HOW DISTINGUISHABLE ARE VOCODER MODELS? ANALYZING VOCODER FINGERPRINTS FOR FAKE AUDIO

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, vocoders powered by deep neural networks (DNNs) have found much success in the task of generating raw waveforms from acoustic features, as the audio generated becomes increasingly realistic. This development however raises a few challenges, especially in the field of forensics, where the attribution of audio to real or generated sources is vital. To our knowledge, our investigation constitutes the first efforts to answer this question on the existence of vocoder fingerprints and to analyze them. In this paper, we present our discoveries in identifying the sources of generated audio waveforms. Our experiments conducted on the multi-speaker LibriTTS dataset show that (1) vocoder models do leave model-specific fingerprints on the audio they generate, and (2) minor differences in vocoder training can result in sufficiently different fingerprints in generated audio as to allow for distinguishing between the two. We believe that these differences are strong evidence that there exist vocoder-specific fingerprints that can be exploited for source identification purposes.

1 INTRODUCTION

Over the past decades, advancements in text-to-speech (TTS) and voice conversion (VC) technologies permit the artificial generation of speech audio. As the last step in TTS or VC pipelines, generating raw waveforms from acoustic features, vocoders are important components in these tasks. With the development of deep learning, deep neural networks (DNNs) have seen increasing usage in both TTS and VC. Recent advancements in these areas (Wang et al., 2017; Shen et al., 2018; Kalchbrenner et al., 2018; Li et al., 2019; Ren et al., 2022) have led to the capacity of generating increasingly realistic and natural audio. While such technological progress is certainly welcomed, it does also admittedly come with challenges, especially when one considers the potential misuse and abuse of TTS and VC technologies, either inadvertently or maliciously. The risks and challenges associated with these technologies mostly arise in the following areas:

Audio forensics. As TTS technologies generate increasingly realistic audio, they also become increasingly prone to being misused, by their very nature, to mislead or even defraud victims who cannot judge the authenticity of audio. The possibility of misuse and abuse evidently gives rise to concerns about security, and public order. Therefore, it is consistent with the public interest to find methods to combat such risk of malicious usage by coming up with effective forensic strategies. While modern audio forensic techniques used in challenges such as ASVspoof (Wu et al., 2015; Wang et al., 2020; Yamagishi et al., 2021) and ADD (Yi et al., 2022) have achieved effective results in separating fake audios from real ones, they focus more on performing binary classification between real and fake sets. There is nonetheless also an interest in surpassing the constraints of binary real/fake classification and actually pinpoint the source responsible for generating any fake audio.

Intellectual property protection. Mainstream TTS and VC systems such as Shen et al. (2018), Arik et al. (2018), and Ren et al. (2022) are built with much effort, as they require careful design and fine-tuning of models and algorithms, as well as extensive training with more data than can reasonably be collected by any one person. As such, it is in the interest of researchers and commercial software developers alike to ensure that they are properly recognized to hold the ownership and copyright of their intellectual properties, the enforcement of which would not be feasible without the

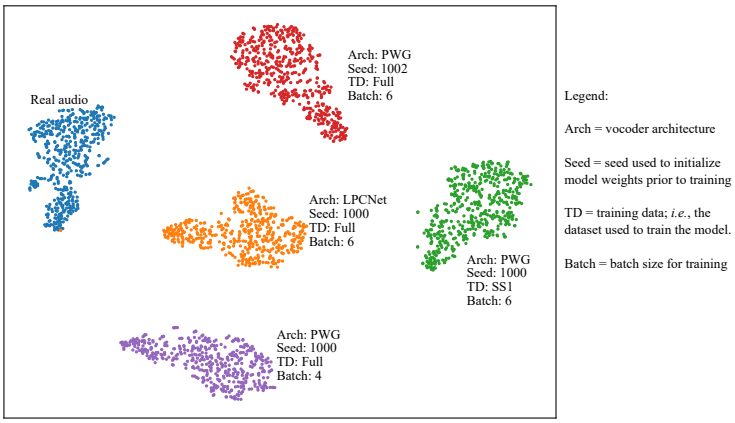


Figure 1: A t-SNE projection of vocoder fingerprints belonging to models of different architectures, initialization seeds, and hyperparameters, obtained in our experiments. ("PWG" refers to the Parallel WaveGAN architecture (Yamamoto et al., 2020); "Full" refers to the clean subset of LibriTTS (see section 4.1); SS1 consists of 50,000 randomly selected samples from the "Full" set)

ability to detect unauthorized usage of a system. While there have been efforts in watermarking a given DNN model (Zhang et al., 2018), these efforts require crafted inputs on models that have been specifically tweaked for watermarking during training, instead of relying on properties intrinsic to the vocoder model in performing waveform generation.

Owing to the reasons above, it is important to be able to identify the source of fake audio. Given the advancement of TTS and VC technologies, and judging by the results obtained in work by other scholars, where generated audios are often judged to be quite natural and realistic, we consider it infeasible to manually perform such scrutiny and source identification. Consequently, this calls for a way to automatically extract relevant information from audio samples and classify fake audio based on its source of generation, with minimal human intervention.

While the idea of source-dependent fingerprints have been explored on GAN-generated images (Yu et al., 2019; Marra et al., 2019), those efforts rely on image properties and are not directly applicable to waveforms generated by vocoders. While there has been some preliminary investigation on the concept of vocoder fingerprints (Yan et al., 2022), the effort of the study rests on an architecture-specific level. We seek to find the existence of fingerprints that allow for distinguishing models sharing the same architecture, separating audios generated by different vocoder models, and identifying their respective sources of generation.

In this paper, we put forth and seek to explain the following questions:

1. Do vocoder models leave model-specific fingerprints on the audio they generate?
2. What is the condition for their existence? In other words, how similar can two vocoder model instances be before it becomes infeasible to distinguish between the two?
3. How do we present these fingerprints for the purposes of fake audio source identification?

To answer these questions, we perform experiments to evaluate our hypothesis that vocoder models used to generate audio leave a fingerprint on the audio they generate that is specific to the model, as well as to explore how these fingerprints may be visualized and exploited for the aforementioned purposes of audio forensics and intellectual property protection. Our experiments confirm the existence of these fingerprints and show that vocoder fingerprints depend on architecture, training data, initialization seeds, and training hyperparameters, as shown in Figure 1.

2 Related Work

As our work deals with finding vocoder-specific features to aid in fake audio source identification, a task that falls in the purview of forensics, it behooves us to reiterate some of the work done by other scholars in the areas of vocoders and audio forensics.

VOCODERS

Vocoders refer to signal processing devices that synthesize speech audio waveform from some more compressed form of data, usually acoustic features, although certain vocoders take linguistic features as inputs (Tan et al., 2021). Apart from some modern end-to-end systems that directly generate waveform from text, vocoders form the last step in TTS and VC. Notable traditional (*i.e.*, non-neural) vocoder implementations include: Griffin-Lim (Griffin & Lim, 1984), STRAIGHT (Kawahara, 2006), and WORLD (Morise et al., 2016).

With the development of deep learning and neural networks came a variety of neural vocoders, of which the notable ones include: WaveNet (van den Oord et al., 2016; 2018), WaveRNN (Kalchbrenner et al., 2018), LPCNet (Valin & Skoglund, 2019a;b). Given the length of waveforms, autoregressive waveform generation takes much inference time; therefore modern neural vocoders have taken a non-autoregressive approach with generative models, among which are FloWaveNet (Kim et al., 2019), Parallel WaveGAN (Yamamoto et al., 2020), HiFiGAN (Kong et al., 2020), Multi-band MelGAN (Yang et al., 2020), Style-MelGAN (Mustafa et al., 2021). These non-autoregressive models are able to achieve faster-than-real-time inference speeds.

AUDIO FORENSICS

One of the primary aspects of audio forensics, known as "integrity verification", deals with the determination of authenticity of a given audio clip, *i.e.*, if the given audio represents real speech and if it has been tempered with (Maher, 2009; Zakariah et al., 2018). Various techniques have been used to analyze details and features contained in the audio, such as the Electronic Network Frequency (Nicolalde-Rodriguez et al., 2010), background noise (Malik, 2013), recording device (Qamhan et al., 2021), or energy (Luo et al., 2018).

In recent years, there has been significant progress in integrity verification of fake audio. Many challenges have accelerated the progress and research in this area, including the ASVspoof challenges (Wu et al., 2015; Kinnunen et al., 2017; Wang et al., 2020; Yamagishi et al., 2021). In addition to these, ADD 2022 (Yi et al., 2022) initiated the first audio deep synthesis detection challenge for more challenging realistic scenarios. In summary, the research on fake audio detection can be broadly divided into two parts: front-end features and back-end classifiers. In the first category, many features (Todisco et al., 2017; Davis & Mermelstein, 1980; Sahidullah et al., 2015) have been widely used for detection. As for the back-end classifier, many effective classifier designs (He et al., 2016; Bagherinezhad et al., 2017; Hu et al., 2018; Tak et al., 2021)have contributed to the detection research.

3 METHODOLOGY

3.1 VOCODER FINGERPRINT

To answer the questions we put forth, we study the impact of each of the following aspects of vocoder models:

1. **Vocoder architecture:** We procure 7 models, each with a different architecture, ranging from traditional signal processing vocoders to more modern GAN-based vocoders, and study the separability of the fingerprints left by these models on the audio they generate. The architectures are chosen to represent both traditional and neural vocoders. Specifically, WORLD is a traditional signal processing vocoder that achieves promising results; LPCNet utilizes linear prediction, and achieves considerable performance despite the limitations of autoregressive vocoders; the GAN-based vocoders are chosen for their performance reported in their respective papers, achieving faster-than-real-time inference speeds given their nonautoregressive nature.
2. **Training data:** In the training process, each training sample may differently influence the direction to optimize the objective function, hence result in slightly different weights when the model converges. To study the impact of training data, we train 4 models of the same architecture with the same initialization seed, but each model is trained with a slightly

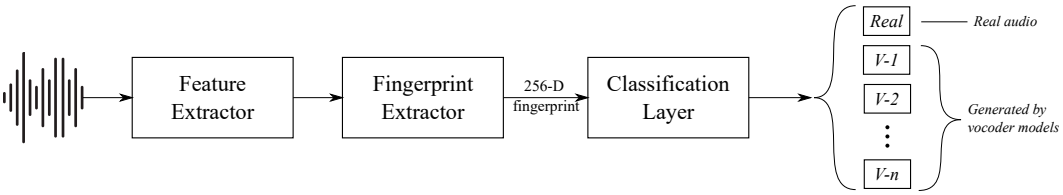


Figure 2: The overall fingerprint extraction and classification pipeline.

different training set that differ from each other in different magnitudes (see Table 4), and analyze the fingerprints left by the models.

3. **Initialization seeds:** Initialization seeds determine the initial weight values before training begins, and thus can potentially subtly influence the resulting weights as the model converges. To examine the impact of initialization seeds, we train 4 models of the same architecture, with the exact same training set, but each model is initialized with a different seed. We then analyze the fingerprints left by the models on the generated audio.

4. **Training hyperparameters:** We train 5 models of the same architecture, with the same training set and initialization seed, but adjust the dropout rate and batch size used in the training of each model. We then analyze the fingerprints left by the models on the audio. The choice of dropout rate and batch size are justified because dropout rate, by "freezing" certain weights, can make the vocoder models less "reactive" to incoming training samples, while batch sizes influence the order in which the samples contribute to the tweaking of model parameters.

3.2 VOCODER MODEL

Investigating the existence of vocoder fingerprint requires focus on vocoder-specific information and features, and as such, all other variables need to be fixed, with the vocoder model used to generate audio waveforms being the only independent variable in our experiments. To accomplish this, we generate fake audio by performing copy synthesis on real audio samples from the LibriTTS dataset. This is done by extracting acoustic features from these real audio samples and passing the extracted features through trained vocoder models to generate new waveforms, which will then be used for fingerprint extraction classification. The purpose behind this design is to eliminate any risk of potential interference that could be induced by acoustic models. Furthermore, since mainstream TTS and VC pipelines like Tacotron 1/2 (Wang et al., 2017; Shen et al., 2018) and FastSpeech 2 (Ren et al., 2022) still rely on vocoder models as the final step to generate audio waveforms, we think that vocoder differences are the most likely to cause fingerprint differences.

In addition, samples in the real audio corpus exhibit some variation in terms of speech content and speaker characteristics, the effect of which should not be relied on for the purpose of source identification; evidently, the text or even phonemic content of sentences being spoken as little to do with vocoders. To account for and preclude the impact of this naturally-occurring variance, the generated audio waveforms from each model mirror and correspond to each other as well as real audio waveforms. More specifically, for every real audio sample x, we generate fake audio samples $\{\hat{x}_m \mid m = 1, \ldots, M\}$ using each of the M models studied in the experiment, so that groups of real-and-fake audio waveforms from the same acoustic features maximally balance out the effect of the naturally occurring variations intrinsic in the real audio corpus.

3.3 FINGERPRINT EXTRACTION PIPELINE

As the motivation behind fingerprint extraction and analysis is essentially to facilitate generation source identification of fake audio, the fingerprints themselves should be such that audio samples from a single source — either real recordings or generated by a single vocoder model — form distinct, preferably linearly-separable, clusters in the feature space that can then be easily distinguished even with a single fully-connected layer (FC layer). As such, we train our fingerprint extraction network essentially as part of a classification model.

As shown in Figure 2, Each audio sample is first passed through a feature extractor to extract its LFCC features, and the features are then fed into the fingerprint extractor, which is a Res2Net network. The network produces a vector for every input, and the vector is then passed into a FC classification layer to be classified into a finite number of labels, representing the source of the audio (either real or the model that generated it).

3.4 EVALUATION METRIC

As discussed in section 3.3, the performance of fingerprint extraction and analysis should be directly related to the ability to classify the audio samples and identify their source of generation, as such, we evaluate our fingerprint extraction model in a way similar to classifier evaluations: by using precision, recall, and F1 scores. While there exists some potential downfalls of F1 scores in evaluating the accuracy of a (binary) classification model especially when the dataset is imbalanced (Chicco & Jurman, 2020) , our audio corpus generated through copy synthesis, as described above in sections 3.2 and 4.1.2, yields a balanced distribution of all classes, mitigating the weaknesses of F1 score as an evaluation metric.

4 EXPERIMENTS AND RESULTS

In this section, we present our experiments as well as the results and observations obtained. We perform four experiments, each of which focuses on one independent variable, while the rest are controlled and kept the same across models.

For each audio waveform sample, we first extract its linear frequency cepstral coefficient (LFCC) representation, with 20 LFCC per frame, using a Hamming window with the window length of $M = 480$ and hop length of $H = 240$. The LFCC matrix is then truncated or 0-filled to contain 500 frames. The LFCC representation of the audio sample is then passed to a Res2Net network (Gao et al., 2021) and projected onto a feature space with a dimensionality of 256. The feature vectors are then passed through a single FC layer classifier to decide between the finite number of sources.[1] The entire network (extraction + classification) is trained using cross-entropy loss.

4.1 DATASET

4.1.1 VOCODER TRAINING DATASET

In our experiments, we use the LibriTTS dataset (Zen et al., 2019) to train the neural vocoder models as well as to generate the audio corpus for vocoder fingerprint extraction network through copy synthesis. LibriTTS is a dataset originally designed for TTS that contains over 585 hours of speech data from over 2400 speakers, with roughly equal male-to-female ratio. In particular, to reduce the effect of unintelligible speech and background noise, we decided to employ only the clean speech subset of the dataset, which contains 262 hours of speech from 1230 speakers, while preserving the roughly equal male–female ratio, as shown in Table 1a.

With the exception of the second experiment described below in section 4.3, the training set (*i.e.*, the union of the subsets labelled "train-clean-100" and "train-clean-360") of LibriTTS is used in its entirety to train the neural vocoder models. This ensures that the training dataset is a controlled variable in these experiments.

4.1.2 EXTRACTION NETWORK TRAINING DATASET

To train our extraction–classification network in each experiment, we generate a dataset specifically for that experiment, containing audio generated by every vocoder model studied in the experiment (and thus the sample can be labeled with the ID of the vocoder model). We obtain the corpora through copy synthesis.

We randomly sample 50,000 samples from the aforementioned LibriTTS training set, 5,000 samples from the validation set, and 4,500 samples from the testing set, as shown in Table 1b, to serve as the real audio portion of the classifier training set the basis to perform copy synthesis. Each audio

[1] At most 7 classes in our experiments.

Table 1: Datasets used in our experiments.

(a) Dataset used in our experiments to train vocoder models: all audio comes from LibriTTS. This table is excerpted from Zen et al. (2019).

		Speakers			
Subset	Hours	Female	Male	Total	Utterances
train-clean-100	53.78	123	124	247	33,236
train-clean-360	191.29	430	474	904	116,500
dev-clean	8.97	20	20	40	5,736
test-clean	8.56	19	20	39	4,837
Total	262.60	592	638	1,230	160,369

(b) Real audios used for copy synthesis to generate training data for fingerprint extraction–classification.

Subset	Speakers	Utterances
train	1129	50,000
dev	40	5,000
test	39	4,500
Total	1208	59,500

Table 2: Results (%) on the impact of vocoder architectures (4.2)

	Precision	Recall	F1 score
REAL	98.30	97.49	97.89
PWG	99.42	99.87	99.65
HIFIGAN	97.73	98.36	98.04
MB_MELGAN	99.62	99.56	99.59
S_MELGAN	99.73	99.82	99.78
WORLD	99.80	99.58	99.69
LPCNET	100.00	99.93	99.97
Average	99.23	99.23	99.23

Table 3: Results (%) on the impact of training data (4.3)

	Precision	Recall	F1 score
REAL	99.78	99.58	99.68
SS0	93.14	92.56	92.84
SS1	92.79	95.58	94.17
SS2	97.76	94.09	95.89
SS3	93.94	95.42	94.68
Average	95.48	95.44	95.45

waveform is passed through each model studied by the experiment to create a balanced corpus for training the fingerprint extraction model. The fake audio part thus mirrors the real audio portion, with each class essentially mirroring the real audio set once by the nature of copy synthesis.

4.2 Impact of vocoder architectures

In our first experiment, we test for the existence of architecture-specific fingerprints in the generated waveforms. We acquire models of the following 6 architectures:

- **WORLD** (Morise et al., 2016): estimates the fundamental frequency (F0), spectral envelope, and aperiodic parameters, as well as generates speech similar to some input speech based on the estimated features.
- **LPCNet** (Valin & Skoglund, 2019a): a neural vocoder combining linear prediction and recurrent neural networks (RNNs) and generates audio waveforms from Mel spectrum.
- **Parallel WaveGAN** (PWG; Yamamoto et al., 2020): a neural vocoder that generates audio from Mel spectrum using a generational adversarial network (GAN).
- **HiFiGAN** (Kong et al., 2020): a GAN-based neural vocoder that employs two discriminators to perform multi-scale and multi-period discriminators.
- **Multi-band WaveGAN** (Yang et al., 2020): a GAN-based neural vocoder with a generator that produces sub-band signals from which the full-band signals are constructed.
- **Style-MelGAN** (Mustafa et al., 2021): a GAN-based neural vocoder that allows for the synthesis of high-fidelity speech with relatively low computational complexity.

We then tag the models as WORLD, LPCNET, PWG, HIFIGAN, MB_MELGAN and S_MELGAN, respectively. Other than WORLD, which does not require training, each model is trained with the entirety of the LibriTTS clean speech subset described in section 4.1.

We then train the extraction–classification network described in section 3.3 to distinguish between audio waveforms from 7 classes: {REAL, WORLD, LPCNET, PWG, HIFIGAN, MB_MELGAN, S_MELGAN}. To this end, we sample 50,000 waveform files from each source for training and 5,000 files for validation.

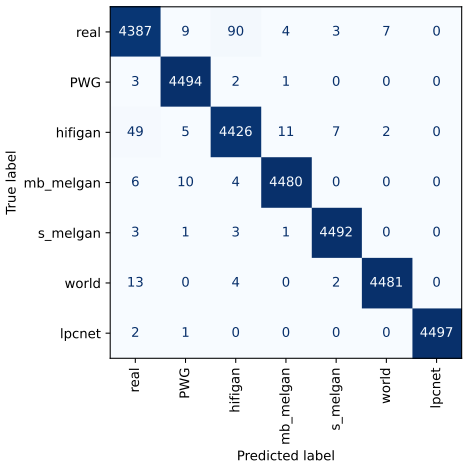


Figure 3: Confusion matrix on the impact of vocoder architectures

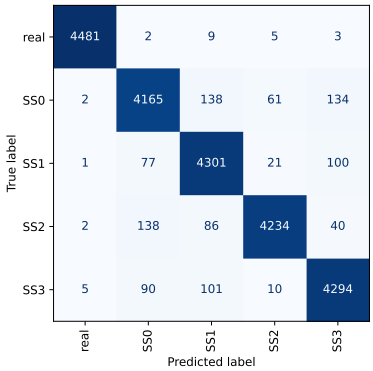


Figure 4: Confusion matrix on the impact of training data

Table 4: Difference (in utterances) of training data used in the experiment

	SS0	SS1	SS2	SS3
SS0	-	1	100	10,000
SS1	1	-	101	10,001
SS2	100	101	-	10,100
SS3	10,000	10,001	10,100	-

Table 5: Vocoder models with different hyper-parameters

	Dropout	Batch size	# Steps trained
BASE	0	6	500,000
1DROPOUT	0.1	6	500,000
2DROPOUT	0.2	6	500,000
4BATCH	0	4	750,000
8BATCH	0	8	375,000

In this experiment, we obtained results reported in Table 2 and Figure 3, showing that models with different architectures produce clearly distinguishable audios. The only relative confusion in this experiment is between the original (real) audio and the audio generated with HiFiGAN, as all other misclassifications constitute less than 1% of the evaluation set. Furthermore, if we were to evaluate the separability between generated audios from different models, no misclassification amounts to more than 0.5% of all audio samples of its real class. As different vocoder architectures take different approaches to construct audio waveforms from spectrograms or other feature representations, this separability is to be expected.

4.3 IMPACT OF TRAINING DATA

In our second experiment, we explore the relationship between the training dataset used to train the vocoder model and its fingerprint on generated waveforms. We fix the models to a single GAN architecture: Parallel WaveGAN (Yamamoto et al., 2020), and use slightly different training datasets to train them. This variance is controlled by sampling 50,000 waveform files from the LibriTTS clean speech set for training and 5,000 waveforms for testing. We tag this base set as SUBSET-0. Three more training subsets of the same sizes are then constructed, each differing from the base set by replacing 1, 100, and 10000 samples, respectively. We use tags $\{\text{SUBSET-}i \mid 1 \leq i \leq 3\}$ to denote them. The four subsets are used as training sets train 4 Parallel WaveGAN models $\{\text{SS}i \mid 0 \leq i \leq 3\}$. The training processes of the four models do however share the same validation set, since validation data has no bearing on model parameters. We then train the extraction–classification network described in section 3.3 to distinguish between audio waveforms from 5 classes: $\{\text{REAL}, \text{SS}i \mid 0 \leq i \leq 3\}$. To this end, we sample 50,000 waveform files from each source for training and 5,000 files for validation.

Table 3 and Figure 4 display the classification results and confusion matrix, respectively. The results obtained from the experiment corroborate our hypothesis that fingerprints are vocoder-specific on a finer granularity than architecture difference. Furthermore, it is shown that even a minor dif-

Table 6: Results (%) of on the impact of initialization seeds (4.4)

	Precision	Recall	F1 score
REAL	99.82	99.84	99.83
SEED-1000	94.51	94.91	94.71
SEED-1001	94.49	94.42	94.45
SEED-1002	94.37	97.16	95.74
SEED-1003	95.52	92.33	93.90
Average	95.74	95.73	95.73

Table 7: Results (%) of on the impact of hyperparameters (4.5)

	Precision	Recall	F1 score
REAL	99.76	99.84	99.80
BASE	97.38	94.93	96.14
1DROPOUT	97.70	97.07	97.38
2DROPOUT	99.12	98.09	98.60
4BATCH	95.66	97.96	96.79
8BATCH	95.08	96.71	95.89
Average	97.45	97.43	97.43

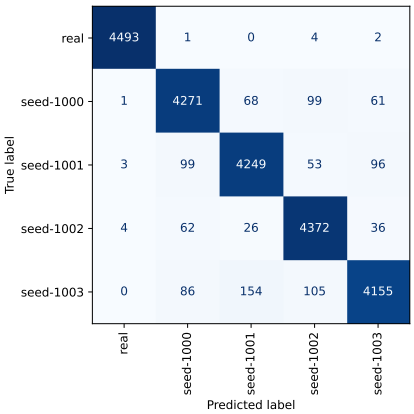


Figure 5: Confusion matrix of Experiment 3 (varied initialization seeds)

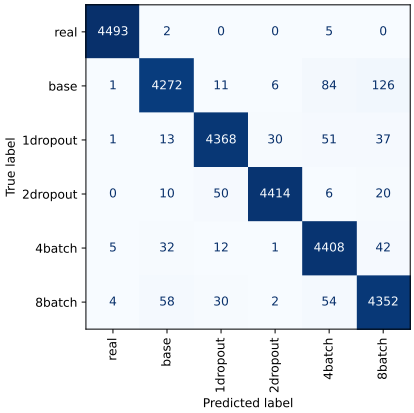


Figure 6: Confusion matrix of Experiment 4 (varied training hyperparameters)

ference in vocoder model training (as is the case in SS0 vs. SS1, where the training sets differ by only one sample) can result models that produce sufficiently different fingerprints as to allow for distinguishing between audio generated by the two models.

Interestingly, if one were to use classification accuracy as the sole metric to judge the separability between the fingerprints, then it would seem that the magnitude of difference in training data has no meaningful impact on separability. Two models with training sets differing by 1 sample (accounting for 0.002% of the training set) still produced easily distinguishable fingerprints, as our results clearly show. This observation could be unfortunate for those seeking to employ source-identification for the purpose of dataset protection, as minimal tweaking of the dataset could invalidate the identification system by producing wildly different fingerprints.

4.4 Impact of Initialization Seeds

In our third experiment, we investigate an even smaller variance factor: that of the initialization seeds. To accomplish this, we limit the models to the Parallel WaveGAN architecture, just as in the second experiment discussed above; we also use the same training data to train all models, and instead manually vary the seeds used in the initialization of model parameters. Specifically, we use seeds $s \in \{1000, 1001, 1002, 1003\}$ to initialize 4 models that we tag as $\{\text{SEED-}s\}$, which are then trained with the entirety of the LibriTTS clean speech subset. We then train the classifier to distinguish between 5 classes: $\{\text{REAL}, \text{SEED-}s \mid s \in \{1000, 1001, 1002, 1003\}\}$. We again sample 50,000 waveform files from each source for training and 5,000 files for validation.

Table 6 and Figure 5 display the classification results and confusion matrix, respectively. These results show that, even when models share the same architecture and are trained with the same data, the variation of initialization seeds can still result in sufficiently different fingerprints for distinguishing between audio generated by the models. Given that the weight parameters of the vocoder models are supposed to converge towards a configuration that minimizes the objective function, our observation

from this experiment entails that differences in vocoder weights can result in different fingerprints being produced by the models.

4.5 IMPACT OF TRAINING HYPERPARAMETERS

Our fourth and final experiment reported in this paper seeks to explore the impact of training hyperparameters that govern the training procedure, while fixing the architecture, training data, and initialization seeds of the vocoder models. We again limit the models to the Parallel WaveGAN architecture, with initialization seed as 1000 for all models, and use the entire clean training set of LibriTTS to train the vocoder models.

For the purposes of this experiment, we are interested specifically in the impact of adjusting the dropout rate and the batch size in training. We copy the PARALLEL-WAVEGAN-SEED-1000 model from the previous experiment and retag it as the baseline. This baseline does not use dropout, and each batch contains 6 samples. We then train four more models with the following parameters:

The last two models are trained for 150% and 75% of the number of steps as the baseline model, respectively, to compensate for the batch size influencing the total amount of data. Any unmentioned hyperparameter is kept the same as in the baseline model.

We then train the classifier to distinguish between 6 classes: {REAL, BASE, 1DROPOUT, 2DROPOUT, 4BATCH, 8BATCH}. As with all other experiments, we again sample 50,000 waveform files from each source for training and 5,000 files for validation.

Table 7 and Figure 6 display the classification results and confusion matrix, respectively. These results show that, even when models share the same architecture and are trained with the same data and initialization seeds, the precise manner of training can result in different models producing fingerprints that are different enough for distinguishing between audio generated by the models. It would therefore seem that the vocoder fingerprint is volatile to training. While identical models should trivially share the same fingerprints, since they deterministically generate identical audio waveforms, any minor variation in the training process will result in different models with different fingerprints.

5 VISUALIZATION

While we observe the effectiveness of fingerprints in distinguishing between audio generated by different vocoder models through evaluating the classification performances, it is also possible to directly see the separability of these fingerprints through the t-SNE method (van der Maaten & Hinton, 2008) that maps from a space with higher dimensionality (in our case, 256) to a 2D surface while making similar points closer and dissimilar points farther on the new surface. Figure 7 (see Appendix) shows the results of t-SNE mapping of the fingerprints extracted in our four experiments. While minor overlaps exist, corresponding to the confusions in classification, the mapped representations still clearly show that fingerprints belonging to different vocoder models are distinguishable. Furthermore, the fact that the t-SNE algorithm naturally clusters fingerprints of the same model together indirectly confirms that vocoder fingerprints are model-specific, in that fingerprints of each model forms a concentrated distribution around different mean values.

6 CONCLUSION

We have presented our study of extracting vocoder fingerprints as an effort to identify fake audio generation source. Our four experiments verify the existence of model-specific vocoder fingerprints and demonstrate that even small differences in the training process of vocoder models can result in their fingerprints being distinct, allowing for fine-grained source identification of fake audio. Despite our results, we have yet to study the impact of different languages (and therefore different phonological feature distributions) and acoustic models can have on the fingerprints of generated audios, and exploring these questions will form the focus of our future studies.

REFERENCES

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. LCNN: Lookup-Based Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7120–7129, 2017.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1): 6, January 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, August 1980. ISSN 0096-3518. doi: 10.1109/TASSP.1980.1163420.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, February 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2938758.

D. Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, April 1984. ISSN 0096-3518. doi: 10.1109/TASSP.1984.1164317.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient Neural Audio Synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, July 2018.

Hideki Kawahara. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353, 2006. doi: 10.1250/ast.27.349.

Sungwon Kim, Sang-Gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon. FloWaveNet : A Generative Flow for Raw Audio. In *International Conference on Machine Learning*, pp. 3370–3378. PMLR, May 2019.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Interspeech 2017*, pp. 2–6. ISCA, August 2017. doi: 10.21437/Interspeech.2017-1111.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv:2010.05646 [cs, eess]*, October 2020.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural Speech Synthesis with Transformer Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6706–6713, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33016706.

Da Luo, Paweł Korus, and Jiwu Huang. Band Energy Difference for Source Attribution in Audio Forensics. *IEEE Transactions on Information Forensics and Security*, 13(9):2179–2189, September 2018. ISSN 1556-6021. doi: 10.1109/TIFS.2018.2812185.

Robert C. Maher. Audio forensic examination. *IEEE Signal Processing Magazine*, 26(2):84–94, March 2009. ISSN 1558-0792. doi: 10.1109/MSP.2008.931080.

Hafiz Malik. Acoustic Environment Identification and Its Applications to Audio Forensics. *IEEE Transactions on Information Forensics and Security*, 8(11):1827–1837, November 2013. ISSN 1556-6021. doi: 10.1109/TIFS.2013.2280888.

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)*, pp. 506–511, San Jose, CA, USA, March 2019. IEEE. ISBN 978-1-72811-198-8. doi: 10.1109/MIPR.2019.00103.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884, 2016. ISSN 0916-8532, 1745-1361. doi: 10.1587/transinf.2015EDP7457.

Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization. *arXiv:2011.01557 [cs, eess]*, February 2021.

Daniel Patricio Nicolalde-Rodriguez, José Antonio Apolinario, and Luiz Wagner Pereira Biscainho. Audio Authenticity: Detecting ENF Discontinuity With High Precision Phase Analysis. *IEEE Transactions on Information Forensics and Security*, 5(3):534–543, September 2010. ISSN 1556-6021. doi: 10.1109/TIFS.2010.2051270.

Mustafa A. Qamhan, Hamdi Altaheri, Ali Hamid Meftah, Ghulam Muhammad, and Yousef Ajami Alotaibi. Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning. *IEEE Access*, 9:62719–62733, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3073786.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*, March 2022.

Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. In *Interspeech 2015*, pp. 2087–2091. ISCA, September 2015. doi: 10.21437/Interspeech.2015-472.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, April 2018. doi: 10.1109/ICASSP.2018.8461368.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-End anti-spoofing with RawNet2. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373, June 2021. doi: 10.1109/ICASSP39728.2021.9414234.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A Survey on Neural Speech Synthesis. *arXiv:2106.15561 [cs, eess]*, July 2021.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45: 516–535, September 2017. ISSN 08852308. doi: 10.1016/j.csl.2017.01.001.

Jean-Marc Valin and Jan Skoglund. LPCNet: Improving Neural Speech Synthesis through Linear Prediction. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, May 2019a. doi: 10.1109/ICASSP.2019.8682804.

Jean-Marc Valin and Jan Skoglund. A Real-Time Wideband Neural Vocoder at 1.6kb/s Using LPCNet. In *Interspeech 2019*, pp. 3406–3410. ISCA, September 2019b. doi: 10.21437/Interspeech.2019-1255.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016.

Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In *International Conference on Machine Learning*, pp. 3918–3926. PMLR, July 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. ISSN 1533-7928.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64: 101114, November 2020. ISSN 08852308. doi: 10.1016/j.csl.2020.101114.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. In *Interspeech 2017*, pp. 4006–4010. ISCA, August 2017. doi: 10.21437/Interspeech.2017-1452.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech 2015*, pp. 2037–2041. ISCA, September 2015. doi: 10.21437/Interspeech.2015-462.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv:2109.00537 [cs, eess]*, September 2021.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. *arXiv:1910.11480 [cs, eess]*, February 2020.

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. An Initial Investigation for Detecting Vocoder Fingerprints of Fake Audio. *arXiv:2208.09646 [cs, eess]*, August 2022. doi: 10.1145/3552466.3556525.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech. *arXiv:2005.05106 [cs, eess]*, November 2020.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. ADD 2022: The first Audio Deep Synthesis Detection Challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220, May 2022. doi: 10.1109/ICASSP43922.2022.9746939.

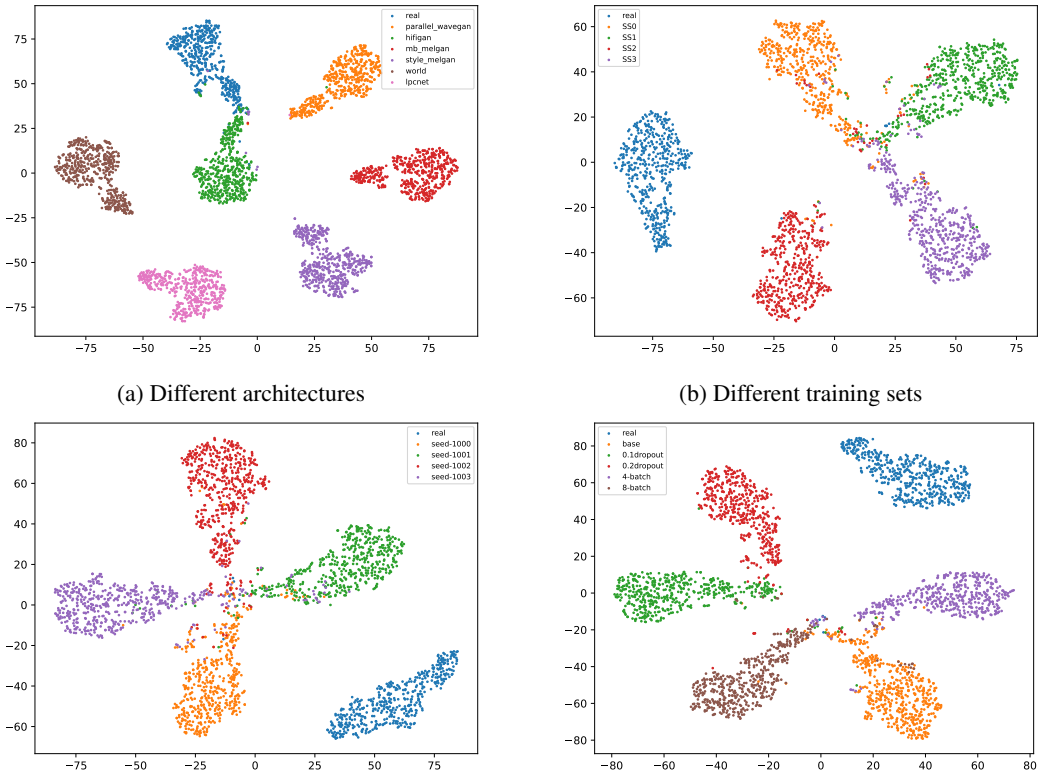
Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7555–7565, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00765.

Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. Digital multimedia audio forensics: Past, present and future. *Multimedia Tools and Applications*, 77(1):1009–1040, January 2018. ISSN 1380-7501, 1573-7721. doi: 10.1007/s11042-016-4277-2.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *arXiv:1904.02882 [cs, eess]*, April 2019.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting Intellectual Property of Deep Neural Networks with Watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pp. 159–172, Incheon Republic of Korea, May 2018. ACM. ISBN 978-1-4503-5576-6. doi: 10.1145/3196494.3196550.

APPENDIX



(a) Different architectures

(b) Different training sets

(c) Different initialization seeds

(d) Different training hyperparameters

Figure 7: t-SNE representation of fingerprints extracted in our experiments.