%) ↓	AUC (%) ↑	EER (%) ↓	AUC (%) ↑
ECAPA-TDNN	[13]	**4.31**	**98.29**	**15.56**	**91.47**	**30.34**	**76.11**	**19.24**	**87.41**	**17.36**	**88.32**
SSL	[14]	16.13	91.14	25.30	81.78	36.34	68.16	32.92	73.53	27.67	78.65
RESNET-TDNN	[15]	8.70	96.28	23.05	84.29	31.46	75.25	21.38	86.41	21.15	85.56

Table 2: Singer identification performance of three models across four different datasets.

	A02	*REAL*	A04	A05	A01	A12	A03	A06	A11	A13	A09	A10	A07	A08
EER (%) ↓	8.83	10.73	11.68	12.01	13.88	14.16	14.91	22.99	24.29	30.36	33.61	33.98	36.02	36.05
AUC (%) ↑	96.94	95.48	95.57	94.21	93.51	93.53	91.90	84.63	83.52	76.33	71.58	71.12	68.67	69.17

Table 3: Singer identification performance (ECAPA-TDNN) on CTRSVDD dataset. High EER for algorithms (A07-A10, A13) is linked to poor cloned voice fidelity, impacting identification accuracy.

	Pipeline	EER (%) ↓	AUC (%) ↑
CTRSVDD	$\mathcal{S}$	30.34	76.11
	$\mathcal{D} \circ \mathcal{S}$	**16.82**	**88.90**
WILDSVDD	$\mathcal{S}$	19.24	87.41
	$\mathcal{D} \circ \mathcal{S}$	**15.55**	**91.55**
Average	$\mathcal{S}$	24.79	81.76
	$\mathcal{D} \circ \mathcal{S}$	**16.19**	**90.23**

Table 4: Singer identification ($\mathcal{S}$) performance of ECAPA-TDNN with ($\mathcal{D} \circ \mathcal{S}$) and without ($\mathcal{S}$) discriminator $\mathcal{D}$. Using the discriminator significantly improves singer identification.

Singing Voice Deepfake Discriminator: We introduced a discriminator $\mathcal{D}$ to handle poor-quality deepfakes, and evaluate it as a binary vocal deepfake classifier. The confusion matrices in Figure 2 show a low False Positive Rate (FPR) across all datasets, indicating high reliability on authentic tracks. This is an important property since, in practice, flagging an authentic track as deepfake may have damaging consequences. The False Negative Rate (FNR) (i.e. deepfake tracks misclassified as authentic) is significantly higher in WILDSVDD than in CTRSVDD. Given our empirical observation of the varying deepfakes quality in CTRSVDD, we hypothesize it may explain the difference in FNR. WILDSVDD's higher FNR would then suggest it contains higher-quality deepfakes that can effectively fool $\mathcal{D}$, whereas CTRSVDD's lower quality deepfakes are easier to detect.

Combining Singing Voice Deepfake Detection and Singer Identification: As a final experiment, we evaluate the benefits of the proposed 2-step pipeline described in Figure 1. We evaluate the framework on the CTRSVDD and WILDSVDD datasets, considering only tracks classified as authentic by the discriminator (we label this condition $\mathcal{D} \circ \mathcal{S}$) and compare it to the performance of the singer identification model only (labeled $\mathcal{S}$). Table 4 shows the results for EER and AUC metrics, which reveal that introducing $\mathcal{D}$ significantly enhances singer identification performance. Combining the results of our experiments and our empirical observation of varied deepfake quality, our interpretation is that ensuring $\mathcal{S}$ operates on realistic vocal likenesses, makes the singer identification task more meaningful and tractable (attempting to identify a singer in the case of an unidentifiable likeness is bound to fail). In future work we propound to cross-reference these results with a study of deepfakes perceptual quality to test our interpretation further.

4 Conclusions

This paper addresses singer identification in authentic and deepfake singing voices recordings. We proposed a novel two-stage pipeline based on the premise that the highest quality deepfakes are those with the greatest potential for harm. Our experiments show that our proposed method outperforms baselines on multiple benchmarks. Combining these results with empirical observation suggests that the performance of singer identification models degrades on low quality deepfakes, where the vocal likeness is not faithfully reproduced. Our interpretation is that the introduction of the discriminator allows the singer identification model to only operate on high quality deepfakes and therefore makes the identification task more meaningful and tractable. For future work we recommend a perceptual study of deepfake quality to further test this interpretation.

References

[1] Reinhard Sonnleitner and Gerhard Widmer. Robust quad-based audio fingerprinting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):409–421, 2015.

[2] Avery Wang et al. An industrial strength audio search algorithm. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, pages 7–13. Washington, DC, 2003.

[3] Joren Six and Marc Leman. Panako-a scalable acoustic fingerprinting system handling time-scale and pitch modification. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 259–264, 2014.

[4] Yuankun Xie, Jingjing Zhou, Xiaolin Lu, Zhenghao Jiang, Yuxin Yang, Haonan Cheng, and Long Ye. FSD: An initial chinese dataset for fake song detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[5] Anmol Guragain, Tianchi Liu, Zihan Pan, Hardik B Sailor, and Qiongqiong Wang. Speech Foundation Model Ensembles for the Controlled Singing Voice Deepfake Detection (CtrSVDD) Challenge 2024. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[6] Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. SingFake: Singing voice deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[7] Mahyar Gohari, Davide Salvi, Paolo Bestagini, and Nicola Adami. Audio Features Investigation for Singing Voice Deepfake Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[8] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[9] Bidisha Sharma, Rohan Kumar Das, and Haizhou Li. On the importance of audio-source separation for singer identification in polyphonic music. In *Interspeech*, 2019.

[10] Dorian Desblancs, Gabriel Meseguer-Brocal, Romain Hennequin, and Manuel Moussallam. From Real to Cloned Singer Identification. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[11] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. STC Antispoofing Systems for the ASVspoof2019 Challenge. *Interspeech 2019*, 2019.

[12] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[13] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech*, 2020.

[14] Bernardo Torres, Stefan Lattner, and Gael Richard. Singer Identity Representation Learning using Self-Supervised Techniques. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[15] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. *Comput. Speech Lang.*, 60(C), March 2020.

[16] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[17] Daniel PW Ellis. Classifying music audio with timbral and chroma features. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[18] Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, and Zhiyao Duan. CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. In *Interspeech*, 2024.

[19] You Zhang, Yongyi Zang, Jiatong Shi, Ryuichi Yamamoto, Tomoki Toda, and Zhiyao Duan. SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[20] Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. Music Source Separation with Band-Split RoPE Transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[21] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[22] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.