Generative Audio Synthesis with a Parametric Model

Krishna Subramani\textsuperscript{1} & Alexandre D’Hooge\textsuperscript{2} & Preeti Rao\textsuperscript{1}
\textsuperscript{1}IIT Bombay, \textsuperscript{2}ENS Paris-Saclay
subramani.krishna97@gmail.com, dhooge@crans.org

Machine Learning & Audio Synthesis?
\begin{itemize}
\item Early analog synthesizers used voltage controlled oscillators, filters, amplifiers to generate the waveform, and ‘envelope generators’ to shape it.
\item Data-driven statistical modeling + computing power \implies Deep Learning for audio synthesis!
\item Rely on ability of algorithms to extract musically relevant information from vast amounts of data.
\end{itemize}

Nearest Neighbours
\begin{itemize}
\item \cite{harer2014first} first to use autoencoders to perform frame-wise reconstruction of short-time magnitude spectra
\item \cite{roche2018} extended this analysis to try out different autoencoder architectures
\item \cite{zhang2019} regularized the VAE latent space in order to effect control over perceptual timbre of synthesized instruments.
\item \cite{saroff2019} inspired by WaveNet \cite{koel2016} autoregressive modeling capabilities for speech extended it to musical instrument synthesis.
\item \cite{nyse2018} also autoregressively modelled the audio, albeit by conditioning the waveform samples on additional parameters like pitch, velocity(loudness) and instrument class.
\end{itemize}

Why Parametric?
\begin{itemize}
\item Better control over the musically relevant attributes such as pitch, dynamics and timbre can be obtained by using a parametric model.
\item A general spectral representation such as the Fourier transform or the time domain representation often fails to offer such attributes.
\item Recognizing this in the context of speech synthesis \cite{harer2014} used a vocoder representation for speech, and then trained a VAE to model the frame-wise spectral envelope.
\end{itemize}

The Parametric Model

1. Frame-wise magnitude spectrum \rightarrow harmonic representation using Harmonic plus Residual (HpR) model \cite{sinclair1999} (currently, we neglect the residual).

\begin{itemize}
\item Output of HpR block \implies log-\text{dB magnitudes} + harmonics
\item log-\text{dB magnitudes} + harmonics \rightarrow TAE algorithm \cite{dickhut2005, roche2018}
\end{itemize}

Generative Models

\begin{itemize}
\item Autoencoders \cite{vincent2008} - Minimize the MSE between input and network reconstruction. Not truly a generative model, as you cannot ‘generate new’ data.
\item Variational Autoencoders \cite{kingma2013} - Inspired from Variational Inference, enforce a prior on the latent space. These can ‘generate new data’ by sampling from the prior.
\item Conditional Variational Autoencoders \cite{kingma2014} \cite{dinh2014} - Same principle as a VAE, however learns the conditional distribution over an additional conditioning variable.
\end{itemize}

Experiments

\begin{itemize}
\item We use a subset of the NSynth \cite{engel2017} dataset in our work. We have implemented the parametric representation and used it to successfully train a CV AE network.
\item Timbre hybridization
\begin{itemize}
\item Trained on two different instruments, the network is capable of generating new sounds with a hybrid timbre.
\item 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!).
\item 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.
\end{itemize}
\item The standard VAE formulation assumes a unimodal Gaussian prior, which is not good enough to 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.
\item We evaluate our approach on a dataset of violin \cite{romani2015}, a popular instrument in Indian music, adopted from the West, due to its human voice-like timbre and ability to produce continuous pitch movements \cite{haigh2019}.
\end{itemize}

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

Literature

\begin{itemize}
\item Deep Learning for audio synthesis!
\item Rely on ability of algorithms to extract musically relevant information from vast amounts of data.
\item Data-driven statistical modeling + computing power \implies Deep Learning for audio synthesis!
\item Rely on ability of algorithms to extract musically relevant information from vast amounts of data.
\end{itemize}

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.