ATTRIBUTABLE WATERMARKING OF SPEECH GENERATIVE MODELS

Yongbaek Cho, Changhoon Kim*, Yezhou Yang, Yi Ren*

Arizona State University, Tempe, AZ, USA

ABSTRACT

Generative models are now capable of synthesizing images, speeches, and videos that are hardly distinguishable from authentic contents. Such capabilities cause concerns such as malicious impersonation and IP theft. This paper investigates a solution for model attribution, i.e., the classification of synthetic contents by their source models via watermarks embedded in the contents. Building on past success of model attribution in the image domain, we discuss algorithmic improvements for generating user-end speech models that empirically achieve high attribution accuracy, while maintaining high generation quality. We show the tradeoff between attributability and generation quality under a variety of attacks on generated speech signals attempting to remove the watermarks, and the feasibility of learning robust watermarks against these attacks. Watermarked speech samples are available at `https://attdemo.github.io/attdemofull.github.io`.

***Index Terms*—** Speech Generation, Voice Impersonation, Speech Watermarking, Model Attribution

1. INTRODUCTION

Generative Adversarial Networks (GANs) [1] have achieved successes in generating artificial contents (e.g., images [2], videos [3], and audios [4]) that are almost indistinguishable from authentic contents. These models and their synthetic contents inevitably pose a variety of threats regarding privacy [5, 6], malicious impersonation [7], and copyright infringement [8]. Existing countermeasures to these threats can be categorized into detection [9] and attribution [10, 11] methods. The detection methods develop binary classifiers to distinguish between generated and authentic contents via intrinsic fingerprints of generative models; the attribution methods develop models from which generated contents are watermarked, so that they can be attributed to their source models via multi-class classification. Recent studies showed that detection may fail when intrinsic fingerprints are removed, e.g., through implicit neural representation [12]. Instead, our focus is attribution, which is much more difficult to spoof.

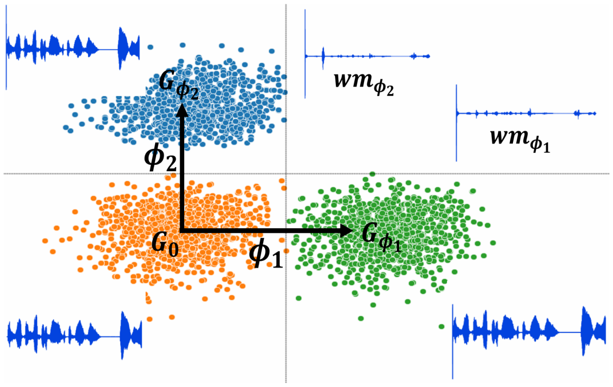


Fig. 1: MelGAN model distribution projected to the space spanned by user-specific keys ϕ_1 and ϕ_2. The default model G_0 is perturbed according to the keys to produce attributable models G_{ϕ_1} and G_{ϕ_2}, which add sparse watermarks (wm_{ϕ_1} and wm_{ϕ_2}) to the beginning of generated speeches.

Protocol: This study assumes the following model distribution and attribution protocol (Fig. 1): Consider a model developer who distributes copies of a generative model to its users (e.g., WaveGAN [13] and MelGAN [14]). Each user-end model $G_\phi : \mathcal{Z} \rightarrow \mathcal{X}$ maps the latent space $\mathcal{Z} \subset \mathbb{R}^{d_z}$ to the content space $\mathcal{X} \subset \mathbb{R}^{d_x}$, and has a key $\phi \in \mathbb{R}^{d_x}$ that defines its unique watermark. A third-party registry (e.g., law enforcement) manages all keys ($\Phi = \{\phi_i\}_{i=1}^N$) and the associated user IDs. The registry accepts contents in question (x), performs attribution via a sequence of binary classification, and returns IDs of the users (i) who generated the contents [10] ($\phi_i^T x > 0$).

Sufficient conditions for model attribution: Within this setting, Kim et al. [10] studied the sufficient conditions of keys to achieve certifiable attribution. Informally, a set of user-end models are attributable if (1) these models are distinguishable from the authentic dataset, and (2) the inner product of any pair of keys is smaller than a data-dependent threshold. These conditions guide the computation of keys.

Contributions: We claim the following contributions: (1) The algorithm proposed in [10] has only been tested on image generation models. This paper extends the domain to speech generation. We present improvements from [10] to address practical challenges encountered in the speech domain. Specifically, enforcing the alignment between user-end models and their corresponding keys is necessary for maintaining high model attributability. (2) We empirically test the trade-

*: Equal Contribution, YC, CK and YY are with the Active Perception Group at the School of Computing and Augmented Intelligence, Arizona State University. YR is with the School for Engineering of Matter, Transport, and Energy, Arizona State University.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Table 1: Evaluation results for various loss designs. When proposed configurations are applied, we achieved the best results. Aug.:augmented key, Dist.: distinguishability, Att.: attributability, FDSD: Fréchet Deep Speech Distance. ↓ / ↑ indicates lower/higher result is desirable. Base FDSD scores for WaveGAN and MelGAN are 25.65 and 4.74, respectively.

	Model	Dist.↑	Att.↑	FDSD.↓
A Baseline [10]	WaveGAN	0.68	0.1	27.28
	MelGAN	0.74	0.0	12.82
B + Aug.	WaveGAN	0.94	0.17	30.87
	MelGAN	0.99	0.68	21.85
C + L_d	WaveGAN	0.97	0.31	**26.67**
	MelGAN	0.99	0.73	7.32
D + L_A	WaveGAN	**0.98**	**0.94**	26.92
	MelGAN	**0.99**	**0.93**	**7.30**

off between generation quality and robust attributability under adversarial post-processing in the speech domain.

2. RELATED WORKS

Speech generative models: GANs [1] were invented to train neural networks which map latent vectors to real distribution via solving real/fake binary classification problem. This development brings great success to realistic data synthesis not only in the image domain but also in the speech domain. Many successful speech synthesis models are established based on GANs (e.g., WaveGAN and MelGAN). These speech models enable ordinary people to generate realistic fake audio, which warns society against misuse [15, 16]. As one possible solution, our model attribution reveals responsible user who synthesized audio.

Detection and attribution of generative models: Fake detection [9] boils down to binary classification between authentic and generated contents. Model attribution, on the other hand, relies on multi-class classification to trace the corresponding models of the generated contents [17, 18, 10, 11]. Among attribution methods are two directions: model structure attribution [17, 18] and user-end model attribution [10, 11]. Model structure attribution is multi-class classification problem of classifying synthesized contents into one of structures of generators (e.g., StyleGAN2 [2]). However, user-end model attribution classifies contents into responsible user's generator even the users' generators have same structure. In this work, we focus on user-end model attribution.

3. METHODS

3.1. Notations and preliminaries

Given an authentic dataset $\mathcal{D} \subset \mathbb{R}^{d_x}$, we assume the existence of a default generator G_0 for which the output distribution P_{G_0} matches with the authentic data distribution $P_\mathcal{D}$. Let the user-specific keys be $\Phi := \{\phi_1, \phi_2, ..., \phi_N\}$ for N users, where $\phi_i \in \mathbb{R}^{d_x}$ and $||\phi_i||_2 = 1$ for $i = 1, ..., N$. G_0 will be fine-tuned according to all $\phi_i \in \Phi$ to produce the set of user-end generators $\mathcal{G} := \{G_{\phi_1}, G_{\phi_2}, ...\}$ (see Sec. 3.4). Let the ith binary classifier be $f_{\phi_i}(x) = sign(\phi_i^T x)$, which returns 1 if and only if $x \in G_{\phi_i}$.

We introduce the following metrics to facilitate the discussion: (1) *Distinguishability* of G_ϕ measures the classification accuracy of $f_\phi(x)$:

$$D(G_\phi) := \frac{1}{2}\mathbb{E}_{\substack{x\sim P_{G_\phi} \\ x_0\sim P_\mathcal{D}}} \left[\mathbb{1}(f_\phi(x) = 1) + \mathbb{1}(f_\phi(x_0) = -1)\right], \tag{1}$$

where P_{G_ϕ} a user-end distribution. (2) *Attributability* measures the averaged classification accuracy of all models of the collection $\mathcal{G} := \{G_{\phi_1}, ..., G_{\phi_N}\}$:

$$A(\mathcal{G}) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{x\sim G_{\phi_i}} \mathbb{1}(\phi_j^T x < 0, \forall\, j \neq i, \phi_i^T x > 0). \tag{2}$$

(3) We measure the *generation quality* of G_ϕ by Fréchet DeepSpeech Distance [19].

3.2. Sufficient conditions for model attribution

We summarize the sufficient conditions for model attribution from [10] in Theorem 1.

Theorem 1. *We say ϕ is data-compliant when $\phi^T x < 0$ for $x \sim P_\mathcal{D}$. Let $d_{min}(\phi) = \min_{x\in\mathcal{D}} |\phi^T x|$, $d_{max}(\phi) = \max_{x\in\mathcal{D}} |\phi^T x|$. Then $A(\mathcal{G}) \geq \max\{0, 1 - N\delta\}$, if $D(G) \geq 1 - \delta$ for all $G_\phi \in \mathcal{G}$, and*

$$\phi^T \phi' \leq \min\left\{\frac{d_{min}(\phi)}{d_{max}(\phi)}, \frac{d_{min}(\phi')}{d_{max}(\phi')}\right\} \tag{3}$$

for any pair of data-compliant keys ϕ, $\phi' \in \Phi$.

From the theorem, model attribution requires keys to be designed in such a way that their corresponding user-end models satisfy (1) data compliance, (2) distinguishability, and (3) the minimal angle constraint in Eq. (3). It should be noted that a sufficient and computationally more feasible angle constraint is $\phi^T \phi' \leq 0$.

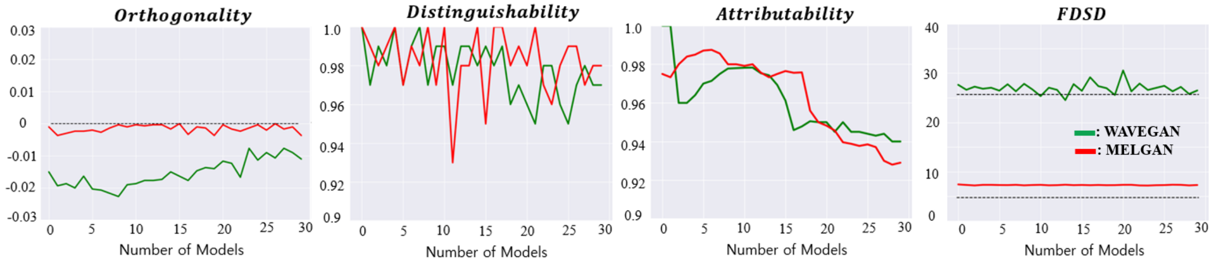
3.3. Key generation

We now discuss the computation of keys. First, we notice that for the tested speech data (SC09 [20], LJSpeech [21]), there does not exist data-compliant keys, i.e., no sub-space classifies the authentic data as one class. This is evident from the low distinguishability in Tab. 1A. We resolve this issue by adding a bias to the binary classifies: $f_{\phi_i}(x) = sign(\phi_i'^T x + b_i)$ for all $i = 1, ..., N$. The resultant distinguishability improves as seen in Tab. 1B. To reduce notational burden, we will denote $[\phi_i', b_i]$ by ϕ_i and the augmented data (with appended 1s) by x.

Each new key is generated by solving the following problem with existing keys ϕ_j for $j = 1, ..., i-1$:

$$\min_\phi \mathbb{E}_{x\sim G_0}[\max\{1 + f_\phi(x), 0\}] + \sum_{j=1}^{i-1} \max\{\phi_j^T \phi, 0\}. \tag{4}$$

The orthogonality penalty (second term of RHS) is omitted for the generation of the first key ($i = 1$).



(a) **(b)** **(c)** **(d)**

Fig. 2: (a-d) Average orthogonality, distinguishability, attributability, FDSD of 30 WaveGAN user-end models on SC09 and 30 MelGAN user-end models on LJSpeech, respectively. The dotted lines depict baselines.

Table 2: Evaluation metrics before (Bfr.) and after (Afr.) robust training against adversarial post-processes. Before robust training, FDSD scores of WaveGAN and MelGAN are 26.92 and 7.30, respectively (Tab. 1D). Dist. = Distinguishability, Att. = Attributability

Metric	Model	Noise		Gain		Speed		Pass filter		Combination	
		Bfr.	Afr.	Bfr.	Afr.	Bfr.	Afr.	Bfr.	Afr.	Bfr.	Afr.
Dist. ↑	WaveGAN	0.91	0.98	0.95	0.98	0.85	0.98	0.94	0.98	0.79	0.92
	MelGAN	0.97	0.99	0.88	0.97	0.60	0.86	0.80	0.99	0.73	0.95
Att. ↑	WaveGAN	0.88	0.96	0.94	0.98	0.71	0.90	0.64	0.91	0.31	0.73
	MelGAN	0.72	0.92	0.63	0.88	0.40	0.70	0.64	0.84	0.23	0.56
FDSD. ↓	WaveGAN	36.54		42.58		46.12		50.85		47.56	
	MelGAN	7.99		8.55		24.48		9.415		27.49	

3.4. Retraining of user-end generator

While Theorem 1 holds when G_ϕ models the perturbed distribution $\{x+\phi | x \in P_\mathcal{D}\}$, this exact match of distributions may not be achieved in practice due to the limited capacity of G_ϕ and the domain-specific boundaries of x (e.g., for speech data, $x \in [-1,1]^{d_x}$). We found through experiments that this mismatch deteriorates the attributability of user-end models. In the following, we describe a practical formulation for retraining the default model G_0 so that the resultant user-end model G_ϕ will (1) be distinguishable from the authentic data, (2) have low generation quality drop, and (3) be attributable.

Distinguishability loss: We use a standard hinge loss to penalize G_ϕ on distinguishability:

$$L_h = \mathbb{E}_{x \in P_{G_\phi}} \max\{1 - f_\phi(x), 0\}. \tag{5}$$

Generation quality loss: To discourage quality degradation, we introduce a loss that computes the expected distance between samples from the user-end and the default models. We utilize l_1 distance which gives better perceptual quality than l_2 distance [22]:

$$L_d = \mathbb{E}_{z \sim P_z}\left[\|G_0(z) - G_\phi(z)\|_1\right]. \tag{6}$$

Angle loss: Through experiments, we notice that the expected perturbation $\mathbb{E}_{z \sim P_z}\left[G_\phi(z) - G_0(z)\right]$ may not align with ϕ, which causes attributability to be lower than expected. See Tab. 1C. Therefore, we propose an angle loss to encourage the alignment:

$$L_A = \max\left\{1 - \frac{(G_0(z) - G_\phi(z)) \cdot \phi}{\|(G_0(z) - G_\phi(z))\|_2 \cdot \|\phi\|_2}, 0\right\}. \tag{7}$$

Tab. 1D shows the effectiveness of the angle loss at improving the empirical attributability of 30 user-end models. The training objective is thus the following:

$$\min_{G_\phi} \; \lambda_1 L_h + \lambda_2 L_d + \lambda_3 L_A, \tag{8}$$

where λ_1,λ_2 and λ_3 are set to 10, 10000, 1000, respectively. We optimize this loss function to create G_{ϕ_i} iteratively.

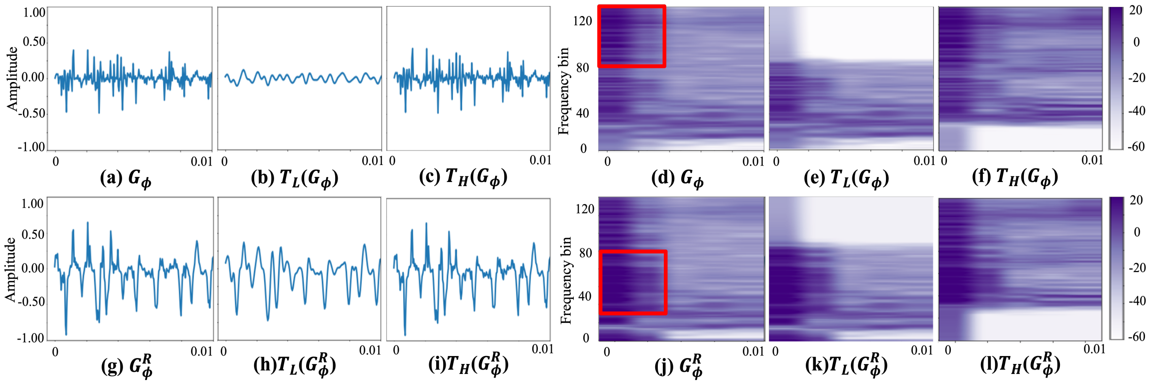
4. EXPERIMENTS

4.1. Experimental setup

Dataset: We tested our model attribution using SC09 [20] and LJspeech [21] datasets. SC09 is a subset of speech commands by a variety of speakers that include spoken ten vocabulary words from zero to ten each of a duration of 1 second. The dataset is split into training, test and validation sets consisting of 18.5k, 2.5k, and 2.5k data points, respectively. LJspeech contains 13.1k audio clips by a single speaker. We split LJspeech into 11.5k, 0.5k, 0.5k for training, test, and validation, respectively. **Model and training**: WaveGAN maps the latent vectors to audio samples and we directly employed SC09 dataset to train. MelGAN takes a mel-spectrograms and we cut each LJspeech clip to 3 seconds in length.

4.2. Experimental results

We report in Fig. 2 the average distinguishability, attributability, and the average generation quality (in terms of FDSD) of a sequence of user-end WaveGAN and MelGAN models being iteratively trained by solving Eq (4) and Eq. (8). Results with all 30 models are reported in Tab. 1D. It should be highlighted that the angle loss significantly improves the attributability of models, achieving 94% and 93% on WaveGAN and MelGAN, respectively. This shows that in practice, it is necessary to align the trained model G_ϕ with the corresponding ϕ.



(a) G_ϕ (b) $T_L(G_\phi)$ (c) $T_H(G_\phi)$ (d) G_ϕ (e) $T_L(G_\phi)$ (f) $T_H(G_\phi)$

(g) G_ϕ^R (h)$T_L(G_\phi^R)$ (i)$T_H(G_\phi^R)$ (j) G_ϕ^R (k)$T_L(G_\phi^R)$ (l)$T_H(G_\phi^R)$

Fig. 3: Results of robust training against pass-filter attack. (a,g): Audio signal from a non-robust generator G_ϕ and the corresponding robust generator G_ϕ^R. (d,j): Corresponding Mel-spectrogram of (a,g). The peak-amplitude regions are highlighted. (b,c,h,i,e,f,k,l): Audio signal and Mel-spectrogram after the attack. T_L / T_H indicates low and high pass-filter, respectively.

4.3. Adversarial post-processing

Lastly, we test the robustness of our method against various post-processes that aim at removing the watermarks from generated contents. Following the experimental setting in [23, 11, 10], we assume that the registry is aware of the distribution of attacks P_T, where $T : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x}$ can represent (1) adding noise, (2) gain, (3) changing speed, (4) combined pass filters, and the combination of these four. To train robust user-end models G_ϕ^R, we propose the following problem formulation for updating user-end models given ϕ:

$$\begin{aligned}\min_{G_\phi} \mathbb{E}_{T\sim P_T,\, x\sim P_{G_\phi}} &\left[\lambda_1 \max\left\{1 - f_{\phi_i}(T(x), 0)\right\}\right. \\ &\left.+\lambda_2 L_d + \lambda_3 L_A\right].\end{aligned} \tag{9}$$

Setup: Details of post-processes are as follows. *Noise*: Noise type is uniformly sampled from Brown, Blue, Violet, and Pink Noise. *Gain*: Gain multiplies a random amplitude to reduce or increase the volume. *Pass filter*: High and low pass filters are both considered. We set the cut off frequency to $[2200, 4000]$ for low pass filter and $[200, 1200]$ for high pass filter, respectively. It should be noted that the frequency ranges are chosen to avoid removing the semantic contents of the generated speeches. *Speed*: Speed perturbations speed up or slow down an audio signal with re-sampling. The speed percentage is uniformly chosen from $[80, 90, 110]$. Lastly, *combination* attacks combine the other four attacks, each with a 50% chance to be applied. **Results**: Tab. 2 reports the average distinguishability, attributability, and generation quality with and without robust training against post-processes. A tradeoff is observed between robust attributability and generation quality. To further understand the effectiveness of robust training, here we pick low/high-pass filters as the attack and compare a non-robust watermark and its corresponding robust version for MelGAN in Fig. 3a,g, as well as their filtered watermarks in Fig. 3b,c,h,i. We focus on the first 0.01 second of the signals where watermarks dominate. From the results, we see that robust training successfully finds watermarks that have frequency ranges in between the low- and high-pass filters. To further support this finding, we average Mel-spectrogram before and after filters over 1000 samples in Fig. 3(d-f, j-l). We reiterate that since attacks should avoid removing the semantic contents of a generated speech, there always exists a frequency window for which robust watermarks can be created.

4.4. Collusion Attack

We define a collusion attack as to merge multiple sources of the same content to produce a new copy that averages source watermarks and potentially makes it difficult to attribute [24]. However, this attack will not be successful with the presented definition of attributability. To explain, attacker(s) download two models of the collection G_1 and G_2, and interpolate the output by $x_{new} = \lambda x_1 + (1-\lambda)x_2$, where $x_i \in G_{\phi_i} \forall i \in \{1, 2\}$, and $\lambda \in [0, 1]$. When G_1 and G_2 are attributable, for any key ϕ that does not belong to G_1 or G_2, we have $\phi^T x_1 < 0$ and $\phi^T x_2 < 0$. Thus, $\phi^T(\lambda x_1 + (1-\lambda)x_2) < 0$.

5. CONCLUSION

We investigated the feasibility of model attribution in the speech domain. Our method is based on a protocol where the model distributor trains attributable user-end generative models by embedding unique watermarks. We showed that in practice, it is necessary to enforce the alignment between user-end models and their designated keys in order to achieve empirically high attributability in the speech domain. This is verified on WaveGAN and MelGAN trained on SC09 and LJSpeech datasets, respectively. Lastly, we revealed the tradeoff between generation quality and robust attributability.

6. ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation under Grant No. 2101052 and by an Amazon AWS Machine Learning Research Award (MLRA). Any opinions, findings, and conclusions expressed in this material are those of the author(s) and do not reflect the views of the funding entities.

7. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. 2014, vol. 27, Curran Associates, Inc.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Aidan Clark, Jeff Donahue, and Karen Simonyan, "Adversarial video generation on complex datasets," *arXiv preprint arXiv:1907.06571*, 2019.

[4] Yang Gao, Rita Singh, and Bhiksha Raj, "Voice impersonation using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[5] Danielle K Citron and Robert Chesney, "Deepfakes and the new disinformation war," *Foreign Affairs*, 2019.

[6] Raphael Satter, "Experts: Spy used ai-generated face to connect with targets," *Experts: Spy used AI-generated face to connect with targets*, Jun 2019.

[7] Jon Bateman, *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*, Carnegie Endowment for International Peace., 2020.

[8] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot, "Not my deepfake: Towards plausible deniability for machine-generated media," *arXiv preprint arXiv:2008.09194*, 2020.

[9] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, "Cnn-generated images are surprisingly easy to spot... for now," *arXiv preprint arXiv:1912.11035*, 2019.

[10] Changhoon Kim, Yi Ren, and Yezhou Yang, "Decentralized attribution of generative models," in *International Conference on Learning Representations*, 2021.

[11] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz, "Artificial gan fingerprints: Rooting deepfake attribution in training data," 2020.

[12] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov, "Image generators with conditionally-independent pixel synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14278–14287.

[13] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[14] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[15] Stupp Catherine, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," Aug 2019.

[16] Metz Rachel, "How a deepfake tom cruise on tiktok turned into a very real ai company," Aug 2021.

[17] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, "Do gans leave artificial fingerprints?," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 506–511.

[18] Michael Albright, Scott McCloskey, and ACST Honeywell, "Source generator attribution via inversion," *arXiv preprint arXiv:1905.02259*, 2019.

[19] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan, "High fidelity speech synthesis with adversarial networks," *arXiv preprint arXiv:1909.11646*, 2019.

[20] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[21] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[22] Ashutosh Pandey and Deliang Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5414–5418.

[23] Ning Yu, Larry Davis, and Mario Fritz, "Attributing fake images to gans: Analyzing fingerprints in generated images," *arXiv preprint arXiv:1811.08180*, 2018.

[24] Jonathan K Su, Joachim J Eggers, and Bernd Girod, "Capacity of digital watermarks subjected to an optimal collusion attack," in *2000 10th European Signal Processing Conference*. IEEE, 2000, pp. 1–4.