Is Quality Enough? Integrating Energy Consumption in a Large-Scale Evaluation of Neural Audio Synthesis Models

Constance Douwes, Giovanni Bindi, Antoine Caillon, Philippe Esling, Jean-Pierre Briot

IS QUALITY ENOUGH? INTEGRATING ENERGY CONSUMPTION IN A LARGE-SCALE EVALUATION OF NEURAL AUDIO SYNTHESIS MODELS

Constance Douwes[⋆] *Giovanni Bindi*[⋆] *Antoine Caillon* [⋆] *Philippe Esling*[⋆] *Jean-Pierre Briot*[†‡]

[⋆]IRCAM, Sorbonne Université, CNRS, UMR 9912 F-75004 Paris, France
[†]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[‡]UNIRIO, Rio de Janeiro, RJ 22290-250, Brazil

ABSTRACT

Deep learning models are now core components of modern audio synthesis, and their use has increased significantly in recent years, leading to highly accurate systems for multiple tasks. However, this quest for quality comes at a tremendous computational cost, which incurs vast energy consumption and greenhouse gas emissions. At the heart of this problem are the standardized evaluation metrics used by the scientific community to compare various contributions. In this paper, we suggest relying on a multi-objective metric based on Pareto optimality, which considers equally the accuracy and energy consumption of a model. By applying our measure to the current state-of-the-art in generative audio models, we show that it can drastically change the significance of the results. We hope to raise awareness on the need to more systematically investigate the energy efficiency of high-quality models, in order to place computational costs at the center of deep learning research priorities.

Index Terms— Neural audio synthesis, Multi-objective evaluation, Energy footprint

1. INTRODUCTION

Deep learning currently holds most of the state-of-the-art results in a wide range of application fields. Despite major advances in manufacturing energy-efficient hardware, the computational cost of deep learning remains humongous and continuously rising [1, 2], significantly contributing to global warming [3, 4]. Although part of the current research effort is concerned with the true cost of deep models [5], taking into account the environmental cost of these models is mostly overlooked against the never-ending quest for accuracy. This aspect, while novel in deep learning research, already emerges in some communities such as natural language processing (NLP) [6]. Nonetheless, those energy footprint measurements have never been addressed in the field of generative neural networks for audio generation. This is an essential question as the use of deep generative models for audio is becoming more and more frequent.

Generating raw audio waveform using neural networks is not a straightforward endeavor. It requires handling high dimensional structures with long-term dependencies, which usually leads to computationally and energetically costly models. Despite these significant limitations, we now count a large variety of deep models that can produce high-quality raw audio [7, 8] with each having its own set of advantages and restrictions. However, measuring the precise energy consumption of a given model remains a complex task [9], which remains mostly neglected both in terms of training and sample generation. The disparities in power consumption between various neural audio synthesis architectures are significant and, therefore, should be integrated into the evaluation process.

In this paper, we focus on neural vocoders, which are a class of generative models used for speech generation based on mel-spectrograms conditioning. We train a wide variety of models from major families of deep generative models and compute the energy footprint of sample generation. We perform audio perceptual tests on the converged networks to obtain precise measures of their quality. We propose the novel use of a *multi-objective* criterion based on Pareto optimality to evaluate the trade-off between energy and quality, and assess our methodology on six state-of-the-art models. We implement three distinct configurations for each of these models, ranging from lighter to larger architectures, to perform an in-depth analysis of the corresponding computational cost. We show that lighter models can produce high-quality samples while maintaining more sustainable energy consumption than bigger models. Through the adoption of this proposed energy-quality efficiency framework, we aim to endow future research proposals with more complete evaluation and, consequently, put the energy footprint on a first-grade level of importance. To summarize, our key contributions are:

- Perform a large-scale evaluation of the energetic cost of deep neural vocoders depending on the size of their architecture.
- Propose a new methodology for incorporating both the energy footprint and generation quality in the evaluation process, by relying on a multi-objective approach.[1]

2. ENERGY EFFICIENCY MEASURES

Measuring the exact energy consumption of any type of computer software is an extremely challenging task, as it is usually intertwined with other processes (e.g. cache hits and misses, memory accesses) [10]. In the context of deep learning, we can divide the energy consumption of a given model between two different modes: the amount of energy required to *train* the model until convergence, and the amount of energy required by the model for a single *inference* (generating a given sample in the case of audio synthesis). To approximate the energy and carbon cost of training models, [6] sampled GPU, CPU, and DRAM power consumption using the NVIDIA System Management Interface and Intel's Running Average Power Limit. Recently, [11] proposed an online tool called the *Machine Learning Impact Calculator*, which estimates carbon emissions produced while training deep leaning models according to the overall time, hardware and geographic position. In the same spirit, [12] developed an open-source Python package called *Carbontracker*, which tracks the energy consumption of a single epoch and predicts the entire training cost.

[1]All of our source code is available in our supporting webpage at https://github.com/ConstanceDws/neural-audio-energy

Another common measure of energy efficiency is the total number of model parameters, as it is easy to determine and partly correlated with computational complexity. Unlike aforementioned measures, this metric provides the advantage of being hardware- and location-independent. Nonetheless, the number of parameters does not accurately reflect power consumption as some operations consume more than others. Hence, the best way to alleviate this issue is to consider the number of Floating Point Operations (FLOPs) of a model [13]. Although this computation is not straightforward as it depends on various hyperparameters of the model (e.g., input size, kernel size, stride, padding, bias), several python packages provide approximations of these calculations, such as the profiler from *Deepspeed*[2] that also computes per-layer values such has the number of MACs (Multiply–Accumulate operation) and the latency of a forward pass. Other methods exist to account for the on-device consumption, like the *pyJoules*[3] python package that monitors GPU, CPU and DRAM energy usage.

3. GENERATIVE MODELS FOR AUDIO

Generative models are a flourishing class of unsupervised learning approaches that deals with learning to generate novel data based on the observation of existing examples. Several methods exist, which we can split in five categories: *auto-regressive* models, *Variational Auto-Encoders (VAE)* [14], *Generative Adversarial Networks (GAN)* [15], *flow-based* models [16] and *diffusion* models [17].

Auto-regressive models attempt to model examples by assuming that a given output element is only related to prior values. Following this formulation, *WaveNet* [18] and *SampleRNN* [19] have tackled direct waveform learning through end-to-end generation. Unfortunately, these methods are based on heavy architectures whose computational complexity incur large energy consumption, especially for inference. Furthermore, these also provide almost no direct control on the generative process. Some approaches use *VAE* [20] that learn a latent space providing a low-dimensional representation of the data while remaining rather simple and fast to train. However, the generated samples tend to be slightly blurry compared to recent adversarial networks, such as *WaveGan* [21] or *GANSynth* [22]. These show impressive reconstruction abilities but usually lack latent expressivity and are difficult to optimize due to unstable training dynamics. The recently proposed *Normalizing Flows (NF)*, used in *WaveFlow* [23] or *FloWaveNet* [24] allow to model highly complex distributions and already yield remarkable results. However, *NF* do not provide any dimensional reduction, thus taking considerable amounts of time to train. Finally, *diffusion* models define a Markov chain of diffusion steps, where the data is increasingly corrupted by noise and the aim is to learn the reverse denoising diffusion process. As a relatively recent class of models, they are yet to be extensively studied for audio generation. However, the seminal works are those of [25] and [26], in which a denoising diffusion model is learned through dilated convolutional architectures.

Despite the successes provided by these models, they still incur large computational costs, only handled by modern accelerators (such as GPUs or TPUs). Moreover, the plurality of models and training time needed for them to converge questions the real effectiveness with regards to the quality of the generated results, and what could be the best compromise in terms of energetic cost.

[2]https://www.deepspeed.ai/tutorials/flops-profiler/

[3]https://github.com/powerapi-ng/pyJoules

4. MUTLI-OBJECTIVE CRITERIA

4.1. Methodolody

Increasing the size of deep models often improve their quality at the expense of large energy costs. As these objectives seem to be conflicting, we propose to introduce the use of multi-objective evaluation criteria, also called Pareto optimization. Formally, we consider a multi-objective optimization problem as

$$\min_{x \in X} \ \{f_1(x), f_2(x), \ldots, f_k(x)\} \tag{1}$$

where k is the number of objective functions and x the feasible solutions. A feasible solution $x_a \in X$ is said to *dominate* another feasible solution $x_b \in X$, notated $x_a \prec x_b$, if :

- $\forall i \in \{1, \ldots, k\}, f_i(x_a) \leq f_i(x_b)$
- $\exists j \in \{1, \ldots, k\}, f_j(x_a) < f_j(x_b)$

A solution $x^* \in X$ is a *Pareto optimal* solution if there are no $\hat{x}$ such that $\hat{x} \prec x^*$. The set of all these optimal solutions is called the *Pareto front*; an illustration is shown in Figure 1.

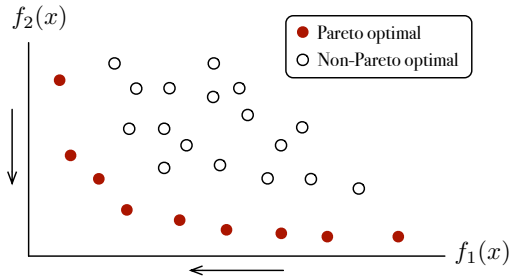

Fig. 1. Example of a Pareto front where we seek to minimize two functions f_1 and f_2. Red points are Pareto optimal solutions while white ones are non optimal.

Now, consider two generative models A and B. If A and B have the same sound quality, but A consumes less than B (so A *dominates* B), then A is Pareto-optimal. Conversely, if A and B have the same energy footprint, but B provides higher quality, then B is Pareto-optimal. Hence, we aim to find the set of all Pareto optimal models to form a Pareto front and measure the dependency of two main factors: quality and energy.

4.2. Experiments

4.2.1. Models

In order to account for energy costs inside the evaluation of neural audio synthesis models, we consider six state-of-the-art approaches belonging to the three major families of generative models discussed earlier: GAN, flow-based and diffusion models. Within these groups, we respectively consider *MelGAN* [27] and *HiFi-GAN* [25], *WaveGlow* [28] and *WaveFlow* [23] and, finally, *WaveGrad* [26] and *DiffWave* [25]. The choice of these models was dictated by their impact in the community and their recent introduction. They are also part of the neural vocoder subfamily, which have the particularity of being conditioned on mel-spectrogams, and are therefore widely used in speech synthesis. At the time of writing this article, all these models have been introduced within the previous three years. Our choice to exclude autoregressive models as well as VAE is respectively due to their prohibitively long inference time, and their lower

quality in the audio domain. Therefore, they are currently rarely used in applications compared to other families. Due to space constraints, we do not provide the per-layer description of the models, but for more details, please refer to the original papers.

For each of these six models, we consider three different configurations : a small (S), medium (M) and large (L) configuration. Most of these are architectures already proposed in the original papers, while others are our own suggestion following the same logic present in the considered papers (e.g. channel or depth variation). For more information, all configurations are available in the code.

4.2.2. Dataset

We train and evaluate all models on one of the reference datasets in speech generation, namely *LJSpeech* [29]. This dataset is composed of 13,100 audio samples of various speakers, ranging from 1 to 10 seconds at 22050 Hz, for a total of around 24 hours. In our experiments, after downsampling the data to 16000 Hz, we apply the preprocessing strategy which is the most commonly used across the tested models (WaveGlow, MelGan and WaveFlow). Hence, we keep only the first $N = 2^{14}$ samples from each clip and then extract an 80-bands mel-spectrogram $\mathbf{s}$ from this audio with a FFT of size 2048 and a hop size of 256. We then perform *min-max* normalization on each spectrogram. Finally, we split the data between *training* (80%) and *testing* (20%) sets.

4.2.3. Training

All models are trained 120 hours a single NVIDIA RTX A5000 GPU, where batch-size is scaled automatically to maximize the GPU memory usage in order to enhance parallelization. Hence, all models are evaluated on the same amount of energy consumption, estimated at $300\text{W} \times 120\text{h} = 36\text{kWh}$ per model by the ML impact calculator [11]. Furthermore, the choice of using a single GPU allows both to simplify the energy consumption cost, and also represent typical computational capacities of public research institutions. For all models, we use the ADAM [30] optimizer and rely on the respective learning rate of the tested models in their original implementations.

4.3. Results

4.3.1. Synthesis quality

In order to provide estimations of perceptual audio quality, we performed a human-based perceptual quality evaluation. Participants were asked to rate the naturalness of sounds, by grading each sample between 1 ("*bad*") and 5 ("*perfect*"). In this analysis we include both the ground truth data (from the test set) and each model reconstruction. A total of 41 participants undertook the complete test, the majority of whom were audio professionals. We present the results of this MOS (Mean Opinion Score) evaluation in Table 1, alongside the spectral distance as defined in [31], denoted as $\mathcal{D}_{\text{STFT}}$.

As we can see, there are large discrepancies between the MOS and $\mathcal{D}_{\text{STFT}}$ when evaluating the generation quality. This underlines the need for human-based evaluation, as slight reconstruction artefacts can have a large perceptual impact. Indeed, although the DiffWave model have the lowest $\mathcal{D}_{\text{STFT}}$ reconstruction error, it exhibits quite low MOS scores. Overall, the WaveGrad and HiFi-GAN models (all configuration included) have largely higher MOS scores than other models. On the other hand, WaveGlow has rather poor MOS results, which could be explained by the lack of sufficient training time when compared to the original paper. Across all models, it appears that increasing the size of the architecture consistently increases the quality of the corresponding generations, which is coherent with the current trend in the scientific literature and was expected at this point of the study.

Model	**MOS**	$\mathcal{D}_{\text{STFT}}$
Ground truth	4.34 ($\pm$0.005)	0
MelGAN* S	1.37 ($\pm$0.004)	0.1496
MelGAN M	2.12 ($\pm$0.007)	0.1199
MelGAN* L	2.22 ($\pm$0.007)	0.1146
HiFi-GAN S	3.90 ($\pm$0.007)	0.0804
HiFi-GAN M	3.59 ($\pm$0.008)	0.0791
HiFi-GAN L	4.12 ($\pm$0.007)	0.0712
WaveGrad* S	3.24 ($\pm$0.007)	0.0758
WaveGrad M	3.66 ($\pm$0.008)	0.0736
WaveGrad L	3.59 ($\pm$0.007)	0.0709
DiffWave S	1.89 ($\pm$0.006)	0.0838
DiffWave M	2.18 ($\pm$0.006)	0.0725
DiffWave L	2.41 ($\pm$0.007)	0.0698
WaveFlow S	1.50 ($\pm$0.005)	0.1192
WaveFlow M	2.44 ($\pm$0.010)	0.1059
WaveFlow L	2.77 ($\pm$0.008)	0.1180
WaveGlow* S	1.07 ($\pm$0.002)	0.1518
WaveGlow M	1.52 ($\pm$0.004)	0.1177
WaveGlow L	1.80 ($\pm$0.006)	0.1136

Table 1. Perceptual (Mean Opinion Score) and reconstruction ($\mathcal{D}_{\text{STFT}}$) qualities of neural vocoders conditioned on mel-spectrogram. (*) indicate configurations that we suggest in addition to those of the original papers.

4.3.2. Energy efficiency

In order to better understand the tradeoff between increased size (and quality) of the models and their corresponding energetic impact, we compute the number of parameters as well as the number of floating point operations per second of generated content. We then record the energy for our models to generate 10 clips of 10 seconds of raw audio on a single NVIDIA RTX A5000 GPU using the *pyJoules* package. For flows models, we remove the weight normalization layers as they slow down the audio generation without impacting the corresponding inference quality. We summarize all of these energy footprint metrics in Table 2.

By analyzing the in-use energy footprint as well as the number of GFLOPs, we notice extremely large differences between the various types of generative models. GAN tend to be really efficient, whereas diffusion models have an inference energy cost around 100 times larger.

4.4. Pareto analysis

In order to fully understand the tradeoff between quality and energy impact, we display our proposed multi-objective analysis in Figure 2. We separate this analysis between the hardware-agnostic metric GFLOPs (left) and the inference energy costs (right). In both cases, we plot different models depending on their corresponding MOS evaluation. We depict the optimal Pareto models, which are circled in red. A first noticeable result of this study is that our multi-objective analysis produces coherent results, since we can directly find optimal models with low energy consumption but high quality

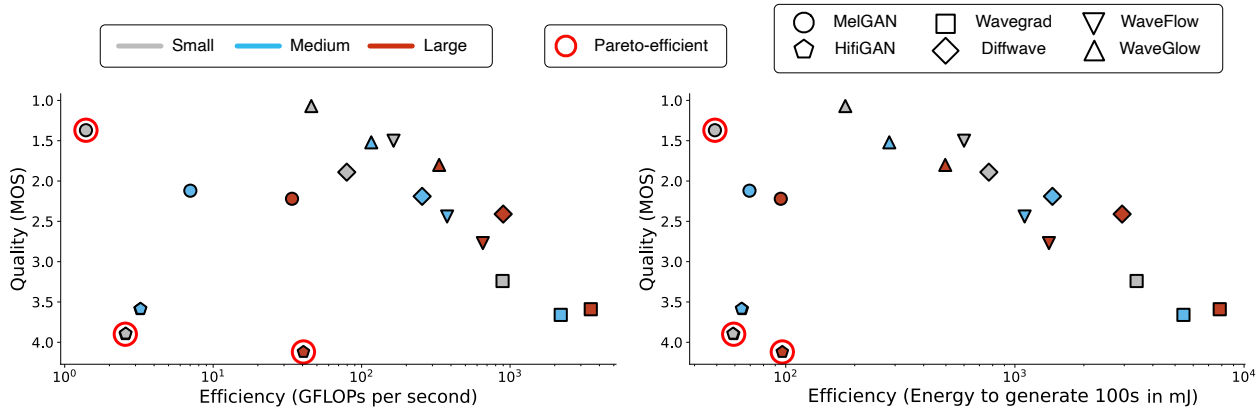

Fig. 2. Representation of Pareto Frontier for energy vs quality. The objective is to maximize the quality (MOS) and minimize the number of GLFOPs (left) and the energy cost of inference (right).

Model	# Param	# GFLOPs	E_{gen} (mJ)
MelGAN* S	1.03M	1.39	49.08
MelGAN M	4.27M	7.02	69.46
MelGAN* L	18.21M	33.98	95.14
HiFi-GAN S	0.928M	2.56	59.28
HiFi-GAN M	1.46M	3.22	64.59
HiFi-GAN L	13.94M	40.57	96.61
WaveGrad* S	4.18M	890.44	3398.11
WaveGrad M	17.12M	3498.64	5439.46
WaveGrad* L	33.91M	8522.08	7833.62
DiffWave S	1.23M	79.43	769.82
DiffWave M	2.62M	255.78	1458.17
DiffWave L	6.89M	899.61	2937.61
WaveFlow S	5.95M	852.85	599.95
WaveFlow M	12.86M	3419.39	1102.88
WaveFlow L	22.39M	6063.60	1408.01
WaveGlow* S	17.56M	45.84	181.98
WaveGlow M	34.76M	116.21	496.72
WaveGlow L	87.73M	333.04	283.07

Table 2. Comparison of computation and energy footprints of various generative models for speech synthesis conditioned on mel-spectrogram. (*) indicate configurations that we suggest in addition to those of the original papers.

score. A second result of this analysis is that only few models are included inside the Pareto front, with the vast majority of the models being dominated in both aspects simultaneously. This means that our proposed multi-objective approach allows to efficiently discriminate between different models on both their audio quality and energy impact. A third key component of this study is that the hardware agnostic metric and the GPU metric are consistent, with slight shifts showing that energy and GFLOPs are not linearly correlated.

If we take a closer look at the per-model inference tradeoff (by considering only the same symbol), we can see that it also forms what we can call sub-Pareto front, from lighter to larger configurations (from light green to dark blue), but it's only when we look at the big picture that it reveals disparities of generative models architectures and configurations. Hence, our analysis allows to raise attention and provide new keys for researchers to evaluate their models within the context of a multi-objective analysis rather than comparing quality and efficiency separately. Furthermore, we believe that it is through this research effort that we will be able to achieve a more sustainable computing.

5. CONCLUSIONS

In this paper, we proposed a large-scale evaluation of neural vocoders while integrating their energy footprint for training and inference. We relied on six stat-of-the-art models and evaluate their quality according to three different configurations from lighter to larger architectures. Then, we proposed a multi-objective analysis of both quality from human-based evaluation (MOS) and energy consumption. Within this framework, we showed that this energy footprint must be linked to the model perceptual quality and that, small models can perform better than larger and more costly models. We believe this is the first attempt to integrate both energy consumption and quality in neural audio synthesis models, taking a step forward against blind evaluations that only take into account audio quality. This, in the future, can become increasingly important, since lightweight models are fundamental for real-time embedded systems. It should be noted that our approach is generic and could be applied to any type of model or input data.

CO2 Emission Related to our Experiments Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.7 kgCO_2eq/kWh. A cumulative of 2160 hours of computation was performed on hardware of type RTX A5000 (TDP of 300W). Total emissions are estimated to be 453.6 kgCO_2eq.

6. REFERENCES

[1] Amodei Dario and Hernandez Danny, "Ai and compute," 2018.

[2] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute trends across three eras of machine learning," 2022.

[3] Emma Strubell, Ananya Ganesh, and Andrew Mccallum, "Energy and Policy Considerations for Modern Deep Learning Research," *Aaai*, 2020.

[4] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-hsin S Lee, Gu-yeon Wei, David Brooks, and Carole-jean Wu,

"Chasing Carbon : The Elusive Environmental Footprint of Computing," 2019.

[5] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," 2020.

[6] Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and policy considerations for deep learning in NLP," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, , no. 1, pp. 3645–3650, 2020.

[7] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[8] Karan Goel, Albert Gu, Chris Donahue, and Christopher Re, "It's raw! Audio generation with state-space models," in *Proceedings of the 39th International Conference on Machine Learning*. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 7616–7633, PMLR.

[9] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn, "Estimation of energy consumption in machine learning," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 75–88, 2019.

[10] Luca Ardito, Riccardo Coppola, Maurizio Morisio, and Marco Torchiano, "Methodological Guidelines for Measuring Energy Consumption of Software Applications," *Scientific Programming*, vol. 2019, pp. 5284645, 2019.

[11] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres, "Quantifying the Carbon Emissions of Machine Learning," 2019.

[12] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems, 2020.

[13] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni, "Green AI," pp. 1–12, 2019.

[14] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, , no. Ml, pp. 1–14, 2014.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, vol. 3.

[16] Danilo Jimenez Rezende and Shakir Mohamed, "Variational inference with normalizing flows," *32nd International Conference on Machine Learning, ICML 2015*, vol. 2, pp. 1530–1538, 2015.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[18] Aaron Van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," pp. 1–15, 2016.

[19] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–11, 2016.

[20] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 1771–1780, 2017.

[21] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–16, 2019.

[22] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts, "GANSynth: Adversarial neural audio synthesis," *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–17, 2019.

[23] Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song, "WaveFlow: A compact flow-based model for raw audio," in *Proceedings of the 37th International Conference on Machine Learning*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7706–7716, PMLR.

[24] Sungwon Kim, Sang Gil Lee, Jongyoon Song, Jaehyeon Kim, and Sungroh Yoon, "FloWaveNet: A generative flow for raw audio," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 5852–5860, 2019.

[25] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *International Conference on Learning Representations*, 2021.

[26] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2021.

[27] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis.," in *NeurIPS*, 2019, pp. 14881–14892.

[28] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 3617–3621, 2019.

[29] Keith Ito and Linda Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[30] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[31] Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach, "Sing: Symbol-to-instrument neural generator," *Advances in Neural Information Processing Systems*, vol. 2018-Decem, no. Nips, pp. 9041–9051, 2018.