Notes on references

Juan Corredor

October 21st, 2024

1 DDSP

[Engel et al., 2020] used as main reference for the research.

1.1 Important bullet points

- Interpretable and modular approach to generative modeling
- Classic signal processing elements with deep learning methods
- Models that rely on strided convolution or windowing (STFT) need to align waveshapes or suffer from spectral leakage.
- DDSP takes the approach of vocoders in using oscillators to synthesize signals.
- According to the DDSP paper [Engel et al., 2020], the DDSP library is capable of extrapolating timbres not seen during training, and independent control over pitch and loudness during synthesis.
- References to Neural Source Filter [Wang et al., 2019] imply this might be a good additional reference for research. [日本音響学会音響標準委員会, 1953]

1.2 Newly learned concepts from this document

- Strided Convolution:

  Convolution that has a hop length, meaning, it skips some information to avoid analyzing redundancies.

- Teacher forcing:

  Feeding back the correct answers into training algorithms to reduce training times and lead the model in the right direction during training.

- Automatic differentiation:

  Also: 'algorithmic differentiation' or 'computational differentiation' means splitting derivatives into simpler operations using the chain rule.

- Deterministic autoencoder
- Adversarial training

- Variational inference
- Jacobian design
- Stochastic latents
- CREPE model
- Latent encoding

1.3 How it relates to my research

I believe the general structure and inner workings of the proposed DDSP synthesizer from the paper are exactly how I need to model my own synthesizer.

The proposed system requires an audio database to train the machine learning model and then use that model to extract the required features to build a faithful additive synthesizer. What's interesting about this system is that it allows for the creation of timbres beyond the training set, meaning that theoretically, if my training database consists of a collection of plucked string samples, I could interpolate timbres between them or exaggerate some of their features beyond what I have already recorded.

Since this model has been tested with longer audio samples, so far a lot of the development has been using audio sampled at relatively low sample rates (in the range of the 16kHz sample rate). However, due to the narrow scope of my project, I believe I should be able to use samples recorded at higher sample rates and generate high-quality output.

Another interesting aspect of the DDSP library is that it allows for flexible scalability. Meaning, I could generate consistent output with very few parameters and relatively low computational power.

2 Latent Space Interpolation

[Le Vaillant and Dutoit, 2024] used as main reference for the research.

2.1 Important bullet points

- Interpolation between timbres for a synthesizer should be more nuanced than a simple crossfade.
- This work was carried out with a Variational Auto-Encoder (VAE) dedicated to preset interpolation.
- This is helpful as it establishes an initial framework for sound design.
- Presets are handled as sequences of parameters using *multi-head attention networks*.
- Includes an objective morphing evaluation method based on audio features.
- Mentions harmonic-percussive source separation from the following article: [Fitzgerald, 2010]
- Existing synthesizers and research either underperform under slightly unexpected inputs, or yield lower quality outputs, or cannot be used in real-life contexts using MIDI input.

- This research focuses on smooth transitions for synthesis.
- In this case synthesizers are treated as black-box de-facto systems, which are not differentiable.

2.2 The proposed model

2.2.1 Synthesizer and Dataset

Presets and synthesized audio:

- 30k presets (80% training, 10% validation and 10% for testing).
- Volume, transpose and filter were not altered.
- Presets are rendered to 16kHz audio with a single note (MIDI 56) and velocity (75).
- 257-band mel-spectrograms are used during training.

Audio features:

- Audio Commons Timbral Models (ACTM) and Timbre Toolbox [Peeters et al., 2011] allow the extraction of 8 and 46 features respectively. Then reduced to 6 and 32 due to high'correlation between some of them.
- "These audio features act as a proxy for the human perception of timbre." [Le Vaillant and Dutoit, p. 4]
- Name of the model introduced: SPINVAE-2.

The structure:

- VAE with an extra mel-spectrogram decoder.

Attribute-based regularization: Timbre Loss

- L_DKL is not enough to guarantee that timbre characteristics are encoded.
- "Latent coefficients have been shown not to easily relate to perception"
- Models are not directly trained for interpolation. Rather they get interpolated latent spaces and must decode those into audio.
- Two regularization methods:
-
  - [Pati and Lerch, 2021b] "minimizes the binary cross-entropy between S(a) and S(u), where S denotes the logistic sigmoid."
  - [Pati and Lerch, 2021a] "enforces monotonic relationships between timbre attributes and some latent dimensions."

Training procedure:

- A bi-modal (mel-spectrograms and presets) VAE has been pre-trained.

- A CNN is added to the model and its oututs are re-shaped and summed to the u and sigma vectors.
- SPINVAE-1 [Vaillant and Dutoit, 2023]
- First only CNN encoder and decoder are trained from a mel-spectrogram dataset from NSynth notes, 2.2k Surge synthesizer pathes and 24k Dexed presets.
- The weights of embedding, encoder and decoder layers are used as initial weights for all models.
- Fine-tuning is performed without CNN encoder.

2.3 Objective analysis

While the evaluation of linear interpolation is relatively simple, it is more complex to evaluate the interpolation of audio.

[Vaillant and Dutoit, 2023] talks about smoothness as a measure for interpolation.

According to [Caetano and Rodet, 2013], "an interpolation is perceptually linear when timbre feature values change linearly".

Therefore *both linearity and smoothness must be measured.*

In this case

2.4 Newly learned concepts from this document

- Multi-head attention networks:
- Regularization term based on timbre attributes
- Latent Timbre Synthesis (LTS) [Tatar et al., 2021]
- Dexed
- Zero-mean, unit-variance normalized using statistics of the training set
- Attention masks, hidden tokens
- Transformer-based VAEs
- Linear projection
- Discretized Logistic Mixture
- Regularized dimensions vs latent dimensions

2.5 How it relates to my research

References

[Caetano and Rodet, 2013] Caetano, M. and Rodet, X. (2013). Musical instrument sound morphing guided by perceptually motivated features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1666–1675.

[Engel et al., 2020] Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*.

[Fitzgerald, 2010] Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering.

[Le Vaillant and Dutoit, 2024] Le Vaillant, G. and Dutoit, T. (2024). Latent space interpolation of synthesizer parameters using timbre-regularized auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3379–3392.

[Pati and Lerch, 2021a] Pati, A. and Lerch, A. (2021a). Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Computing and Applications*, 33:4429–4444.

[Pati and Lerch, 2021b] Pati, A. and Lerch, A. (2021b). Is disentanglement enough? on latent representations for controllable music generation.

[Peeters et al., 2011] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916.

[Tatar et al., 2021] Tatar, K., Bisig, D., and Pasquier, P. (2021). Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications. *Neural Computing and Applications*, 33:67–84.

[Vaillant and Dutoit, 2023] Vaillant, G. L. and Dutoit, T. (2023). Synthesizer preset interpolation using transformer auto-encoders. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[Wang et al., 2019] Wang, X., Takaki, S., and Yamagishi, J. (2019). Neural source-filter-based waveform model for statistical parametric speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5916–5920. IEEE.

[日本音響学会音響標準委員会, 1953] 日本音響学会音響標準委員会 (1953). 音響標準較正法について. 日本音響学会誌, 9(2):72–82.