Distinguishing Neural Speech Synthesis Models Through Fingerprints in Speech Waveforms

Chu Yuan Zhang[1,2], **Jiangyan Yi**[1], **Jianhua Tao**[3,4], **Chenglong Wang**[1], **Xinrui Yan**[1,2]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Department of Automation, Tsinghua University, Beijing, China
[4] Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China
{zhangchuyuan2021, yanxinrui2021}@ia.ac.cn,
{jiangyan.yi, chenglong.wang}@nlpr.ia.ac.cn,
jhtao@tsinghua.edu.cn

Abstract

Recent advancements in neural speech synthesis technologies have brought about widespread applications but have also raised concerns about potential misuse and abuse. Addressing these challenges is crucial, particularly in the realms of forensics and intellectual property protection. While previous research on source attribution of synthesized speech has its limitations, our study aims to fill these gaps by investigating the identification of sources in synthesized speech. We focus on analyzing speech synthesis model fingerprints in generated speech waveforms, emphasizing the roles of the acoustic model and vocoder. Our research, based on the multi-speaker LibriTTS dataset, reveals two key insights: (1) both vocoders and acoustic models leave distinct, model-specific fingerprints on generated waveforms, and (2) vocoder fingerprints, being more dominant, may obscure those from the acoustic model. These findings underscore the presence of model-specific fingerprints in both components, suggesting their potential significance in source identification applications.

CCL 2024

1 Introduction

Recently, neural speech synthesis (also known as text-to-speech, or TTS) systems have made substantial advancements in generating highly realistic speech waveforms. These neural TTS systems can mainly be divided into two categories. The first category, pipeline systems, generally comprise an acoustic model such as Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2021) and GradTTS (Popov et al., 2021), alongside neural vocoders such as Parallel WaveGAN (Yamamoto et al., 2020), HiFiGAN (Kong et al., 2020), Style MelGAN (Mustafa et al., 2021) and Multiband MelGAN (Yang et al., 2020). The second category, end-to-end systems, also have been gaining traction in recent years. These systems, such as FastSpeech 2s (Ren et al., 2021) and VITS (Kim et al., 2021), are capable of generating speech waveforms directly from text.

While these advancements have found widespread application, they raise concerns about misuse in fraudulent activities, threatening security and privacy. Consequently, researchers have aimed to develop robust methods for differentiating genuine human speech from synthetic counterparts (Yi et al., 2023b). Initiatives like the ASVspoof challenges (Wu et al., 2015; Kinnunen et al., 2017; Wang et al., 2020; Yamagishi et al., 2021) and the ADD challenges (Yi et al., 2022) showcase advancements in audio anti-spoofing, including the use of front-end spectral features like linear frequency cepstral coefficients (LFCCs) (Todisco et al., 2018) as well as backend classification models like Res2Net (Gao et al., 2021) in anti-spoofing.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

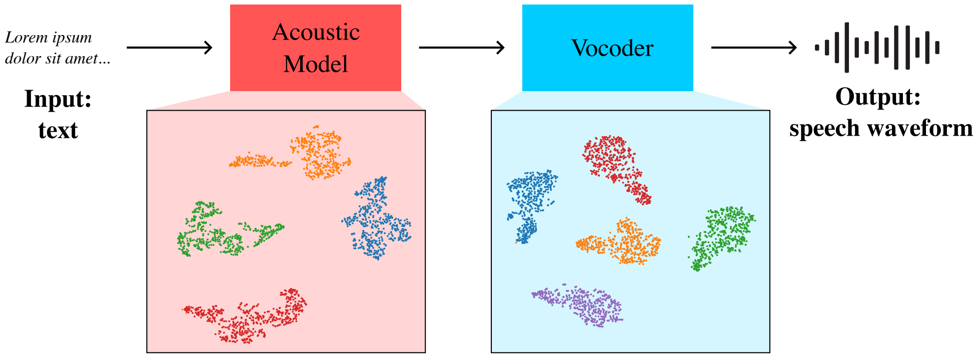


Figure 1: TTS pipeline and t-SNE projection of acoustic model and vocoder fingerprints.

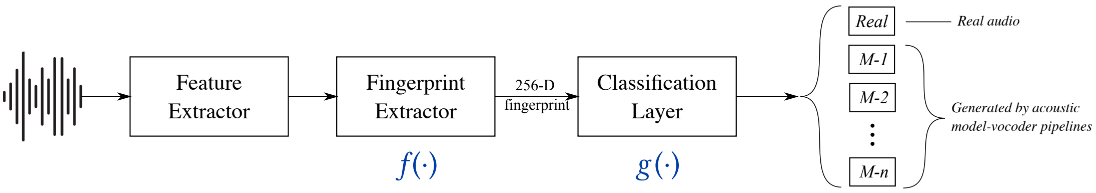


Figure 2: Fingerprint extraction pipeline.

CCL 2024

There is a also growing interest in identifying the source of synthesized speech, especially in forensics and intellectual property protection. The ADD 2023 challenge (Yi et al., 2023a) includes a deepfake algorithm recognition track, addressing the recognition of synthesis tools, where various methods have been proposed (Lu et al., 2023; Tian et al., 2023). It is however worth noting that deepfake source attribution is a research topic that has not been extensively explored. Pons *et al.* (Pons et al., 2021) focused on upsampling artifacts but limited their investigation to visual representation without exploring exploitability for attribution. Similarly, Yan *et al.* (Yan et al., 2022) studied vocoder fingerprints but overlooked finer parameter differences within vocoders of the same architecture. This underscores the need for more comprehensive investigations covering various factors to enhance our understanding of deepfake source attribution. Notably, the traditional approach to TTS usually involves a pipeline system, where an acoustic model generates mel spectrograms from text, which are then converted into speech waveforms by a vocoder; mainstream approaches to source attribution have largely overlooked the potential of these components to leave fingerprints on the synthesized speech by performing attribution on a pipeline-level basis.

In the interest of filling this gap and fostering a more robust and nuanced comprehension of deepfake source attribution, we present our study. We aim to extract fingerprints from synthesized speech to identify their sources, focusing on acoustic model and vocoder fingerprints in pipeline TTS systems. In particular, we put forth and seek to explain the following questions: (1) Do acoustic models and vocoders leave model-specific fingerprints on the speech waveforms they generate? (2) When taking into consideration both the acoustic model and the vocoder, does one overshadow the other in terms of fingerprints?

To answer these questions, we train various acoustic models and vocoders on the LibriTTS (Zen et al., 2019) dataset to form TTS pipelines. Through waveform fingerprint extraction from these pipelines, we assess our hypothesis that both the acoustic model and vocoder leave model-specific fingerprints on the generated waveforms. Our experiments confirm these fingerprints' existence, highlighting the vocoder's more prominent role, potentially masking acoustic model fingerprints. These findings suggest the potential utility of model-specific fingerprints in source identification applications.

In this paper, we present our methodology in Section 2, followed by our experimental setup and results in Section 3, and then conclude with a discussion of our findings in Section 4.

(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Methodology

In this section we present the methodology employed for the verification of fingerprints within synthesized speech generated by TTS pipelines. Abstractly, a speech synthesis pipeline composed of a given acoustic model $a(\cdot)$ and a given vocoder $v(\cdot)$ can be viewed as a function that maps a text sequence x to a speech waveform y, i.e., $y = v(a(x))$. Suppose there theoretically exist a "perfect" acoustic model a^* and a "perfect" vocoder v^* that generate "perfect" speech waveforms that are mathematically indistinguishable from genuine human speech. As current TTS pipelines are imperfect, with imperfections arising from both the theoretical foundation and the practical implementation, we can therefore expect that the speech waveforms generated by these pipelines will deviate from perfection; i.e.,

$$\begin{aligned} a(x) &= a^*(x) + \epsilon_a(x) \\ v(a(x)) &= v^*(a(x)) + \epsilon_v(a(x)) \end{aligned} \tag{1}$$

where ϵ_a and ϵ_v are residuals that contain the fingerprints of the acoustic model and the vocoder, respectively.

Inspired by the assumption in Marra et al. (2019), we assume that the fingerprints of the acoustic model and the vocoder are independent of each other, and that each of the residuals ϵ_a and ϵ_v contains the fingerprint plus a Gaussian noise, such that, with a sufficiently large dataset, the Gaussian noise can be ignored. We therefore seek to verify the existence of such fingerprints through an extraction network $f(\cdot)$ that extracts the fingerprints from the speech waveform y, i.e., either

$$\begin{aligned} f_a(y) &= \hat{\epsilon}_a(x) \\ f_v(y) &= \hat{\epsilon}_v(a(x)) \end{aligned} \tag{2}$$

as the case may be. We then train a classifier $g(\cdot)$ to classify the fingerprints extracted by $f(\cdot)$ into the corresponding acoustic model or vocoder, i.e.,

$$\begin{aligned} g_a(f(y)) &= a \\ g_v(f(y)) &= v \end{aligned} \tag{3}$$

The entire extraction and classification pipeline is shown in 2. We then investigate the performance of the classifier $g(\cdot)$ to determine whether the fingerprints of the acoustic model and the vocoder are distinguishable.

Following the methodology outlined above, we investigate the fingerprints of the acoustic model and the vocoder separately, and then investigate the relationship between the two fingerprints. We also investigate the robustness of the fingerprints against perturbations in the input speech samples.

2.1 Vocoder fingerprint

For vocoders, each model is characterized not only by the model architecture, but also by the precise weights in the neural network. This is especially true with generative adversarial networks (GANs) due to the adversarial nature of their training, meaning that the precise weights are sensitive to training setups (Brock et al., 2019). For this reason, we investigate the impact of both model architectures and training setups in vocoder fingerprints.

To study the impact of model architectures, we train vocoders of 4 architectures: Parallel WaveGAN (Yamamoto et al., 2020), HiFiGAN (Kong et al., 2020), Multiband MelGAN (Yang et al., 2020), and Style MelGAN (Mustafa et al., 2021), and investigate the fingerprints of the generated waveforms to determine whether the architecture leaves a fingerprint. To investigate the impact of training setup, we train several vocoders of the 4 aforementioned architectures, each with varying training setups (see Table 1), and then investigate the fingerprints of various vocoders trained with different setups to analyze the influence of training setup on the fingerprint.

CCL 2024

Architecture	Model ID	Seed	Batch size
Parallel WaveGAN (PWG)	P0	1000	6
	P1	1001	6
	P2	1000	8
	P3	1002	6
	P4	1003	6
HiFiGAN (HFG)	H0	1000	16
	H1	1001	16
	H2	1000	8
Multiband MelGAN (MMG)	M0	1000	16
	M1	1001	16
	M2	1000	8
StyleMelGAN (SMG)	S0	1000	16
	S1	1001	16
	S2	1000	8

Table 1: Vocoder model setups.

Exp.	Training set	Validation set	Test set
V1	P0, H0, M0, S0	(same methods as training set)	(same methods as training set)
V2	P0, P1, P3, P4	(same methods as training set)	(same methods as training set)
V3	P0, H0, M0, S0	P1, P2, H1, H2, M1, M2, S1, S2	P1, P2, H1, H2, M1, M2, S1, S2
A1	F2+H0, GD+H0, T2+H0	(same methods as training set)	(same methods as training set)
A2	F2+H0, F2+P0, F2+S0, GD+H0, GD+P0, GD+S0, T2+H0, T2+P0, T2+S0	(same methods as training set)	(same methods as training set)
R1	F2+H0, GD+H0, T2+H0	F2+P1, F2+S1, GD+P1, GD+S1, T2+P1, T2+S1	F2+P1, F2+S1, GD+P1, GD+S1, T2+P1, T2+S1
R2	T2+P0, T2+H0, T2+M0, T2+S0	F2+P0, F2+H0, F2+M0, F2+S0	F2+P0, F2+H0, F2+M0, F2+S0
N1	(R2 training set)	(R2 validation set)	(R2 test set with noise at 10 dB SNR)
N2	(R2 training set)	(R2 validation set)	(R2 test set with reverberation)
N3	(R2 training set)	(R2 validation set)	(R2 test set with speed adjustment)

CCL 2024

Table 2: Fingerprint analysis experiment setups. (Exp. = Experiment ID)
Acoustic models used: FastSpeech 2 (F2), Grad-TTS (GD), Tacotron 2 (T2)
Vocoders used: see Table 1
In Experiments V1, V2, A1, and A2, different inputs are used with the same set of vocoders/acoustic models to generate samples for the training, validation and test sets.

2.2 Acoustic model fingerprint

Acoustic models are paired with a vocoder to generate waveforms, therefore, we study two aspects of the acoustic model fingerprint: when one vocoder is used to generate waveforms from the mel spectrograms produced by the acoustic model, and when multiple vocoders are used to generate waveforms from the mel spectrograms produced by the acoustic model. We train acoustic models of 3 architectures: Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2021), and GradTTS (Popov et al., 2021), and use our trained models in conjunction with the previously trained vocoders to generate waveforms from the mel spectrograms produced by the acoustic models. We then investigate the fingerprints of the generated waveforms to determine whether the acoustic model leaves a fingerprint.

2.3 Relationship between fingerprints

When it comes to the relationship between vocoder and acoustic model fingerprints, there are two competing hypotheses, both needing to be verified: either (1) vocoder fingerprints can visibly interfere with acoustic model fingerprints, or (2) acoustic model fingerprints can visibly interfere with vocoder fingerprints. To verify both hypotheses, we design two experiments with complementary setups, aimed to verify the two aforementioned, somewhat contradictory hypotheses. By setting one module — acoustic model or vocoder, as the case may be — to be known in the testing set, and testing the fingerprint against unknown vocoders or acooustic models, respectively, we can verify the two hypotheses. The setups of the two experiments are shown in Table 2, in rows R1 and R2.

2.4 Perturbations

We furthermore seek to investigate the robustness of these fingerprints against perturbations in speech samples, namely noise, reverberation, and speed adjustment. To this end, we perform three experiments, for each of the aforementioned perturbations, where we add the specific perturbation to the speech samples in the test set, and then investigate the fingerprints of the generated waveforms to determine the extent of impact such perturbation would have on the original fingerprints. The setups of the three experiments are shown in Table 2, in rows N1, N2 and N3.

3 Experiments and results

In this section, we outline our conducted experiments and the subsequent findings. We have undertaken a series of ten distinct experiments. The first five experiments are conducted to verify the existence of vocoder (V1–V3) and acoustic model (A1–A2) fingerprints in synthesized speech, and the next two (R1–R2) are conducted to verify the influence of the acoustic model and the vocoder on the fingerprints. Finally, three experiments (N1–N3) are conducted to investigate the robustness of the fingerprints against perturbations in the input speech samples. The setups of the experiments are shown in Table 2.

3.1 Experimental setup

In the process of analyzing speech waveforms, each sample undergoes a detailed extraction method to obtain its linear frequency cepstral coefficients (LFCCs). This process begins with the segmentation of the speech waveform into frames, from which we extract 20 LFCCs for each frame. This extraction uses a Hamming window, which is a specific type of function used to smooth out the signal, minimizing the edge effects during the analysis. The parameters for this window are carefully chosen, with a window length set at $M = 480$ and hop length of $H = 240$.

We standardize the size of the resulting LFCC matrix across all samples, since the matrix might vary in the number of frames depending on the duration of the speech sample. To address this variance, we either truncate the matrix to contain exactly 500 frames if it exceeds this number, or we pad it with zeros to reach the frame count of 500 if it contains fewer frames. This step is crucial for maintaining consistency across the dataset, allowing the neural network to process the data more effectively.

Inspired by Marra et al. (2019), the LFCC representation of the speech sample is then passed to a Res2Net network (Gao et al., 2021) and projected onto a 256-dimension feature space, then passed through a single fully-connected (FC) layer classifier to decide between the finite number of sources. The entire network, i.e., both the extraction network and the classifier, is trained using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^{(j)} \log(p_i^{(j)}) \tag{4}$$

where N is the number of samples, C is the number of classes, y and p denote the ground truth label and the predicted probability, respectively.

3.2 Dataset

In our experiments, we use the multi-speaker dataset LibriTTS (Zen et al., 2019) to both train acoustic models and vocoders as well as to generate the waveform corpus for fingerprint

CCL 2024

Variable	Exp.	Precision	Recall	F_1
Vocoder	V1	98.46	98.44	98.44
	V2	95.74	95.73	95.73
	V3	91.57	91.19	91.20
Acoustic model	A1	99.79	99.79	99.79
	A2	98.99	98.99	98.99
Relationship	R1	23.75	25.01	10.05
	R2	98.46	98.42	98.42
Perturbation	N1	98.21	98.16	98.17
	N2	87.82	86.36	86.03
	N3	91.24	91.12	91.07

Table 3: Summary of experiment results (%)

extraction. The “clean” training set of LibriTTS, containing 160,369 genuine utterances from 1,230 speakers, totaling 262 hours of speech, is used to train all models until convergence.

In departure from the traditional setup where the validation set and training set are from the same distribution, we synthesize the validation set using the same setups as the test set (see Table 2), since we are interested in the performance of the classifier on the test set, which is synthesized using different setups from the training set. The training, validation and test sets are, however, disjoint, to help avoid overfitting.

CCL 2024

3.3 Evaluation metric

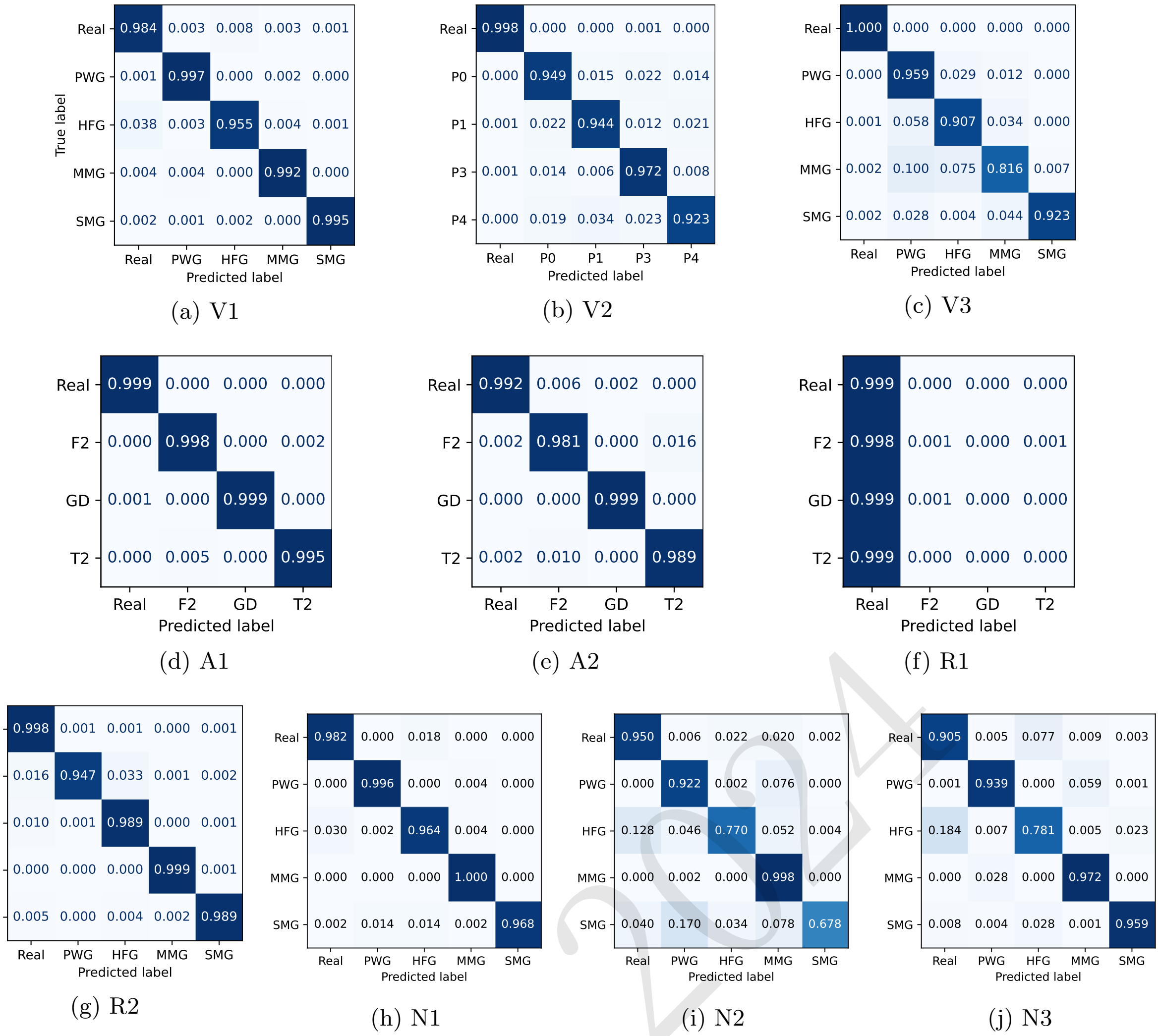
The performance of fingerprint extraction and analysis should align with the ability to classify the audio samples and identify their source of generation. To thoroughly evaluate the performance of our fingerprint extraction model, we employ a trio of established metrics commonly utilized in classification tasks: precision, recall, and the F1 score. These metrics offer a comprehensive view of the model's accuracy and its ability to distinguish between different sources of generated speech. Beyond these numerical metrics, we delve into the analysis of the confusion matrix associated with our classifier, to gain insight into the performance of the fingerprint extraction model, as it visualizes the performance of a classifier by identifying the number of instances where the classifier misclassified the speech samples and in turn evaluate the performance of the fingerprint extraction model, on a per-category basis. This allows us to go beyond the global metrics and gain a more nuanced understanding of the model's performance.

3.4 Results and analysis

Following the methodology outlined in Section 2, and the setup of the experiments in Table 2, we conduct a series of experiments, aimed at verifying the existence of vocoder and acoustic model fingerprints in synthesized speech, and to investigate the relationship between the two fingerprints. We also investigate the robustness of the fingerprints against perturbations in the input speech samples. The results of these experiments are summarized in Table 3, and the confusion matrices of the experiments are shown in Figure 3.

3.4.1 Vocoder fingerprints

For vocoder fingerprint analysis, we use copy synthesis to emulate a fixed acoustic model, minimizing variables. We train 3 vocoders for each of the four architectures: Parallel WaveGAN (PWG), HiFiGAN (HFG), Multiband MelGAN (MMG), and Style MelGAN (SMG), each with varying seeds and batch sizes (see Table 1). We additionally train two more PWG vocoders with different initialization seeds to investigate the impact of training setup on fingerprints. Three experiments assess vocoder architecture and training setup impact on fingerprints. In V1 and V3, we apply copy synthesis on the training set using P0, H0, and M0 models, labeling results by vocoder architecture for Res2Net classifier training. The validation and testing sets differ in vocoder models used for copy synthesis (see Table 2). In V2 we investigate training setup



(a) V1

(b) V2

(c) V3

(d) A1

(e) A2

(f) R1

(g) R2

(h) N1

(i) N2

(j) N3

Figure 3: Confusion matrices of experiments in our study.

CCL 2024

impact by classifying along initialization seed lines within the same architecture (i.e., Parallel WaveGAN).

Results are summarized in the V1–V3 rows of Table 3 as well as the confusion matrices in Figures 3a to 3c. The high F1 score of V1 and the highly diagonal confusion matrix in Figure 3a confirm architecture-specific fingerprints in the generated waveforms serving as strong indicators of source. In contrast, the slight degredation of classification performance in Experiment V3, as shown in Figure 3c, suggests vocoder training setups weaken fingerprints, an observation corroborated by Experiment V2 (Figure 3b). Overall, results confirm the existence of finer-grained vocoder fingerprints based on parameter differences, but with less pronounced impact than architectural ones.

3.4.2 Acoustic model fingerprints

We investigate and analyze the fingerprint of the acoustic model. Three models, Tacotron 2, FastSpeech 2 and Grad-TTS, are trained on the LibriTTS dataset. These models generate labeled speech waveforms from the LibriTTS training set text corpus. We use the entire speech corpus, encompassing both real and synthesized samples, to train the Res2Net classifier. The same approach applies to validation and test sets. To thoroughly examine the detectability and characteristics of the acoustic model fingerprint, we conduct two distinct experiments. Experiment A1 serves as a baseline, where we limit the diversity of the vocoders used in the synthesis process. Experiment A2, on the other hand, introduces a wider variety of vocoders into the mix (see Table 2). This variation allows us to assess the impact of vocoder diversity on the ability of the Res2Net classifier to learn and identify acoustic model fingerprints, and more importantly, to ascertain that the fingerprint features learned by our model are indeed acoustic model-specific.

The results of our study are encapsulated in confusion matrices, as referenced in our report (Figures 3d to 3e), and the detailed analysis is provided in rows A1–A2 of Table 3. These results shed light on our question, demonstrating that the acoustic model fingerprint is indeed detectable with a high degree of accuracy when the vocoder used during the synthesis process is either fixed or known. This revelation underscores the effectiveness of a fixed vocoder setup in enhancing the learning of acoustic model fingerprints. However, it is also noteworthy that introducing a variance in the vocoders, as done in Experiment A2, does not significantly impede the learning process. This is particularly true when the vocoder variance is evenly distributed across the categories under investigation. The insights gained from Experiment A2 highlight the resilience of the acoustic model fingerprint learning process, even in the face of increased diversity in vocoder usage.

3.4.3 Relationship between fingerprints

As outlined in Section 2.3, we investigate the relationship between vocoder and acoustic model fingerprints with a pair of complementary experiments, R1 and R2. In R1, we fix the acoustic model and vary the vocoder; the acoustic models used to generate the testing set are known, but the vocoder (H0) is swapped for two unknown models (P1 and S1). In R2, we fix the vocoder architectures, but vary the acoustic models used to generate the testing set (TN being replaced by FS). The results of these experiments are summarized in the R1–R2 rows of Table 3 and the confusion matrices in Figures 3f to 3g. Results in Figure 3f demonstrate that when vocoders vary, the acoustic model fingerprint becomes nearly undetectable, aligning with our hypothesis that vocoder fingerprints overshadow acoustic model fingerprints. This is further supported by the results in Figure 3g, where the vocoder fingerprint is still detectable when the acoustic model varies.

The experiment outcomes align with our expectations, considering the distinct roles of the vocoder and acoustic model in the process of speech synthesis. The vocoder directly generates the waveform by translating the spectrogram produced by the acoustic model into speech. Given its role as the last step in the synthesis chain, the vocoder's involvement in the generation of the speech waveform is direct and substantial. It takes the more compact, frequency-based information from the spectrogram and transforms it into the waveform that we ultimately hear as speech. This direct engagement implies that it has a significant influence on the final characteristics of the speech waveform. Consequently, it's plausible to suggest that the vocoder leaves a more distinguishable fingerprint on the generated speech than the acoustic model does.

This distinction between the vocoder and acoustic model's contributions is essential for understanding how different components of the speech synthesis process affect the final product. It implies that when analyzing synthesized speech for the purpose of identifying its source or assessing its quality, the vocoder's characteristics may be a more telling factor than those of the acoustic model. This insight is crucial for the development of speech synthesis systems and the evaluation of their performance.

3.5 Fingerprint robustness against speech perturbation

Given the analyses above, demonstrating the existence of distinct fingerprints for both vocoders and acoustic models, as well as the overshadowing of acoustic model fingerprints by vocoder fingerprints, we further investigate the robustness of the fingerprints against perturbations in the input speech samples. We analyze the robustness from three aspects:

1. Noise: We add white noise to the speech samples in the test set of R2, with various signal-to-noise ratios (SNR) of 0–10 dB.
2. Reverberation: We add reverberation to the speech samples in the test set of R2, with various reverberation times (RT) of 0.5–1.5 seconds.

CCL 2024

Variable	Param.	Precision	Recall	F_1
Noise	SNR=10dB	98.21	98.16	98.17
	SNR=5dB	97.04	96.98	97.00
Reverberation	RT=0.5s	87.82	86.36	86.03
	RT=1.0s	85.12	83.68	83.35
Speed adjustment	0.9x	91.24	91.12	91.07
	1.1x	90.12	89.98	89.94

Table 4: Summary of perturbation experiment results (%)

3. Speed adjustment: We adjust the speed of the speech samples in the test set of R2, with various speed factors of 0.9x–1.1x.

We conduct an experiment where we add noise to the speech samples in the test set. We use SoX[0] to add white noise to the speech samples in the test set of R2, with a signal-to-noise ratio (SNR) of 10 dB. We then investigate the fingerprints of the generated waveforms to determine the extent of impact such perturbation would have on the original fingerprints, as summarized in Row N1 of Table 3.

Similar setups are used to investigate the robustness of the fingerprint extraction model against reverberations and speed adjustments (varying from 0.9x to 1.1x), with results listed in rows N2 and N3 of Table 3, respectively. The results of these experiments show that the fingerprint extraction model is somewhat robust against perturbations in the input speech samples, with a slight degradation in performance. Among the three perturbations tested, reverberations seem to have the most adverse effect on the fingerprints. This is likely due to the fact that reverberations are more likely to affect the spectral characteristics of the speech samples, which are the basis of the fingerprints, than noise and speed adjustments.

We furthermore conduct experiments with various signal-to-noise ratios (SNR) and reverberation times (RT) to investigate the robustness of the fingerprints against noise and reverberation. The results of these experiments are summarized in Table 4. The results reinforce our analysis of the fingerprint robustness against perturbations, with speed adjustments and reverberation having a more pronounced impact on the fingerprints than noise, with reverberation being the most detrimental to the fingerprints.

CCL 2024

3.6 Visualization

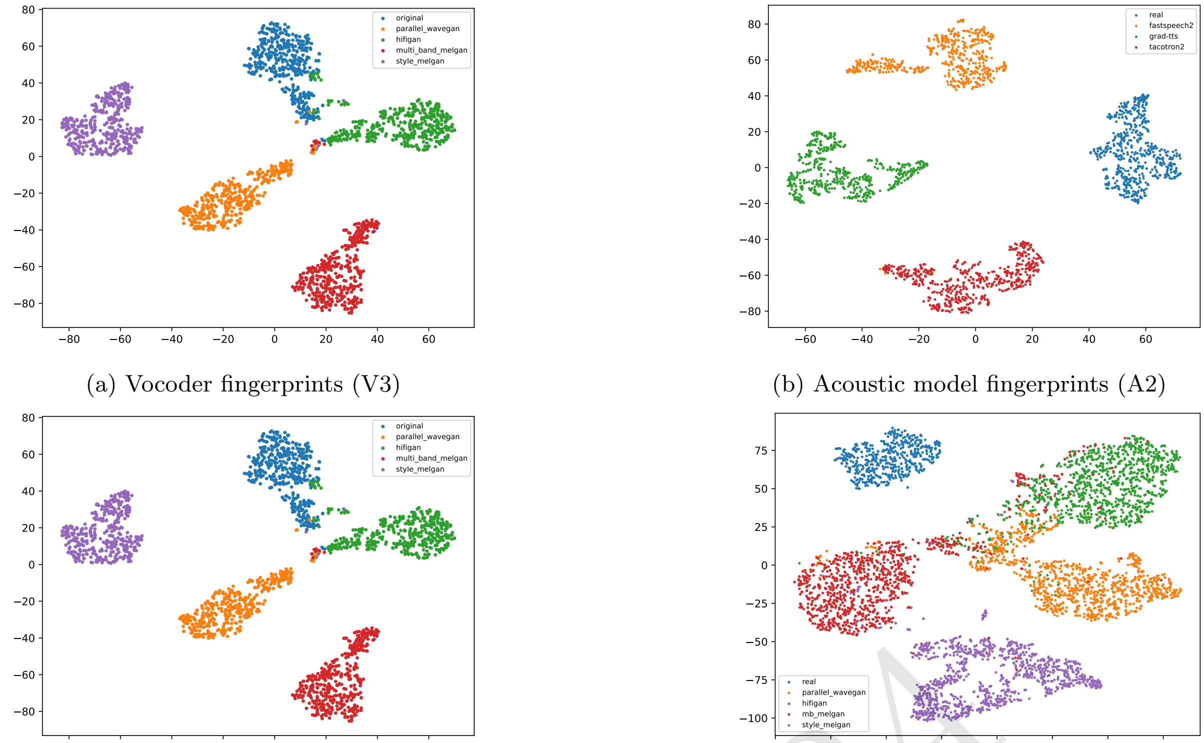
Given the results reported in the previous section, to better and more intuitively understand the fingerprints of the vocoders and acoustic models, we visualize the fingerprints of the vocoders and acoustic models using t-SNE (van der Maaten and Hinton, 2008).

We focus first on the vocoders. By applying t-SNE, we project the high-dimensional fingerprints of various vocoders onto a two-dimensional plane. This visualization is reflected in Figure 4a. The resulting projection clearly illustrates how the fingerprints of the vocoders are dispersed, forming clearly distinguishable clusters on the plane. Each of these clusters represents a unique vocoder architecture, unequivocally demonstrating the presence of distinct, identifiable fingerprints for each vocoder.

Following the analysis of vocoders, we extend the same t-SNE visualization technique to the acoustic models. The outcomes of this process are captured in Figure 4b, where, akin to the vocoders, the fingerprints of the acoustic models are also neatly organized into discernible clusters. This pattern reinforces the premise that acoustic models, much like vocoders, possess their own unique fingerprints. The clarity with which these fingerprints can be distinguished affirms the potential for precise identification and differentiation of acoustic models based on their inherent characteristics.

Figure 4c shows the t-SNE projection of the fingerprints of the vocoders and acoustic models

[0] Available at `https://sourceforge.net/projects/sox/`



(a) Vocoder fingerprints (V3)

(b) Acoustic model fingerprints (A2)

(c) Relationship between vocoder and acoustic model fingerprints (R2)

(d) Perturbation experiment: audio speed readjustment (N3)

Figure 4: t-SNE representation of fingerprints extracted in our experiments.

CCL 2024

in Experiment R2. As with the previous visualizations, the fingerprints of the vocoders are clearly separated into distinct clusters, despite the presence of different and unknown acoustic models. This result confirms that the vocoder fingerprints are more prominent than the acoustic model fingerprints, potentially overshadowing the acoustic model fingerprints.

Lastly, Figure 4d shows the t-SNE projection of the fingerprints of the perturbation experiment N3. The fingerprints of the perturbed speech samples are still distinguishable, albeit with a slight overlap, indicating that the fingerprint extraction model is somewhat robust against perturbations in the input speech samples.

4 Conclusion

In conclusion, our study centers on extracting fingerprints from acoustic models and vocoders to identify the sources of synthesized speech. Our experiments confirm the existence of distinct, model-specific fingerprints for both vocoders and acoustic models. Notably, we observe that vocoders tend to leave more prominent fingerprints, potentially overshadowing those of the acoustic models. This discovery holds promise, indicating the feasibility of employing fingerprint extraction for source attribution in synthesized speech. This insight is particularly relevant for researchers and developers in the field of computational linguistics and speech synthesis, as it highlights the importance of carefully selecting or designing vocoders to achieve the desired qualities in synthesized speech. Nonetheless, our current research leaves room for further exploration. We have yet to consider the potential impact of different languages, which entail diverse phonological feature distributions, on the fingerprints of generated waveforms. It will also be interesting and important to devise methods for extracting fingerprints from end-to-end TTS systems, which are gaining traction in recent years. Lastly, even in the context of pipeline TTS systems, we have yet to come up with a reliable method for extracting acoustic-model-specific fingerprints despite the masking presence of vocoder fingerprints. These unexplored dimensions will form the focal point of our future investigations, seeking to provide a more comprehensive understanding of the nuances in the fingerprints of synthesized speech.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No. 62322120, No. 62306316, No U21B2010, No. 62206278), as well as the ANSO Scholarship for Young Talents.

References

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2021. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, page 5530–5540. PMLR.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Interspeech 2017*, page 2–6. ISCA.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033.

Jingze Lu, Yuxiang Zhang, Zhuo Li, Zengqiang Shang, WenChao Wang, and Pengyuan Zhang. 2023. Detecting unknown speech spoofing algorithms with nearest neighbors. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do GANs Leave Artificial Fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE Computer Society.

Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. 2021. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà. 2021. Upsampling artifacts in neural audio synthesis. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3005–3009.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, page 8599–8608. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*.

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. 2018. Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Ye Tian, Yunkun Chen, Yuezhong Tang, and Boyang Fu. 2023. Deepfake algorithm recognition through multi-model fusion based on manifold measure. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

CCL 2024

Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. 2018. Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion. In *Proc. Interspeech 2018*, pages 77–81.

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

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. 2020. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101–114.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech 2015*, page 2037–2041. ISCA.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

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

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. 2022. An initial investigation for detecting vocoder fingerprints of fake audio. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22)*, page 61–68.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2020. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. *arXiv:2005.05106 [cs, eess]*.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. ADD 2022: the first audio deep synthesis detection challenge. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. 2023a. ADD 2023: the second audio deepfake detection challenge. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023b. Audio deepfake detection: A survey. *arXiv 2308.14970*.

Heiga Zen, 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. In *Proc. Interspeech 2019*, pages 1526–1530.

CCL 2024