GENERATIVE AUDIO SYNTHESIS WITH A PARAMETRIC MODEL

Krishna Subramani
IIT Bombay
krishna.subramani@iitb.ac.in

Alexandre D'Hooge
ENS Paris-Saclay
dhooge@crans.org

Preeti Rao
IIT Bombay
prao@ee.iitb.ac.in

EXTENDED ABSTRACT

Audio synthesis is of interest to people with different backgrounds such as music enthusiasts, performers, composers and researchers. With the advent of data-driven statistical modeling and abundant computing power, researchers are turning increasingly to deep learning for audio synthesis. Various approaches such as autoregressive modeling, GANs and VAEs have been proposed with varying degrees of success given the ultimate goal of modeling complex instrument sound sources, possibly while also achieving the flexible control of musically relevant attributes. Recently, Roche et al. [5] studied different autoencoder architectures, like variational (VAE) and LSTM based autoencoders, to perform the frame-wise reconstruction of audio spectra. They also presented an analysis of the lower dimensional representation of the input spectrum, or 'latent space', which a musician can explore to generate new sounds. Esling et al. [4] took this idea further and tried to incorporate structure into this latent space to match the perceptual timbre space of the corresponding instruments.

In the interest of more flexible control over the generated sound however, it could be useful to somehow decompose the latent representation into relevant musical attributes such as pitch, dynamics and timbre. Such a possibility is more likely to be available with a parametric signal model rather than a general spectral representation such as the Fourier transform. Recognizing this in the context of speech synthesis, Blaauw et al. [1] used a vocoder representation for speech, and then trained a VAE to model the frame-wise spectral envelope. Thus, by incorporating a powerful signal processing approach to separate the distinct perceptual attributes of the sound class, the network can be expected to learn a much more meaningful representation.

The parametric representation we adopt is inspired from the analysis pipeline of [2], where "perceptually motivated" representations of audio are proposed for sound morphing. A Source Filter decomposition is applied to the harmonic component of the spectrum extracted by the Harmonic model [6]. The filter is estimated as the envelope of the harmonic spectrum and represented via low-dimensional cepstral coefficients. Therefore, as opposed to simply training a VAE on the full magnitude spectrum (upper branch in Figure 1), we train a conditional VAE (CVAE) on the real cepstral coefficients (CCs) conditioned on the pitch and velocity (lower branch in Figure 1) assuming that a suitably labeled dataset is available for the supervised training of the network. Our motivation to use a CVAE comes from the fact that the Source and Filter are not entirely decoupled for musical instruments. By conditioning on the pitch and velocity, our model is expected to generate the spectral envelope more accurately.

We use a subset of the NSynth [3] dataset in our work. We have implemented the parametric representation and used it to successfully train a CVAE network. Trained on two different instruments, the network is capable of generating new sounds with a hybrid timbre. We have been able to test the validity of our model by training on sounds restricted to the odd MIDI pitches and examining the generation of the sounds with the missing pitches. The network indeed learned how to fill in the missing pitches, returning sounds with the expected timbre. Furthermore, the network could even produce perceptually acceptable continuous sounds when conditioned with a semi-continuous pitch frequency sweep (one octave divided into small steps). Finally, we observed that using a parametric representation allows us to use a very small neural network that can be trained in under an hour (on a mobile Nvidia GeForce GTX 1050Ti).

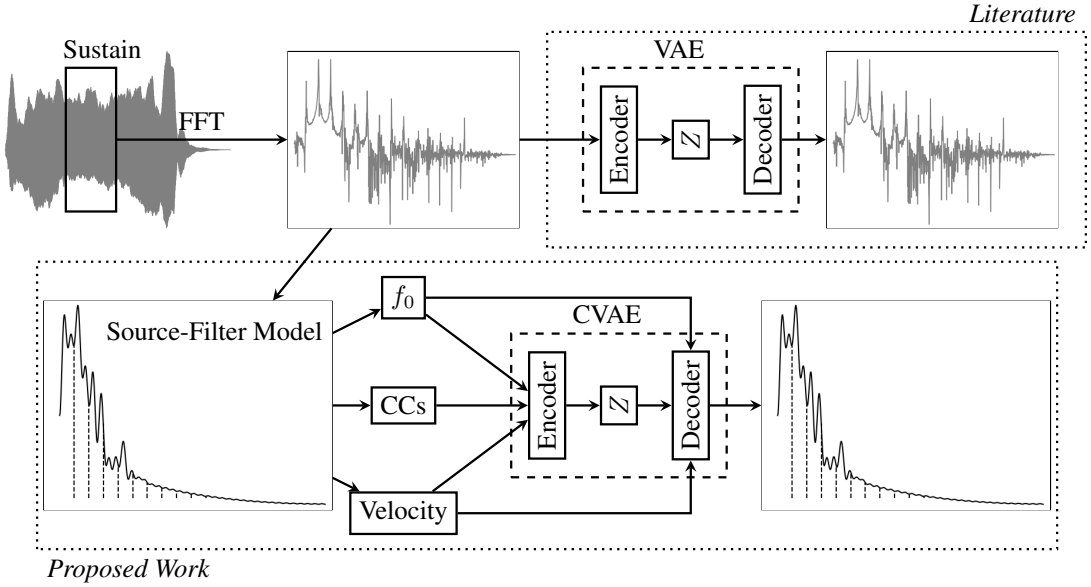

Figure 1. Flowchart of the state of the art audio synthesis pipeline (upper branch) and our proposed work (lower branch). Z represents the latent space learned by the (C)VAE.

Our network is currently trained only on the sustained segments of selected instrument sounds. By applying instrument-specific temporal segmentation, we plan to augment the network toward generating realistic instrument sounds with the components of attack, sustain and decay. Moreover, we are only training on the harmonic part of sounds while the residual also contains important information identifying the instrument (consider the bowing sound of a violin). Future work shall address this issue either by modifying the current CVAE or by using two networks that would exchange information during synthesis to produce perceptually better sounds.

ACKNOWLEDGMENTS

The authors thank Prof. Xavier Serra for insightful discussions on the problem.

REFERENCES

[1] Merlijn Blaauw and Jordi Bonada. Modeling and transforming speech using variational autoencoders. In *Interspeech*, pages 1770–1774, 2016.

[2] Marcelo Caetano and Xavier Rodet. Musical Instrument Sound Morphing Guided by Perceptually Motivated Features. *IEEE Transactions on Audio, Speech and Language Processing*, 21(8):1666–1675, 2013.

[3] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. *arXiv:1704.01279 [cs]*, April 2017. arXiv: 1704.01279.

[4] Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics. *arXiv:1805.08501 [cs, eess]*, May 2018. arXiv: 1805.08501.

[5] Fanny Roche, Thomas Hueber, Samuel Limier, and Laurent Girin. Autoencoders for music sound modeling: a comparison of linear, shallow, deep, recurrent and variational models. *arXiv:1806.04096 [cs, eess]*, June 2018. arXiv: 1806.04096.

[6] Xavier Serra et al. Musical sound modeling with sinusoids plus noise. *Musical signal processing*, pages 91–122, 1997.