Article Review
Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, Mohammad Norouzi, 2017

Student: Matteo MARENGO - ENS Paris-Saclay - MVA

Teacher: Emmanuel BACRY - École Polytechnique

March 26, 2024

1 Abstract

This review of the paper "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders" will study and explain many topics related to it. First, the paper is introduced, coming back on some theoretical definitions and history of audio synthesis. Then, the methodology of the paper is explained by highlighting the two main contributions that are; **a WaveNet-style autoencoder that can learns temporal hidden codes without external conditioning and NSynth that is a large dataset of musical notes**. Then the results of this approach will be presented. In the second part, I will present some of my own experiments that I made to understand and to explore results from a more personal point of view. Finally, a critical assessment of the paper will be made, the avenues of improvement and outlooks discussed before concluding on this review.

Keywords: Audio Synthesis, WaveNet, TTS, Autoregressive models, spectral autoencoders, CQT spectrogram

2 Introduction

This review is part of the evaluation for the course "Audio Signal Processing" given by Pr. Emmanuel BACRY in the MSc MVA *(Mathématiques, Vision, Apprentissage)* at ENS Paris-Saclay. I have chosen to study the paper **"Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders"** [13] by Engel et al. as I found that the final application of this paper is relevant and impactful. Furthermore, I have previously worked a lot with Autoencoders applied to the images and I wanted to delve deeper in Autoencoders related to signal processing. This review will be like this; first a summary of the paper, the problem it addresses, its relation to the State-of-the-art, its methodology, and results. Then, a part will be dedicated to my own understanding and experiments on the paper. Finally, a part will be dedicated to its critical assessment to end-up with the outlooks and conclusions.

3 Paper summary

3.1 Problem definition

The paper "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders" [13] by Engel et al. strategically situates itself within the burgeoning field of generative audio synthesis by introducing **a novel autoencoder architecture and the NSynth dataset**, addressing key limitations in existing methodologies. The main challenge addressed is Audio Synthesis and more precisely Audio Generation algorithms. **It has always been a difficult task as audio has a very high temporal resolution**, at least 16000 samples per second and high-temporal long-term dependency. Indeed, compared to the field of images, audio lacks a solid reference dataset that could enhance the performance, research and relevant benchmarking between different proposed architectures. Regarding the novel autoencoder structure, the one proposed removes the need for external conditioning when modeling long-term signals.

3.2 Standard related methods

To synthesize audio has always been a challenge. It is known as **vocoders for text-to-speech systems and synthesizers for music**. Vocoders were born in 1939 [11] where voice was synthesized by applying a series of frequency filters to a source sound. Later, in 1960 formant synthesis was introduced focusing on the role of resonant frequencies in speech production [14]. The development of Linear Predictive Coding (LPC) in 1971 [1] significantly advanced TTS as LPC models the vocal tract as a series of digital filter coefficients.

Regarding the field of music generation, one of the first electronic synthesizers was developed in 1959 [25] allowing first electronic tracks to be created. In 1964, the Moog synthesizer was introduced [24]. This invention was able to produce a wide range of sounds through voltage-controlled oscillators, filters, and amplifiers. In 1973, **Frequency -Modulation (FM)** was introduced [6] and it led to the development of the Yamaha DX7 synthesizer. It would allow creating complex sounds with rich timbres. Finally, recent advances in music synthesis and TTS has seen the days with the rise of digital methods. In music synthesis for example, Physical modeling Synthesis in 1992 [32] allows stimulating the physical properties of musical instruments to generate

sound. In audio synthesis more broadly, seminal works such as van den Oord et al.'s original WaveNet (2016) [33] and Mehri et al.'s SampleRNN (2016) [23] demonstrated significant advancements.

This work is distinct in its approach to overcoming the challenges of **long-term dependency modeling in audio signals**, a limitation previously noted in but relied heavily on external conditioning to capture longer-term structures, a dependency Engel et al. seek to eliminate with their autoencoder design that infers temporal embeddings directly from raw audio.

Further distinguishing their contribution, Engel et al. develop **NSynth**, a comprehensive dataset for musical notes, addressing the scarcity of large-scale, high-quality datasets for audio synthesis. It has always been a challenge in the field, as highlighted by prior efforts such as the Million Song Dataset by Bertin-Mahieux et al. (2011) [2] and smaller scale datasets used by Goto et al. (2003) [15]. In contrast to these earlier works, NSynth provides a much-needed platform for consistent and comparative evaluation of generative audio models, similar to the role of ImageNet in the vision domain as described by Deng et al. (2009) [8].

By directly comparing their method to baseline approach; traditional spectral autoencoders, Engel et al. showcase the superiority of their model in capturing the intricate dynamics and timbres of musical notes. This comparison not only underscores the limitations of existing models' reliance on spectral features and external conditioning but also demonstrates the efficacy of their temporal embedding approach in generating more expressive and realistic sounds.

3.3 Main contributions

Engel et al.'s study makes significant contributions to neural audio synthesis. They introduce a powerful model that can capture the long-term structure of music without external help, and they present NSynth, a vast dataset that sets new standards in the field. Their work allows for the creation of musical notes with intricate dynamics and variations in timbre and pitch. Timbre describes the unique quality or "color" of a sound that lets us tell different sounds apart, even if they're the same pitch and loudness. Pitch, meanwhile, is about whether a note sounds high or low and depends on the sound wave's frequency. Essentially, timbre is about the sound's character, and pitch is about its frequency. The NSynth dataset represents a significant step forward for research in synthesizing musical sounds.

3.4 Methodology

3.4.1 Spectral AutoEncoder

The baseline autoencoder is composed of convolutional structures where both the encoder and the decoder are each 10 layers deep with 2×2 strides and 4×4 kernels. At the end, a single 1984 dimensional hidden vector is created. It is inspired by models in computer vision [10] [19]. The training was made on the log magnitude of the power spectra.

3.4.2 WaveNet AutoEncoder

The autoencoder model is a significant advancement over previous approaches in generative audio modeling. Unlike the original WaveNet and SampleRNN, which require external conditioning for modeling long-term structure, this autoencoder learns temporal embeddings directly from the raw audio waveform. The methodology consists of a WaveNet-like encoder that captures hidden embeddings distributed in time, and a WaveNet decoder that uses these embeddings to reconstruct the original audio without the need for external inputs. **This approach allows the model to encode and reproduce longer sequences of audio with high fidelity, addressing a key limitation in prior models.**

First, it is important to fully understand what WaveNet [33] is about. WaveNet operates directly on the raw audio waveform. As a generative model, the next sample of the audio is predicted from a fixed-size input

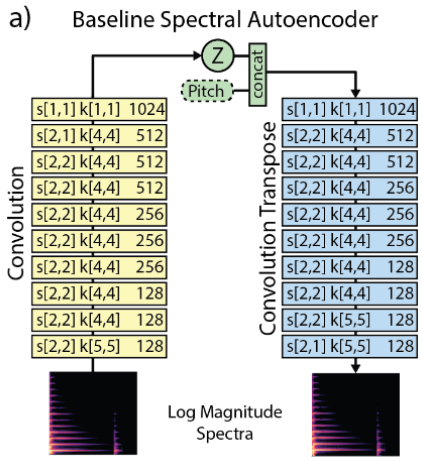

Figure 1: Baseline spectral autoencoder. Retrieved from [13].

of prior sample values. The joint probability of the audio x is therefore:

$$p(x) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{N-1})$$

Each audio sample is conditioned on the samples of the previous time steps. It is indeed an autoregressive (AR) model [18] [3]. As a reminder, the AR(p) process is defined like this:

$$x[n] = -\sum_{i=1}^{p} a_i x[n-i] + b[n]$$

where:

- p: order of the model
- $a_1, \ldots, a_p$: AR coefficients
- $b[n]$: white noise (often called innovation)

As in PixelCNNS [34] the conditional probability is modeled by a stack of convolutional layers. They form causal convolution. An other aspect to understand in WaveNet are dilated causal convolutions as seen in Fig 2. This dilated convolution is a convolution where certain input values are skipped with a certain step.

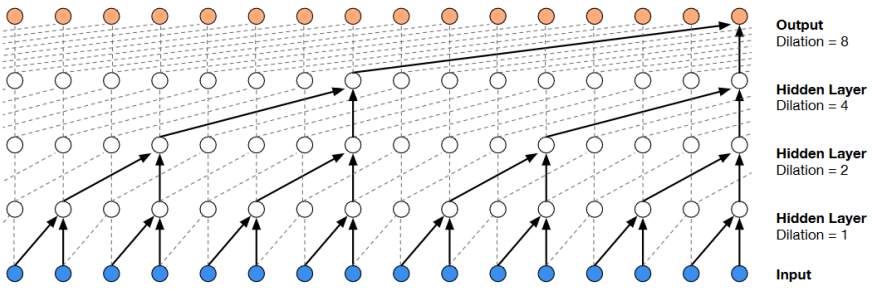

Figure 2: A stack of dilated causal convolutional layers. Retrieved from [33].

Therefore, in the studied architecture, the encoder employs a **30-layer residual network of dilated convolutions followed by 1x1 convolutions**, generating a sequence of hidden codes that represent the audio temporally. The decoder is conditioned on these temporal embeddings. **The autoencoder removes**

the need for that external conditioning, it takes the embedding as conditioning. Now the joint probability is:

$$p(x) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{N-1}, f(x))$$

The model predicts output samples with a softmax over quantized mu-law encoded values. This architecture is designed to scale with the input size, enabling it to handle longer audio sequences effectively. The wavenet architecture is shown in Fig 4 next to the wavenet autoencoder that is being studied in Fig 3.

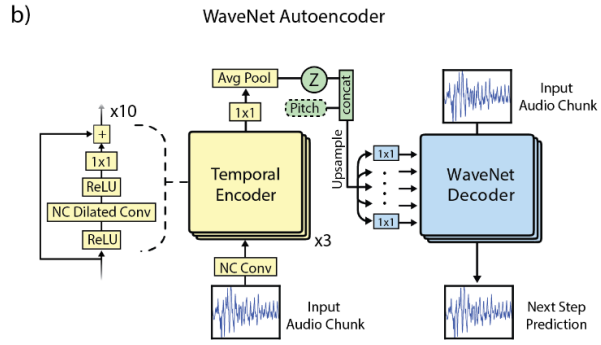

Figure 3: WaveNet autoencoder. Retrieved from [13].

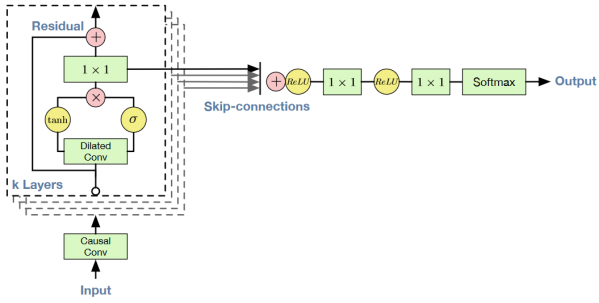

Figure 4: Architecture of the WaveNet. Retrieved from [33].

3.4.3 NSynth

Recognizing the lack of large-scale, high-quality datasets for audio synthesis comparable to those in image processing (e.g., ImageNet, MNIST), the authors introduce NSynth. NSynth is a dataset consisting of approximately 300k (306 403) annotated musical notes, each with unique pitch, timbre, and envelope characteristics, from 1,006 instruments. The generated samples are four seconds, monophonic 16kHz audio snippets. Every pitch of a standard MIDI piano (21-108) was played. As a reminder, MIDI was created by electronic music manufacturers as a digital representation of the notes [28]. Some MIDI files can be found on this website. The notes are annotated with three additional pieces, **Source, Family, Qualities**. This dataset is an order of magnitude larger than existing public datasets for audio synthesis, facilitating improved model training and evaluation. The split of the dataset organized per family and per instrument is displayed in Fig 5 and in Fig 6. We remark that the dataset is not balanced between instruments, where it is however more or less balanced for the sources.

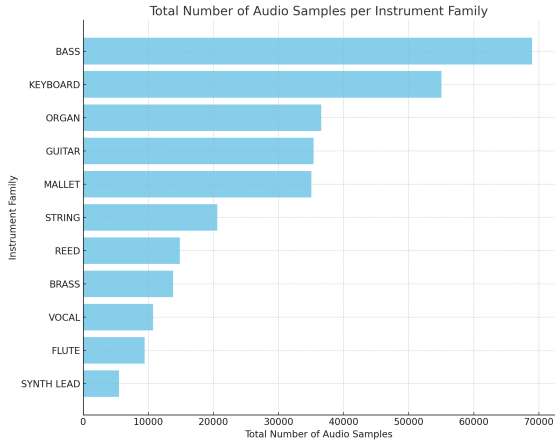

Figure 5: Total Number of Audio Samples per Instrument Family. CC Marengo Matteo.

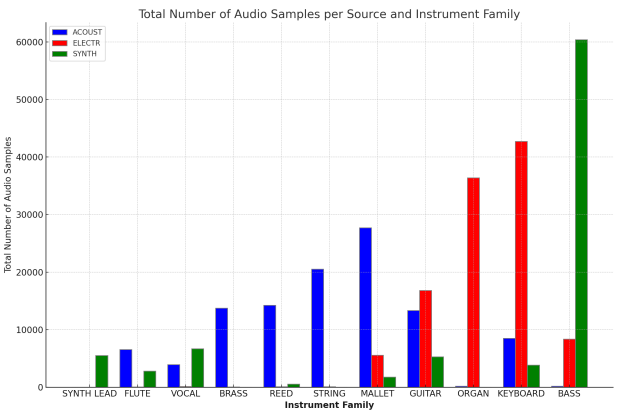

Figure 6: Total Number of Audio Samples per Source and Instrument Family. CC Marengo Matteo

3.4.4 Evaluation and Experiments

The authors conduct extensive evaluations comparing the WaveNet autoencoder with a baseline spectral autoencoder on tasks like note reconstruction and timbre interpolation. To study the performances, plots of the constant-q transform (CQT) [4] will be done. In comparison with the traditional spectrogram where a time-frequency representation is given using the Short-time Fourier Transform, CQT spectrogram uses filters spaced logarithmically in frequency, offering a variable resolution that better matches human auditory perception; **greater frequency resolution at lower frequencies and better temporal resolution at higher frequencies**. Therefore, a big window size will be used in low frequencies and as frequency increases, the size of the window will diminish. For example, at the bottom of the piano scale (30 Hz) a difference of 1 semitone is a difference of 1.5 Hz whereas at the top of the musical scale (about 5 kHz) a difference of 1 semitone is a difference of approximately 200 Hz (as shown in Fig 7). The formula to compute it is:

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{\frac{-j2\pi Qn}{N[k]}}$$

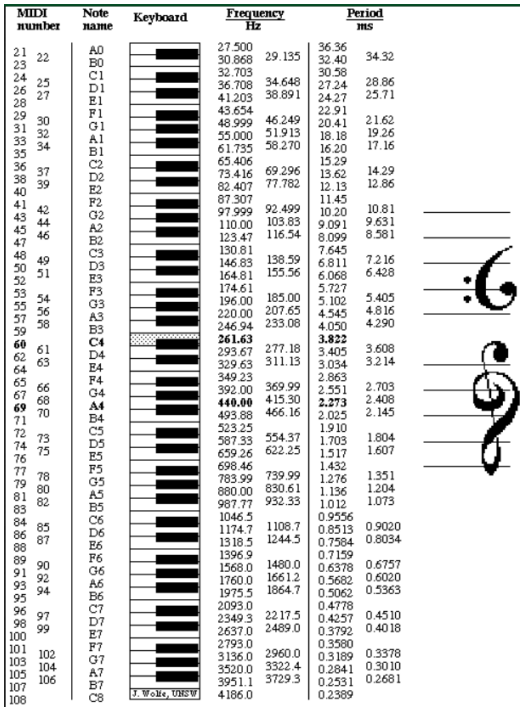

Figure 7: Graphics of piano notes Retrieved from this link

To complete their visual assessment, authors have developed a multi-task classification network inspired by the Inception Score for images [31]. It predicts the pitch and quality labels.

The paper also explores **the learned embedding space**, illustrating how the model can interpolate between different instruments to create new, expressive sounds. This capability to morph between timbres while preserving audio quality showcases the model's potential for music generation and sound design applications.

3.5 Validation and Results

3.5.1 Reconstruction

In the sound reconstruction, each rainbowgram matches the general contour of the original note. The wavenet autoencoder effectively captures the key characteristics of the notes, including the **fundamental frequencies and noise on attack**, despite slight pitch inaccuracies at onset. The baseline model, however,

adds percussive sounds and suffers from noisy phase estimation, resulting in less accurate reproductions. For complex sounds like the flügelhorn, the autoencoder nearly replicates the original's vibrato and richness, while the baseline cannot capture the detailed phase structures, leading to a less expressive sound. Indeed, as shown in Fig 8, the WaveNet autoencoder replicates the oscillations across the harmonics even if there are some discontinuities as underlined by the vertical lines whereas the baseline has correct harmonics, but the phase is more random; the oscillations are not present. The reconstruction is also outlined with the quantitative comparison and then the use of a multi-task classification network.

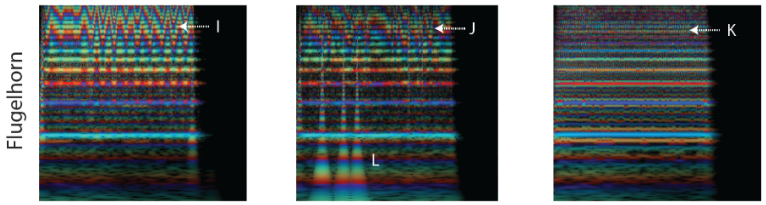

Figure 8: Reconstruction of notes for the flugerhorn. Left, original, Middle, WaveNet, Right, Baseline. Retrieved from [13].

3.5.2 Interpolation in Timbre and Dynamics

Successful reconstructions should represent the dataset's range of **timbre and dynamics**. The experiment compares interpolations in an embedding space by generative models with simple audio space superpositions. The generative models blend characteristics of different instruments, creating new, distinct sounds. Specifically, the WaveNet autoencoder exhibits more realistic and perceptually interesting blends, dynamically mixing overtones in time, unlike the baseline model, which introduces phase distortion. The WaveNet model also tends to amplify subtle modulations, adding complexity to the sounds. This tendency is discussed in the 9.1.

3.5.3 Entanglement of Pitch and Timbre

In this part, the authors have generated different pitches from a single vector representation while preserving the characteristics of timbre and dynamics. Initial models trained with pitch conditioning didn't effectively separate pitch from the timbre and dynamics in the embeddings. Further analysis using a linear pitch classifier on embeddings from models trained with and without pitch conditioning showed **a decrease in pitch classification accuracy with conditioning**. This suggests some success in disentangling pitch from other characteristics. This effect was more pronounced in models with smaller embedding sizes, indicating that larger models might not rely as heavily on pitch conditioning. All the results are displayed in Fig 9.

Practical demonstrations with a baseline model of a specific embedding size (128) managed to balance reconstruction quality with response to pitch conditioning. The chord reconstructions showed that the harmonic structure of an original note was partially preserved across pitches but introduced sub-harmonics when shifted upwards, aligning with pitch classification errors—mainly pitches being confused with those one octave apart.

Correlation analysis of WaveNet embeddings among pitches for a single instrument revealed unique partitioning into registers where notes of different pitches had similar embeddings. This suggests a broad distinction between high and low registers even when averaged across all instruments, which corresponds to the natural variation in an instrument's timbre and dynamics across its range.

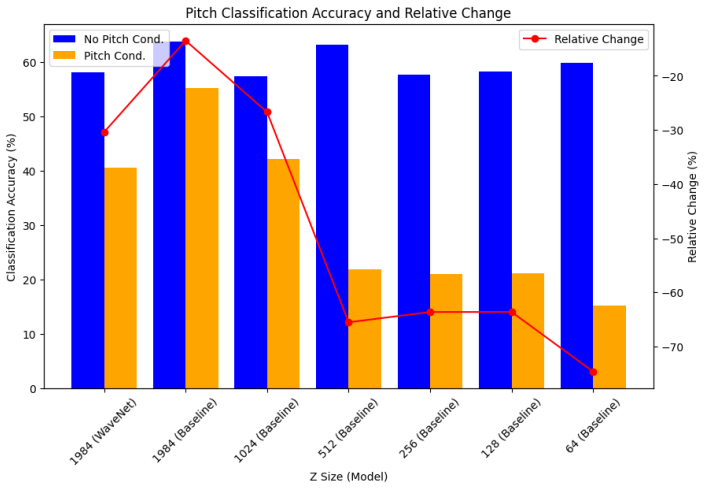

Figure 9: Classification accuracy of a linear pitch classifier trained on learned embeddings. Adapted from [13] by CC Marengo Matteo.

4 Experiments

To understand more deeply what this WaveNet autoencoder is about, I have conducted some experiments to have my own results and compare them to the ones presented in the paper.

4.1 Understand the used tools

First to understand the tools authors have used I decided to plot the CQT spectrogram on the simplest example that can be; the A4 note (*'La' in french*). This sound has a fundamental frequency of f = 440 Hz. It is confirmed on Fig 11. What we indeed observe when comparing the spectrogram with the CQT spectrogram (Fig 14 vs Fig 13) is that the frequency appears way more easily and can be more precisely identified.

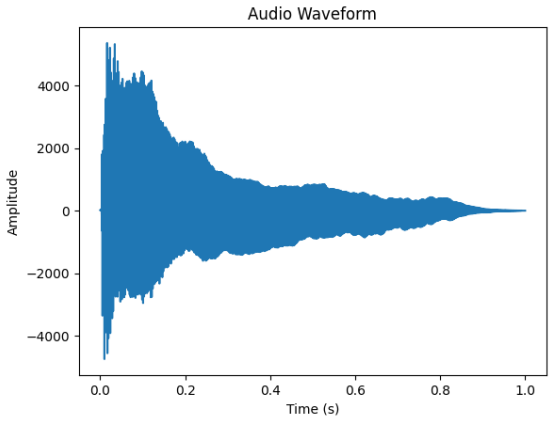

Figure 10: Shape of the signal pianoA4. CC Marengo Matteo.

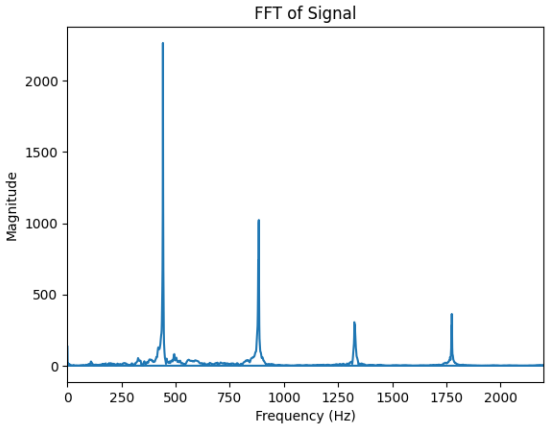

Figure 11: Fourier transform of pianoA4. CC Marengo Matteo

The behaviour is even more obvious when comparing CQT spectrogram of A4 piano with A3/4 piano, the A3 fundamental frequency is 220 Hz. We observe this frequency really distinct from the A4 CQT spectrogram.

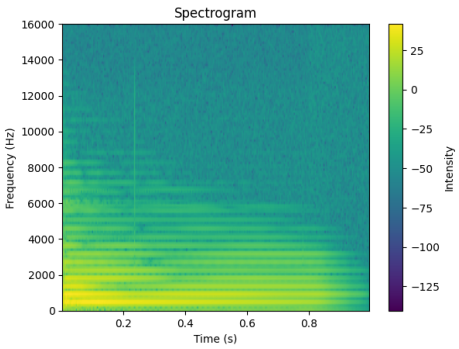

Figure 12: Spectrogram of pianoA4. CC Marengo Matteo

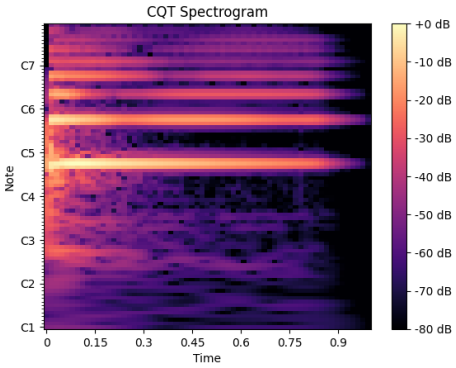

Figure 13: CQT spectrogram of pianoA4. CC Marengo Matteo.

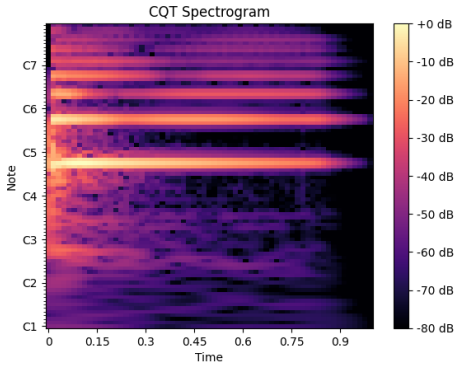

Figure 14: Spectrogram of pianoA4. CC Marengo Matteo

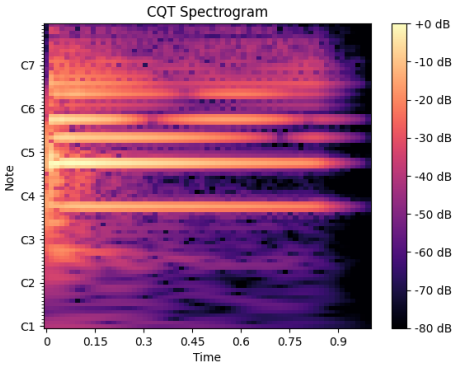

Figure 15: CQT spectrogram of pianoA3A4. CC Marengo Matteo.

4.2 Understand the WaveNet architecture

To understand the WaveNet architecture and how it works I will adapt the code from this GitHub associated to this Medium Blog "Wavenet: A Generative Model for Raw Audio Synthesis" [7]. When training the WaveNet model on A3A4 we obtain such a loss in Fig 16. It learns rapidly even if there is a unexpected peak at the epoch 120.

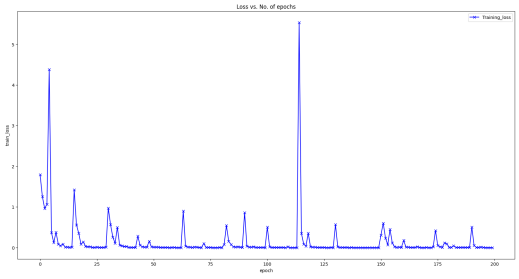

Figure 16: Wavenet loss training evolution. CC Marengo Matteo

When comparing the original vs the reconstructed sound (Fig 17 and Fig 18) there seems to be more frequencies involved but the global shape and amplitude stay the same. Then when comparing the original CQT spectrogram with the reconstructed CQT spectrogram (Fig 19 vs Fig 20) we observe that the one at the output is nearly the same than the original one. We can then conclude that on a simple sound like that the WaveNet works well.

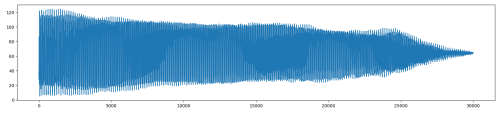

Figure 17: Original signal. of pianoA3A4. CC Marengo Matteo

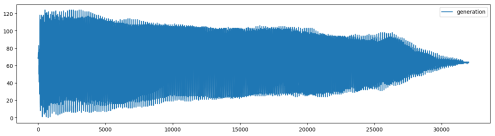

Figure 18: Reconstructed signal. CC Marengo Matteo.

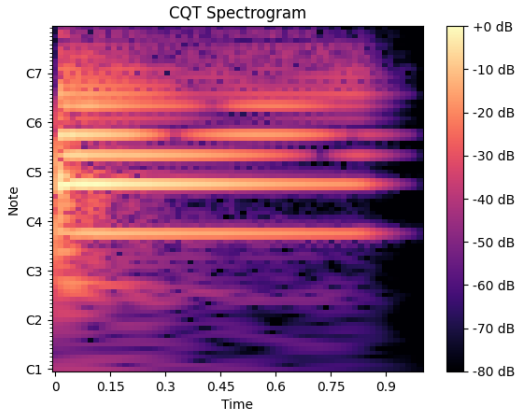

Figure 19: CQT Spectrogram of pianoA3A4. CC Marengo Matteo

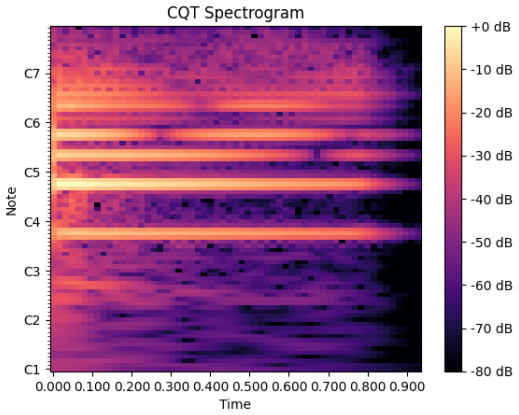

Figure 20: CQT spectrogram of pianoA3A4 reconstructed with wavenet. CC Marengo Matteo.

5 Critical assessment

5.1 Paper strengths

The methodology employed in this paper is thorough and well-structured. The authors detail the construction of the WaveNet autoencoder and the NSynth dataset with clarity, allowing for reproducibility and further research. Engel et al. conducted comprehensive experiments to evaluate the performance of the WaveNet autoencoder against a well-tuned spectral autoencoder baseline, employing both **qualitative and quantitative measures**. The experiments included tasks like note reconstruction, instrument interpolation, and pitch interpolation, providing a detailed analysis of the learned embeddings and the quality of the generated audio. Overall, the methodology is rigorous and comprehensive. They discussed well their motivations, training strategy, evaluation and listed all the hyperparameters.

5.2 Paper weaknesses

The authors acknowledge a significant limitation in their methodology: the WaveNet autoencoder's difficulty in capturing global context due to memory constraints. This limitation suggests that while the model excels at encoding local audio features and short to medium-term structure, it may struggle with longer audio sequences or compositions that require understanding broader musical contexts. The paper discusses challenges in **disentangling pitch and timbre** in the learned embeddings, with experiments showing mixed results in pitch classification accuracy. This indicates potential areas for improvement in how the model handles the nuances of musical pitch and the independence of pitch and timbre. They only compare their WaveNet Autoencoder with a baseline Autoencoder. Therefore, it should have been more complete with more models such as more complex convolutional autoencoders, adding dropout or batch normalization for example.

Given the extensive size and diversity of the NSynth dataset, there is an implicit challenge in ensuring that models trained on this dataset generalize well to unseen data and do not overfit to the specific characteristics of the dataset. The paper does not explicitly address this concern or present strategies employed to mitigate overfitting. An other issue regarding the dataset being that this is unbalanced between classes (between family especially). It would have been interesting to understand their motivations for such a behavior and whether it

can be fixed to obtain more accurate and fair results. **There is no real benchmark as there is no training with other datasets as the one by Google in their WaveNet paper.**

More weaknesses are that there is no part on the **computational part** telling which GPU is being used, how long it takes. Loss functions could be plotted to show how it evolves. This part lacks of reproducibility. On the same subject, they do not compare different batch sizes during training, it could have been interesting to visualize it. **Figure 4 and 5** lacks of clarity and they could have been put elsewhere as it is not easy to read the results and at the same understand where the figures are. Finally, there are no ablation studies or discussions on different parts of the WaveNet Autoencoders that could have been modified to study their effects.

5.3 Recommendations for improvement

To overcome the limitation of capturing global context, future work could explore more **memory-efficient neural network architectures** or training strategies that allow the model to process longer sequences without significant memory overhead. Techniques such as attention mechanisms or recurrent neural network layers could be integrated to enhance the model's ability to understand and generate longer audio sequences with coherent musical structures. Some examples of improvements can be transformer models with efficient attention mechanisms (Sparse Transformers, 2019, [5]). These models utilize sparse attention mechanisms to reduce the computational complexity from quadratic to linear concerning sequence length. This approach allows the model to handle much longer sequences. Efficient Chunking strategies such as the Reformer [20] incorporate techniques such as locality-sensitive hashing to reduce the complexity of attention, allowing for efficient processing of long sequences with less memory. Other techniques are Adaptive Computation time for recurrent neural networks [16] or Memory Augmented Neural Networks (MANNs) [17].

Further research could focus on improving the model's ability to separate pitch and timbre in the learned embeddings. This could involve experimenting with alternative conditioning strategies, incorporating additional input features, or designing more sophisticated loss functions that explicitly encourage disentanglement in the latent space.

To address potential concerns about overfitting and ensure that models generalize well to new musical contexts, future versions of the paper could include more extensive validation on diverse external datasets. Additionally, implementing and discussing regularization techniques or data augmentation methods would provide insight into how the model's robustness and generalization capabilities were enhanced.

6 Outlooks

Since 2017 and the release of this paper many research groups have tried to tackle the challenges that the paper encounters. This is especially true for the improvement of human speech synthesis and music synthesis that can have a wide impact on the public and the society. Some examples are therefore presented.

Models such as OpenAI's Jukebox *(JukeBox, 2020 [9])* and MuseNet *(MuseNet, OpenAI, 2019 [26])* demonstrate the capability of AI to compose music in various styles and genres, even emulating the styles of specific artists. These models analyze vast datasets of music to learn patterns, harmonies, and rhythms, enabling them to generate complex compositions.

Complementary to WaveNet [33], Google's Tacotron *(Tacotron, 2017, Google [35])* , Facebook's MelGAN *(MelGAN, 2019, Facebook [22])* , Deep Voice 3 [27], FastSpeech [30], WaveGlow [29] or NeMo [21] are capable of capturing the nuances of human speech, including intonation, emotion, and accent, making synthesized speech increasingly indistinguishable from real human voices.

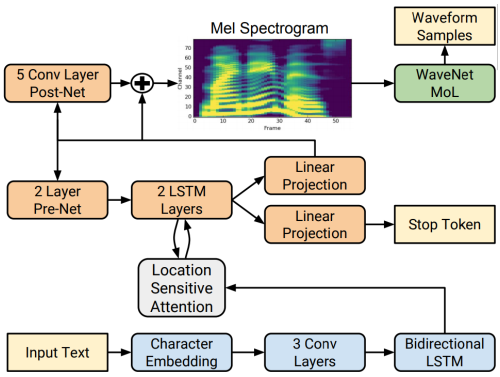

Figure 21: Tacotron diagram. Retrieved from [35].

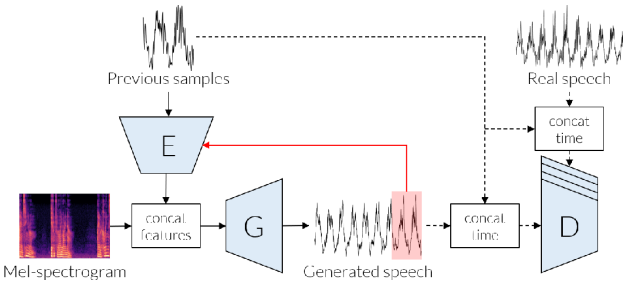

Figure 22: Melgan diagram. Retrieved from [22].

7 Conclusion

To sum-up, this paper achieved to create a large and diverse audio dataset that can indeed be used for further research and benchmarking to cope with the computer vision field that have already plenty of them. One nice example of its use is in the paper GanSynth: Adversarial Neural Audio Synthesis in 2019 [12] that introduces GAN in the field leveraging from NSynth. In addition, they leverage the architecture of the WaveNet to create a WaveNet Autoencoders that can do audio synthesis. Their robust experiments shown the superiority of their method on the baseline spectral autoencoder architecture. If the field is blooming with many impactful research papers presented in the Outlooks, it also the fuel for worries. Indeed, one drawback of audio synthesis has been the surge of deep fakes as shown in Fig 23 and in Fig 24.

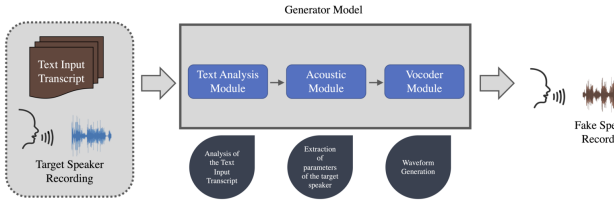

Figure 23: Synthetic based approach

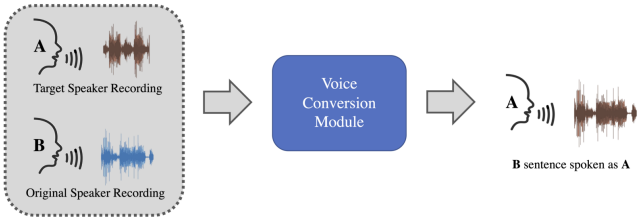

Figure 24: Imitation based approach

The domain of audio synthesis offers ample opportunity for innovation, yet it necessitates cautious progress to avoid the ethical pitfalls and potential misuse associated with deepfake technologies.

8 Acknowledgements

Thank you to Pr. Emmanuel BACRY for this course that I enjoyed to follow as I learned many tools for audio signal processing.

References

[1] Hanauer S.L. Atal, B.S. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 1971.

[2] Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The million song dataset. pages 591–596, 01 2011.

[3] Jenkins G. M. Reinsel G. C. Box, G. E. Time series analysis: forecasting and control. *John Wiley Sons.*, 2011.

[4] Judith C. Brown. Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89:425–434, 1991.

[5] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

[6] J Chowning. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 1973.

[7] Prantosh Das. Wavenet: A generative model for raw audio synthesis. *Medium*, 2021.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: a large-scale hierarchical image database. pages 248–255, 06 2009.

[9] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020.

[10] Jianfeng Dong, Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Learning deep representations using convolutional auto-encoders with symmetric skip connections, 2017.

[11] H. Dudley. The vocoder. *Bell Labs Technical Journal*, 1939.

[12] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis, 2019.

[13] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[14] G. Fant. Acoustic theory of speech production. *Mouton Co.*, 1960.

[15] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Music genre database and musical instrument sound database. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, Vol. 3, 01 2003.

[16] Alex Graves. Adaptive computation time for recurrent neural networks, 2017.

[17] Alex Graves, Greg Wayne, and Reynolds et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538, 10 2016.

[18] J. D. Hamilton. Time series analysis. *Princeton University Press*, 1994.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[20] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer, 2020.

[21] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. Nemo: a toolkit for building ai applications using neural modules, 2019.

[22] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis, 2019.

[23] 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, 2017.

[24] R.A. Moog. Voltage-controlled electronic music modules. *Journal of the Audio Engineering Society.*, 1964.

[25] Belar H. Olson, H.F. Electronic music synthesizer". *Journal of the Audio Engineering Society.*, 1959.

[26] Christine Payne. Musenet. *openai.com/blog/musenet*, 2019.

[27] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning, 2018.

[28] Graham Poliner and Daniel Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 01 2007.

[29] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis, 2018.

[30] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[32] J.O. Smith. Physical modeling using digital waveguides. *Computer Music Journal.*, 1992.

[33] 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, 2016.

[34] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016.

[35] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

9 Appendix

9.1 Interpolation in Timbre and Dynamics

As shown in Fig 25, baseline introduces phase distorsion whereas WaveNet captures accurately the harmonic structure of the two notes.

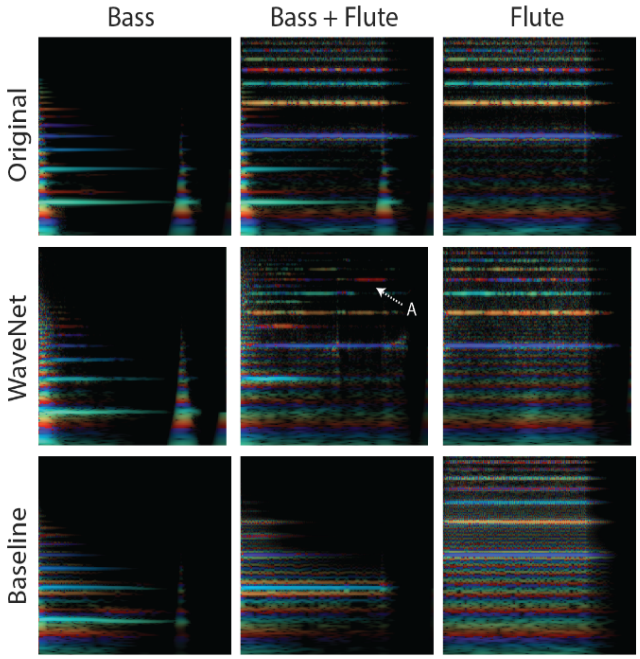

Figure 25: Linear interpolation between Bass and Flute note. Retrieved from [13].