VaPar Synth - A Variational Parametric Model for Audio Synthesis

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

[1]Indian Institute of Technology Bombay, India

[2]ENS Paris-Saclay, France

ICASSP 2020

Audio Synthesis?

- What comes to your mind when you hear 'Audio Synthesis'?

Audio Synthesis?

Figure: One of the early Moog Modular Synthesizers

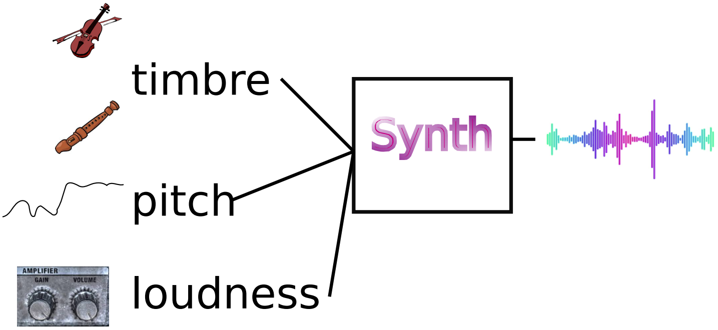
Audio Synthesis?

- Analog synths (Moog!) $\rightarrow$ voltage controlled oscillators, filters, amplifiers to generate, and envelope generators to shape waveforms

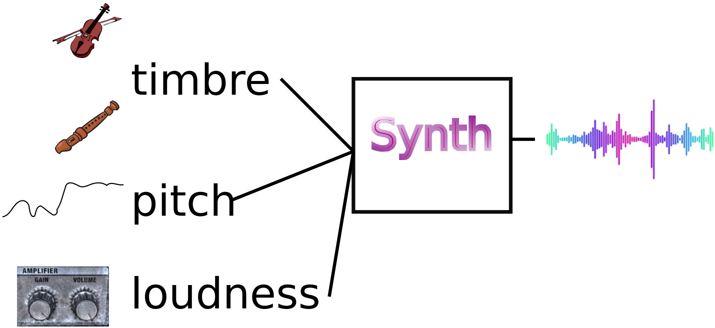
Audio Synthesis?

- Analog synths (Moog!) $\rightarrow$ voltage controlled oscillators, filters, amplifiers to generate, and envelope generators to shape waveforms
- Data-driven statistical modeling + computing power
  $\implies$ Deep Learning for audio synthesis!

Generative Models for Audio Synthesis

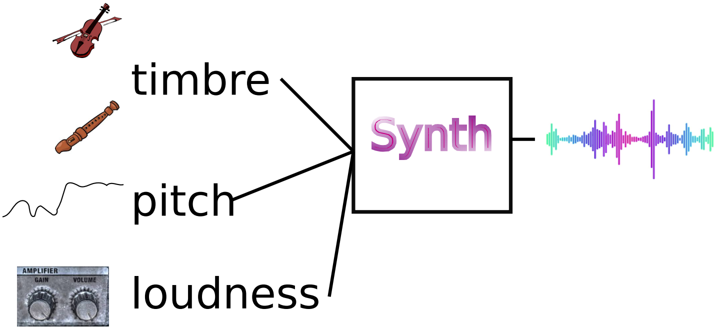


Generative Models for Audio Synthesis



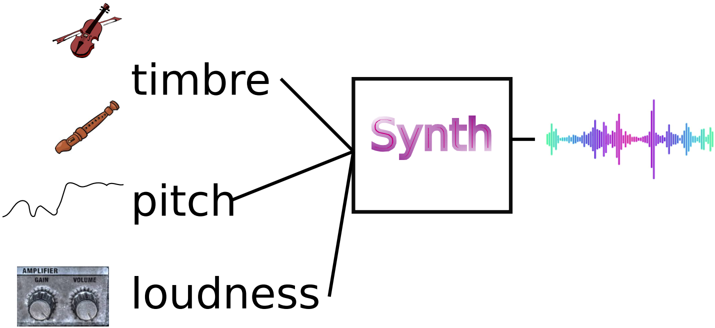
- timbre $\rightarrow$ "difference" between a violin and flute A4

Generative Models for Audio Synthesis



- timbre $\rightarrow$ “difference” between a violin and flute A4
- pitch $\rightarrow$ fundamental frequency

Generative Models for Audio Synthesis



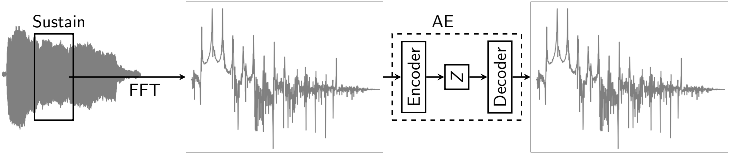
- timbre → “difference” between a violin and flute A4
- pitch → fundamental frequency
- loudness → intensity (energy)

Our Nearest Neighbours

- [Sarroff and Casey, 2014] frame-wise reconstruction of short-time magnitude spectra with autoencoders

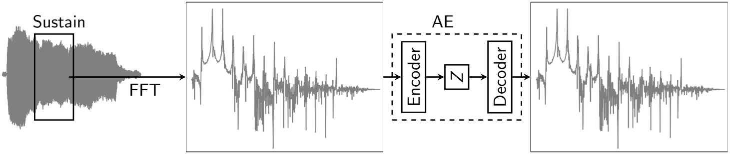
Our Nearest Neighbours

- [Sarroff and Casey, 2014] frame-wise reconstruction of short-time magnitude spectra with autoencoders



Our Nearest Neighbours

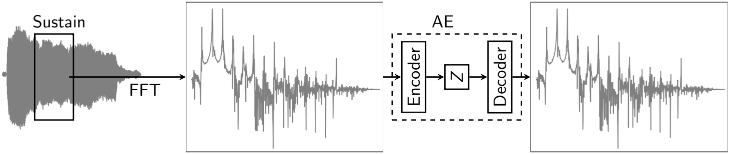
- [Sarroff and Casey, 2014] frame-wise reconstruction of short-time magnitude spectra with autoencoders



- [Roche et al., 2018] tried out autoencoder architectures, analysis of ‘audio latent space’

Our Nearest Neighbours

- [Sarroff and Casey, 2014] frame-wise reconstruction of short-time magnitude spectra with autoencoders



- [Roche et al., 2018] tried out autoencoder architectures, analysis of ‘audio latent space’
- [Esling et al., 2018] regularized this latent space for better control over timbre of synthesized instruments

Our Nearest Neighbours

- Frame-wise analysis-synthesis based reconstruction
  $\rightarrow$ no temporality and phase estimation issues

Our Nearest Neighbours

- Frame-wise analysis-synthesis based reconstruction
  $\rightarrow$ no temporality and phase estimation issues
- [Engel et al., 2017] inspired by Wavenets [Oord et al., 2016] autoregressive modeling capablities for speech extended it to musical instrument synthesis

Our Nearest Neighbours

- Frame-wise analysis-synthesis based reconstruction
  $\rightarrow$ no temporality and phase estimation issues
- [Engel et al., 2017] inspired by Wavenets [Oord et al., 2016] autoregressive modeling capablities for speech extended it to musical instrument synthesis
- [Wyse, 2018] proposed generating audio samples with RNN's, albeit by conditioning the waveform samples on additional parameters like pitch, velocity (loudness) and instrument class

Our Nearest Neighbours

- Frame-wise analysis-synthesis based reconstruction
  $\rightarrow$ no temporality and phase estimation issues
- [Engel et al., 2017] inspired by Wavenets [Oord et al., 2016] autoregressive modeling capablities for speech extended it to musical instrument synthesis
- [Wyse, 2018] proposed generating audio samples with RNN's, albeit by conditioning the waveform samples on additional parameters like pitch, velocity (loudness) and instrument class
- [Défossez et al., 2018] proposed frame-by-frame waveform generation with LSTMs

Why Parametric?

- Consider synthesis of a given instrument sound with flexible control over the pitch

Why Parametric?

- Consider synthesis of a given instrument sound with flexible control over the pitch
- Pitch shifting without timbre modification $\implies$ source-filter model with the filter (spectral envelope) being kept constant [Roebel and Rodet, 2005]

Why Parametric?

- Consider synthesis of a given instrument sound with flexible control over the pitch
- Pitch shifting without timbre modification $\Longrightarrow$ source-filter model with the filter (spectral envelope) being kept constant [Roebel and Rodet, 2005]
- A powerful parametric representation over raw waveform or spectrogram has the potential to achieve high quality with less training data + better generalization

Why Parametric?

- Consider synthesis of a given instrument sound with flexible control over the pitch
- Pitch shifting without timbre modification $\implies$ source-filter model with the filter (spectral envelope) being kept constant [Roebel and Rodet, 2005]
- A powerful parametric representation over raw waveform or spectrogram has the potential to achieve high quality with less training data + better generalization
  1. [Blaauw and Bonada, 2016] used a vocoder representation to train a generative model for speech synthesis

Why Parametric?

- Consider synthesis of a given instrument sound with flexible control over the pitch
- Pitch shifting without timbre modification $\implies$ source-filter model with the filter (spectral envelope) being kept constant [Roebel and Rodet, 2005]
- A powerful parametric representation over raw waveform or spectrogram has the potential to achieve high quality with less training data + better generalization
    1. [Blaauw and Bonada, 2016] used a vocoder representation to train a generative model for speech synthesis
    2. [Engel et al., 2020] (DDSP) recently proposed the control of a parametric model based on a deterministic autoencoder

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments
- We work with the **Violin**

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments
- We work with the **Violin**
  - Mezzo-forte loudness, 4^{th} octave (MIDI 60-71)

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments
- We work with the **Violin**
  - Mezzo-forte loudness, 4^{th} octave (MIDI 60-71)
  - Played by 4 violinists on a single violin

Dataset

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments
- We work with the **Violin**
  - Mezzo-forte loudness, 4^{th} octave (MIDI 60-71)
  - Played by 4 violinists on a single violin
  - Trained on 1000 frames (duration 21.3ms)

Dataset

Why we chose Violin?
Popular in Indian Music, Human voice-like timbre,
Ability to produce continuous pitch!

- **Good-sounds** dataset [Romani Picas et al., 2015]
  - Individual note/scale recordings for 12 instruments
- We work with the **Violin**
  - Mezzo-forte loudness, 4^{th} octave (MIDI 60-71)
  - Played by 4 violinists on a single violin
  - Trained on 1000 frames (duration 21.3ms)

Non-Parametric Reconstruction

- Setup:
  1. Include/Exclude MIDI 63, train with neighbours
  2. **Reconstruct** MIDI 63

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	✓/×	✓	✓	✓

Non-Parametric Reconstruction

- Setup:
  1. Include/Exclude MIDI 63, train with neighbours
  2. **Reconstruct** MIDI 63

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	✓/×	✓	✓	✓

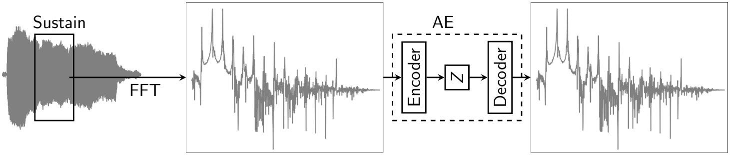


Figure: [Sarroff and Casey, 2014, Roche et al., 2018]

Non-Parametric Reconstruction

- Setup:
  1. Include/Exclude MIDI 63, train with neighbours
  2. **Reconstruct** MIDI 63

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	✓/×	✓	✓	✓

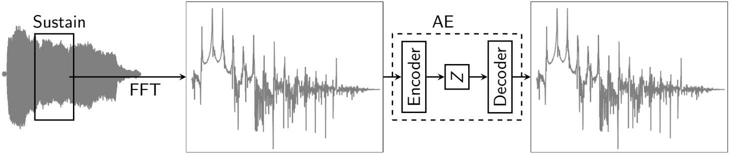


Figure: [Sarroff and Casey, 2014, Roche et al., 2018]

- Framewise autoencoding + inversion with Griffin-Lim [Griffin and Lim, 1984]

Non-Parametric Reconstruction

Figure: Input MIDI 63, $\boxed{1}$ [1]

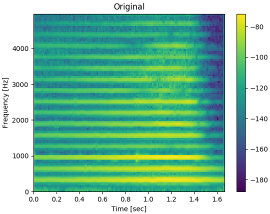


Figure: Including MIDI 63, $\boxed{2}$ [2]

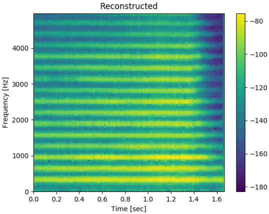
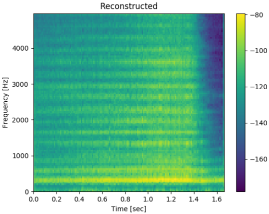


Figure: Excluding MIDI 63, $\boxed{3}$ [3]

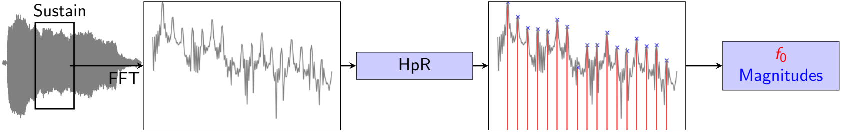


Parametric Model

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

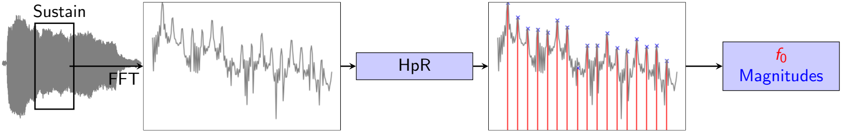
Parametric Model

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



Parametric Model

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

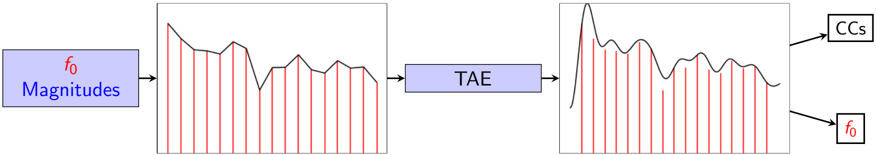


- Output of HpR block $\implies$ log-dB magnitudes + harmonics

Parametric Model

2. log-dB magnitudes + harmonics → TAE algorithm
   [Roebel and Rodet, 2005, IMAI, 1979]

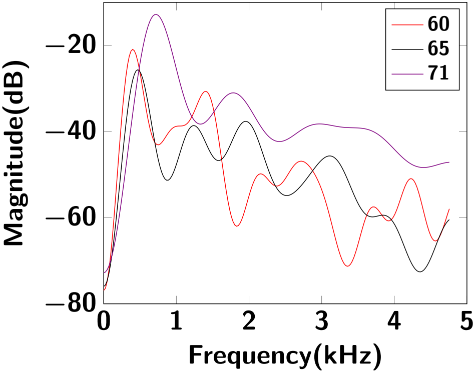
$K_{CC} \leq \frac{F_s}{2f_o}$



* No open source Python implementation of TAE, we implement it following procedure highlighted in [Roebel and Rodet, 2005, Caetano and Rodet, 2012]

1 [4] 2 [5]

Parametric Model



- Spectral envelope shape varies across pitch
  1. Dependence of envelope on pitch [Slawson, 1981, Caetano and Rodet, 2012]
  2. Variation due the TAE algorithm
- Envelope → smooth function to estimate harmonic amplitudes

Generative Models

- Autoencoders (AE) [Hinton and Salakhutdinov, 2006] - Optimal (MSE) lower dimensional representation of input

Generative Models

- Autoencoders (AE) [Hinton and Salakhutdinov, 2006] - Optimal (MSE) lower dimensional representation of input
- Variational AEs (VAE) [Kingma and Welling, 2013] - Enforce a prior on the lower dimensional representation

Generative Models

- Autoencoders (AE) [Hinton and Salakhutdinov, 2006] - Optimal (MSE) lower dimensional representation of input
- Variational AEs (VAE) [Kingma and Welling, 2013] - Enforce a prior on the lower dimensional representation
- Conditional VAEs (CVAE) [Doersch, 2016, Sohn et al., 2015] - Enforce a 'conditional' prior

Generative Models

- Why VAE over AE?

Generative Models

- Why VAE over AE?
  - Continuous latent space from which we can sample points (and synthesize the corresponding audio)

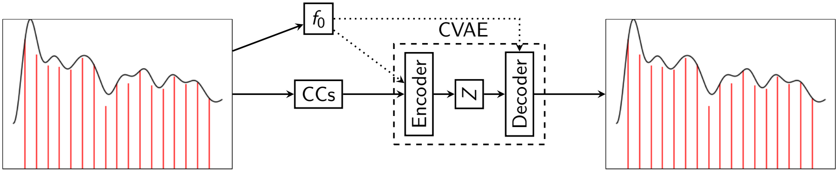
Generative Models

- Why VAE over AE?
  - Continuous latent space from which we can sample points (and synthesize the corresponding audio)
- Why CVAE over VAE?

Generative Models

- 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 $\implies$ Network captures dependencies between the timbre and the pitch $\implies$ More accurate envelope generation + Pitch control

Network Architecture



- Network input is CCs $\rightarrow$ MSE represents perceptually relevant distance in terms of squared error between the input and reconstructed log magnitude spectral envelopes

Network Architecture

- Main hyperparameters -
    1. β - tradeoff between reconstruction and prior enforcement

Network Architecture

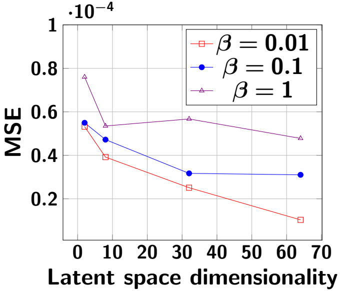
$L \propto \text{MSE} + \beta.\text{KLD}$

- Main hyperparameters -
  1. β - tradeoff between reconstruction and prior enforcement
  2. Dimensionality of latent space - networks reconstruction ability

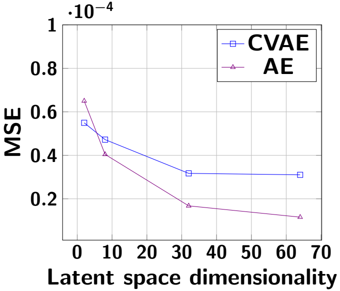
Network Architecture

$L \propto MSE + \beta.KLD$

- Main hyperparameters -
  1. β - tradeoff between reconstruction and prior enforcement
  2. Dimensionality of latent space - networks reconstruction ability



(a) CVAE, varying β



(b) CVAE($\beta = 0.1$) vs AE

Figure: MSE plots to decide hyperparameters

Experiments

- Two kinds of experiments to demonstrate networks capabilities

Experiments

- Two kinds of experiments to demonstrate networks capabilities
  1. **Reconstruction** - Omit pitch instances during training and see how well model reconstructs notes of omitted target pitch

Experiments

- Two kinds of experiments to demonstrate networks capabilities
  1. **Reconstruction** - Omit pitch instances during training and see how well model reconstructs notes of omitted target pitch
  2. **Generation** - How well model 'synthesizes' note instances with new unseen pitches

Reconstruction

- Two training contexts -

Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it 1 [6] 2 [7] 3

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

  2. Octave endpoints

Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

  2. Octave endpoints

MIDI	60	61	62	63	64	65
Kept	✓	×	×	×	×	×
MIDI	66	67	68	69	70	71
Kept	×	×	×	×	×	✓

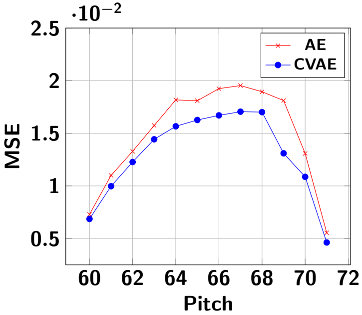
Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

  2. Octave endpoints

MIDI	60	61	62	63	64	65
Kept	✓	×	×	×	×	×
MIDI	66	67	68	69	70	71
Kept	×	×	×	×	×	✓



- Conditioning captures the pitch dependency of the spectral envelope more accurately

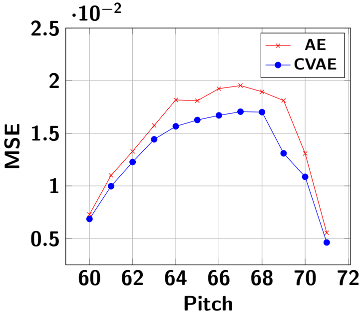
Reconstruction

- Two training contexts -
  1. Train excluding MIDI 63; reconstruct it

MIDI	60	61	62	63	64	65	66
Kept	✓	✓	✓	×	✓	✓	✓

  2. Octave endpoints

MIDI	60	61	62	63	64	65
Kept	✓	×	×	×	×	×
MIDI	66	67	68	69	70	71
Kept	×	×	×	×	×	✓



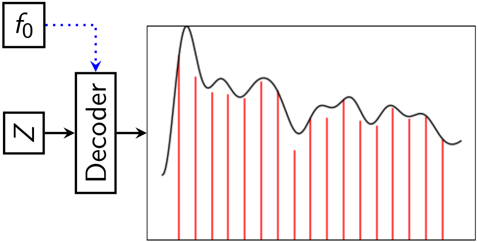
- Conditioning captures the pitch dependency of the spectral envelope more accurately [1][8] [2][9]

Generation

- Generate ('Synthesize') unseen/untrained pitch

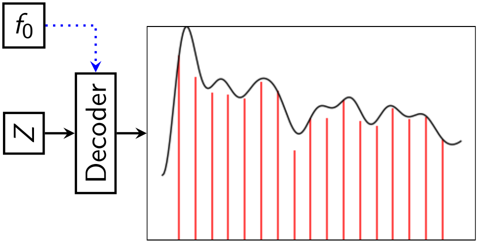
Generation

- Generate ('Synthesize') unseen/untrained pitch



Generation

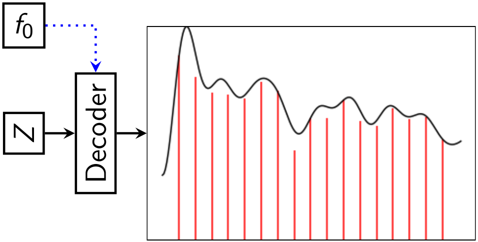
- Generate ('Synthesize') unseen/untrained pitch



- Random walk in latent space to coherently sample envelopes [Blaauw and Bonada, 2016]

Generation

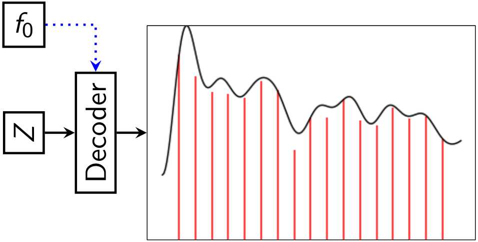
- Generate ('Synthesize') unseen/untrained pitch



- Random walk in latent space to coherently sample envelopes [Blaauw and Bonada, 2016]
- Skip MIDI 65 $\rightarrow$ Generate MIDI 65

Generation

- Generate ('Synthesize') unseen/untrained pitch



- Random walk in latent space to coherently sample envelopes [Blaauw and Bonada, 2016]
- Skip MIDI 65 → Generate MIDI 65
  1 [10] 2 [11] 3 [12]

Putting it all together

✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples 'timbre' and 'pitch', network models inter-dependencies

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples ‘timbre’ and ‘pitch’, network models inter-dependencies
- ✓ Pitch conditioning allows generation of spectral envelope for that pitch, thus enabling us to vary the pitch contour continuously and obtain coherent envelopes (and thus audio!)

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples ‘timbre’ and ‘pitch’, network models inter-dependencies
- ✓ Pitch conditioning allows generation of spectral envelope for that pitch, thus enabling us to vary the pitch contour continuously and obtain coherent envelopes (and thus audio!)

But . . .

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples 'timbre' and 'pitch', network models inter-dependencies
- ✓ Pitch conditioning allows generation of spectral envelope for that pitch, thus enabling us to vary the pitch contour continuously and obtain coherent envelopes (and thus audio!)

But . . .

- × No residual modeling

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples 'timbre' and 'pitch', network models inter-dependencies
- ✓ Pitch conditioning allows generation of spectral envelope for that pitch, thus enabling us to vary the pitch contour continuously and obtain coherent envelopes (and thus audio!)

But . . .

- × No residual modeling
- × No dynamics (timbre might change with loudness as well!)

Putting it all together

- ✓ Autoencoder frameworks in generative models for audio synthesis of instrumental tones
- ✓ A parametric representation decouples 'timbre' and 'pitch', network models inter-dependencies
- ✓ Pitch conditioning allows generation of spectral envelope for that pitch, thus enabling us to vary the pitch contour continuously and obtain coherent envelopes (and thus audio!)

But . . .

- × No residual modeling
- × No dynamics (timbre might change with loudness as well!)
- × No temporality

Concluding Thoughts

- Would like to build a system that can synthesize melodic elements from Carnatic Music

Concluding Thoughts

- Would like to build a system that can synthesize melodic elements from Carnatic Music $\boxed{1}$[13]

Concluding Thoughts

- Would like to build a system that can synthesize melodic elements from Carnatic Music
- To the best of our knowledge, we have not come across any work using a parametric model for musical tones in the neural synthesis framework, especially exploiting the conditioning function of the CVAE!

Concluding Thoughts

- Would like to build a system that can synthesize melodic elements from Carnatic Music
- To the best of our knowledge, we have not come across any work using a parametric model for musical tones in the neural synthesis framework, especially exploiting the conditioning function of the CVAE!
- All of our code/audio examples are available `https://github.com/SubramaniKrishna/VaPar-Synth`


рахмат
danke
謝謝
ngiyabonga
spas
Баярлалаа
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
tapadh leat
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràcie
hvala
köszönöm
akun
dankon
aciu
murakoze
djiere dieuf
tau
mochchakkeram
enkosi
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
terima kasih
kam sah hamnida
rahmat
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci

References I

[Blaauw and Bonada, 2016] Blaauw, M. and Bonada, J. (2016).
Modeling and transforming speech using variational autoencoders.
In *Interspeech*, pages 1770–1774.

[Caetano and Rodet, 2012] Caetano, M. and Rodet, X. (2012).
A source-filter model for musical instrument sound transformation.
In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 137–140. IEEE.

[Défossez et al., 2018] Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., and Bach, F. (2018).
Sing: Symbol-to-instrument neural generator.
In *Advances in Neural Information Processing Systems*, pages 9041–9051.

[Doersch, 2016] Doersch, C. (2016).
Tutorial on variational autoencoders.
arXiv preprint arXiv:1606.05908.

[Engel et al., 2020] Engel, J., Hantrakul, L., Gu, C., and Roberts, A. (2020).
Ddsp: Differentiable digital signal processing.
arXiv preprint arXiv:2001.04643.

References II

[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.

[Esling et al., 2018] Esling, P., Bitton, A., et al. (2018).
Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics.
arXiv preprint arXiv:1805.08501.

[Griffin and Lim, 1984] Griffin, D. and Lim, J. (1984).
Signal estimation from modified short-time fourier transform.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2):236–243.

[Hinton and Salakhutdinov, 2006] Hinton, G. E. and Salakhutdinov, R. R. (2006).
Reducing the dimensionality of data with neural networks.
science, 313(5786):504–507.

[IMAI, 1979] IMAI, S. (1979).
Spectral envelope extraction by improved cepstrum.
IEICE, 62:217–228.

References III

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013).
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114.

[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.

[Roche et al., 2018] Roche, F., Hueber, T., Limier, S., and Girin, L. (2018).
Autoencoders for music sound modeling: a comparison of linear, shallow, deep, recurrent and variational models.
arXiv preprint arXiv:1806.04096.

[Roebel and Rodet, 2005] Roebel, A. and Rodet, X. (2005).
Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation.
In *International Conference on Digital Audio Effects*, pages 30–35, Madrid, Spain. cote interne IRCAM: Roebel05b.

[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 IV

[Sarroff and Casey, 2014] Sarroff, A. M. and Casey, M. A. (2014).
Musical audio synthesis using autoencoding neural nets.
In *ICMC*.

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

[Slawson, 1981] Slawson, W. (1981).
The color of sound: a theoretical study in musical timbre.
Music Theory Spectrum, 3:132–141.

[Sohn et al., 2015] Sohn, K., Lee, H., and Yan, X. (2015).
Learning structured output representation using deep conditional generative models.
In *Advances in neural information processing systems*, pages 3483–3491.

[Wyse, 2018] Wyse, L. (2018).
Real-valued parametric conditioning of an rnn for interactive sound synthesis.
arXiv preprint arXiv:1805.10808.

Audio examples description I

1. Input MIDI 63 to Spectral Model
2. Spectral Model Reconstruction(trained on MIDI63)
3. Spectral Model Reconstruction(not trained on MIDI63)
4. Input MIDI 60 note to Parametric Model
5. Parametric Reconstruction of input note
6. Input MIDI 63 Note
7. Parametric CVAE reconstruction of input
8. Input MIDI 65 note(endpoint trained model)
9. Parametric CVAE reconstruction of input(endpoint trained model)
10. CVAE Generated MIDI 65 Violin note
11. Similar MIDI 65 Violin note from dataset
12. CVAE Generated MIDI 65 Violin note with vibrato
13. Carnatic Violin Melody