NEURAL SYNTHESIS OF FOOTSTEPS SOUND EFFECTS WITH GENERATIVE ADVERSARIAL NETWORKS

Marco Comunità, Huy Phan, Joshua D. Reiss

Centre for Digital Music, Queen Mary University of London, UK

ABSTRACT

Footsteps are among the most ubiquitous sound effects in multimedia applications. There is substantial research into understanding the acoustic features and developing synthesis models for footstep sound effects. In this paper, we present a first attempt at adopting neural synthesis for this task. We implemented two GAN-based architectures and compared the results with real recordings as well as six traditional sound synthesis methods. Our architectures reached realism scores as high as recorded samples, showing encouraging results for the task at hand.

Index Terms— footsteps, neural synthesis, sound effects, GAN

1. INTRODUCTION

When sound designers are given the task of creating sound effects for movies, video games or radio shows, they have essentially two options: pre-recorded samples or procedural audio. In the first case, they usually rely on large libraries of high quality audio recordings, from which they have to select, edit and mix samples for each event they need to sonify. For a realistic result, especially in video games where the same actions are repeated many times, several samples are selected for every event and randomised during action. This creates challenges in terms of memory requirements, assets management and implementation time.

Alternatively, procedural audio [1] aims to synthesise sound effects in real time, based on a set of input parameters. In the context of video games, these parameters might come from the specific interaction of a character with the environment. This approach presents challenges in terms of development of procedural models which can synthesise high quality and realistic audio, as well as finding the right parameters values for each sound event.

Footsteps sounds are a typical example of the challenges of sound design. It is an omnipresent sound, which is generally fairly repetitive, but on which the human ear is being constantly trained. In fact, it is possible for a subject to identify from recorded or synthesised footsteps, things like: gender [2], emotions [3], posture [4], identity [5], ground materials [6], and type of locomotion [7].

Thus, it is not surprising that researchers put substantial efforts into understanding the acoustic features [8] and developing synthesis models of footsteps sounds. Cook [9] made a first attempt at synthesising footsteps on different surfaces, based on his previous work on physically-informed stochastic models (PhISM) [10]. Another physically-informed model - based on using a stochastic controller to drive sums of microscopic impacts - was proposed by Fontana and Bresin [11]. DeWitt and Bresin [12] proposed a model that included the user's emotion parameter. Farnell [13] instead, developed a procedural model by studying the characteristics of locomotion in primates. In [14] Turchet *et al.* developed physical and physically-inspired models coupled with additive synthesis and signals multiplication.

To this day, there has not yet been an attempt at exploring the use of neural networks for the synthesis of footsteps sounds although there is substantial literature exploring neural synthesis of broadband impulsive sounds, such as drums samples, which have some similarities to footsteps. One of the first attempts was in [15], where Donahue *et al.* developed WaveGAN - a generative adversarial network for unconditional audio synthesis. Another example of neural synthesis of drums is [16], where the authors used a Progressive Growing GAN. Variational autoencoders [17] and U-Nets [18] have also been used for the same task. But, the application of recent developments to neural synthesis of sound effects is yet to be explored. We could only find one other work [19], related to the present study, where the authors focused on synthesis of knocking sounds with emotional content using a conditional WaveGAN.

In this paper we describe the first attempt at neural synthesis of footsteps sound effects. In Section 2, we propose a hybrid architecture that improves the quality and better approximates the real data distribution, with respect to a standard conditional WaveGAN. Objective evaluation of this architecture is provided in Section 3. Section 4 reports on the first listening test that compares "traditional" and neural synthesis models on the task at hand. Discussion and conclusions are given in Section 5.

2. ARCHITECTURE

2.1. Data

To train our models we curated a small dataset (81 samples) using free footsteps sounds available on the Zapsplat website[1]. These are high quality samples, recorded by Foley artists using a single type of shoes (women, high heels) on seven surfaces (carpet, deck, metal, pavement, rug, wood and wood internal). The samples were converted to WAV file format, resampled at 16kHz, time aligned and normalised to -6dBFS.

2.2. Generator

The original WaveGAN generator [15] was designed to synthesise 16384 samples ($\sim$1s at 16kHz sampling frequency); and afterwards extended to 32768 or 65536 samples[2] ($\sim$2s and $\sim$4s at 16kHz). We adapted the architecture to synthesise 8192 samples, which is sufficient to capture all footsteps samples in our dataset. As shown in Fig. 1, the generator expands the latent variable z to the final audio output size. After reshaping and concatenation with the conditioning label [19, 20], the output is synthesised by passing the input through 5 upsampling 1-D convolutional layers. Upsampling is obtained by: zero-stuffing or nearest neighbour, linear or cubic interpolation. Zero-stuffing plus 1-D convolution is equivalent to using 1-D transposed convolution with stride equal to the upsampling rate. In the other cases, we split the operation into upsampling and 1-D

[1] https://www.zapsplat.com/sound-effect-packs/footsteps-in-high-heels

[2] https://github.com/chrisdonahue/wavegan

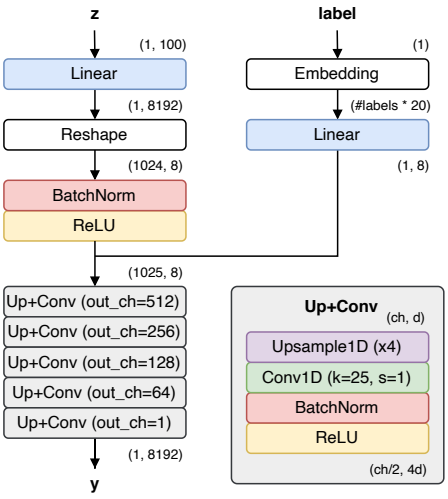

Fig. 1: WaveGAN generator.

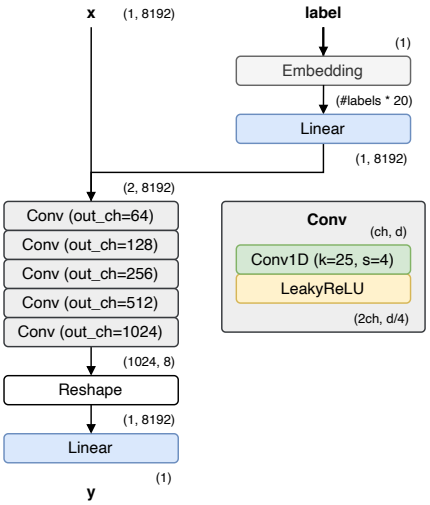

Fig. 2: WaveGAN discriminator.

convolution with stride of 1. Every convolutional layer uses a kernel size of 25, as in the original design. Differently from the original, we included batch normalisation layers after each convolution. With the exception of the last layer, the number of output channels is halved at each convolution layer.

2.3. Discriminator

Recent results in the field of neural speech synthesis [21–26] have shown how GAN-based vocoders are capable of reaching state-of-the-art results in terms of mean opinion scores. You *et al.* [27] hypothesised that this success is not related to the specific design choices or training strategies; they identified it in the multi-resolution discriminating framework. In their work, the authors trained 6 different generators using the same discriminator (Hifi-GAN - [26]) reaching very similar performance independently of the specific generator.

We experimented with a similar approach by implementing conditional versions of WaveGAN and HiFi-GAN discriminators. Our WaveGAN discriminator was again adapted to 8192 samples of the original architecture, based on 5 1-D convolutional layers with stride of 4 (see Fig. 2). The HiFi-GAN discriminator is made of two separate discriminators (multi-scale and multi-period) each of which is made of several sub-discriminators that work with inputs of different resolutions. The multi-scale discriminator works on raw audio, $\times 2$ average-pooled and $\times 4$ average-pooled audio (i.e., a downsampled and smoothed version of the original signal). The multi-period discriminator works on "equally spaced samples of an input audio; the space is given as period p". The periods are set in the original model to 2, 3, 5, 7 and 11.

2.4. Loss and Training Procedure

We trained the two architectures - WaveGAN generator and discriminator (referred to as WaveGAN), WaveGAN generator and HiFi-GAN discriminator (referred to as HiFi-WaveGAN) - using different paradigms.

In the first case we opted for a Wasserstein GAN with gradient penalty [28] (WGAN-GP), which has been shown to improve training stability and help convergence towards a minimum which approximates better the real data distribution as well as the synthesised samples' quality. When training a WGAN-GP the discriminator's weights were updated several times for each update of the generator; we followed the standard approach of 5 to 1 updates ratio.

For HiFi-WaveGAN we followed the approach suggested in [26], opting for least squares GAN (LS-GAN) [29], where the binary cross-entropy terms of the original GAN [30] are replaced with least squares losses. [26] included additional losses for the generator; specifically, a mel-spectrogram loss and a feature matching loss. The mel-spectrogram loss measures the $\ell 1$-distance between the mel-spectrogram of a synthesised and a ground truth waveforms. We discarded this term since, differently from HiFi-GAN, our generator was not designed and trained to synthesise waveforms from ground-truth spectrograms. However, we kept the feature matching loss, which measures the $\ell 1$-distance of the features extracted at every level of each sub-discriminator, between real and generated samples. See [26] for a more detailed description of each loss term. The final losses for our HiFi-WaveGAN were:

$$\mathcal{L}_G = \mathcal{L}_{Adv}(G;D) + \lambda_{fm}\mathcal{L}_{FM},$$
$$\mathcal{L}_D = \mathcal{L}_{Adv}(D;G),$$

where $\mathcal{L}_G$ and $\mathcal{L}_D$ are the generator and discriminator total losses with $\mathcal{L}_{Adv}(G;D)$ and $\mathcal{L}_{Adv}(D;G)$ the adversarial loss terms for generator and discriminator and $\lambda_{fm}\mathcal{L}_{FM}$ the feature matching loss. Both architectures were trained for 120k batches, with a batch size of 16 and a learning rate of 0.0001. WaveGAN used Adam while HiFi-WaveGAN used AdamW optimisers.

3. OBJECTIVE EVALUATION

There are no formalised methods to reliably evaluate the quality and diversity of synthesised audio, but there are several metrics which are commonly adopted to analyse and compare neural synthesis models. We followed a similar approach to [16] for objective evaluation, where the authors relied on Inception Score (IS), Kernel Inception Distance (KID) and Fréchet Audio Distance (FAD). We also relied on the maximum mean discrepancy (MMD) [31] as a measure of similarity between real and synthesised samples, using the same formulation adopted in [32], and computed the MMD using the $\ell 1$-distance between OpenL3 embeddings [33] (env, mel128, 512)[3].

[3]https://github.com/torchopenl3/torchopenl3

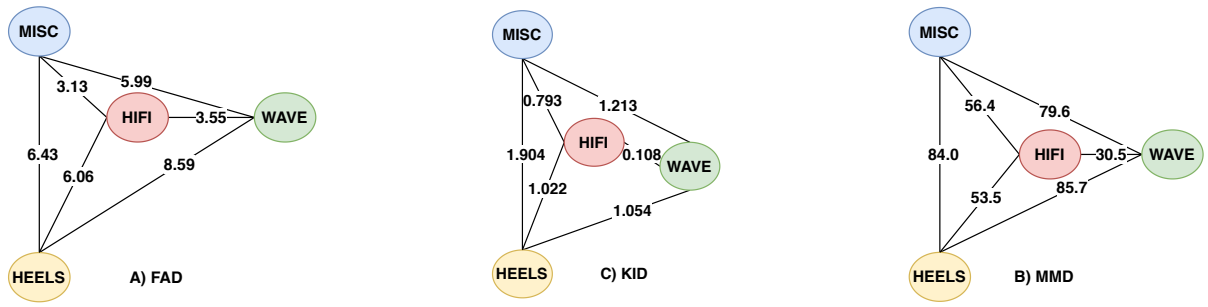

Fig. 3: Graphs representing Fréchet Audio Distance (A), Kernel Inception Distance (B) and Maximum Mean Discrepancy (C) for our models.

To analyse whether our models were capable of learning the real data distribution - and to verify that they were not affected by overfitting or mode collapse - we used principal components analysis (PCA) on the OpenL3 embeddings of real and synthesised samples. We generated 1000 samples for each class and plot the results of PCA in Fig. 4. It can be seen that HiFi-WaveGAN reached a better approximation of the real data without collapsing into synthesising only samples that were seen during training. However, there is currently no way to establish a correlation between distance in the embedding space with distance in terms of human perception. It is therefore not possible to comment on the diversity of synthesised samples from these results.

Instead, we used IS, which gave us a way to measure diversity and semantic discriminability. This measure ranges from 1 to n (with n number of classes) and it is maximised for models which can generate samples for all possible classes and that are classified with high confidence. As classifier, we used the same Inception Net variant developed in [16] and adapted it to our domain. We trained the model on a separate dataset (*Zapsplat Misc*) obtained by scraping all the freely available footsteps samples on the Zapsplat website [4] and organised them into 5 classes depending on the surface material (*carpet/rug, deck/boardwalk, metal, pavement/concrete, wood/wood internal*). Likewise, we merged the training (*Zapsplat Heels*) and generated data into the same 5 classes by grouping *carpet/rug* and *wood/wood internal* samples together.

We trained the Inception Net for 100 epochs, reaching 86% validation accuracy. The results are shown in Table 1. It is interesting to notice how the IS for HiFi-WaveGAN is slightly higher then it is for the training data (*Zapsplat Heels*). Assuming that the synthesised data cannot reach a higher semantic discriminability than the data it learned from, the score should be related to diversity. Which in turn might depend on two factors. First, we evaluated IS on a number of Hifi-WaveGAN samples orders of magnitudes greater than the training data (3500 versus 81). Excluding the case of mode collapse, this will inherently carry higher diversity. And second, Hifi-WaveGAN is actually able to synthesise samples not seen during training, and could therefore lead to a more diverse set. However, it is not possible to comment on the "perceptual" diversity of the generated samples based on IS only.

Fig. 3 shows the other metrics we adopted to compare real (*Misc* and *Heels*) and synthesised (*HiFi* and *Wave*) data. For clarity, we represent FAD, KID and MMD on 3 graphs, where the distance between nodes is roughly proportional to each metric values (written on the edges). The 3 metrics, which are based on comparing embeddings distributions (VGG-ish for FAD, Inception for KID and OpenL3 for MMD), all depict a similar picture where HiFi-WaveGAN seems to better approximate the training data. FAD and

[4]https://www.zapsplat.com

Dataset	IS
Zapsplat Misc	4.139
Zapsplat Heels	3.221
HiFi-WaveGAN	3.411
WaveGAN	3.232

Table 1: Inception Score for training and generated data.

KID also place HiFi-WaveGAN samples nearer to the *Misc* dataset, again suggesting that the model is capable of synthesising samples with greater diversity than the training data. FAD, correlating with human judgement, is also a measure of the perceived quality of individual sounds (the lower the FAD, the higher the quality). In our case, HiFi-WaveGAN also seems to score higher in terms of quality.

4. SUBJECTIVE EVALUATION

For the subjective evaluation we followed a similar paradigm to [34]. The Web Audio Evaluation Tool [35] was adopted to run an audio perceptual evaluation (APE) [36]; a multi-stimulus test in which a participant was presented with a series of samples to be compared and rated on a continuous scale from 0 to 1. On this scale, 4 reference values (very unrealistic, somewhat unrealistic, somewhat realistic, very realistic) were given at the 0, 0.33, 0.66 and 1 points, respectively. Eight synthesis methods were compared to real recordings:

1. Procedural model 1 (PM1) by Fontana and Bresin [11]
2. Procedural model 2 (PM2) by Farnell [13] [5]
3. Procedural model 3 (PM3) from Nemisindo [37] [6]
4. Sinusoidal plus stochastic (SPS) by Amatriain *et al.* [38] [7]
5. Statistical modelling (STAT) by McDermott *et al.* [39] [8]
6. Additive synthesis (ADD) by Verron *et al.* [40] [9]
7. WaveGAN (WAVE)
8. HiFi-WaveGAN (HIFI)

To present a more realistic and reliable scenario we prepared 10 s long walks by concatenating single samples. We started from real recordings and chose - through informal listening - the time interval between samples that gave the most realistic result. The same pace was then replicated for all the other synthesis methods. We prepared a total of 10 series of 9 walks (1 for each synthesis method plus the real recordings); and presented each participant with 5 of these series

[5]http://aspress.co.uk/sd/practical26.html
[6]https://nemisindo.com/
[7]https://www.dafx.de/DAFX_Book_Page_2nd_edition/chapter10.html
[8]https://mcdermottlab.mit.edu/downloads.html
[9]http://www.charlesverron.com/spad.html

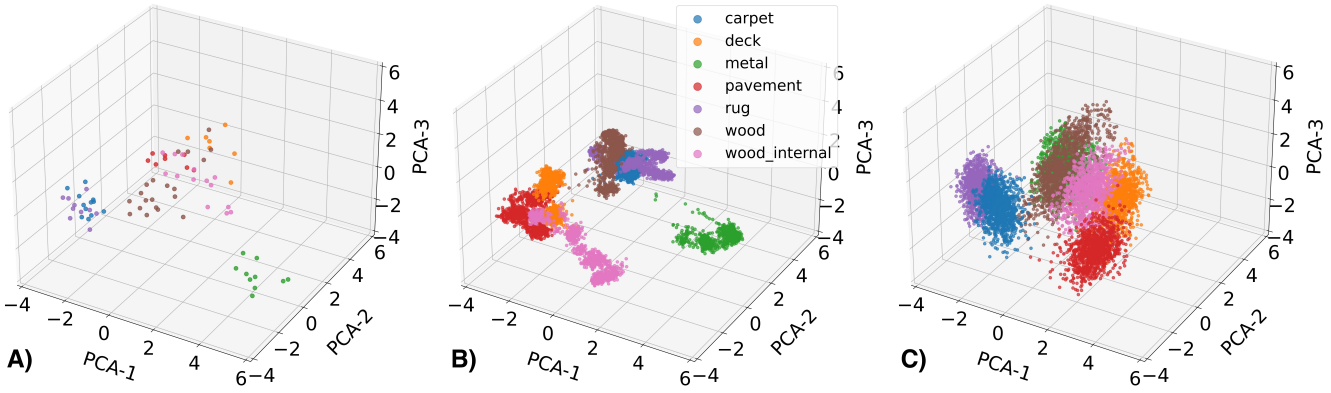

Fig. 4: Scatter plots of principal components analysis on OpenL3 embeddings for: A) Training Data, B) HiFi-WaveGAN and C) WaveGAN.

to be able to compare many different conditions (i.e., shoe types and surface materials) while keeping the test short.

4.1. Results

A total of 19 participants took part in the online test[10]. Of these, 10 identified as male, 6 as female and 3 preferred not to indicate their gender. 3 participants were excluded since they had no previous experience with critical listening tests. Of the remaining participants, all but 1 had experience as: musicians (15 out of 16, $\mu = 9.7$, $\sigma = 7.6$ and max $= 23$ years), sound engineers (11 out of 16, $\mu = 4.6$, $\sigma = 7.6$, max $= 15$ years) and sound designers (7 out of 16, $\mu = 1.7$, $\sigma = 2.9$, max $= 10$ years). We also enquired about the headphone models and verified that all participants used good quality devices.

Two criteria were adopted to judge the reliability of each rating. We used one of the procedural models (PM1) as an anchor, and excluded all cases where this model was rated above 0.1. Also, we considered real recordings to be the reference, and excluded all cases where they were rated below 0.5.

Final results are shown in Fig. 5. Together with the anchor (PM1), Farnell's model (PM2), as well as statistical and sinusoidal modelling, are the lowest rated. This result is not surprising since Farnell's procedural model is a fairly basic implementation, which lacks the necessary output quality. Statistical modelling is usually adopted for texture synthesis, and also in those situations, it shows good results only for very specific cases (see [34, 39]). Also, sinusoidal modelling is more suitable for harmonic sounds (e.g. musical instruments) where the broadband, noisy components play a minor role in the overall spectrum. Nemisindo (PM3), being a more recent and advanced procedural model, scores higher than the other two; and additive synthesis, by synthesising sounds as a sum of five core elements (modal impact, noisy impact, chirped impact, band-limited noise, equalised noise), is rated even higher. Both WaveGAN and HiFi-WaveGAN score as high as the reference. This result shows that the two neural synthesis models manage to capture all the important details of the training data, and allow the generated samples to reach synthesis quality and realism comparable to recorded audio.

5. DISCUSSION

We presented a first attempt at neural synthesis of footsteps - one of the most common and challenging sound effects in sound design. In this work, two GANs architectures were implemented: a standard conditional WaveGAN, and a hybrid consisting of a conditional WaveGAN generator and a conditional HiFi-GAN discriminator. The hybrid architecture improves the audio quality of generated samples and better approximates the training data distribution. Differently from what is commonly suggested in the literature, upsampling with zero-stuffing gave us better results than nearest neighbour interpolation. The same is true for linear or cubic interpolation.

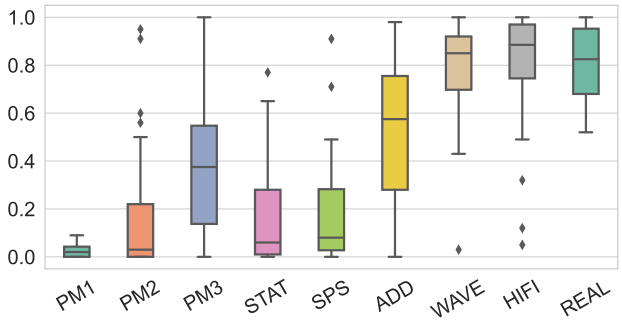

Fig. 5: Results of the subjective evaluation.

Both objective and subjective evaluation of results were conducted. It is not common for neural synthesis methods to be compared with "traditional" synthesis algorithms in a listening test, which makes it impossible to establish whether a neural synthesis method can actually reach state-of-the-art results. In this work, we compared the two architectures with 6 other methods as well as recorded samples. The two architectures reached "realism" ratings as high as real sounds. Although, from informal listening tests, we would have expected a greater difference in the ratings. In fact, samples synthesised by WaveGAN are affected by a perceivable amount of background noise and distortion, while the opposite is true for HiFi-WaveGAN [11].

This opens questions about the definition of realism of sound effects, to what extent audio quality correlates with perceived realism, and what other aspects play a significant role in such judgements. Following work will focus on increasing the degrees of control and diversity of synthesised samples, while retaining the audio quality.

6. ACKNOWLEDGEMENTS

Funded by UKRI and EPSRC as part of the "UKRI CDT in Artificial Intelligence and Music", under grant EP/S022694/1.

[10] http://webprojects.eecs.qmul.ac.uk/mc309/FootEval/test.html?url=tests/ape_footsteps.xml

[11] https://mcomunita.github.io/hifi-wavegan-footsteps_page/

7. REFERENCES

[1] A. Farnell, *Designing sound*, Mit Press, 2010.

[2] X. Li, R.J. Logan, et al., "Perception of acoustic source characteristics: Walking sounds," *J. of the Acoustical Society of America*, vol. 90, no. 6, 1991.

[3] B. Giordano and R. Bresin, "Walking and playing: What's the origin of emotional expressiveness in music," in *Int. Conf. Music Perception and Cognition*, 2006.

[4] R.E. Pastore, J.D. Flint, et al., "Auditory event perception: The source—perception loop for posture in human gait," *Perception & Psychophysics*, vol. 70, no. 1, 2008.

[5] K. Makela, J. Hakulinen, et al., "The use of walking sounds in supporting awareness," in *Int. Conf. on Auditory Display*, 2003.

[6] R. Nordahl, S. Serafin, et al., "Sound synthesis and evaluation of interactive footsteps for virtual reality applications," in *IEEE Virtual Reality Conference*, 2010.

[7] R. Bresin and S. Dahl, "Experiments on gestures: walking, running, and hitting," *The sounding object*, 2003.

[8] L. Turchet, D. Moffat, et al., "What do your footsteps sound like? an investigation on interactive footstep sounds adjustment," *Applied Acoustics*, vol. 111, 2016.

[9] P.R. Cook, "Modeling bill's gait: Analysis and parametric synthesis of walking sounds," in *AES Conf.: Virtual, synthetic, and entertainment audio*, 2002.

[10] P.R. Cook, "Physically informed sonic modeling (phism): Synthesis of percussive sounds," *Computer Music J.*, vol. 21, no. 3, 1997.

[11] F. Fontana and R. Bresin, "Physics-based sound synthesis and control: crushing, walking and running by crumpling sounds," in *Colloquium on Musical Informatics*, 2003.

[12] A. DeWitt and R. Bresin, "Sound design for affective interaction," in *Int. Conf. on Affective Computing and Intelligent Interaction*, 2007.

[13] A.J. Farnell, "Marching onwards: procedural synthetic footsteps for video games and animation," in *Pure Data Conv.*, 2007.

[14] L. Turchet, "Footstep sounds synthesis: design, implementation, and evaluation of foot–floor interactions, surface materials, shoe types, and walkers' features," *Applied Acoustics*, vol. 107, 2016.

[15] C. Donahue, J. McAuley, et al., "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[16] J. Nistal, S. Lattner, et al., "Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," *arXiv preprint arXiv:2008.12073*, 2020.

[17] C. Aouameur, P. Esling, et al., "Neural drum machine: An interactive system for real-time synthesis of drum sounds," *arXiv preprint arXiv:1907.02637*, 2019.

[18] A. Ramires, P. Chandna, et al., "Neural percussive synthesis parameterised by high-level timbral features," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020.

[19] R.A. Barahona and S. Pauletto, "Synthesising knocking sound effects using conditional wavegan," in *Sound and Music Computing Conf.*, 2020.

[20] C.Y. Lee, A. Toffy, et al., "Conditional wavegan," *arXiv preprint arXiv:1809.10636*, 2018.

[21] K. Kumar, R. Kumar, et al., "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[22] R. Yamamoto, E. Song, et al., "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020.

[23] W. Jang, D. Lim, et al., "Universal melgan: A robust neural vocoder for high-fidelity waveform generation in multiple domains," *arXiv preprint arXiv:2011.09631*, 2020.

[24] E. Song, R. Yamamoto, et al., "Improved parallel wavegan vocoder with perceptually weighted spectrogram loss," in *IEEE Spoken Language Technology Workshop*, 2021.

[25] J. Yang, J. Lee, et al., "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.

[26] J. Kong, J. Kim, et al., "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[27] J. You, D. Kim, et al., "Gan vocoder: Multi-resolution discriminator is all you need," *arXiv preprint arXiv:2103.05236*, 2021.

[28] I. Gulrajani, F. Ahmed, et al., "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[29] X. Mao, Q. Li, et al., "Least squares generative adversarial networks," in *IEEE Int. Conf. on Computer Cision*, 2017.

[30] I. Goodfellow, J. Pouget-Abadie, et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[31] A. Gretton, K. M Borgwardt, et al., "A kernel two-sample test," *J. of Machine Learning Research*, vol. 13, no. 1, 2012.

[32] Joseph Turian, Jordie Shier, George Tzanetakis, Kirk McNally, and Max Henry, "One billion audio sounds from gpu-enabled modular synthesis," *arXiv preprint arXiv:2104.12922*, 2021.

[33] J. Cramer, H. Wu, et al., "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019.

[34] D. Moffat and J.D. Reiss, "Perceptual evaluation of synthesized sound effects," *ACM Trans. on Applied Perception*, vol. 15, no. 2, 2018.

[35] N. Jillings, B. De Man, et al., "Web audio evaluation tool: A browser-based listening test environment," 2015.

[36] B. De Man and J.D. Reiss, "Ape: Audio perceptual evaluation toolbox for matlab," in *AES Conv. 136*, 2014.

[37] P. Bahadoran, A. Benito, et al., "Fxive: A web platform for procedural sound synthesis," in *AES Conv. 144*, 2018.

[38] X. Amatriain, J. Bonada, et al., "Spectral processing," *DAFX-Digital Audio Effects*, 2002.

[39] J.H. McDermott and E.P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, 2011.

[40] C. Verron, M. Aramaki, et al., "A 3-d immersive synthesizer for environmental sounds," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 6, 2009.