SINGFAKE: SINGING VOICE DEEPFAKE DETECTION

Yongyi Zang*, You Zhang*, Mojtaba Heydari, Zhiyao Duan

Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA

ABSTRACT

The rise of singing voice synthesis presents critical challenges to artists and industry stakeholders over unauthorized voice usage. Unlike synthesized speech, synthesized singing voices are typically released in songs containing strong background music that may hide synthesis artifacts. Additionally, singing voices present different acoustic and linguistic characteristics from speech utterances. These unique properties make singing voice deepfake detection a relevant but significantly different problem from synthetic speech detection. In this work, we propose the singing voice deepfake detection task. We first present SingFake, the first curated in-the-wild dataset consisting of 28.93 hours of bonafide and 29.40 hours of deepfake song clips in five languages from 40 singers. We provide a train/validation/test split where the test sets include various scenarios. We then use SingFake to evaluate four state-of-the-art speech countermeasure systems trained on speech utterances. We find these systems lag significantly behind their performance on speech test data. When trained on SingFake, either using separated vocal tracks or song mixtures, these systems show substantial improvement. However, our evaluations also identify challenges associated with unseen singers, communication codecs, languages, and musical contexts, calling for dedicated research into singing voice deepfake detection. The SingFake dataset and related resources are available[1].

Index Terms— singing voice deepfake detection, anti-spoofing, dataset, singing voice separation

1. INTRODUCTION

"I mean really, how do you fight with someone who is putting out new albums in the time span of minutes."
— Stefanie Sun [1]

Quoted from a prominent Singaporean singer, this remark underscores a rapidly emerging challenge in the modern music industry and cultural landscape: the surge of AI-generated singing voices. With the development of singing voice synthesis techniques, AI-generated singing voices sound increasingly natural, align well with the music scores, and can clone any singer's voice with a small amount of training data. Such techniques have been made more accessible with open-source singing voice synthesis and voice conversion projects, such as VISinger [2] and DiffSinger [3], raising concerns for artists, record labels, and publishing houses. For example, unauthorized synthetic productions mimicking a singer could potentially undermine the singer's commercial value, leading to potential copyright and licensing disputes. The ever-increasing societal

apprehensions accentuate the urgency for developing methods to accurately detect deepfake singing voices.

As singing voice is a type of human vocalization, it is intuitive to explore solutions from an analogous research domain - speech deepfake detection, often referred to as voice spoofing countermeasures (CM). Existing research has been investigating different methods to discern speech spoofing attacks from bonafide human speech. Significant progress has been made in recent years. Contemporary state-of-the-art systems have showcased commendable performance, with some [4, 5, 6] achieving Equal Error Rate (EER) below 1% on ASVspoof2019 [7] test partitions. However, CM systems still suffer from generalization issues to unseen attacks and diverse acoustic environments, having shown strong degradation when evaluated on in-the-wild data [8, 9].

Singing voice deepfake detection, on the other hand, poses a distinct set of challenges not presented in speech. First, singing voices typically follow a specific melody or rhythmic pattern, which significantly affects the pitch and duration of different phonemes. Second, singing voices have more artistic voicing traits and a wider range of timbre compared to speech, and are prone to influence by musical context. Lastly, singing voices often undergo extensive editing, digital signal processing and are mixed with musical instrumental accompaniments. Recognizing these unique properties of singing voices, we question whether countermeasures developed for speech can be directly applied to singing voice spoof detection.

In this paper, we propose the Singing Voice Deepfake Detection (SVDD) task. As a first step, we curate the first in-the-wild dataset named SingFake to support this task. The SingFake dataset contains 28.93 hours of bonafide and 29.40 hours of deepfake song clips gathered from popular user-generated content platforms. Spanning five languages, we collect clips from 40 distinct singers and their AI counterparts. Additionally, we use a source separation model (Demucs [10]) to extract singing vocals from song mixtures, allowing us to examine the effects of singing vocals and song mixtures for SVDD systems separately. We also provide a train/validation/test split, where the test set contains a diverse set of scenarios, including unseen singers, languages, communication codecs, and musical contexts. With SingFake, we evaluate four types of leading speech countermeasure systems. We first use their models pretrained on speech utterances, and test them on the test split of SingFake. Results show a notable performance degradation compared to their performance on the ASVspoof2019 benchmark, on both song mixtures and separated vocals. We then retrain these systems on the training split of SingFake in two conditions, on separated vocal tracks and song mixtures, and test them on the test split. Results show significant improvement over the models trained on speech data. More detailed analyses of the results reveal challenges associated with unseen singers, communication codecs, languages, and musical contexts, underscoring the need for more focused research on crafting robust singing voice deepfake detection systems.

During the process of writing this manuscript, we discovered an-

[1] https://www.singfake.org/

Table

Splits	Description	# Si
Train	Training set	
Val	Validation set (unseen singers)	
T01	Test set for seen singer Stefanie Sun	
T02	Test set for unseen singers	
T03	T02 over 4 communication codecs	
T04	Test set for Persian musical context	

other very recent study [11] that assesses the performan
CMs on clean singing voices and mixtures of them with
music for Chinese songs. The study offers insight into s
capacity to learn deepfake cues under controlled idea
while our work focuses on more challenging in-the-wild

2. SINGFAKE DATASET CURATION

2.1. Data collection and annotation

We source deepfake singing samples from popular us
content websites where users upload both bonafide a
samples of singing. For every deepfake song sample, v
annotate its metadata: AI singer, language, website, an
a "deepfake" song. The deepfake generation models,
are also annotated. We then collect bonafide samples fro
sponding real singer and annotate them as "bonafide". I
samples, we identify many AI-generated performances
languages not sung by the original artists. Also, as th
mimic celebrities, we verify during annotation that tho
do not actually do the performances. These steps ensu
racy of our annotations. During the annotation process
that most people use SoftVC-VITS[2] with different ve
same uploader usually uses the same model to generate AI singer(s).
As manual annotation is a tedious and error-prone process, to ensure
the accuracy of metadata labels and correct potential inaccuracies,
we employ GPT4 [12] to verify the annotations against song titles
and descriptions. We then manually review any discrepancy found
by GPT4. This process yields our SingFake dataset, capturing real-
istic deepfake nuances prevalent in online communities. SingFake
contains 40 genuine singer and their corresponding AI singers, with
634 bonafide songs and 671 deepfake songs across multiple lan-
guages. The SingFake dataset has been open-sourced under the CC
BY-NC 4.0 license[3].

2.2. Dataset splits

Our primary guideline was to ensure that the singers were dis-
tinct across each section; therefore, we partition the data into
train/validation/test splits. Also, to form a comprehensive evalu-
ation on the robustness of SVDD systems, we employ test subsets
T01-T04 with increasing difficulties. Notably, there are many sam-
ples from the singer "Stefanie Sun", so we set aside a portion of
them, creating the T01 testing scenario to evaluate the system's per-
formance on a seen-in-training singer. T02 testing set contains six
unseen-in-training singers, while the T03 testing condition simulates
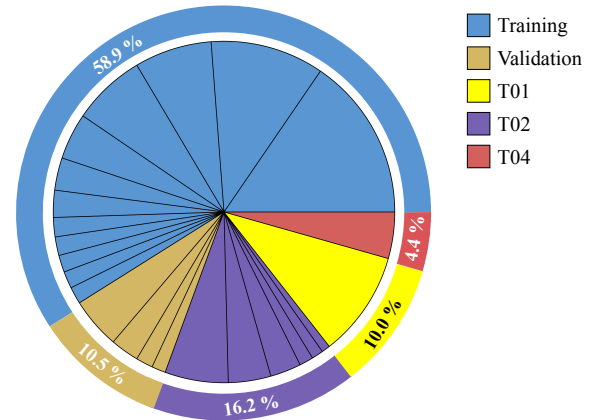the effect of media compression codecs by encoding T02 through

---

Fig. 1. SingFake dataset partition. Each color represents a subset, and each slice denotes an AI singer. T03 is excluded here since it contains the same song clips as T02 but is repeated 4 times through 4 different codecs.

MP3 128 Kbps, AAC 64 Kbps, OPUS 64 Kbps, and Vorbis 64 Kbps. All song clips in T02 are passed through all 4 codecs to form T03. The Persian singers stand out as they contain mostly persian language while showing different musical styles. To investigate the effects of potential disparities in language and musical style, we allocate the Persian singers to a separate T04 test set. As T04 is collected from different platforms from other testing conditions, we believe T04 also contains unseen codecs.

The rest of the dataset is split between training and validation, maintaining a rough song ratio of 6:1:1:2 between training, valida-tion, T01, and T02 subsets. Together, these subsets offer a compre-hensive evaluation of the singing voice deepfake detection systems. The final partitioning is illustrated in different colors in Figure 1.

2.3. Data processing

Much of the data we collected comprises pop songs, which typically feature both instrumental sections, such as Intro, Outro and Inter-lude; and sections containing vocals, such as Verse and Chorus. For the purpose of singing voice deepfake detection, our primary inter-est lies in segments containing vocals, as only the singing voices are synthesized, rendering pure instrumental sections irrelevant for the SVDD task. Consequently, we narrowed our focus to regions with active singing, treating each as a distinct song clip.

To extract active singing regions, we first employed the state-of-the-art music source separation model, Demucs [10][4], which is

---

built on a U-Net convolutional architecture with an attention-based bottleneck. This model is adept at separating vocals from musical accompaniments; the particular checkpoint we selected was trained on the MusDB [13] dataset with extra training data, and secured the 2nd position on track B of the MDX challenge [14]. We use the separated vocals from Demucs as the source of separated song clips.

Next, the separated vocals are processed through the Voice Activity Detection (VAD) pipeline from PyAnnote [15], which provides us with the timecodes for segmentation. These timecodes are subsequently used to segment both mixtures and vocals into individual song clips. All clips are resampled to 16 kHz during training and inference. For those songs originally in stereo, we maintained the stereo quality, but chose a random channel for each clip during training. The average length for clips in the dataset is 13.75 seconds.

The final statistics for all subsets including the splits at clip-level are shown in Table 1. We open-source the datasheet including original user-uploaded media links and our metadata annotations, dataset split generation and data processing code[5].

3. EXPERIMENTS

In this section, we first evaluate existing speech spoofing countermeasure systems using SingFake. Subsequently, we retrain these systems from scratch using the SingFake training set and assess their performance across various test scenarios with our dataset splits.

3.1. Experimental setup

We construct four state-of-the-art systems that have demonstrated remarkable performance on speech datasets, representing different levels of input feature abstraction. This allows us to assess these features on both speech and singing spoof detection tasks.

Model architectures: **AASIST** [4] uses raw waveform as feature, leverages graph neural networks and incorporates spectro-temporal attention. **Spectrogram+ResNet** uses a linear spectrogram extracted with 512-point FFT, with a hop size of 10 ms. We feed the extracted spectrogram into the ResNet18 [16] architecture. **LFCC+ResNet** [17] uses Linear-Frequency Cepstral Coefficients (LFCC) as speech features, then feeds the LFCC into the ResNet18 model. The 60-dim LFCCs are extracted from each frame of the utterances, with frame length set to 20ms and hop size 10ms. **Wav2vec2+AASIST** [18] is a model leveraging Wav2Vec2 [19], a self-supervised front-end trained on large-scale external speech datasets. Note that we removed the RawBoost data augmentation module from the original paper [18] for fair comparisons between methods, since no other method has such augmentation.

Evaluation metric: Each system produces a score for each utterance, indicating the confidence that the given utterance is bonafide. The Equal Error Rate (EER) is determined by setting a threshold on the produced scores where the false acceptance rate matches the false rejection rate. EER is widely used as an indicator for biometric verification systems' performance, and we think it is a good metric for SVDD as well.

3.2. Speech CM heavily degrades on SVDD task

We train and validate all speech CM systems on the speech dataset ASVspoof 2019 logical access (LA) [7] for 100 epochs. The model checkpoint with the best validation performance is selected for evaluation. We use the same train/dev/eval splits as ASVspoof 2019 LA.

Table 2. Test results on speech and singing voice with CM systems trained on speech utterance from ASVspoof2019LA (EER (%)).

Method	ASVspoof2019 LA - Eval	SingFake-T02 Mixture	Vocals
AASIST	0.83	58.12	37.91
Spectrogram+ResNet	4.57	51.87	37.65
LFCC+ResNet	2.41	45.12	54.88
Wav2Vec2+AASIST	7.03	56.75	57.26

To form batches, we use 4 seconds of audio, following [4]. We use repeat padding for shorter trials, and we randomly choose consecutive 4 seconds for longer trials. All of the CM systems achieve good performance on ASVspoof 2019 LA evaluation data, as shown in Table 2. The results of AASIST and LFCC+ResNet are also comparable to the published ones [4, 17] while the wav2vec2+AASIST are not since we did not apply the data augmentation as they did.

We then test them on the T02 condition of SingFake to evelute their performance on singing data. All systems show heavy degradation as shown in Table 3. The EERs are near 50% on song mixtures, indicating that the speech deepfake detection systems are not able to distinguish real singers and their corresponding AI singers in the existence of accompanying music. Interestingly, both spectrogram-based and raw-waveform-based systems achieved around 38% EER on the separated singing vocals, much better than the results on song mixtures. This might be due to the fact that singing vocals are more similar to speech compared to song mixtures since there would be nearly no music accompaniment presented after separation. However, the LFCC and Wav2Vec2-based systems are still performing near 50% EER, indicating that these speech features tend to overfit more to the speech data and cannot generalize to singing voices.

3.3. Training on singing voices improves SVDD performance

To investigate whether training on our curated SingFake dataset improves singing voice deepfake detection (SVDD) performance, we trained models using either full song mixtures (labeled as 'Mixture') or separated singing vocals (labeled as 'Vocals'). Training on mixtures provides raw information, while training on separated vocals reduces instrumental distraction but may introduce separation artifacts that mask deepfake cues.

As shown in Table 3, SVDD performance declined from the training set (all seen) to T01 (seen singers, unseen songs) to T02 (unseen singers, unseen songs), indicating increasing task difficulty. All systems achieved good training set performance, showing that SingFake is helpful in learning the SVDD task. We also observed that the LFCC+ResNet system achieved the lowest training set performance on mixtures and the second-best performance on separated vocals, suggesting that instrumental interference may heavily hurt the spectral envelope. However, the noticeable decline in T02 performance highlights the challenge of generalizing SVDD to new singers. T01 performance fell between the training set and T02, suggesting that deepfakes of seen singers are easier to detect in new songs than those of unseen singers.

Compared to CM systems trained on speech, those trained on SingFake have better performance in terms of EER on T02, suggesting that the systems trained on SingFake are better at detecting singing deepfakes. The systems trained on separated vocals in general achieve better performance than those trained on mixtures except Wav2Vec2+AASIST. This suggests that separated singing voices could highlight artifacts for detecting singing deepfakes.

12158

Table 3. Evaluation results for SVDD systems on all testing conditions in our SingFake dataset (EER (%))

Method	Setting	Train	T01	T02	T03	T04
AASIST	Mixture	4.10	7.29	11.54	17.29	**38.54**
	Vocals	3.39	8.37	10.65	13.07	43.94
Spectrogram+ResNet	Mixture	4.97	14.88	22.59	24.15	48.76
	Vocals	5.31	11.86	19.69	21.54	43.94
LFCC+ResNet	Mixture	10.55	21.35	32.40	31.85	50.07
	Vocals	2.90	15.88	22.56	23.62	39.27
Wav2Vec2+AASIST (Joint-finetune)	Mixture	**1.57**	**4.62**	**8.23**	13.62	42.77
	Vocals	1.70	5.39	9.10	**10.03**	42.19

Our results indicate that the Wav2Vec2+AASIST model excels in learning directly from song mixtures, delivering the most superior performance and robustness among all tested systems, similar to results reported for other tasks [18, 20]

3.4. SVDD systems show limited robustness to unseen scenarios

While training set, T01 and T02 represents more and more out-of-distribution sets at singer/song clips level, T03 and T04 sets are designed to evaluate performance in two challenging real-world situations: unseen communication codecs and unseen languages/musical contexts. Significant performance degradation is observed and well-studied under varying transmission and telecommunication codecs for speech CM systems [21, 9]. However, when testing our system on the T03 condition, the performance drop was not as much as anticipated. As social media platforms typically employ a diverse set of audio compression codecs to more efficiently stream and deliver user-uploaded content, we believe the SingFake data we collected already utilizes codecs. Thus, when training the SVDD system, the model inherently learns to form a more robust representation that generalizes well across lossy audio compression algorithms.

At the same time, we observe significant performance degradation across all SVDD systems on T04, which is noticeably more pronounced than on both T02 and T03. T04 and T03 vary by unseen language and musical context, hinting that challenges posed by these attributes are still prominent for SVDD systems.

4. DISCUSSIONS

The ability of AI to synthesize highly realistic singing voices demonstrate major technological progress, and also understandably cause public distrust, sometimes prompting calls to ban such technologies entirely. However, stopping advancement is rarely the answer. We believe transparency around content origins is key for establishing public trust, and more research into SVDD systems can allow users to make informed decisions about synthesized content. In this section, we summarize our findings on the strengths and weaknesses of SVDD systems.

Unseen communication codecs. We observed robust performance of SVDD systems against unseen compression codecs, as shown by the T03 example. This robustness differs from what we have seen with speech countermeasures. We hypothesize that this robustness stems from the SVDD systems being exposed to various compression codecs during training. Unlike speech deepfakes, most singing deepfakes are created for entertainment and are frequently posted on social media platforms where compression codecs are commonly applied. This entertainment focus and social media presence presumably leads to singing deepfakes incorporating more real-world compression compared to speech deepfakes.

Interference from backing tracks. SVDD systems need to work on mixtures containing vocals and instrumental tracks, where the prominence of the instrumental tracks can make it challenging to detect deepfake vocals, since they may mask deepfake artifacts and introduce new artifacts that cause the systems to fail. While using source separation might mitigate this problem, as discussed in Section 3.3, any less-than-perfect separation result could inadvertently introduce new artifacts or mask deepfake cues, which may then confound the deepfake detection algorithms. As an example, we found that both the Demucs model and PyAnnote VAD pipeline tend to classify string instruments as active voice. This misclassification may have contributed to performance degradation on T04, as Persian music is rich in string instrumentation. This vulnerability calls for developing interference-resilient SVDD systems and identifying more robust representations for this task.

Diverse musical genres. Singing voices in different genre follows significantly different musical context, exhibiting vastly different patterns of pitch, timbre and rhythm. By manual inspection, we discovered that the T04 subset contains many songs with heavy Hip-Hop influence, while most of the songs in other sets are rock and ballads. We believe this also contributes to the performance degradation seen on T04, indicating that SVDD systems fail to generalize to unseen musical genres. Since music reflects diverse cultural backgrounds, varying musical genres are likely to present in real-world SVDD situations. This vulnerability calls for future research efforts to disentangle musical genre effects from deepfake cues, enabling more genre-agnostic SVDD systems.

As SVDD systems advance, we anticipate them to help enhance confidence in AI technologies within the music industry, restoring trust that have been eroded by the rise of unauthorized deepfakes.

5. CONCLUSIONS

In this paper, we proposed the Singing Voice Deepfake Detection (SVDD) task and presented the Singfake dataset, containing a substantial collection of in-the-wild bonafide and deepfake song clips in various languages and singers. We demonstrated that state-of-the-art speech CM systems trained on speech show strong degradation when evaluated on singing voice, while re-training on singing voice leads to substantial improvements, highlighting the necessity of specialized SVDD systems. Additionally, we assessed the strengths and weaknesses associated with unseen singers, communication codecs, different languages and musical contexts, underscoring the need for robust SVDD systems. Through releasing the SingFake dataset and benchmarking systems on the SVDD task, we aim to catalyze more research focused on developing specialized techniques for detecting deepfakes in singing voices.

6. REFERENCES

[1] https://www.makemusic.sg/blog/wodeai, "My AI," 2023, Accessed: 2023-08-06.

[2] Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi, "VISinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7237–7241.

[3] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proc. AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 11020–11028.

[4] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[5] Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, and Zhao Lv, "Learning from yourself: A self-distillation method for fake speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] Siwen Ding, You Zhang, and Zhiyao Duan, "SAMO: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[8] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does Audio Deepfake Detection Generalize?," in *Proc. Interspeech*, 2022, pp. 2783–2787.

[9] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[10] Simon Rouard, Francisco Massa, and Alexandre Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[11] Yuankun Xie, Jingjing Zhou, Xiaolin Lu, Zhenghao Jiang, Yuxin Yang, Haonan Cheng, and Long Ye, "FSD: An initial chinese dataset for fake song detection," *arXiv preprint arXiv:2309.02232*, 2023.

[12] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[14] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, and Fabian-Robert Stöter, "Music demixing challenge 2021," 2021.

[15] Hervé Bredin and Antoine Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021, pp. 3111–3115.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] You Zhang, Fei Jiang, and Zhiyao Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[18] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2022, pp. 112–119.

[19] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 12449–12460.

[20] Mojtaba Heydari and Zhiyao Duan, "Singing beat tracking with self-supervised front-end and linear transformers," in *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, 2022.

[21] You Zhang, Ge Zhu, Fei Jiang, and Zhiyao Duan, "An empirical study on channel effects for synthetic voice spoofing countermeasure systems," in *Proc. Interspeech*, 2021, pp. 4309–4313.