# Generative Audio Synthesis with a Parametric Model

**Krishna Subramani**[1] **& Alexandre D'Hooge**[2] **& Preeti Rao**[1]

[1]IIT Bombay, [2]ENS Paris-Saclay

subramani.krishna97@gmail.com, dhooge@crans.org

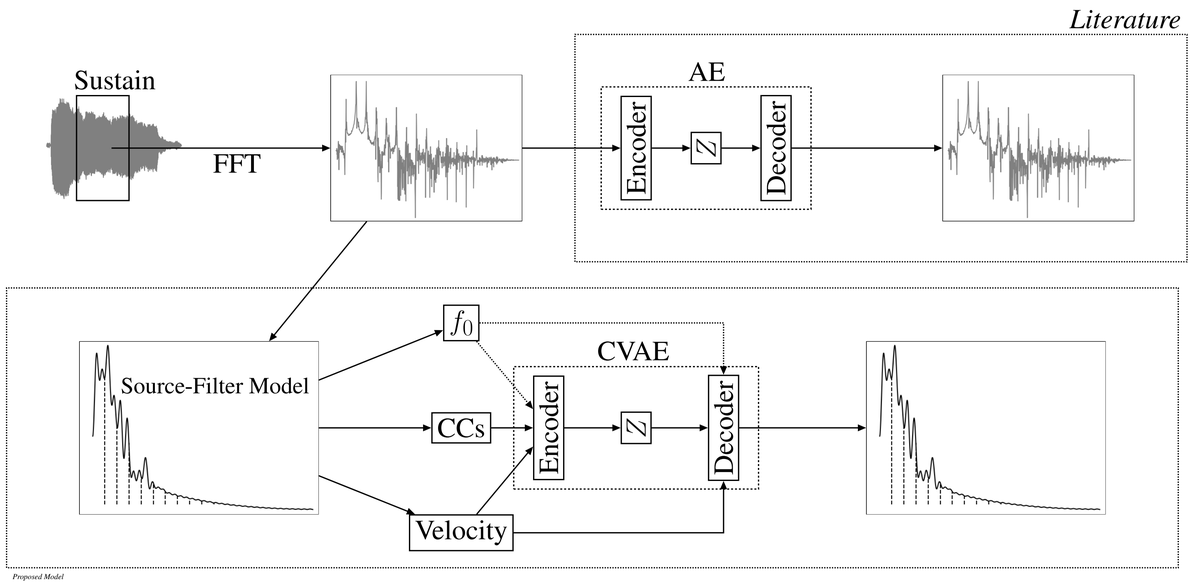
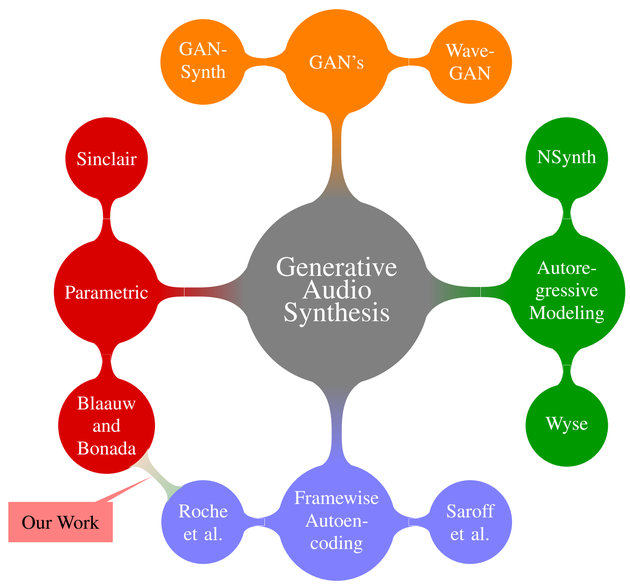



**Figure 1:** Flowchart of the state of the art frame-wise audio synthesis pipeline (upper branch) and our proposed model (lower branch). $Z$ represents the latent space learned by the (CV)AE.

## Nearest Neighbours

- [Sarroff and Casey, 2014] first to use autoencoders to perform frame-wise reconstruction of short-time magnitude spectra
- [Roche et al., 2018] extended this analysis to try out different autoencoder architectures



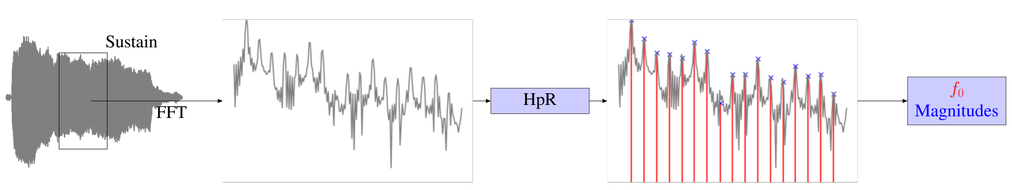

- [Esling et al., 2018] regularized the VAE latent space in order to effect control over perceptual timbre of synthesized instruments.
- [Engel et al., 2017] inspired by Wavenets [Oord et al., 2016] autoregressive modeling capablities for speech extended it to musical instrument synthesis.
- [Wyse, 2018] also autoregressively modelled the audio, albeit by conditioning the waveform samples on additional parameters like pitch, velocity(loudness) and instrument class.

## Why Parametric?

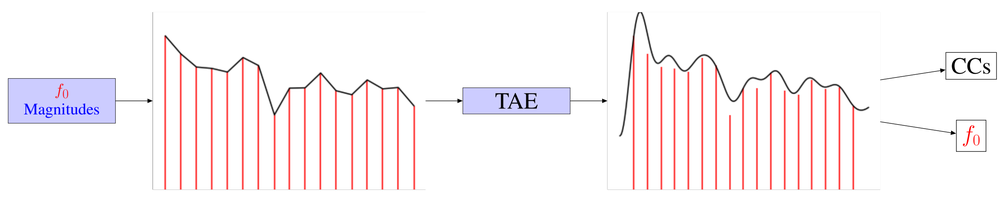
- Better control over the musically relevant attributes such as pitch, dynamics and timbre can be obtained by using a parametric model.
- A general spectral representation such as the Fourier transform or the time domain representation often fails to offer such attributes.
- Recognizing this in the context of speech synthesis [Blaauw and Bonada, 2016] used a vocoder representation for speech, and then trained a VAE to model the frame-wise spectral envelope.

### The Parametric Model

1. Frame-wise magnitude spectrum → harmonic representation using Harmonic plus Residual(HpR) model [Serra et al., 1997](currently, we neglect the residual).




- Output of HpR block ⟹ log-dB magnitudes + harmonics
- log-dB magnitudes + harmonics → TAE algorithm [Roebel and Rodet, 2005, IMAI, 1979]




## Generative Models

- Autoencoders [Hinton and Salakhutdinov, 2006] - Minimize the MSE between input and network reconstruction. Not truly a generative model, as you cannot 'generate new' data.
- Variational Autoencoders [Kingma and Welling, 2013] - Inspired from Variational Inference, enforce a prior on the latent space. These can 'generate new data' by sampling from the prior.
- Conditional Variational Autoencoders [Doersch, 2016, Sohn et al., 2015] - Same principle as a VAE, however learns the conditional distribution over an additional conditioning variable.

☐ Why VAE over AE?

■ Continuous latent space from which we can sample points(and synthesize the corresponding audio).

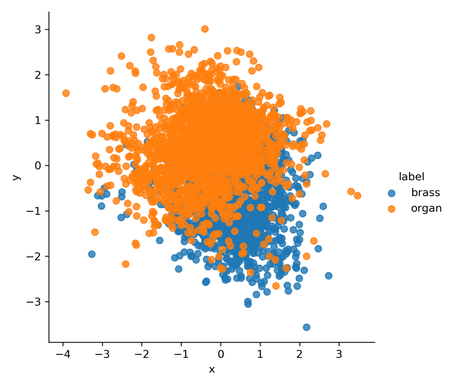
☐ Why CVAE over VAE?

■ Conditioning on pitch ⟹ Network captures dependencies between the timbre and the pitch ⟹ More accurate envelope generation + Pitch control.

## Experiments

- We use a subset of the NSynth [Engel et al., 2017] dataset in our work. We have implemented the parametric representation and used it to successfully train a CVAE network.

### Timbre hybridization




- Trained on two different instruments, the network is capable of generating new sounds with a hybrid timbre.
- The figure shows an example of a 2-D latent space we obtained when training on brass and organ instances(the reconstruction is not good enough due to the low dimensionality!).
- We try to generate hybrid timbres by sampling intermediate points from the latent space. We do indeed observe the audio timbre changing subtly as you move from one cluster to the other.
- The standard VAE formulation assumes a unimodal Gaussian prior, which is not good enough to model multiple instruments. Future work will include using a mixture of Gaussians as a prior.

### Interpolation/Extrapolation over pitch

- Rather than generating new timbres, we consider the problem of synthesis of a given instrument's sound with flexible control over the pitch and loudness dynamics.
- The motivation is the desire to synthesize realistic sounds of an instrument at pitches that may not be available in the training data. Such a context can arise in styles such as Indian art music where continuous pitch movements are integral parts of the melody.
- We evaluate our approach on a dataset of violin [Romani Picas et al., 2015], a popular instrument in Indian music, adopted from the West, due to its human voice-like timbre and ability to produce continuous pitch movements [Haigh, 2019].

**Sound examples can be found at `https://www.ee.iitb.ac.in/student/~krishnasubramani/ismir_LBD_poster.html`** (QRcode)

### Machine Learning & Audio Synthesis?

- Early analog synthesizers used voltage controlled oscillators, filters, amplifiers to generate the waveform, and 'envelope generators' to shape it.
- Data-driven statistical modeling + computing power ⟹ Deep Learning for audio synthesis!
- Rely on ability of algorithms to extract musically relevant information from vast amounts of data.