Generative Models for Audio Synthesis

Krishna Subramani
Supervised by Prof. Preeti Rao

Electrical Engineering
Indian Institute of Technology Bombay, India

What is Audio Synthesis?

- Very informally, can be thought of 'using a digital computer as a musical instrument' [Mathews, 1963].
- Motivated by the early analog synthesizers like the Moog modular synthesizers which primarily used components like voltage controlled oscillators, filters, amplifiers. They were also equipped with 'envelope generators'. audio[1] audio[2]

What is Audio Synthesis?

- However with the advent of computing, it became easier to perform all the processing using a digital computer [Haynes, 1982].
- Enables everyone who has a computer(effectively everyone today!) to play with music and compose their own orchestras.(Example: Google's Magenta Library)

A History of Audio Synthesis

- The following methods of audio synthesis have been widely researched in the past,
  1. Physical Modeling Synthesis.
  2. Spectral Modeling Synthesis.
- This Link here titled '120 years of Music' is an interesting read for the electronic music enthusiast.
- Recently however, there has been very interesting work in the use of Generative Models using Deep Learning, which has been labelled as **"Neural Audio Synthesis"**.
- Generative models primarily rely on the ability of algorithms today to extract musically relevant information from tons of data.

A History of Audio Synthesis

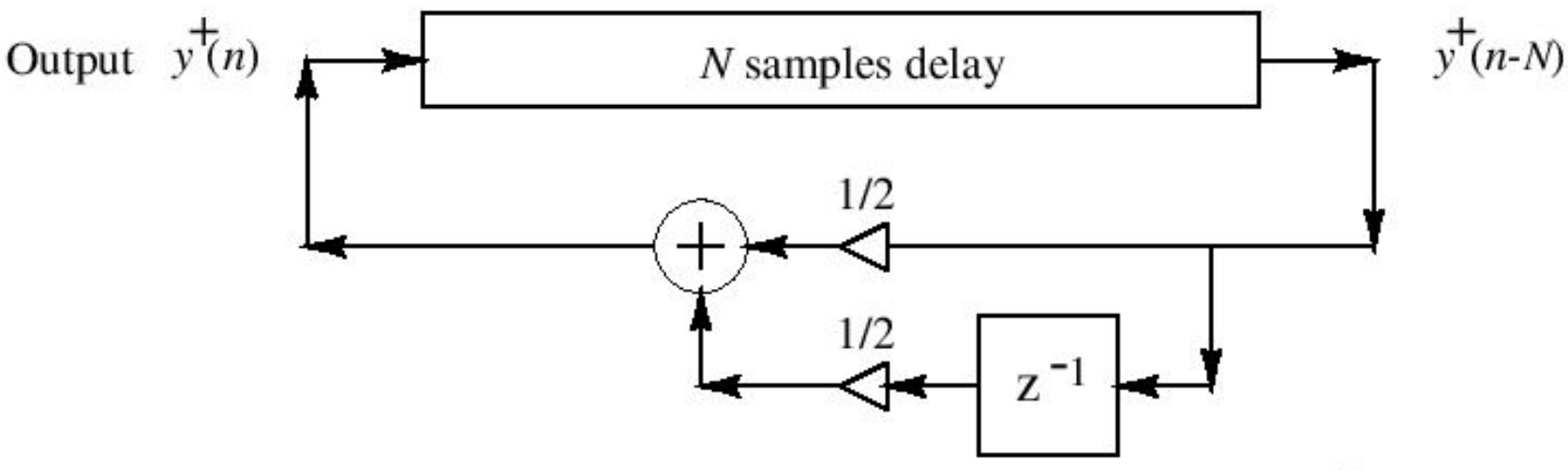
- The flow of this presentation will be as follows,
  1. Introduction and discussion on Physical Modeling and Spectral Modeling Synthesis, with some more emphasis on the latter.
  2. Current work on Generative Modeling of audio.
  3. Presenting a new framework for 'Sound Transformation and Synthesis' which combines aspects discussed above, and discussion on how to proceed ahead.
- What we would like to achieve would be a system which has 'knobs' which can control the kind of audio we want to synthesize.

Physical Modeling Synthesis

- The main motivation behind Physical Modeling is **to model the underlying physics of the source which generates the sound**.
- An example is solving the constrained wave equation to obtain the sound of strings.
- Some physical system models,
  1. State Space models.
  2. Digital Waveguide models.

An example - The Karplus Strong Algorithm

- Precursor to the digital waveguide models which attempt to solve discretized differential equations using delay lines and filters.
- The Karplus strong algorithm attempts to mimic the sound made by plucked strings. audio [3]



Karplus Strong Flow Diagram [Smith, 2010]

Advantages and Disadvantages

- Advantages,
  1. Have the freedom to control musically relevant aspects in terms of modeling the 'sound source' via its underlying physics.
- Disadvantages,
  1. If you cannot accurately model the underlying physics you have to approximate/simplify the model.
  2. Fast computation needed for real-time generation.

Spectral Modeling Synthesis

- Why move ahead from Physical Modeling?
  1. Physical Models try mimicking the behaviour of sound sources. However, we are more interested in how we 'perceive' sound.
  2. Physical Modeling is not general enough i.e. you can only describe sounds whose generation dynamics are known apriori. Thus, it fails to 'generalize' to any sound in particular.
  3. Another interesting thing to study is how sounds 'morph' with each other, and this can be best studied with spectral models of sound.
- Spectral Modeling Synthesis(SMS) is a technique that aims at modeling the spectral characteristics of sound with the aim to obtain 'musically useful' representations
  [Serra, 1989, Serra et al., 1997].

Spectral Modeling Synthesis

- The major assumption - signals($\boldsymbol{x}$) can be represented as $\boldsymbol{x} = \boldsymbol{x}_{\boldsymbol{sine}} + \boldsymbol{x}_{\boldsymbol{stochastic}}$ where $\boldsymbol{x}_{\boldsymbol{sine}}$ is the sinusoidal component and $\boldsymbol{x}_{\boldsymbol{stochastic}}$ is the stochastic/random component.
- Why analyze in the spectral domain? - Motivated by perceptual models of our ear[1], which say that the ear acts similar to a harmonic analyzer which might perform some operation analogous to the Fourier Transform.

[1] Hearing and Perception, Link

Spectral Modeling Synthesis

- Short Time Fourier Transform(STFT) is commonly used to analyze the spectrum for non-stationary signals,

$$X_l(k) := \sum_{n=0}^{N-1} w(n)x(n + l.H)e^{-j\omega_k n}.$$

- Motivated by this, the following three models for SMS were proposed [Serra, 1989, Serra et al., 1997],
  1. Sinusoidal Modeling.
  2. Deterministic + Residual Modeling.
  3. Deterministic + Stochastic Modeling.

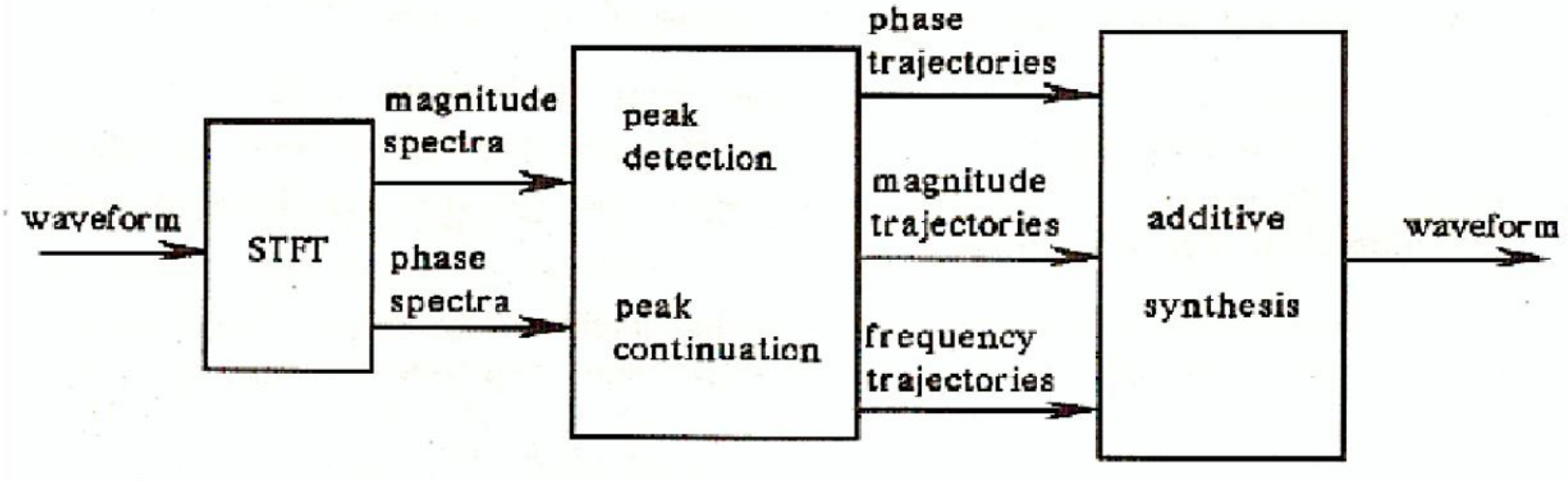
  Each of these methods shall briefly be discussed ahead.

Sinusoidal Model

- The signal is modelled as a sum of time varying sinusoidal components, audio[4] audio[5] audio[6]

$$s(t) = \sum_{r=1}^{R} A_r(t) cos(\theta_r(t)) \; ; \; \theta_r(t) = \int_0^t \omega_\tau(\tau) d\tau + \theta_r(0) + \phi_r.$$

  Here, R is the number of sinusoidal components, $A_r(t)$ is the instantaneous amplitude and $\theta_r(t)$ is the instantaneous phase.
- The main steps in the algorithm are,
  1. Peak picking - At each time frame, the maxima in the spectra have to be chosen.
  2. Peak continuation - Across time frames, the peaks have to be connected to obtain a 'smooth' variation across time.

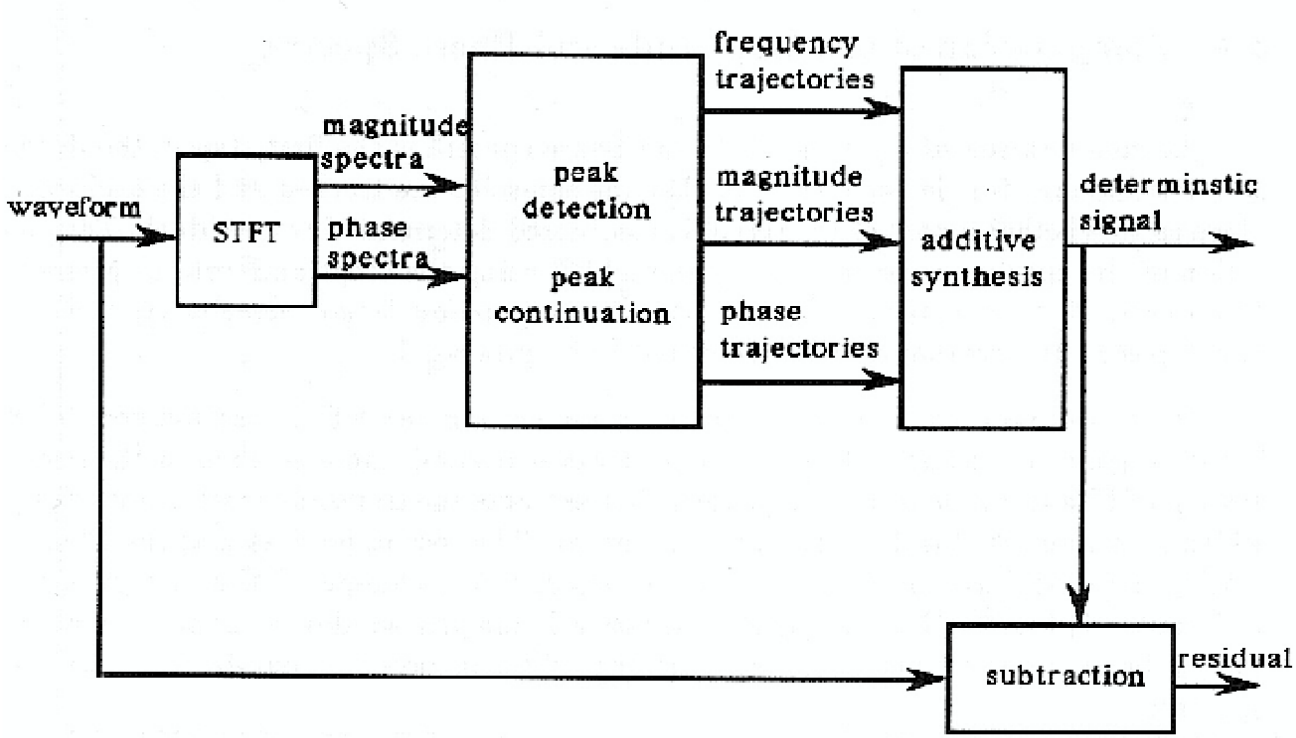


Sinusoidal Model Analysis and Synthesis [Serra, 1989]

- Major Drawbacks,
  1. Difficult to model noise with sinusoidal components(need very large number of sine waves).
  2. Because of this, the reconstructed sound seems a bit artificial. audio[7] audio[8]

Deterministic + Residual Model

- Unlike the sinusoidal model where the sinusoids model all components of the signal, you enforce the condition that the sinusoids only model the quasi-sinusoidal(or the 'partials') components. audio[9] audio[10]
- The remaining 'non-sinusoidal' portion of the signal is called the residual. audio[11]
- Another relaxation introduced by this model is to ignore the modeling of the phase and only consider the magnitude spectra.
- At the same time, phase continuity is maintained to prevent artifacts from forming in the sound.
- This is motivated by the fact that the ear is primarily sensitive to the magnitude and not phase of the spectrum.

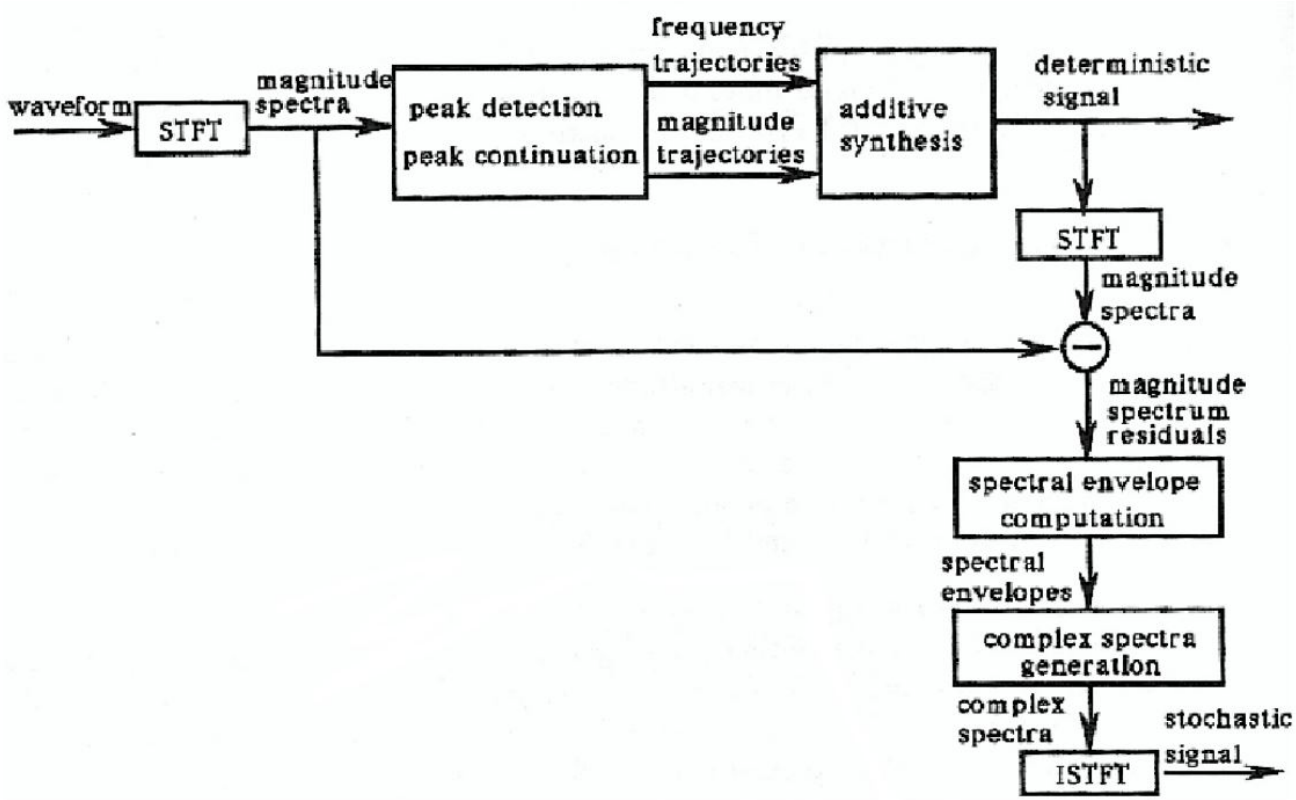


Deterministic + Residual Model Analysis and Synthesis [Serra, 1989]

- Major Drawbacks,
  1. Residual lacks the flexibility to be transformed.
  2. If the signal is not mainly composed of quasi-sinusoidal components, the above method will perform poorly.

Deterministic + Stochastic Model

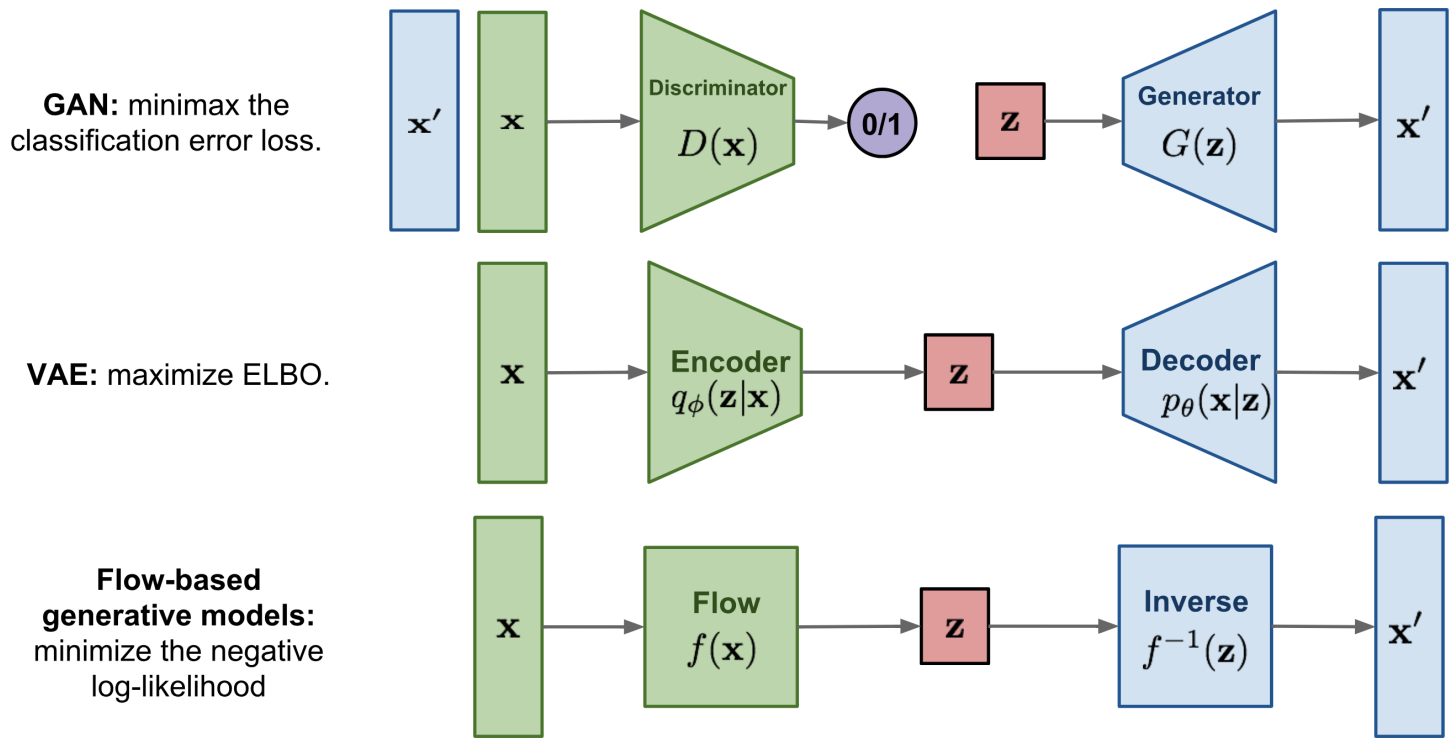
- You take the residual from the previous signal, and treat it as the output of a linear time variant system acting on white noise. The filter is chosen to approximate the envelope of the residual. audio[12] audio[13]



Deterministic + Stochastic Model Analysis and Synthesis [Serra, 1989]

Generative Models

- What are Generative Models? - Given data, tries to generate samples from the 'same distribution' as the original data.



Generative Models, Image Link

Audio Synthesis

- Generative models for audio synthesis are currently following two different methodologies,
  1. Generating the time domain waveform samples using autoregressive models - WaveNet [Oord et al., 2016], Wavenet Autoencoders [Engel et al., 2017] and WaveGAN [Donahue et al., 2018].
  2. Generating a time-frequency(TF) representation of the audio, and inverting it to obtain the audio - GANSynth [Engel et al., 2019] and TimbreTron [Huang et al., 2018].
- [Sinclair, 2018] consider a parametric model for sound, namely the physical model, and use a generative model to generate the parameters from which the sound is re-constructed.

Autoregressive Generative Models

- First such work in this line(originally for speech synthesis) was the WaveNet [Oord et al., 2016], which was heavily inspired by previous work(by the same author), the PixelCNN [Van den Oord et al., 2016]. audio[14]
- Model the joint probability distribution in an autoregressive fashion i.e. each sample depends on all the samples previous to it,

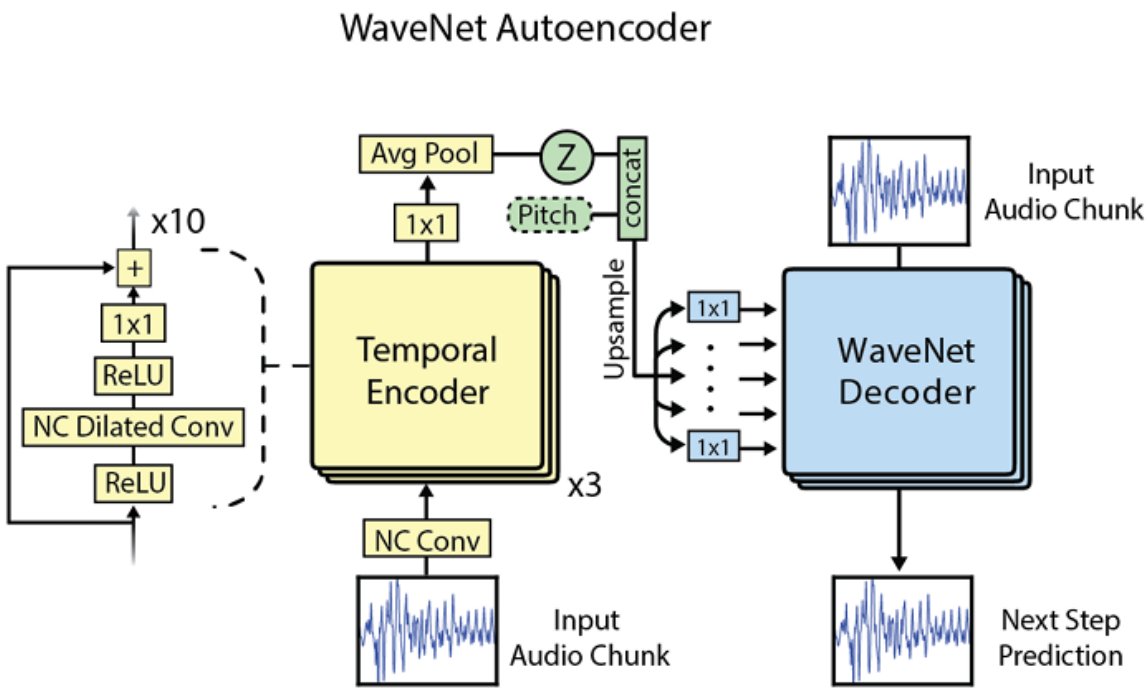
$$P(\bar{x}) = \prod_{t=1}^{T} P(x_t | x_{t-1}, x_{t-2} \dots x_1).$$

  They use the idea of Dilated Convolutions[2] with a 'Receptive Field' to achieve this
- The major issue with music is that it has longer temporal range as compared to speech and hence it became necessary to use large Receptive Fields to capture long term structure.

[2] WaveNet

Autoregressive Generative Models

- [Engel et al., 2017] were inspired by WaveNet, and to overcome some of the shortcomings they felt in wavenet, they came up with their own model which combines an Encoder with a WaveNet model as shown below, audio[15]



Wavenet Encoder, [Engel et al., 2017]

Autoregressive Generative Models

- The advantage of using this Autoencoder is that it can learn an efficient representation via the Latent Space of the audio which can capture structural aspects of the audio. This so called "Manifold of Embeddings" [Engel et al., 2017] can be further analyzed and studied.
- Few disadvantages of the Autoregressive models are -
  1. Takes very long to generate samples from the model(1 second of audio takes a few minutes to generate!).

Generative Adversarial Modeling

- Inspired by the success of GAN's in generating highly realistic images [Goodfellow et al., 2014], [Donahue et al., 2018] were motivated to use GAN's to generate time domain audio by modifying the architecture to generate samples(WaveGAN). audio[16]
- [Donahue et al., 2018] also mentions using GAN's to generate audio using a TF representation(SpecGAN).
- One of the major issues with the TF representations is the issue of invertibility to obtain the time domain waveform. As you do not preserve the phase information, the generated waveform is noisy. audio[17]
- To deal with this, [Engel et al., 2019] propose modifications to the TF representations.

Generative Adversarial Modeling

- [Huang et al., 2018] presents two very interesting modifications to the generation approach, audio[18] audio[19]
  1. Use of a Cycle-GAN([Zhu et al., 2017]) to 'learn' mappings between inputs and outputs
  2. Use of a Wavenet decoder to invert the TF representation
- The advantages of adversarial modeling,
  1. In principle, they can model all the timescales involved in audio synthesis, and thus can generate much 'richer' audio.
- The disadvantages are,
  1. GAN's are very unstable to train. Modifications have been proposed to the architecture(WGAN's,[Arjovsky et al., 2017]), inspite of which they take extremely long to train.
  2. The issue of invertibility of the TF representation still remains.

Parametric Modeling

- [Sinclair, 2018] has proposed to use a parametric model namely the physical model as an input to the model.
- As opposed to generating waveforms, the above model generates the parameters from which the waveform is constructed.
- Thus, the above model avoids the invertibiliy issue as the physical model can reconstruct the audio 'exactly' given the parameters.
- This work is one of the major motivations behind the proposed model, which will be discussed ahead.

Proposed Work

- A few issues observed,
  1. The dataset NSynth [Engel et al., 2017] is an artificially generated dataset, and lacks the expressivity in real life audio.
  2. Not much analysis done on how the latent space variables affect the synthesized audio.
  3. Generation time(in autoregressive models) and model complexity(in adversarial models).
  4. The issue of invertibility of the obtained TF representation(specially in adversarial modeling).
- Thus, the following are proposed,
  1. Use the dataset by [Romani Picas et al., 2015] instead of Nsynth as they are actual recordings of instruments as opposed to generated audio.
  2. Try out a parametric model like the Sinusoidal or the Deterministic + Stochastic model. This can give us better control over the generated audio, as well as avoid the invertibility issue.
  3. Study the Latent Space more closely

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds	Model any general sound

Generative Modeling

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds	Model any general sound

Generative Modeling
Depending on available data can model anything

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds Sound is expressive and natural	Model any general sound Sound is not that expressive

Generative Modeling
Depending on available data can model anything

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds Sound is expressive and natural	Model any general sound Sound is not that expressive

Generative Modeling
Depending on available data can model anything Data-driven

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds Sound is expressive and natural Several Equations to solve	Model any general sound Sound is not that expressive Several operations to perform

Generative Modeling
Depending on available data can model anything Data-driven

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds Sound is expressive and natural Several Equations to solve	Model any general sound Sound is not that expressive Several operations to perform

Generative Modeling
Depending on available data can model anything Data-driven Computationally Intensive

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds	Model any general sound
Sound is expressive and natural	Sound is not that expressive
Several Equations to solve	Several operations to perform
Represents sound source	Represents sound receiver

Generative Modeling
Depending on available data can model anything
Data-driven
Computationally Intensive

[3] Link

Comparing the discussed synthesis methods

- Inspired by Professor Julius Smith's "Projections for the Future" [3], we were inspired to extend the table for Generative modeling.

Physical Modeling	Spectral Modeling
Model only restricted sounds	Model any general sound
Sound is expressive and natural	Sound is not that expressive
Several Equations to solve	Several operations to perform
Represents sound source	Represents sound receiver

Generative Modeling
Depending on available data can model anything
Data-driven
Computationally Intensive
Represents either depending on model and data

[3] Link

Putting it all together

- The presentation was started with a discussion on Physical and Spectral Models for audio.
- Then, generative models were introduced, along with the concept of the Z-space.
- What we would like to achieve is a model which has good generated audio quality and has a good amount of control over the different perceptual attributes of the generated audio.
- To achieve the above, we combine the two models to obtain a 'Parametric Generative Model' for sound synthesis, thus the proposed system.
- It would be good to have an interface with a front-end(like the synthesizer shown at the start) which can easily control the audio generation, and a back-end which is driven by the above proposed model.

Concluding Thoughts

1. It was really exciting to go through and analyze the current work going on this area. At the same time, it becomes important to clearly define the end-goal. In our case, the end goal is to generate audio(and not music!).
2. Considering how unstable GAN's are, it seems like a better option to start working with the autoregressive models first, and then move on to more complicated architectures.
3. For those interested in doing research or exploring this area, this is my github repository, which contains a list of all the papers I have gone through, and also briefly summarizes them.

References I

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein gan.
arXiv preprint arXiv:1701.07875.

[Donahue et al., 2018] Donahue, C., McAuley, J., and Puckette, M. (2018).
Adversarial audio synthesis.
arXiv preprint arXiv:1802.04208.

[Engel et al., 2019] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019).
Gansynth: Adversarial neural audio synthesis.
arXiv preprint arXiv:1902.08710.

[Engel et al., 2017] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. (2017).
Neural audio synthesis of musical notes with wavenet autoencoders.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1068–1077. JMLR. org.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
In *Advances in neural information processing systems*, pages 2672–2680.

References II

[Haynes, 1982] Haynes, S. (1982).
The computer as a sound processor: A tutorial.
Computer Music Journal, 6(1):7–17.

[Huang et al., 2018] Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., and Grosse, R. B. (2018).
Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer.
arXiv preprint arXiv:1811.09620.

[Mathews, 1963] Mathews, M. V. (1963).
The digital computer as a musical instrument.
Science, 142(3592):553–557.

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

[Romani Picas et al., 2015] Romani Picas, O., Parra Rodriguez, H., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., and Serra, X. (2015).
A real-time system for measuring sound goodness in instrumental sounds.
In *Audio Engineering Society Convention 138*. Audio Engineering Society.

References III

[Serra, 1989] Serra, X. (1989).
A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.

[Serra et al., 1997] Serra, X. et al. (1997).
Musical sound modeling with sinusoids plus noise.
Musical signal processing, pages 91–122.

[Sinclair, 2018] Sinclair, S. (2018).
Sounderfeit: Cloning a physical model using a conditional adversarial autoencoder.
arXiv preprint arXiv:1806.09617.

[Smith, 2010] Smith, J. O. (accessed ¡2010¿).
Physical Audio Signal Processing.
`http://ccrma.stanford.edu/~jos/pasp/`.
online book, 2010 edition.

[Van den Oord et al., 2016] Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016).
Conditional image generation with pixelcnn decoders.
In *Advances in Neural Information Processing Systems*, pages 4790–4798.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017).
Unpaired image-to-image translation using cycle-consistent adversarial networks.
In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Audio examples description I

1. Synthesizer envelope sound example(instant attack), Link
2. Synthesizer envelope sound example(slow attack), Link
3. Karplus Strong synthesized audio example, Link
4. Original Piano Audio
5. Reconstructed piano using sinusoidal model with 5 sinusoids
6. Reconstructed piano using sinusoidal model with 150 sinusoids
7. Original Rain sound
8. Reconstructed rain sound using sinusoidal model with 100 sinusoids
9. Original Carnatic sound
10. Reconstructed Carnatic sound using Deterministic + Residual model with 250 sinsuoids, the deterministic component
11. Reconstructed Carnatic sound using Deterministic + Residual model with 250 sinsuoids, the residual component

Audio examples description II

12. Reconstructed Carnatic sound using Deterministic + Residual model with 250 sinsuoids, the residual component
13. Stochastic approximation to the Carnatic residual
14. WaveNet generated audio, Link
15. WaveNet Autoencoder(Nsynth) generated audio, Link
16. WaveGAN generated audio, Link
17. SpecGAN generated audio, Link
18. TimbreTron source audio, Link
19. TimbreTron timbre transferred audio, Link