ADVERSARIAL AUDIO SYNTHESIS USING A HARMONIC-PERCUSSIVE DISCRIMINATOR

Jihyun Lee[⋆] *Hyungseob Lim*[⋆] *Chanwoo Lee*[⋆] *Inseon Jang*[†] *Hong-Goo Kang*[⋆]

[⋆] Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
[†]Electronics and Telecommunications Research Institution, Daejeon, South Korea

ABSTRACT

In this paper, we propose a discriminator design scheme for generative adversarial network-based audio signal generation. Unlike conventional discriminators that take an entire signal as input, our discriminator separates the audio signal into harmonic and percussive components and analyzes each component independently. The rationale behind this idea is that conventional discriminators cannot reliably capture subtle distortions in audio signals, which have complicated time-frequency characteristics. By considering the time-frequency resolution of audio signals, our proposed method encourages the generator to better reconstruct harmonic and percussive features, both of which are critical for the quality of the generated signals. Listening tests show that our framework significantly enhances the stability of pitches and generates clearer piano samples compared to a baseline.

Index Terms— Audio synthesis, generative adversarial networks, harmonic–percussive source separation

1. INTRODUCTION

Neural audio synthesis, which involves the generation of audio waveforms using a neural network, has been applied to various tasks such as music synthesis, audio coding, and sound effect generation. Since the quality of synthesized audio signals is crucial to system performance in these target tasks, there has been much interest in developing more advanced neural networks for the generation of higher-quality waveforms. Recent research has focused on the potential of deep generative models to synthesize high-quality audio [1, 2].

Such audio synthesis methods can be categorized as either unconditional or conditional, depending on whether an external condition is given. Unconditional methods attempt to generate audio signals without using any conditional features [3–5]. On the other hand, conditional methods use extra information such as musical chords [6], instruments [7], or sound classes [8] in order to generate audio signals with certain desired characteristics. In neural vocoders, spectrograms, which represent the time-frequency (T-F) domain of audio signals, have often been used as effective acoustic features for generating high-quality waveforms [9–11]. Spectrograms have also been employed as intermediate features, instead of using high-level conditions directly [8].

In generating realistic audio waveforms from spectrograms, generative adversarial networks (GANs) have shown particularly successful performance among various types of deep generative models. Here, one core architectural approach to improving synthesis quality is to design a discriminator in such a way that it can capture the key properties of audio signals in detail. Various GAN-based speech synthesis networks have been designed based on this idea [9–11].

However, it is still challenging to generalize this synthesis process to general audio signals, which tend to have more complicated T-F structures than speech signals. Compared to speech signals, which only have a simple harmonic structure, there is no limitation to the number and types of harmonic structures in audio signals. In addition, audio signals often contain dominant impulsive sounds. Since most discriminators proposed for GAN-based speech synthesis use an entire signal as their input, it is very difficult for them to capture the complicated T-F structures of audio signals. Therefore, there is a need to design alternative types of discriminators to reliably produce high-quality audio signals.

In this paper, we propose a new design framework for the discriminator network of a GAN-based audio signal generation model conditioned on T-F representations. Our proposed discriminator consists of two separate networks that are designed to separately capture the quality of harmonic and percussive components in audio signals. We use a signal processing-based harmonic–percussive source separation (HPSS) algorithm [12] to faithfully extract harmonic and percussive components from audio signals, and use these components as inputs to the discriminators. Harmonic sounds are quasi-stationary for certain periods, and thus require a finer frequency resolution. On the other hand, percussive sounds, which can occur suddenly and attenuate quickly, require better time resolution. By considering these time-frequency properties of both components, our discriminator design differentiates the receptive fields for each component and precisely detects the distortion of audio signals produced by the generator. In doing so, the generation network is able to better synthesize waveforms based on the conditioning information without any modifications to the generator architecture. We verify the performance of our discriminator

framework in a waveform generation network conditioned on linear spectrograms of piano samples from the MAESTRO dataset [13]. Subjective evaluation results demonstrate that our method outperforms a Parallel WaveGAN baseline [11] in terms of the pitch stability and sound clarity of generated piano samples.

2. RELATED WORK

2.1. GAN-based audio synthesis using T-F features

Many conditional GAN-based models have been proposed to generate high-quality audio signals with T-F spectral features as the auxiliary condition. Some recent works have focused on improving audio quality by utilizing a variety of distinctive discriminator architectures. MelGAN [9] and VocGAN [14] utilize one or a number of multi-scale discriminators (MSD), which discriminates the originality of an input waveform in various frequency ranges. HiFi-GAN [10] proposed a multi-period discriminator (MPD), which considers the periodicity of speech signals in different sample resolutions. FreGAN [15] used the modified MSD and MPD in order to discriminate signals in multiple frequency subbands.

Parallel WaveGAN [11] proposed a simpler discriminator architecture than the methods mentioned above. Under the Parallel WaveGAN framework, [16] designed two discriminators to separately analyze voiced and unvoiced speech signals, based on the fact that the characteristics of both segments are different. However, to the best of our knowledge, there has not yet been a discriminator design for effectively synthesizing generic audio signals conditioned on T-F features. Our approach is unique in that we use two networks to separately discriminate harmonic and percussive components of the audio signal, taking into consideration that the T-F characteristics and the perception of these components are very different.

2.2. Harmonic–percussive source separation (HPSS)

Harmonic–percussive source separation (HPSS) is the task of separating harmonic and percussive components from an audio signal. Fitzgerald *et al.* [17] performed decomposition in the T-F domain by masking the T-F features with binary or soft masks. Specifically, they obtained harmonic-enhanced and percussion-enhanced magnitude spectrograms by performing median filtering on the magnitude spectrogram of the signal along the time and frequency axes, respectively. Then, masking values were determined using the ratio between the two enhanced magnitude spectrograms. Finally, harmonic and percussive waveforms were generated by taking an inverse short-time Fourier transform (STFT) on the masked STFT.

Considering the difference in the T-F resolution between the harmonic and percussive components, Driedger *et al.* [12] proposed a method to iteratively extract harmonic and percussive components. At the first separation stage, harmonic components are extracted from a long temporal window, which offers a high frequency resolution. In the second stage, percussive components are extracted from a remaining signal using a short temporal window to grant a high time resolution. In our method, we apply this algorithm to divide a generated or target signal into its harmonic and percussive components, which are provided to our discriminator.

3. PROPOSED MODEL

In this section, we describe how our harmonic–percussive discriminator works in a GAN-based audio generation framework. Figure 1 illustrates our overall system design, which consists of a generator and the proposed discriminator. The generator utilizes a magnitude spectrogram as an auxiliary conditional feature and synthesizes the corresponding audio waveform. The discrimination process consists of two modules: a harmonic–percussive separator and a discriminator.

3.1. Harmonic–percussive discriminator

Harmonic–percussive separator. In the first module, we separate the waveforms into harmonic and percussive components. Since it is costly to perform HPSS during training, we prepare beforehand the soft masks that are required to extract each component using the iterative algorithm described in Section 2.2. During training, we element-wise multiply the soft masks with the STFT magnitude and phase components of the generated or target audio signals. Finally, we obtain the waveforms of the harmonic and percussive components by taking the inverse STFT of the masked STFT.

Since the masking values for the harmonic and percussive components are estimated using different parameters for each component, the separated waveforms have distinct characteristics; harmonic waveforms contain periodic flows and percussive waveforms contain transient attacks. The two components are then fed into the second discriminator module.

Discriminator. The discriminator module consists of two separate discriminators: a harmonic discriminator D_H and a percussive discriminator D_P. Each discriminator is specifically designed to provide an appropriate time resolution corresponding to the characteristics of its input component.

We adopt the design of the discriminator used in [16] as the base architecture for both the harmonic and percussive discriminators. The overall architecture is almost identical to that of a conventional discriminator—i.e., a stack of CNN layers and nonlinear activations, as illustrated in Fig. 2. However, we modify the number of layers of the network to clearly differentiate the receptive fields for the harmonic and percussive components. Based on the fact that harmonic signals need longer time intervals for analysis, we first increase the network layers of the harmonic discriminator to enlarge the receptive field. We set dilation factors of the harmonic discriminator to increase exponentially by a factor of 2, which

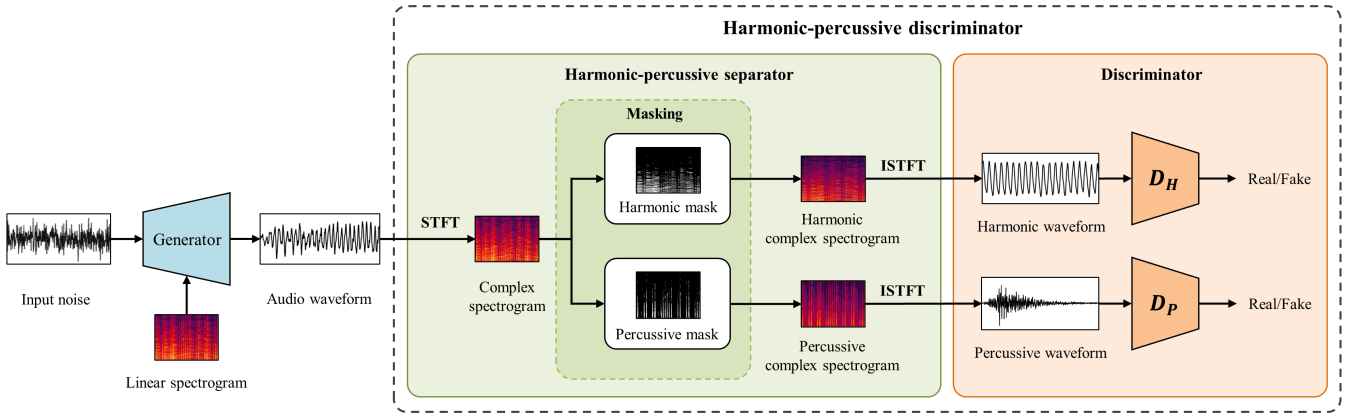

Fig. 1. Proposed GAN-based audio synthesis framework. D_H and D_P denote the harmonic discriminator and the percussive discriminator, respectively.

also contributes to a larger receptive field. A smaller receptive field is sufficient for percussive sounds because their characteristics vary rapidly in time. Therefore, we fix the dilation factors in all layers of the percussive discriminator to 1, maintaining the small receptive field even with the stack of layers for higher model complexity.

3.2. Model details

Generator. For the generator, we used the non-autoregressive WaveNet-like model used in [11]. It transforms random Gaussian noise into an audio waveform using a magnitude spectrogram as an auxiliary feature. As our proposed discriminator is applicable to any type of generator, we could choose other generators, such as the one using transposed convolutions in the up-sampling block [10]. However, these generators have been found to frequently produce undesirable artifacts such as spectral replicas and tonal artifacts [18]. We found these artifacts to become especially noticeable when we synthesized generic audio signals in our internal experiments; therefore, we adopted the WaveNet-like model to avoid this issue.

Discriminators. We used $N = 8$ convolutional layers for each discriminator. Accordingly, the receptive fields for the harmonic and the percussive discriminators were calculated as 511 and 17 samples each. These values correspond to approximately 32 ms and 1 ms under the 16 kHz sampling rate.

Loss functions. We used least square GAN [19] for adversarial training. We defined the same training loss for both the harmonic and percussive discriminators. Training losses for discriminators ($\mathcal{L}_D$) and the generator ($\mathcal{L}_G$) are as follows:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{data}}[(D(x)-1)^2] + \mathbb{E}_{z\sim N(0,I)}[D(G(z,s))^2] \quad (1)$$

$$\mathcal{L}_G = \lambda_{adv} \times \mathbb{E}_{z\sim N(0,I)}[(D(G(z,s))-1)^2] + \lambda_{stft} \times \mathcal{L}_{MultiSTFT} \quad (2)$$

where x and p_{data} denote the target waveform and its distribution, z is random Gaussian noise, and s is the magnitude spectrogram. λ_{adv} and λ_{stft} denote weights for an adversarial loss and a multi-resolution STFT loss, denoted by

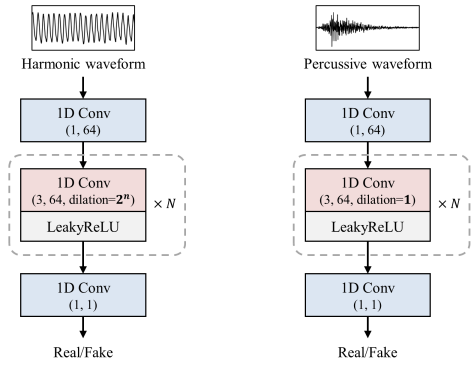

Fig. 2. Architecture of harmonic and percussive discriminators. Parameters in parentheses denote kernel size and output channel size, sequentially. $n = 1, ..., N$ denotes the layer number out of N total layers.

$\mathcal{L}_{MultiSTFT}$, respectively. The multi-resolution STFT loss is an auxiliary loss widely used in neural vocoders [11, 15, 16]. It considers various T-F resolutions of input signals by calculating errors between STFTs with different parameters and summing them all. We used the same parameters for computing $\mathcal{L}_{MultiSTFT}$ as in [11].

4. EXPERIMENTS

4.1. Dataset and preprocessing

To evaluate our proposed method, we trained our model using the MAESTRO [13] dataset, which is composed of piano performances. This dataset is useful for our task, as it consists of not only harmonic sounds during melodies, but also percussive sounds when notes are hit. All audio samples are downsampled to 16 kHz for our experiments. We note that direct generation of a full-band signal is still a challenging task, and the downsampling is unlikely to remove the harmonic and percussive components, which are of our interest. We randomly extracted 6 hours of samples for training and 30

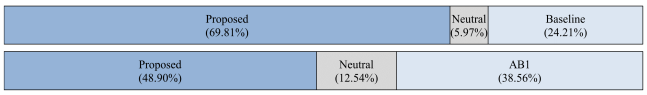

Fig. 3. ABX results on comparison with the baseline (*upper*) and the ablation model (AB1) (*lower*).

minutes for evaluation. Linear spectrograms, which are used as conditional features for the generator, were computed using a Hanning window having a length of 1,024 samples and a hop length of 256 samples.

To prepare soft masks for the harmonic–percussive separator (Section 2.2), we performed STFTs with different parameters for the two stages. In the first stage, we used a window length of 4,096 samples to extract harmonic components. In the second stage, for extracting percussive components, we used a shorter window length of 256 samples. We set the hop length to be 25% of the filter length in both stages. Filter lengths used for median filtering were fixed to 0.2 seconds for the time-axis and 500 Hz for the frequency-axis.

4.2. Training details

We used an Adam optimizer [20] to train the entire model. The learning rate was initially set to 0.0001 for the generator as well as the discriminators, and was scheduled to decay exponentially every epoch with a decay parameter of 0.999. We set λ_{adv} to 4 and λ_{stft} to 1. To stabilize the training, we pretrained the generator for 200K iterations without the discriminator. Afterwards, we trained the generator with the proposed discriminators for an additional 300K iterations. We trained the original Parallel WaveGAN [11] as a baseline.

5. EVALUATION

5.1. Evaluation metrics

To verify the performance of our proposed model, we conducted ABX listening tests on synthesized piano sounds. We randomly extracted 20 samples from the evaluation dataset and asked 16 listeners to assess which sample sound is closer to a reference between the proposed model and the baseline model in terms of pitch stability and sound clarity.

5.2. Evaluation results

Comparison with the baseline. We compared our method with the baseline in order to evaluate the effectiveness of the proposed discriminator framework. As shown in Fig. 3, our model outperforms the baseline, which means that it qualitatively generates more stable and clearer sounds than the baseline. This demonstrates that the harmonic–percussive separator and discriminator contribute to improved waveform generation of audio signals corresponding to the given conditions. We observed that our model is able to generate stable pitches,

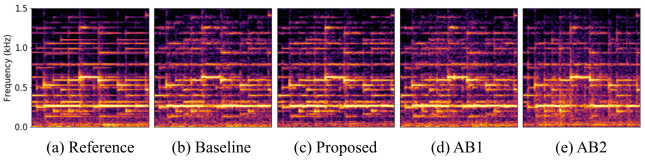

Fig. 4. Spectrograms of (a) reference and (b)-(e) generated waveforms in low frequency band up to 1.5 kHz.

especially in lower frequency bands, as can be seen in Fig. 4. Considering that the generator in GAN-based audio synthesis models has the limitation of not being able to capture characteristics outside its receptive field, this observation shows that our proposed discriminators allow the generator to utilize global characteristics of conditional features without modifications to the generator architecture.

Ablation study. We conducted an ablation study of constituent modules in the proposed discriminator to verify the effect of each module. The harmonic–percussive separator contributed to sound quality when comparing the proposed model to one without the module (AB1), as shown in Fig. 3 and Fig. 4. To see whether the harmonic and percussive discriminators in the discriminator module analyze input signals differently as we expected, we also tried switching the roles of the harmonic and percussive discriminators (AB2). We found that samples generated from this model configuration were very noisy and distorted, as shown in Fig. 4, so we did not perform ABX tests on them. We also observed that percussive components are overemphasized compared to the reference signal in the result of AB2. This observation demonstrates that the signal characteristic-dependent discriminators work appropriately for different types of signals.

6. CONCLUSION

We proposed a harmonic–percussive discriminator design for GAN-based audio signal generation, which separates the audio signal into its harmonic and percussive components. Our discriminator estimates the quality of output signals from the generator by analyzing the characteristics of each component independently. Subjective listening tests showed that an audio generation network trained with our discriminator was able to generate piano signals with more stable pitches and clearer sound compared to a baseline. Analysis of the results also showed that our proposed discriminator design allows the generator to utilize more global characteristics of its conditional features without any modifications to its architecture.

7. ACKNOWLEDGEMENTS

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [21ZH1200, The research of the basic media contents technologies]

8. REFERENCES

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[3] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," in *International Conference on Learning Representations*, 2019.

[4] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] S. Dieleman, A. v. d. Oord, and K. Simonyan, "The challenge of realistic music generation: modelling raw audio at scale," in *32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018.

[6] L. Yang, S. Chou, and Y. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[7] A. Defossez, N. Zeghidour, N. Usunier, L. Bottou, and F. Bach, "Sing: Symbol-to-instrument neural generator," in *32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[8] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *arXiv preprint arXiv:2107.09998*, 2021.

[9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, and A. de Brébisson, Y. Bengio, and A. C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[10] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[11] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[12] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals.," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 611–616.

[13] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations (ICLR)*, 2019.

[14] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "VocGAN: A High-Fidelity Real-Time Vocoder with a Hierarchically-Nested Adversarial Network," in *Proc. Interspeech 2020*, 2020.

[15] J. Kim, S. Lee, J. Lee, and S. Lee, "Fre-GAN: Adversarial Frequency-Consistent Audio Synthesis," in *Proc. Interspeech 2021*, 2021.

[16] R. Yamamoto, and E. Song, and M. Hwang, J. Kim, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[17] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2010.

[18] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling artifacts in neural audio synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[19] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.