Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts

Ehab AlBadawy[1]*,Chengzhe Sun[2]*, Timothy F Davison[3], Sarah R Robinson[3], Ming-Ching Chang[1], Siwei Lyu[2]
[1] Department of Electrical and Computer Engineering, University at Albany, SUNY, Albany, USA
[2] Department of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, USA
[3] Applied Physics Lab, The Johns Hopkins University, Baltimore, USA

***Abstract*—The advancements of AI-synthesized human voices have introduced a growing threat of impersonation and disinformation. It is therefore of practical importance to develop detection methods for synthetic human voices. This work proposes a new approach to detecting synthetic human voices based on identifying artifacts of *neural vocoders* in audio signals. A neural vocoder is a specially designed neural network that synthesizes waveforms from temporal-frequency representations, *e.g.*, mel-spectrograms. The neural vocoder is a core component in most deepfake audio synthesis models. Hence the identification of neural vocoder processing implies that an audio sample may have been synthesized. To take advantage of the vocoder artifacts for synthetic human voice detection, we introduce a binary-class RawNet2 model that shares the front-end feature extractor with the one for vocoder identification. We employ a self-supervised representation learning (SSRL) approach. We treat the vocoder identification as a pretext task to constrain the front-end feature extraction module to build the final binary classifier. Our experiments show that the RawNet2 model SSRL based on the vocoder artifacts achieves an overall high classification performance.**

I. Introduction

Recent years have seen the proliferation of synthetic media, riding the waves of the rapid advancement of AI technologies, in particular, deep learning. These synthetic media are more commonly known as the “DeepFakes”, a portmanteau of deep learning and fake media. State-of-the-art AI media synthesis methods can now create highly realistic still images and videos that can challenge the viewer’s ability to distinguish them from real media [1]. While the AI-synthesized still images and videos are currently in the spotlight of public attention, synthetic human voices have also undergone considerable developments and are reaching unprecedented perceptual quality and generation efficiency. The AI-synthesized human voices can facilitate new capacities in voice-based user interfaces for smart home assistants and wearable devices and can be used to help patients whose speech abilities are damaged by strokes or ALS to gain back voice. However, synthetic human voices could also be misused for deceptions and scams. A case in point is a recent incident in which a scammer used a synthetic voice created with AI algorithms to impersonate the CEO of a UK company in a phone call, which misled an employee to wire transfer a substantial amount of money to the scammer’s bank account [2].

While the detection of AI-synthesized still images and videos have been avidly studied in the recent years [3], methods to detect synthetic human voices have received relatively less attention and are under-developed. One reason is that audio signals have different characteristics that hinder the direct application of image-based detection methods. Existing detection methods usually examine signal statistical features that are particular to audio signals, for instance, the work in [4] compares the higher-order statistics in the bi-spectral domain that capture the local phase inconsistencies in synthetic human voices.

In this work, we propose a new approach to detecting synthetic human voices based on exploiting artifacts introduced by the *neural vocoders* in the synthesized voice signals. The neural vocoder is a specially-designed neural network that synthesizes audio waveforms from temporal-frequency representations such as mel-spectrograms. Because neural vocoders are the last step in most AI-based audio synthesis models, we think that they can provide cues to expose synthetic human voices. In particular, it is highly unlikely to process real audio signals with neural vocoders.

Hence, the foremost objective of our work is to highlight the distinct signal artifacts left by neural vocoders in synthetic audio signals. As there exist no datasets of audio signals generated with different vocoders, we first construct a dataset termed as *LibriVoc*. LibriVoc includes a total of 10.8 hours of “self-vocoding” derived from a subset of the human audio samples in the LibriTTS dataset [5]. Specifically, each input audio is transformed into the corresponding melspectral representation, and then the audio waveform is reconstructed using neural vocoders. We use six neural vocoders in collecting the LibriVoc dataset, to reflect the diversity in the architecture and mechanisms of neural vocoders. Because the “self-vocoding” samples are sourced from the same original audio signals, they highlight the artifacts introduced by the vocoders. While these artifacts are subtle to visualize, they can be captured by a trained classifier. As the vocoder artifacts are often subtle, we choose the recent RawNet2 model [6] as the backbone, which is designed to work with audio waveforms directly. Our experiments show that a multi-class classifier based on RawNet2 can identify vocoders from synthetic human voices with high accuracy.

* suggests equal contribution. This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1822190.

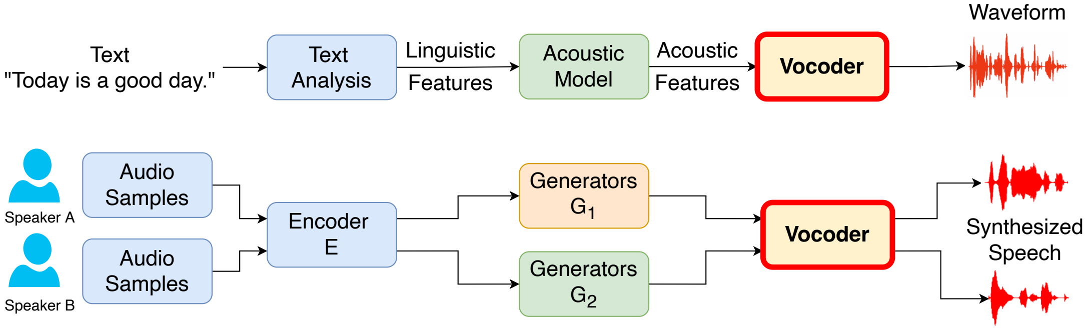


Fig. 1: Overall pipeline for deepfake audio synthesis, including (top) text-to-speech and (bottom) voice conversion. Note that the vocoder is the common component in both types of methods.

To take advantage of the vocoder artifacts in the detection of synthetic human voices, we use a binary classifier that shares the front-end RawNet2 feature extractor in the vocoder identification. This is to accommodate the insufficient number of existing real and synthetic human voice samples by including the self-vocoding samples in LibriVoc as additional training data. We further employ a self-supervised representation learning (SSRL) approach [7], where we treat the vocoder identification as a pretext task to constrain the front-end feature extraction module to build the final binary classifier. Our experiments show that the RawNet2 model SSRL trained on LibriVoc achieves an overall high classification performance, corresponding to 1.61% Equal Error Rate (EER) on the TTS dataset that we created with the state-of-the-art TTS algorithm Tacotron 2 [8]. Our method is also tested on a dataset in the DARPA SemaFor Hackathon 3 of synthetic human voices and achieves satisfactory classification performance.

The main contributions of our work are as follows:

- We are the first to identify neural vocoders as a source of features to expose synthetic human voices;
- We provide LibriVoC as a dataset of self-vocoding samples created with six state-of-the-art vocoders to highlight and exploit the vocoder artifacts;
- We propose a new approach to detect synthetic human voices based on exposing signal artifacts left by neural vocoders and trained with self-supervised representational learning;
- Experimental evaluations of the proposed method on a large neural TTS dataset and the DARPA SemaFor Hackathon dataset demonstrate the effectiveness of our method.

II. Related Works

In this section, we briefly review existing related works to our method. Note that it is a common practice to work with audio signals in their temporal-frequency representations, or spectrograms, which are 2D representations obtained from the short-time Fourier transform of the 1D audio waveform. For human voices, it is often advantageous to use the mel-spectrogram, which is obtained by redistributing the spectrogram of an audio signal on the mel scale using a filter bank. The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The scale was obtained experimentally, with listeners providing a set of points that create an equal frequency with the perceived pitch. The distinct features of human voices can be better revealed with the mel-spectrogram representation.

A. Human Voice Synthesis

Human voice synthesis is a major problem in artificial intelligence and has many practical applications such as voice-driven smart assistants and accessible user interfaces with voice-over. There are two general categories of human voice synthesis techniques, namely, Text-to-speech (TTS) and Voice conversion (VC), Fig.1. In this work, we focus on the more recent TTS and VC methods that are based on the deep neural network models. TTS models convert an input text to audio using the target voice, which usually consists of three components: a text analysis module that transforms the input text sequence into linguistic features; the acoustic model that generates acoustic features that are usually in the form of mel-spectrogram from the linguistic features; and the vocoder (Fig.1 top row). Recent deep neural network-based TTS models include WaveNet [9], Tacotron [10], Tacotron 2 [8], ClariNet [11], and FastSpeech 2s [12]. VC models, on the other hand, take a sample of one subject's voice as input and create output audio of another subject's voice of the same utterance. Recent VC models (*e.g.*, [13]–[15]) usually work with the melspectral domain, and employ deep neural network models to map between the mel-spectrograms of the input and output voice signals. More specifically, neural style transfer models such as variational auto-encoder (VAE) or generative adversarial network (GAN) models are often used to capture the capture the utterance elements in the input voice, and then combines them with the style of the output voices. The resulting mel-spectrogram will be reconstructed into audio waveforms using a neural vocoder. The deep neural network models in both the TTS and VC models are trained over large-scale human voice corpora.

B. Neural Vocoders

A vocoder is a common and essential component in the majority of human voice synthesis models, be it for TTS or VC, to synthesize the final output audio waveforms of the synthesized human voices from mel-spectrogram representations. As the transformation from audio waveforms to mel-spectrograms losses information due to binning and filtering, it is not a trivial task to recover the audio waveform from a mel-spectrogram, as it entails an inference problem. Recent years have seen active development of deep neural network-based vocoders, which significantly improve the training efficiency and synthesis quality. Existing neural vocoders can be divided into three main categories as the autoregressive models, the diffusion models, and the GAN-based models.

Autoregressive models are probabilistic models that predict the distribution of each audio waveform sample based on all previous samples. However, since this process undergoes a linear sample-by-sample generation, the output speed of the model is slower than that of other methods. WaveNet [9] is the first autoregressive neural vocoder, which can also serve as a TTS or VC model depending on the input. WaveRNN [16] is another autoregressive vocoder that uses a single-layer recurrent neural network for audio generation, which is designed to efficiently predict 16-bit raw audio samples from mel-spectrogram slices.

Diffusion models are probabilistic generative models, which run *diffusion* and *reverse* as two main processes [17]. The diffusion process is characterized by a Markov chain, which gradually adds Gaussian noise to an original signal until that noise is eliminated. The reverse process is a de-noising stage that steadily removes the added Gaussian noise and converts a sample back to the original signal. Two notable examples of the diffusion-based vocoder models are WaveGrad [18] and DiffWave [19]. Generally speaking, diffusion models are the most time-efficient vocoders but their reconstruction qualities are inferior to the autoregressive models, and the generated samples may contain higher levels of noises and artifacts.

GAN-based models follow the generative adversarial network (GAN) architecture [20], which employs a deep neural network generator to model the waveform signal in the time domain and a discriminator to estimate the quality of the generated speech. The two most commonly used GAN-based neural vocoders are Mel-GAN [21] and Parallel WaveGAN [22]. GAN-based vocoders have demonstrated extraordinary performance in recent works. They have been shown to outperform autoregressive and diffusion models in both generation speed and generation quality.

C. Synthetic Human Voice Detection

Because of the potential misuse of synthetic human voices, recent years have also seen rapid developments on the detection of synthetic human voices. One of the earliest detection methods is based on the bi-spectral analysis [4] of the audio signals. The bi-spectral analysis can capture the subtle inconsistencies in local phases of the synthetic human voices against the real signals. Real human voice signals have random local phases as the audio waves transmit and bounce around in the physical environment, while synthetic human voices do not have such characteristics. Such local phase inconsistencies cannot be heard by the human auditory system but can be picked up by the bi-spectral analysis. The other work known as DeepSonar [23] leverages network responses of audio signals as the feature to detect synthetic audios. The state-of-the-art synthetic voice detection methods are evaluated in the ASVspoof Challenge 2021, where four primary baseline algorithms, namely, the Gaussian mixture models CQCC-GMM [24] and LFCC-GMM [24], a light convolutional neural network model LFCC-LCNN [24], and RawNet2 [6], show the most reliable classification performance.

III. Method

In this work, we approach the problem of synthetic human voice detection based on the vocoder artifacts left in the synthetic audio signals. As it is highly unlikely for a real human voice signal to have vocoder artifacts other than our specifically designed self-vocoding signals, capturing the vocoder artifacts can be used as an important feature to detect synthetic human voices.

To be more specific, let $\mathbf{x}$ be the waveform of a human voice signal that has a label $y \in \{-1, +1\}$ with $+1$ corresponding to a real human voice while -1 being the synthetic human voice. Our aim is to build a parameterized classifier $\hat{y} = F_\theta(\mathbf{x})$ that predicts the label of an input $\mathbf{x}$. We choose the recent RawNet2 model [6] as the backbone for our classifier. The reason is that RawNet2 was designed to work directly on raw waveforms. This helps by reducing any possible information loss associated with neural vocoder artifacts, as compared to working with pre-processed features *e.g.*, mel-spectrograms or linear frequency cepstral coefficients (LFCCs).

Therefore, our classifier is constructed as a cascade of neural networks $F_\theta(\mathbf{x}) = B_{\theta_B}(R_{\theta_R}(\mathbf{x}))$, where $R_{\theta_R}(\mathbf{x})$ is the front-end RawNet2 model for feature extraction with its own set of parameters θ_R, B_{θ_B} is a back-end binary classifier and θ_B are its specific parameters, with $\theta = (\theta_R, \theta_B)$. We can train this classifier directly, which is done in the previous work [25], by solving

$$\min_{\theta} \sum_{(\mathbf{x},y)\in T} L_{\text{binary}}(y, F_\theta(\mathbf{x})),$$

where $L_{\text{binary}}(y, \hat{y})$ could be any loss function for binary classification, for instance, the cross-entropy loss. T stands for the training dataset with labeled real and synthetic human voice samples. However, this scheme predicates on the existence of a large number of synthetic human voice samples. However, this condition is becoming harder to satisfy in practice as it is difficult to keep up the pace with the fast development of the synthesis technology. More importantly, this model does not consider the distinct statistical characteristics of neural vocoders as an important cue for synthetic audio signals.

In this work, we reformulate this problem as one following the self-supervised representation learning [7]. In particular, we augment the model with a vocoder identifier M_{θ_M}, which classifies a synthetic human voice signal into one of the $c \in \{1, \cdots, k\}$ $(k \geq 2)$ possible neural vocoder models. Our purpose here is to ensure the feature extractor to be sensitive

TABLE I: The number of hours of audio synthesized by each neural vocoder.

Model	train-clean-100	train-clean-360	dev-clean	test-clean
WaveNet (A01)	4.28	15.49	0.75	0.76
WaveRNN (A02)	4.33	14.92	0.67	0.72
MelGAN (G01)	4.36	15.26	0.71	0.76
Parallel WaveGAN (G02)	4.37	15.54	0.68	0.75
WaveGrad (D01)	4.19	15.81	0.76	0.74
DiffWave (D02)	4.16	15.37	0.62	0.66
Total	25.69	92.39	4.19	4.39

to the statistical features in the vocoders. To this end, we form a new classification objective, as

$$\min_{\theta_B,\theta_R} \sum_{(\mathbf{x},y)\in T} L_{\text{binary}}(y, B_{\theta_B}(R_{\theta_R}(\mathbf{x}))) \\ + \min_{\theta_M,\theta_R} \lambda \sum_{(\mathbf{x},c)\in T'} L_{\text{mult}}(c, M_{\theta_M}(R_{\theta_R}(\mathbf{x}))).$$

In this equation, L_{mult} is a multi-class loss function, and we use the soft-max loss in our experiments. T' is a dataset containing only synthetic human voices but created with different neural vocoders as corresponding labels. This dataset is actually much easier to create by performing "self-vocoding", *i.e.*, creating synthetic human voices by running real samples through the process of mel-spectrogram transform and inverse, the latter performed with neural vocoders. We created such a dataset, LibriVoc, which will be described in detail in Section IV-A.

Note that the two terms in the new learning objective function serve different roles. The first one is the original binary classification term, while the second one focuses on vocoder identification, which can be regarded as a pretext task in a self-supervised representation learning framework. The two terms share the feature extraction component so that the distinct features of the vocoders can be captured and transferred to the binary classification task. λ is an adjustable hyper-parameter that controls the trade-off between the two loss terms. We start with equal weights ($\lambda = 1$) at the early epochs of training and then gradually reduce λ to reflect the increasing importance of the main task against the pretext task. The overall optimization of the learning objective is achieved with a stochastic gradient descent algorithm.

IV. Experiments

A. Datasets

Our experiments are based three different datasets. The first one, LibriVoc, is constructed for the task of vocoder identification and will be described in detail subsequently. The second dataset was constructed by us and contains real and synthetic speech created with a state-of-the-art TTS algorithm Tacotron 2 [8], combined with different vocoders in the synthesis process. The TTS dataset contains a total of 140 audio files, in which 120 audio files are generated by the tts model and six different vocoders, including WaveNet, WaveRNN, MelGAN, Parallel WaveGAN, WaveGrad, and DiffWave, and the last 20 original audio files as ground truth data. The third dataset is the DARPA SemaFor Hackathon3 Challenge Problem 1 (HK3CP1). This challenging dataset includes a training set of $7,000$ audio clips generated with 8 synthetic models (fastpitch, glowtts, gtts, tacotron, talknet, tacotron2, fastspeech2, Riva), and a testing set of $10,000$ clips with three additional models (mixertts, vits, speedyspeech) that are not in the training dataset. This dataset will be released to the public.

1) LibriVoc Dataset: As the statistical features of neural vocoders have not been extensively studied previously[1], there is no large-scale dataset for the task of vocoder identification and synthetic audio detection. To this end, we construct LibriVoc as a new open-source, large-scale dataset for the study of neural vocoder artifact detection. LibriVoc is derived from the LibriTTS speech corpus [5]. The LibriTTS corpus [5] itself is derived from the Librispeech dataset [27], wherein each sample is extracted from LibriVox audiobooks.[2]. The LibriTTS corpus has been widely used in text-to-speech research [28]–[30].

We use six state-of-the-art neural vocoders to generate synthesized speech samples in the LibriVoc dataset, namely, WaveNet and WaveRNN from the autoregressive vocoders, Mel-GAN and Parallel WaveGAN from the GAN-based vocoders, and WaveGrad and DiffWave from the diffusion-based vocoders. Specifically, we have 126.41 hours of real samples and 118.08 hours of synthesized, self-vocoded samples in the training set. Table I shows a breakdown of the number of hours of synthesized samples allocated to each neural vocoder in our experiments. Each vocoder is trained to synthesize waveform samples from a given mel-spectrogram extracted from an original sample; this process is referred to as "self-vocoding." By providing each vocoder with the same mel-spectrogram, we ensure that any unique artifacts present in the synthesized samples are attributable to the specific vocoder used to reconstruct the audio signal. We withhold a set of real samples to use as a validation set in the training process. By doing so, we also ensure that input samples will always be new to the vocoder, regardless of the training split.

To demonstrate the artifacts introduced by the neural vocoder, we show in Fig.3 the differences in the melspectra of the original and self-vocoded voice signals. Such differences are the basis of the subsequent detection algorithm.

B. Implementation Details

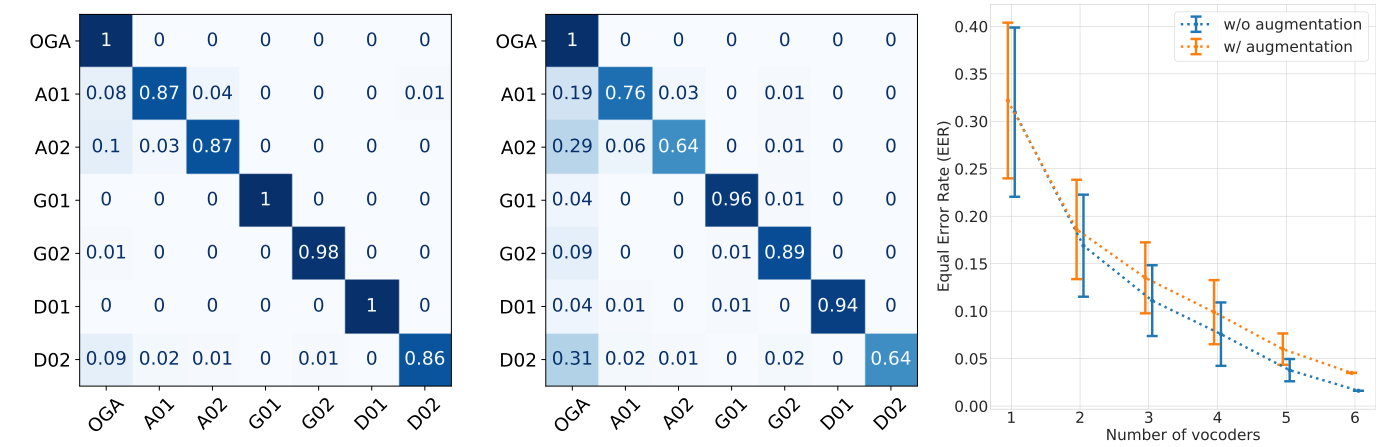
We use the model architecture of RawNet2 in [6]. The original RawNet2 is an end-to-end model designed for speaker verification, which consists of three main components: fixed sink filters, a residual network, and a gated recurrent unit (GRU). To accommodate to our task, we remove the final classification head and only use the feature extraction part of the RawNet2 model as described in Section III.

C. Results

We evaluate the performance of the synthetic human voice detector as described in Section III.

[1] The work of [26] is the only existing work to our best knowledge that qualitatively compares neural vocoders for waveform reconstruction qualities. Yet, that work only involves a small dataset and a limited number of vocoder models.

[2] https://librivox.org/



(a) Confusion matrix for the classification of individual neural vocoders and the original samples on the test-clean subset.

(b) Confusion matrix for the classification of individual neural vocoders and the original samples on the augmented test-clean subset.

(c) Mean Equal Error Rate (EER) for detecting vocoded speech on the test-clean subset, for different models trained with different numbers of vocoders.

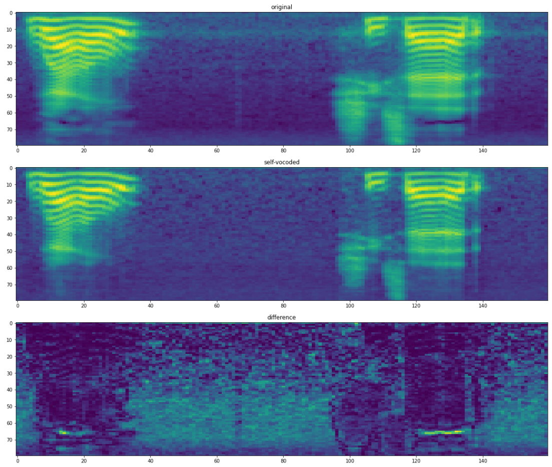


Fig. 3: The artifacts introduced by the neural vocoders to a voice signal. Specifically, the melspectra of the original (top) and the self-vocoded (middle) voice signal. Their differences are shown in the bottom panel. The differences correspond to the artifacts introduced by the vocoder.

Synthetic human voice detection. We first report the performance on the main task, *i.e.*, classification of real and synthetic human voices. On the TTS dataset, the SSRL-trained RawNet2 classifer with vocoder identification as a pretext task achieves a 0.40% equal error rate (EER), while a straightforward binary classifier based on the same RawNet2 model has a much higher EER of 1.61%.

This suggests that the self-supervised representation learning is beneficial to the overall classification task by identifying artifacts in the vocoders. In addition, our method has an overall EER of 11.36 on the DARPA SemaFor Hackathon3 Challenge Problem 1 testing dataset.

Classification of neural vocoder artifacts. We also analyze the performance of vocoder identification to confirm the existence of distinct vocoder artifacts. We use the RawNet2 model with vocoder labels ($M_{\theta_M}(R_{\theta_R}(\mathbf{x}))$), wherein the first label is for original samples and the other six for each of the neural vocoders we selected for our study. Figures 2a shows the confusion matrices evaluated on LibriVoc data subset. We refer to the two **autoregressive models** WaveNet and WaveRNN as A01 and A02, the two **GAN-based models** Mel-GAN and Parallel WaveGAN as G01 and G02, and the two **diffusion models** WaveGrad and DiffWave as D01 and D02. Original samples will be referred to as OGA. The experiment yielded an EER of 2.69%, which shows that the RawNet2 classifier can robustly detect vocoder artifacts, and each neural vocoder can produce unique artifacts, akin to a signature or vocoder fingerprint.

Robustness and Ablation Studies. To test the robustness of our detection method under some common data post-processing steps, we further construct a degraded dataset from the LibriVoc test dataset. First, we resample the input speech to intermediate sampling rates (8kHz, 16kHz, 22.05kHz, 32kHz, and 44.1kHz), and then resample back to the original sampling rate (24 kHz). In addition, we add background noise drawn from a single pre-recorded sample of crowd noise corresponding to three SNR values (*i.e.*, 8dB, 10dB, and 20dB). The probabilities of choosing between the original, re-sampled, or noisy speech segments are 40%, 40% and 20% respectively. The confusion matrix in Fig.2b corresponds to the vocoder identification performance on the augmented dataset, which corresponds to an EER of 3.50%. This shows that our detection method is robust to common data post-processing steps.

To study the effect of the number of vocoders in the training, we further trained detectors following the method of Section III in a leave N out setting, where N represents the number of excluded vocoders (ranging from 0 to 5 excluded). For reliable results, we tested all possible combinations of which neural vocoder to be included in N. This resulted in 63 possible combinations to use for all N values. Figure 2c shows the EER value on the y-axis versus the number of vocoders included in the training set on the x-axis. The error bar reports the mean and and standard deviation of all possible combinations for the same N value on both the augmented and non-augmented test-clean subset. As shown in Figure 2c, both the mean and standard deviation of EER decrease as more vocoders are added to the training set. We also observed that the effect of augmentation on overall performance is more noticeable on experiments with low EER values, as compared to those with higher ones. These results confirm that using fewer vocoders in the training set reduces the efficacy of the RawNet2 classifier, when detecting artifacts from unseen vocoders.

V. Conclusion

The advancements of AI-synthesized human voices have introduced a growing threat of impersonation and disinformation. It is therefore of practical importance to develop detection methods for synthetic human voices. In this work, we propose a new approach to detecting synthetic human voices based on identifying traces of *neural vocoders* in audio signals. A neural vocoder is a neural network that synthesizes waveforms from temporal-frequency representations, e.g., mel-spectrograms. The neural vocoder is a core component in most deepfake audio synthesis models, hence the identification of neural vocoder processing implies that an audio sample may be synthesized. To take advantage of the vocoder artifacts for synthetic human voice detection, we introduce a binary-class RawNet2 model that shares the front-end feature extractor with the one for vocoder identification. we employ a self-supervised representation learning (SSRL) approach [7], where we treat the vocoder identification as a pretext task to constrain the front-end feature extraction module to build the final binary classifier. Our experiments show that the RawNet2 model SSRL based on the vocoder artifacts achieves an overall high classification performance.

There are still rooms for improvement for this work and we will consider a few extensions as future work. First, we would like to augment the LibriVoc dataset to include more diverse real audio signals and environments. Second, there are more neural vocoders developed in recent years, and it is important to continue augmenting the model zoo to keep pace with latest development. Third, we will further explore more tailored solutions to the vocoder identification problem. Last, identification of vocoders is only indirect evidence of voice synthesis. It is our interest to further develop effective methods that can directly differentiate real and synthetic audios by combining cues from vocoders and other signal features.

References

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfake: A survey," *ACM Computing Surveys*, vol. 54, no. 1, 2021.

[2] Forbes, "A Voice Deepfake Was Used To Scam A CEO Out Of $243,000," https://www.cnn.com/2020/02/20/tech/fake-faces-deepfake/index.html, 11 2019.

[3] S. Lyu, "DeepFake detection: Current challenges and next steps," in *International Workshop on Media-Rich Fake News (MedFake) in conjunction with ICME*, London, UK, 2020.

[4] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting AI-synthesized speech using bispectral analysis." in *CVPR Workshops*, 2019, pp. 104–109.

[5] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[6] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[7] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances and challenges," *IEEE Signal Processing Magazine*, vol. 32.

[8] Z. Wang, Y. Liu, and L. Shan, "CE-Tacotron2: End-to-end emotional speech synthesis," in *2021 60th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 2021, pp. 48–52.

[9] A. van den Oord, S. Dieleman, H. Zen *et al.*, "WaveNet: A generative model for raw audio," in *arXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[10] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[11] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[13] E. A. AlBadawy and S. Lyu, "Voice conversion using speech-to-speech neuro-style transfer," *Proc. Interspeech 2017*, 2020.

[14] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[15] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 19–23.

[16] N. Kalchbrenner, E. Elsen, K. Simonyan *et al.*, "Efficient neural audio synthesis," in *ICML*, 2018.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[18] N. Chen, Y. Zhang, H. Zen *et al.*, "WaveGrad: Estimating gradients for waveform generation," in *ICLR*, 2020.

[19] Z. Kong, W. Ping, J. Huang *et al.*, "DiffWave: A versatile diffusion model for audio synthesis," in *ICLR*, 2020.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[21] K. Kumar, R. Kumar, T. de Boissiere *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *arXiv*, 2019.

[22] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020.

[23] R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, and Y. Liu, "Deepsonar: Towards effective and robust detection of ai-synthesized fake voices," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1207–1216.

[24] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[25] G. Yang, S. Yang, K. Liu *et al.*, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *SLT workshop*, 2021, pp. 492–498.

[26] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 7–12.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[28] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[29] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," *arXiv preprint arXiv:2005.05957*, 2020.

[30] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.