Attributable Watermarking of Speech Generative Models

by

Yongbaek Cho

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2021 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Yi Ren
Ni Trieu

ARIZONA STATE UNIVERSITY

December 2021

ABSTRACT

Generative models in various domain such as images, speeches, and videos are being developed actively over the last decades and recent deep generative models are now capable of synthesizing multimedia contents are difficult to be distinguishable from authentic contents. Such capabilities cause concerns such as malicious impersonation, Intellectual property theft(IP theft) and copyright infringement.

One method to solve these threats is to embedded attributable watermarking in synthesized contents so that user can identify the user-end models where the contents are generated from. 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. Existing studies showed the feasibility of model attribution in the image domain and tradeoff between attribution accuracy and generation quality under the various adversarial attacks but not in speech domain.

This work discuss the feasibility of model attribution in different domain and algorithmic improvements for generating user-end speech models that empirically achieve high accuracy of attribution while maintaining high generation quality. Lastly, several experiments are conducted show the tradeoff between attributability and generation quality under a variety of attacks on generated speech signals attempting to remove the watermarks.

DEDICATION

This thesis is dedicated to my parents who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve.

ACKNOWLEDGMENTS

I would like to express deepest gratitude to Prof. Yezhou Yang for welcoming me into Active Perception Group(APG) and providing me with the tools that I needed to choose the right direction and successfully complete my thesis. His invaluable supervision and continuous support have encouraged me in all the of my Master journey and engage me to actively participate in research culture with others. Next, I am extremely grateful to Prof. Yi (Max) Ren for his invaluable advise and patience during my Master milestone. His immense knowledge and plentiful experience have always helped me to formulate the research question and understand the good research. His invaluable advice led me in the right direction when I am stuck in my research. A special thanks to my constant partner Changhoon Kim in my MS journey. He always helped me to keep up my motivation and showed great enthusiasm.

I would like to thank all the members in the adversarial group - Sheng Cheng, Prasanth Buddareddygari, Travis Zhang, Richard Cheng and others for a cherished time spent together in the lab. It is their kind help and support that have made my study and life in the ASU a wonderful time. Finally, I thank my roommate Chunghwan Kim for being supportive during my MS journey.

Without their tremendous supporting and encouragement in the past few years, it would be impossible for me to complete my study.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Overview

Generative Adversarial Networks (GANs) have been significant improvement in recent years through contributions in image successfully. The improvements of GANs can lead to examine the scope of domain extensions and it is capable of generating the indistinguishable audio. Especially, speech is one of the most widely transmitted data in mobile devices and applications. Unlike the image, user can perceive not only the the content of the message but also other characteristics such as pitch, rhythm, genre etc.

In recent years, due to the advancements of speech generation, synthetic contents are almost indistinguishable from authentic contents. Although nobody wants to be fooled by deep fake voice, synthetic contents are used on social media platforms and there are numerous issues and societal awareness such as threatening intellectual property rights, fake news etc. These models and their artificial contents inevitably pose a variety of threats regarding privacy (Citron and Chesney 2019; Harris 2018; Satter 2019), malicious impersonation Bateman 2020, and copyright infringement B. Zhang et al. 2020.

Existing countermeasures to these issues are detection and attribution and these methodology can prevent the societal problems from deepfake. The detection methods develop binary classifiers to distinguish between generated and authentic contents via intrinsic fingerprints of generative models. On the other sides, the attribution

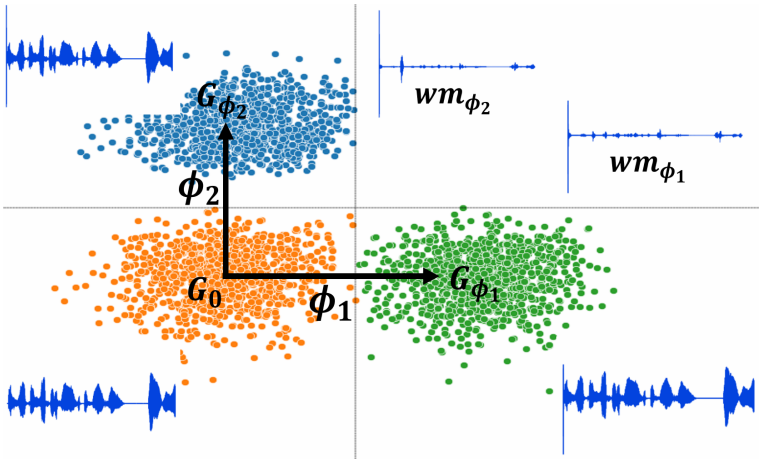


Figure 1. MelGAN model distribution projected to the space spanned by user-specific keys ϕ_1 and ϕ_2. The default model G_0 is perturbed according to the keys to produce attributable models G_{ϕ_1} and G_{ϕ_2}, which add sparse watermarks (wm_{ϕ_1} and wm_{ϕ_2}) to the beginning of generated speeches.

methods develop multi-class classification to identify models from which generated contents are watermarked, so that they can be attributed to their source models.

1.2 Motivation

GANs (Goodfellow et al. 2014) were invented to use neural networks which map latent vectors to real distribution via adversarial training. This development brings success to generate realistic data synthesis not only in the image domain (Karras et al. 2020) but also in the speech domain (Gao, Singh, and Raj 2018). Many successful speech synthesis models are established based on GANs(e.g. WaveGAN (Donahue, McAuley, and Puckette 2018) and MelGAN (Kumar et al. 2019)) and ordinary people enable to generate realistic fake audio which is indistinguishable from real audio. This advancement leads to the warnings against misuse.

Existing countermeasures to these threats can be categorized into detection (Wang

et al. 2019) and attribution (Kim, Ren, and Yang 2021) (Yu, Davis, and Fritz 2018) as possible solutions to prevent misuse. Fake detection boils down to binary classification between authentic and generated fake contents. However, recent works showed that fake detection may fail when intrinsic fingerprints are removed e.g., through implicit neural representation (Anokhin et al. 2021). For this reason, model attribution may address this problem. Model attribution is multi-class classification to detecting traces of specific models of the generated fake contents. There are two directions among the model attribution which are user-end model attribution and model structure attribution. User-end model attribution is a multi-class classification to identify contents into responsible user's generator even the users' generators have same structure. However, model structure attribution is to classify synthesized contents into one of structures of generators. This work aims to focus on user-end model attribution, which is difficult to spoof.

1.3 Challenges

The algorithm of model attribution proposed in (Kim, Ren, and Yang 2021) has only been tested on image domain, it has not been examined on speech generative models. Even though existing methods on model attribution seems practical, there are few practical challenges encountered in speech domain that need to be addressed to achieve high accuracy of attribution while maintaining the quality. Current methodology of attribution is not distinguishable and attributable in speech domain. Specifically, there are not aligned between user-end-models and their corresponding specific keys which cause the low attribution. Therefore, improving the accuracy of attributability corresponding specific keys and model is considered to be a key challenge.

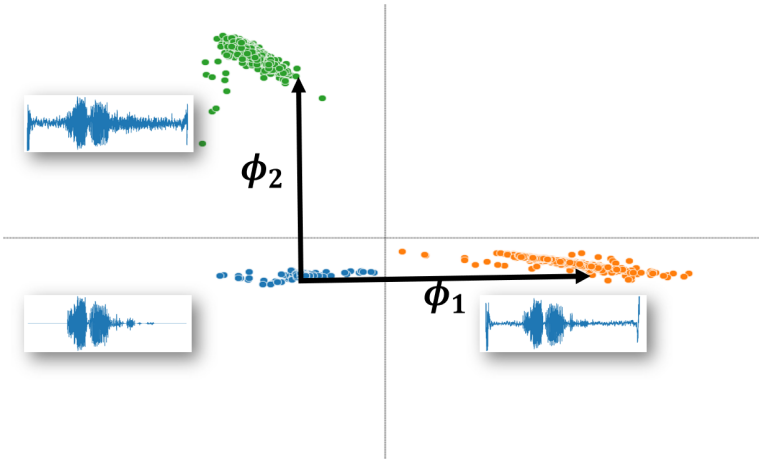


Figure 2. WaveGAN model distribution projected to the space spanned by user-specific keys ϕ_1 and ϕ_2. The default model G_0 is perturbed according to the keys to produce attributable models G_{ϕ_1} and G_{ϕ_2}.

Another key challenge in adopting existing method to speech domain is that generation quality does not guarantee the generation quality with robustness. Since image dataset (Y. Zhang et al. 2020; Cohen et al. 2017; F. Yu et al. 2015) is different to speech signal (Panayotov et al. 2015; Zen et al. 2019), expected perturbation may not align with designated keys. Hence, enforcing alignment between user-end-models and corresponding keys in order to enables the model distributor trains attributable user-end models which contain its unique watermarking is considered to be another challenging task.

1.4 Contribution

Since this research aims to algorithmic improvements of model attribution on speech generative models, this study extends the domain to speech generation. To solve our proposed practical challenges encountered in the speech domain, it is assumed that the attackers can post-process the generated contents to possibly decline the

model attribution and the registry is aware of the distribution of attacks. Overall, this research claims the contributions as follows:

1. Algorithmic development with new constraints is performed to enforce the alignment between user-end-models and their corresponding keys for maintaining the high model attributability.

2. When the adversarial post-processes are known, our proposed method can counter measure through robust training. However, this countermeasure will produce the degradation of generation quality . Further, the trade-off between generation quality and robust attributability under a variety of adversarial post-processing by adversaries in the speech domain is empirically tested to show the robustness of our proposed method.

Chapter 2

BACKGROUND

2.1 Speech Generative Models

Audio generative model (Oord et al. 2016) is introduced by using feedforward deep convolution neural networks can be trained with autoregression setting. Such approach takes the raw waveform as input and this architecture directly generates a raw audio waveform , showing reasonable results in text-to-speech and general audio generation. However, the model structure is computationally heavy for real-time speech synthesis to achieve best results. Whereas, lightweight GAN models e.g. MelGAN and Parallel WaveGAN (Yamamoto, Song, and Kim 2020) are presented and showed high perceptual quality on text-to-speech synthesis.

Generative Adversarial Networks(GAN) have been successfully demonstrated the fidelity at synthesizing not only images (Karras, Laine, and Aila 2019) but also in speech. (Engel et al. 2019) Donahue *et al.* (Donahue, McAuley, and Puckette 2018) are first demonstrated the adversarial audio synthesis with Deep convolution neural networks GANs. WaveGAN learns to produce the one words when trained on small vocabulary speech dataset. Their investigation focuses on unsupervised learning and it can learn from mapping the low-dimensional latent vector to high-dimensional audio signals with the architecture of traditional GAN. On criteria of their evaluation for sound quality, it can generate intelligible spoken digits to humans.

2.2 Threats of DeepFake

Advanced generative models in audio led to new societal problems and pose the particular security and privacy threats, voice impersonation to bio-metric authentication system on machine control (e.g., Apple Siri, Amazon Alexa and Google Home etc). Such high perceptual quality results from text-to-speech synthesis(TTS) and other approaches can successfully deceive both users and speaker verification system (Catherine 2019). At the same time, this has created the need for improving the system that can detect tempered contents and protect the real voice from malicious attacks (Citron and Chesney 2019). However, existing voice interfaces are vulnerable to attacked voice with tempered content and cloned voice that mimicking the authentic user (Bateman 2020; Rachel 2021). Therefore,there is a need for a model that can show robustness against these attacks (Liu et al. 2020) and not only detect forgery but also trace the responsible users. One promising solution to tackle these problems is to disclose the responsible user of misused contents.

2.3 Detection and Attribution

Fake audio are generated by manipulation with DeepFake methods (Satter 2019) and the public concern of misuse e.g. voice impersonation have become growing. Typically, recent studies presented two methods that are detection and model attribution to prevent these significant threats.

2.3.1 Detection

A detector for distinguishing between real and fake contents is introduced by (Wang et al. 2019) for many Convolution Neural Networks(CNN) image generative models. In general, Fake Detection (Zhou and Lim 2021) has been commonly based on intrinsic fingerprints and relies on binary classification between authentic and generated contents. However, this methodologies are highly dependent on specific training scenario and therefore detection can be easily removed the watermarking by some attacks. In other words, traditional fake detection is not robust against a variety of attacks and is vulnerable to content spoofing.

2.3.2 Attribution

Model attribution is dependent on multi-class classification to trace the corresponding models of the generated contents. Yu *et al.* (Yu, Davis, and Fritz 2018) studied empirical feasibility of attribution through a classifier trained on user-end models. However, this approach does not guarantee an attributability in reality when the number of user-end models grows.

Thus, Kim *et al.* studied the sufficient conditions with decentralized scheme to achieve the certifiable attribution on image dataset. They proposed the decentralized attribution scheme as follows: After training speech generative model using our method, the model owner is able to release different copies of the model which include different unique keys but the similar quality of generations. When a user request to download the model, the owner register user's profile with the unique key into the database. Then, each user can download the copy of the model with or without knowledge that

the key is embedded. When misused contents of models are reported, the model owner can collect the misused contents and trace responsible users. Informally, if (1) user-end models are distinguishable from the authentic dataset, and (2) the inner product of any pair of keys is smaller than a data-dependent threshold, a set of user-end models can be attributable.

Existing model attribution based on user-end models on image domain showed the robust against a variety of attacks and achieved high attributability while maintaining generation outputs quality with additional loss of generation quality. Since all of the previous research is implemented and studied only in the image domain, our proposed work is the initial approach that is practical and attributable in the speech domain.

2.3.3 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. Each user-end model $G_\phi : \mathcal{Z} \to \mathcal{X}$ 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 (Kim, Ren, and Yang 2021) ($\phi_i^T x > 0$).

2.3.4 Sufficient conditions for model attribution

Within this setting, Kim et al. (Kim, Ren, and Yang 2021) 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.

Chapter 3

METHODS

In this chapter, the methods on model attribution in speech domain will be mentioned in details.

3.1 Notation and Preliminaries

For a given an authentic dataset $\mathcal{D} \subset \mathbb{R}^{d_x}$, the model owner train a default generator G_0 for which the output distribution P_{G0} that matches with the authentic data distribution P_D. Since the purpose of this work is not improving the quality, the assumption for attribution is that P_D and P_0 are almost same. Before training User-end generators, the owner firstly need to produce user-specific keys. Let the user-end 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 using Φ to produce user-end generators $\mathcal{G} := \{G_{\phi_1}, G_{\phi_2}, ...G_{\phi_N}\}$ that are denoted by $G_\phi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$, where d_z and d_x are dimension of generator input and data in P_D, respectively. Let the ith binary classifier as $f_{\phi_i}(x) = sign(\phi_i^T x)$ that returns 1 only if $x \in G_{\phi_i}$, which is similar to the one versus all classification.

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

$$D(G_\phi) := \frac{1}{2}\mathbb{E}_{x\sim P_{G_\phi}, x_0 \sim P_{\mathcal{D}}} \left[1(f_\phi(x) = 1) + 1(f_\phi(x_0) = -1)\right], \tag{3.1}$$

where P_{G_ϕ} a user-end distribution.

(2) *Attributability* measures the averaged classification accuracy of all models of the collection $\mathcal{G} := \{G_{\phi_1}, ..., G_{\phi_N}\}$:

$$A(\mathcal{G}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{x \sim G_{\phi_i}} 1(\phi_j^T x < 0, \forall\ j \neq i, \phi_i^T x > 0). \tag{3.2}$$

For measuring the generation quality of attributed G_ϕ, Fréchet DeepSpeech Distance (Bińkowski et al. 2019) is computed in this study.

3.2 Sufficient conditions for model attribution

The summary of the sufficient conditions for model attribution from (Kim, Ren, and Yang 2021) in Theorem 1.

Theorem 1 *Let 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

$$\phi^T \phi' \leq \min \left\{ \frac{d_{min}(\phi)}{d_{max}(\phi)}, \frac{d_{min}(\phi')}{d_{max}(\phi')} \right\} \tag{3.3}$$

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

From the theorem, Distinguishability is achieved under two sufficient conditions of keys that the orthogonality of each key ($\phi_i^T \phi_j = 0, i \neq j$) and data complaint ($\phi_i^T x < 0, \forall i$) should be guaranteed. Additionally, the minimal angle constraint between any pair of keys should be orthogonal. Specifically, it should be noted that a sufficient and computationally more feasible angle constraint is $\phi^T \phi' \leq 0$.

3.3 Key Generation

Data compliance, i.e., $\phi_i^T x < 0, \forall i, \forall \in P_{G0}$, is one of the sufficient conditions for the model attribution. However, there does not exist data-complinat keys, i.e., no sub space classifies the authentic data as one class for the tested speech dataset(SC09, LJSpeech). which cause both low distinguishability and low attributability. 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 $j = 1, ..., i-1$:

$$\min_{\phi} \mathbb{E}_{x \sim G_0}[\max\{1 + f_\phi(x), 0\}] + \sum_{j=1}^{i-1} \max\{\phi_j^T \phi, 0\}. \tag{3.4}$$

The orthogonality condition apply from the generation of second key. In other words, The orthogonality penalty (second term of RHS) is omitted for the generation of the first key ($i = 1$). By solving this equation, trained keys are mutually orthogonal and satisfies data compliance. The set of keys ϕ is not fixed. If the owner of model needs more, keys can be trained based on eq. (3.4)

3.4 Retraining of user-end generator

After training user-specific key ϕ, the owner's generator G_0 should be fine-tuned based on corresponding key ϕ. This essential training step enables each generators'

contents to be attributable. 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 Model attribution does to train the binary classifiers $f_\phi(x)$ parameterized by the key ϕ using a standard hinge loss to classify each user-end generative models by the decision boundary given by the key. Standard hinge loss used in this study to penalize G_ϕ on distinguishability:

$$L_h = \mathbb{E}_{x \in P_{G_\phi}} \max\{1 - f_\phi(x), 0\}. \tag{3.5}$$

Generation quality loss To discourage quality degradation, Mean Absolute Error(MAE) as a loss function that computes the expected distance between samples from the user-end and the default models. Therefore, the distance between the user-end generative models G_ϕ and the default model G_0 is computed for the generation quality. Pandey *et al.* (Pandey and Wang 2018) showed that utilizing l_1 distance gives better perceptual quality than l_2 distance.

Angle loss Through experiments, the expected perturbation $\mathbb{E}_{z \sim P_z} [G_\phi(z) - G_0(z)]$ may not align with ϕ, which causes attirbutability to be lower than expected. In a nutshell, our angular loss adjusts the angle between G_ϕ and specific key ϕ so that make the high attributability be attainable. Thus, angle loss function encourages to align with and user-end generator G_ϕ to improve attributability. The following loss function is angular loss:

$$L_A = \max\left\{1 - \frac{(G_0(z) - G_\phi(z)) \cdot \phi}{\|(G_0(z) - G_\phi(z))\|_2 \cdot \|\phi\|_2}, 0\right\}. \tag{3.6}$$

The training objective is thus the following:

$$\min_{G_\phi} \ \lambda_1 L_h + \lambda_2 L_d + \lambda_3 L_A, \tag{3.7}$$

where λ_1,λ_2 and λ_3 are set to 10, 10000, 1000, respectively. For satisfying the all metrics, optiming this loss function to create $G_{\phi i}$ iteratively.

Algorithm 1 Retraining of G_ϕ

Input : ϕ, G_0
Output G_ϕ

1: Train the $G_{\phi 1}$ with corresponding ϕ_1 by solving

$$\min_{\phi} \mathbb{E}_{x \sim G_0}[\max\{1 + f_\phi(x), 0\}] \tag{3.8}$$

2: **if** $D(G_{\phi 1}) < 0.97$ **then**
3: goto step 1 ;
4: **end if**
5: Train new keys by solving the Eq. (3.4)
6: **if** $\cos(\phi_1, \phi_i) > 0.015$ **then**
7: goto step 5 ;
8: **end if**
9: Train new user-end generators by solving the Eq. (3.7)
10: **if** $D(G_{\phi i}) < 0.97$ **then**
11: goto step 9 ;
12: **end if**

Chapter 4

EXPERIMENTS

In this section, the result of our proposed method are introduced to evaluate and show the attributable watermarking. First, we show the experimental setting on the tested dataset and models in speech domain. Finally, we propose few evaluation metrics to quantify the model attribution on speech generative models.

4.1 Experimental setup

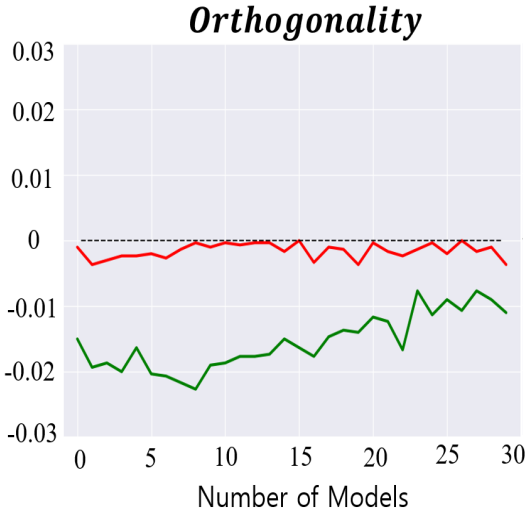
Dataset We tested our model attribution using SC09 (Warden 2018) and LJspeech datasets (Ito and Johnson 2017). SC09 is a subset of speech commands by a variety of speakers that include spoken ten vocabulary words from zero to ten each of a duration of 1 second. The dataset is split into training, test 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 and has been widely used in text-to-speech synthesis model and speech synthesis model. We split LJspeech into 11.5k, 0.5k, 0.5k for training, test, and validation, respectively.

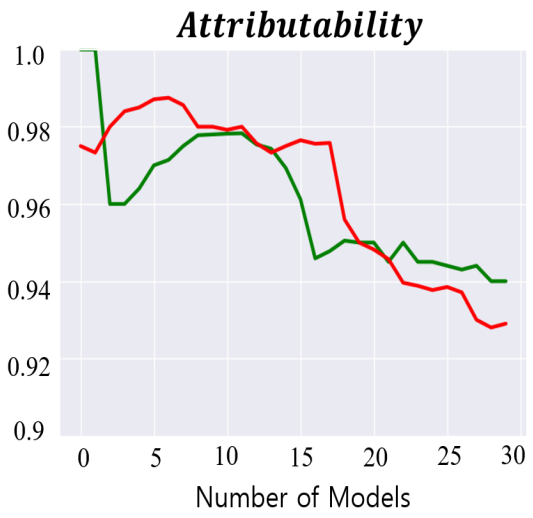
Model and training: In our experiments, all models were trained on NVIDIA TESLA V100 GPUs for the LJSpeech at 22,050 Hz and SC09 at 16,384 Hz. WaveGAN maps the latent vectors to audio samples and we directly employed SC09 dataset to train. MelGAN takes a mel-spectrograms and we cut each LJspeech clip to 3 seconds in length.

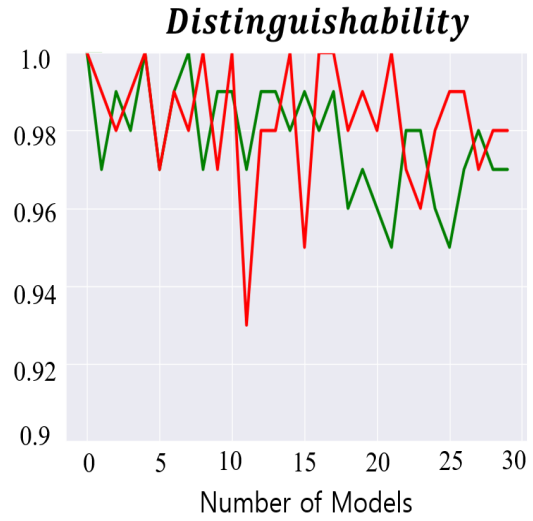
Table 1. Evaluation results for various loss designs. When proposed configurations are applied, we achieved the best results. Aug.:augmented key, Dist.: distinguishability, Att.: attributability, FDSD: Fréchet Deep Speech Distance. ↓ / ↑ indicates lower/higher result is desirable. Base FDSD scores for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

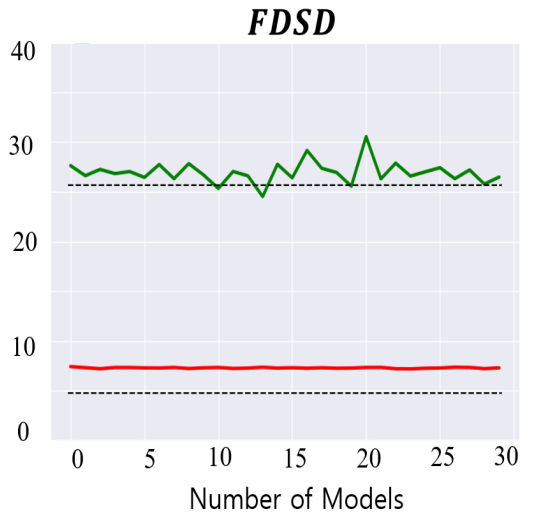
	Model	Dist.↑	Att.↑	FDSD.↓
A + Baseline	WaveGAN	0.68	0.1	27.28
	MelGAN	0.74	0.0	12.82
B + Aug.	WaveGAN	0.94	0.17	30.87
	MelGAN	0.99	0.68	21.85
C + L_d	WaveGAN	0.97	0.31	**26.67**
	MelGAN	0.99	0.73	7.32
D + L_A	WaveGAN	**0.98**	**0.94**	26.92
	MelGAN	**0.99**	**0.93**	**7.30**

Figure 3. (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.




Attributability
1.0
0.98
0.96
0.94
0.92
0.9
0
5
10
15
20
25
30
Number of Models


Distinguishability
1.0
0.98
0.96
0.94
0.92
0.9
0
5
10
15
20
25
30
Number of Models


FDSD
40
30
20
10
0
0
5
10
15
20
25
30
Number of Models

4.2 Experimental results

All models are trained following steps proposed in the Sec. (Kim, Ren, and Yang 2021). We report in Fig. 3 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 (3.4) and Eq. (3.4). Results with all 30 models are reported in Tab. 1D. We experimentally demonstrate that improving the baseline configuration by adding a bias, angular loss, and l_1 distance metric instead of l_2 (Tab. 1). First, Augmented dimension with using vectors as a bias input enables to keys to be achieved the data compliant. As can be observed from Fig. 3, the orthogonality of keys learned by augmented feature vector as a bias input helps to be linearly separable and distinguishable by decision boundary given by keys. We found that the distinguishability was remarkably improved in WaveGAN and MelGAN almost 26% and 25%, respectively (Tab. 1(B)). Since trained user-end models G_ϕ are not aligned with the specific key, trained user-end models with baseline configuration is not attributable. 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.

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 (Yu, Davis, and Fritz 2018; N. Yu et al. 2020; Kim, Ren, and Yang

2021), we assume that the registry is aware of the distribution of attacks P_T, where $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ 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:

$$\min_{G_\phi} \ \mathbb{E}_{T\sim P_T,\ x\sim P_{G_\phi}} \left[\lambda_1 \max\left\{1 - f_{\phi_i}(T(x), 0)\right\} \quad + \lambda_2 L_d + \lambda_3 L_A\right]. \tag{4.1}$$

4.3.1 Setup

In our experiments, we consider five types of post-processes T: noise (Yakura and Sakuma 2018), gain (Das et al. 2018), combined pass-filter (Wen et al. 2009), speed change (Xie et al. 2020), and combination of these four. *Noise*: Most works considers only white noise for post-processing. However, we consider the noise diversity and randomly chosen from 4 noise types. Noise type is uniformly sampled from Brown, Blue, Violet, and Pink Noise.

Gain: Gain multiplies a random amplitude factor to reduce or increase the volume. Gain randomly performs with gain in dB [-18, 6].

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.

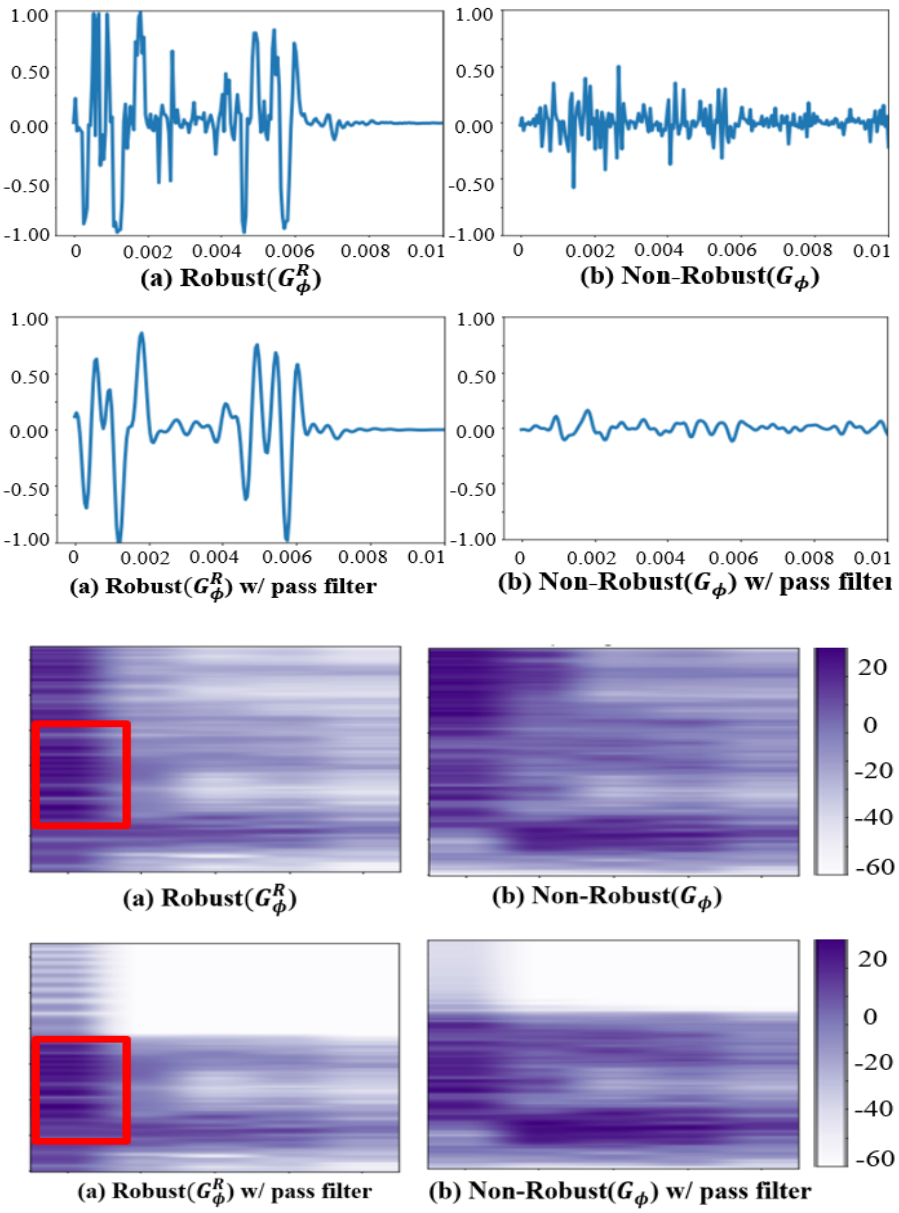


Figure 4. One sample from robust training against low pass-filter attack.

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 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	

Fréchet DeepSpeech Distance(FDSD): Recent work proposed the Fréchet Audio Distance (Kilgour et al. 2019). However, this metric has been designed for only music dataset (Sturm 2012) and uses a classifier as a feature extraction. Thus, it is not suitable to evaluate text-to-speech models or speech generative models. Binkowski *et al.* (Bińkowski et al. 2019) presented the Fréchet DeepSpeech Distance(FDSD) that evaluate the perceptual quality of synthesized speech samples based on their distance to a reference set. FDSD is conceptually similar to Fréchet Inception Distance(FID) (Heusel et al. 2017) that is commonly used for evaluating the output content from GANs. The only difference is that FDSD are computed on the activation function of the Deep Speech 2 (Hannun et al. 2014). In general, Mean Opinion Score (MOS) (Leng et al. 2021) used to judge the quality of synthesized audio assigned by human. However, this metric is subjective test so that it is not well-suited to evaluate in our case. Therefore, MOS is not tested in this research work.

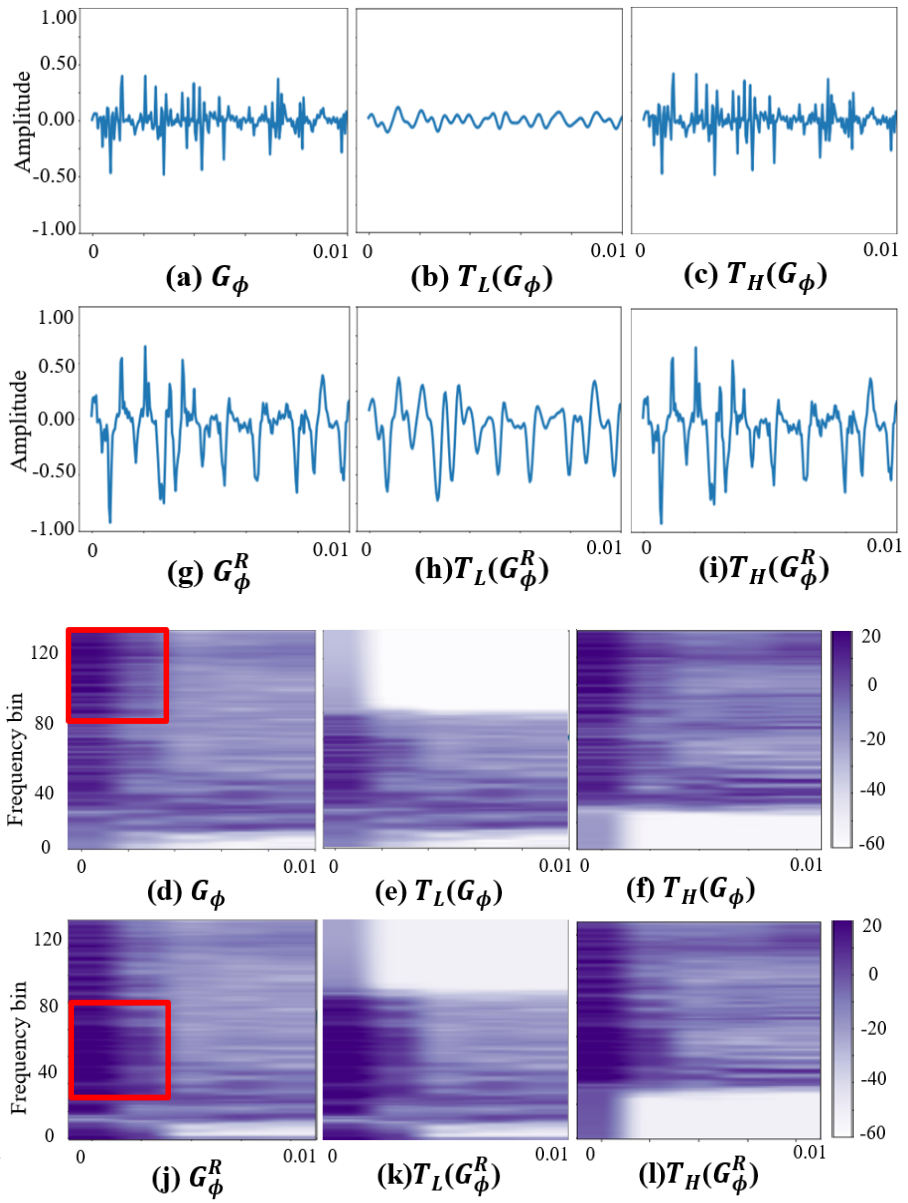


Figure 5. Results of robust training against pass-filter attack. (a,g): Audio signal from a non-robust generator G_ϕ and the corresponding robust generator G_ϕ^R. (d,j): Corresponding Mel-spectrogram of (a,g). The peak-amplitude regions are highlighted. (b,c,h,i,e,f,k,l): Audio signal and Mel-spectrogram after the attack. T_L / T_H indicates low and high pass-filter, respectively.

4.4 Robustness

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. 5a,g, as well as their filtered watermarks in Fig. 5b,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. 5(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.

Chapter 5

CONCLUSION AND FUTURE WORK

Motivated by model attribution in the image domain, we investigated the feasibility of model attribution in the speech domain. Same with previous works, 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 using SC09 and LJSpeech datasets, respectively. Lastly, we revealed the tradeoff between generation quality and robust attributability.

We achieved an algorithmic improvements for embedding attributable watermarking on speech generative models and showed the robustness of model attribution against several adversarial post-processing. Further, there are mainly two important future directions that can be evolved from this work. Model attribution has been successfully adopted on both image and speech generative models. However, the video generative models clark2019adversarial and multi-modal generative models ma2019m3d has been growing rapidly and the model attribution on these generative models have not been explored. In terms of multi-modal model, the generator takes image and audio as input and generate the synthetic content of combined image and audio. We can assume that attacker can easily use an audio extractor to split audio or image from video and Graphics Interchange Format (GIF). In this case, the attributable key can be broken and it means that the watermark is possible to be easily removed. The improvements could be studying the existence of our model attribution in different multimedia content. Another interesting in further improvements is to maintain the

attributability in large set of keys and it is necessary to find approximated solutions to this problem. For real applications, the capacity of attributable models should be explored so that this problem may be solved with sphere packing problems which is also known open challenges.

REFERENCES

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

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

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

Catherine, Stupp. 2019. "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case" (August). https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

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

Cohen, Gregory, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. 2017. "EMNIST: Extending MNIST to handwritten letters." In *2017 International Joint Conference on Neural Networks (IJCNN),* 2921–2926. IEEE.

Das, Nilaksh, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E Kounavis, and Duen Horng Chau. 2018. "Adagio: Interactive experimentation with adversarial attack and defense for audio." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* 677–681. Springer.

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

Engel, Jesse, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. "Gansynth: Adversarial neural audio synthesis." *arXiv preprint arXiv:1902.08710.*

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

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial

Nets." In *Advances in Neural Information Processing Systems,* edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567.*

Harris, Douglas. 2018. "Deepfakes: False pornography is here and the law cannot protect you." *Duke L. & Tech. Rev.* 17:99.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems* 30.

Ito, Keith, and Linda Johnson. 2017. *The LJ Speech Dataset.* https://keithito.com/LJ-Speech-Dataset/.

Karras, Tero, Samuli Laine, and Timo Aila. 2019. "A style-based generator architecture for generative adversarial networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4401–4410.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. "Analyzing and Improving the Image Quality of StyleGAN." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* June.

Kilgour, Kevin, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms." In *INTERSPEECH,* 2350–2354.

Kim, Changhoon, Yi Ren, and Yezhou Yang. 2021. "Decentralized Attribution of Generative Models." In *International Conference on Learning Representations.* https://openreview.net/forum?id=_kxlwvhOodK.

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

Leng, Yichong, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. 2021. "MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 391–395. IEEE.

Liu, Xiaolei, Kun Wan, Yufei Ding, Xiaosong Zhang, and Qingxin Zhu. 2020. "Weighted-sampling audio adversarial example attack." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34:4908–4915. 04.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499.*

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "Librispeech: an asr corpus based on public domain audio books." In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP),* 5206–5210. IEEE.

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

Rachel, Metz. 2021. "How a deepfake Tom Cruise on TikTok turned into a very real AI company" (August). https://www.cnn.com/2021/08/06/tech/tom-cruise-deepfake-tiktok-company/index.html.

Satter, Raphael. 2019. "Experts: Spy used AI-generated face to connect with targets." *Experts: Spy used AI-generated face to connect with targets* (June). https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

Sturm, Bob L. 2012. "An analysis of the GTZAN music genre dataset." In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies,* 7–12.

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

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

Wen, Xiumei, Xuejun Ding, Jianhua Li, Liting Gao, and Haoyue Sun. 2009. "An audio watermarking algorithm based on fast fourier transform." In *2009 International*

Conference on Information Management, Innovation Management and Industrial Engineering, 1:363–366. IEEE.

Xie, Yi, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. 2020. "Enabling fast and universal audio adversarial attack using generative model." *arXiv preprint arXiv:2004.12261.*

Yakura, Hiromu, and Jun Sakuma. 2018. "Robust audio adversarial example for a physical attack." *arXiv preprint arXiv:1810.11793.*

Yamamoto, Ryuichi, Eunwoo Song, and Jae-Min Kim. 2020. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6199–6203. IEEE.

Yu, Fisher, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop." *arXiv preprint arXiv:1506.03365.*

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

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

Zen, Heiga, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. "LibriTTS: A corpus derived from LibriSpeech for text-to-speech." *arXiv preprint arXiv:1904.02882.*

Zhang, Baiwu, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. 2020. "Not My Deepfake: Towards Plausible Deniability for Machine-Generated Media." *arXiv preprint arXiv:2008.09194.*

Zhang, Yuanhan, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. 2020. "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations." In *European Conference on Computer Vision,* 70–85. Springer.

Zhou, Yipin, and Ser-Nam Lim. 2021. "Joint Audio-Visual Deepfake Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 14800–14809.

APPENDIX A

AUDIO RESULTS

Watermarked speech samples are available at