Variational Parametric Models for Audio Synthesis

Krishna Subramani
Guide: Prof. Preeti Rao

Department of Electrical Engineering
IIT Bombay, India

DDP Presentation

Audio
Synthesis

Audio
Synthesis

Audio
Synthesis

Audio Synthesis

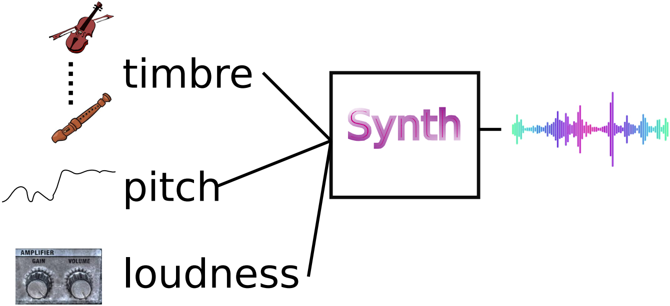
?

Audio
Synthesis

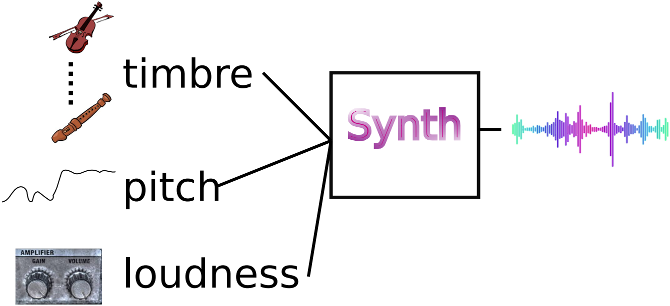
?

Data-driven Statistical Modeling
Abundant Computing Power
DL for Audio Synthesis!

Generative Synth

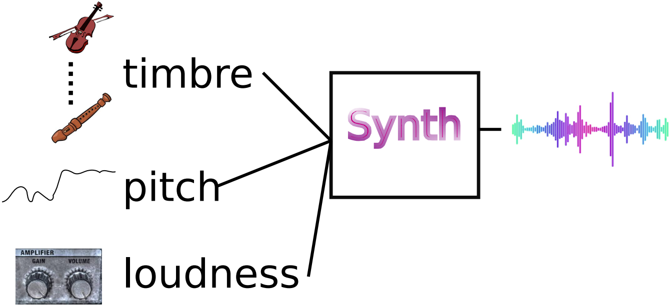


Generative Synth



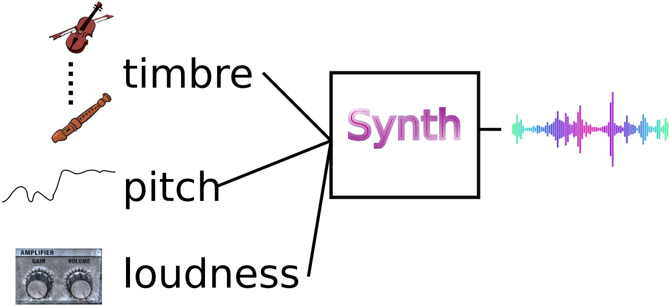
- timbre $\rightarrow$ "difference" between a violin and flute A4

Generative Synth



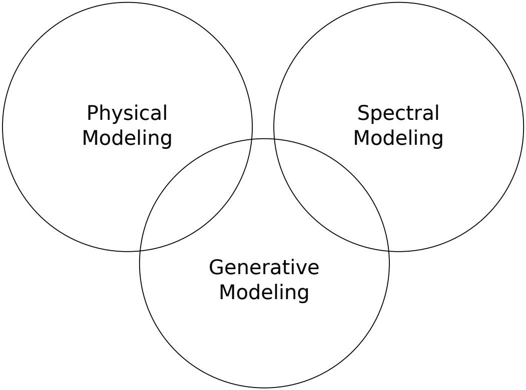
- timbre → “difference” between a violin and flute A4
- pitch → fundamental frequency

Generative Synth

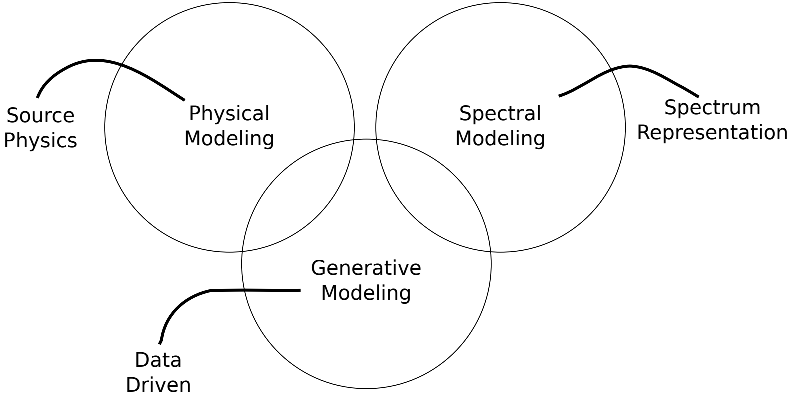


- timbre → "difference" between a violin and flute A4
- pitch → fundamental frequency
- loudness → intensity (energy)

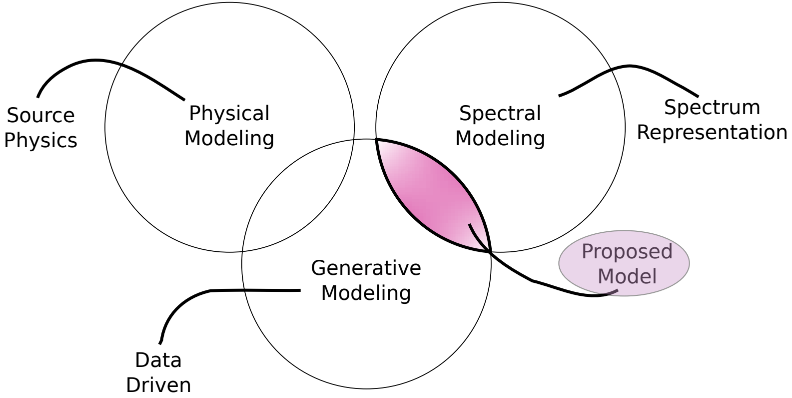
Audio Synthesis



Audio Synthesis



Audio Synthesis



Spectral Modeling Synthesis

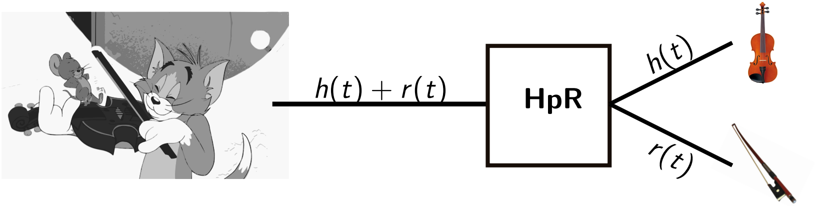
- Introduced by [Serra, 1989, Serra et al., 1997]

Spectral Modeling Synthesis

- Introduced by [Serra, 1989, Serra et al., 1997]
- Idea is: $\boldsymbol{x} = \boldsymbol{x_{sine}} + \boldsymbol{x_{noise}} = h(t) + r(t)$

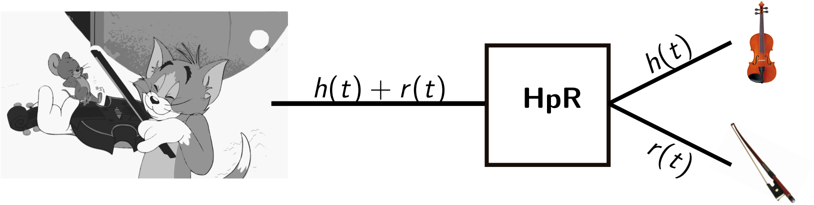
Spectral Modeling Synthesis

- Introduced by [Serra, 1989, Serra et al., 1997]
- Idea is: $\boldsymbol{x} = \boldsymbol{x_{sine}} + \boldsymbol{x_{noise}} = h(t) + r(t)$



Spectral Modeling Synthesis

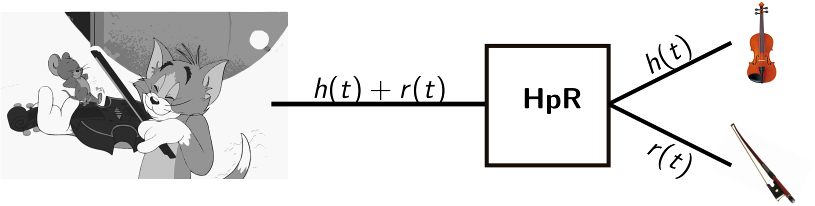
- Introduced by [Serra, 1989, Serra et al., 1997]
- Idea is: $\boldsymbol{x} = \boldsymbol{x_{sine}} + \boldsymbol{x_{noise}} = h(t) + r(t)$



$\boxed{1}$ [1] $\boxed{2}$ [2] $\boxed{3}$ [3]

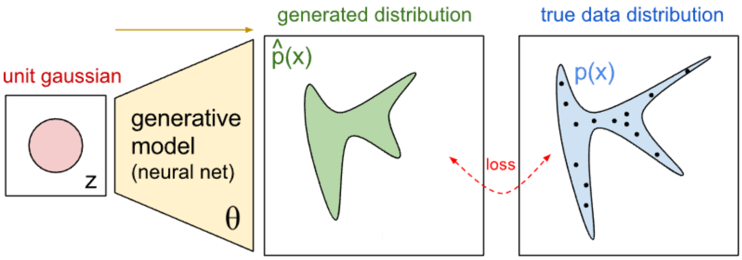
Spectral Modeling Synthesis

- Introduced by [Serra, 1989, Serra et al., 1997]
- Idea is: $\boldsymbol{x} = \boldsymbol{x}_{\boldsymbol{sine}} + \boldsymbol{x}_{\boldsymbol{noise}} = h(t) + r(t)$



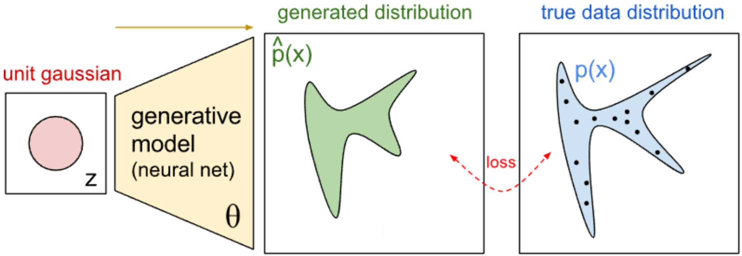
- Our parametric representation is a Source-Filter inspired representation, building on top of the HpR model [Caetano and Rodet, 2012, Caetano and Rodet, 2013]

Generative Models for Audio



[1] https://openai.com/blog/generative-models/

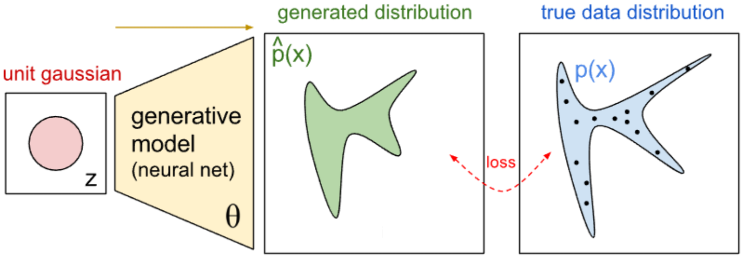
Generative Models for Audio



- Compact representation of data space[1], simultaneously allowing us to sample from it

[1] https://openai.com/blog/generative-models/

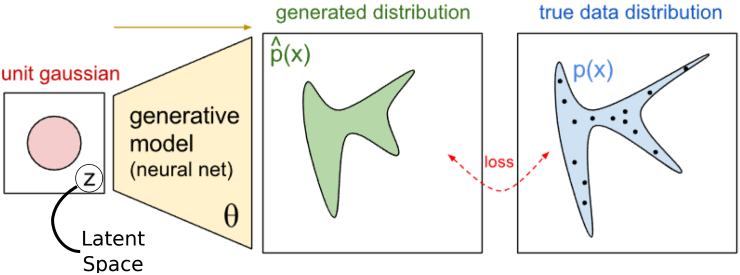
Generative Models for Audio



- Compact representation of data space[1], simultaneously allowing us to sample from it
- This 'Compact' representation → Latent Space

[1] https://openai.com/blog/generative-models/

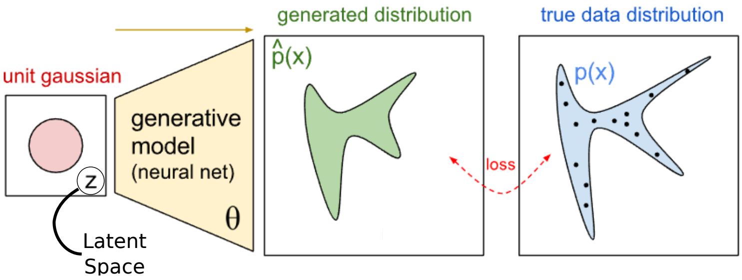
Generative Models for Audio



- Compact representation of data space[1], simultaneously allowing us to sample from it
- This 'Compact' representation → Latent Space

[1] https://openai.com/blog/generative-models/

Generative Models for Audio



- Compact representation of data space[1], simultaneously allowing us to sample from it
- This 'Compact' representation → Latent Space
- **Neural Audio Synthesis** [Engel et al., 2017]

[1] https://openai.com/blog/generative-models/

Generative Models

- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]

Generative Models

- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]
- Generative Adversarial Networks [Goodfellow et al., 2014]

Generative Models

- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]
- Generative Adversarial Networks [Goodfellow et al., 2014]
- Framewise Autoencoding

Generative Models

- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]
- Generative Adversarial Networks [Goodfellow et al., 2014]
- Framewise Autoencoding
  - Autoencoders (AE) [Hinton and Salakhutdinov, 2006]
    Optimal (MSE) lower dimensional representation of input

Generative Models

- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]
- Generative Adversarial Networks [Goodfellow et al., 2014]
- Framewise Autoencoding
  - 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

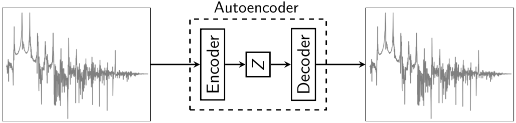
- Sequential/Autoregressive Modeling: LSTMs, WaveNet [Hochreiter and Schmidhuber, 1997, Oord et al., 2016]
- Generative Adversarial Networks [Goodfellow et al., 2014]
- Framewise Autoencoding
  - Autoencoders (AE) [Hinton and Salakhutdinov, 2006]
    Optimal (MSE) lower dimensional representation of 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 ...

Our Nearest Neighbors

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

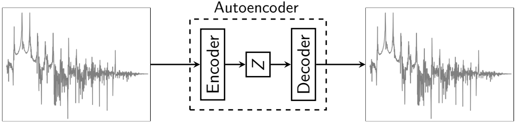
Our Nearest Neighbors

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



Our Nearest Neighbors

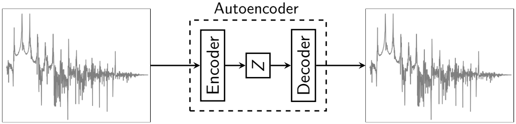
- [Sarroff and Casey, 2014] frame-wise reconstruction of short-time magnitude spectra with autoencoders



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

Our Nearest Neighbors

- [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 Neighbors

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

Our Nearest Neighbors

- 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 Neighbors

- 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 Neighbors

- 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?

- Mostly use audio in raw form (waveform/spectrum)
  - waveform: Complicated architectures, lots of training data, long training times
  - spectrum: Phase estimation

Why Parametric?

- Mostly use audio in raw form (waveform/spectrum)
  - waveform: Complicated architectures, lots of training data, long training times
  - spectrum: Phase estimation
- Parametric?

Why Parametric?

- Mostly use audio in raw form (waveform/spectrum)
  - waveform: Complicated architectures, lots of training data, long training times
  - spectrum: Phase estimation
- Parametric?
  - model a reduced parameter space over waveform/spectrum
  - neural network to generatively model parametric space $\rightarrow$ simple architecture, less data, better generalizability and high quality audio!

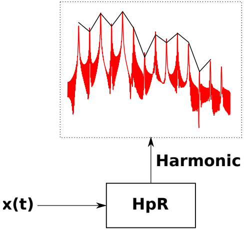
Why Parametric?

- Mostly use audio in raw form (waveform/spectrum)
  - waveform: Complicated architectures, lots of training data, long training times
  - spectrum: Phase estimation
- Parametric?
  - model a reduced parameter space over waveform/spectrum
  - neural network to generatively model parametric space → simple architecture, less data, better generalizability and high quality audio!
    - [Blaauw and Bonada, 2016] used a vocoder representation to train a generative model for speech synthesis
    - [Engel et al., 2020] (DDSP) recently proposed the control of a parametric model based on a deterministic autoencoder

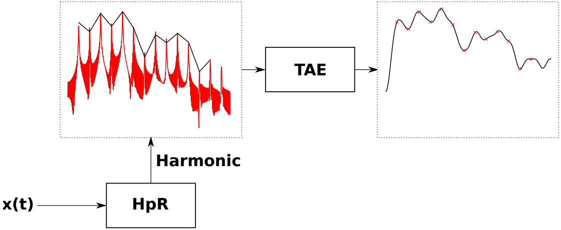
VaPar Synth

x(t)

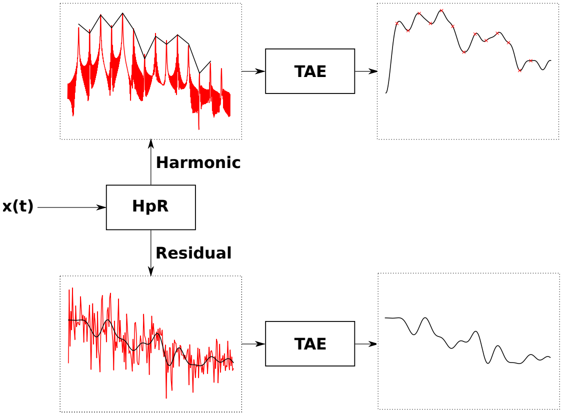
VaPar Synth



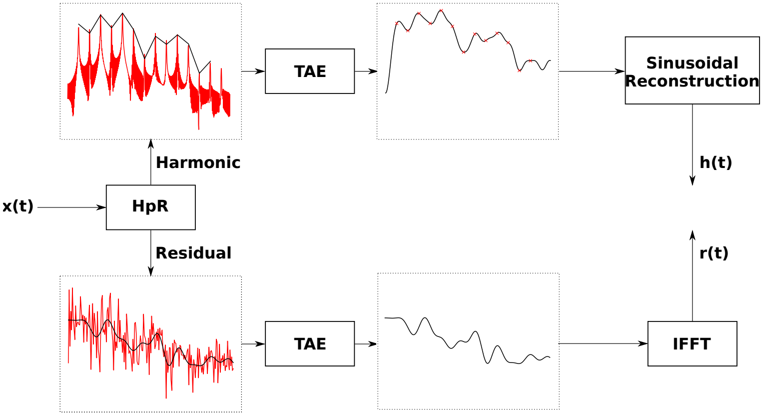
VaPar Synth



VaPar Synth

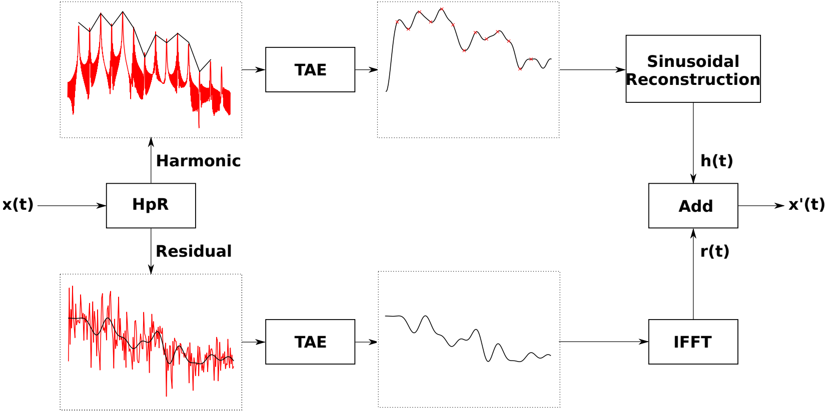


VaPar Synth



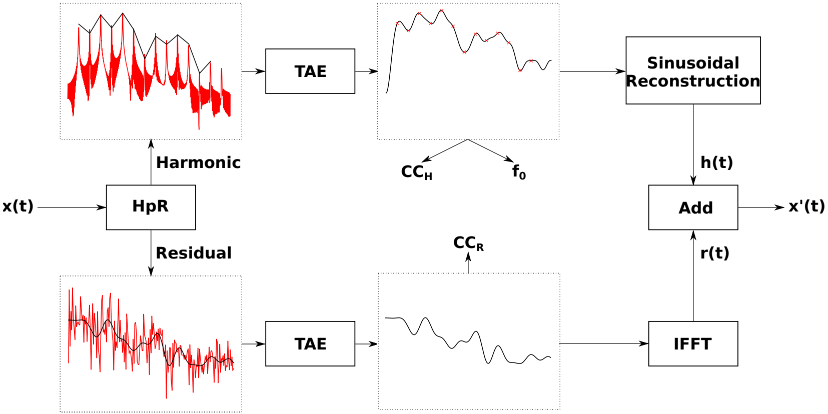
TAE
Sinusoidal Reconstruction
Harmonic
h(t)
x(t)
HpR
Residual
r(t)
TAE
IFFT

VaPar Synth



TAE
Sinusoidal Reconstruction
Harmonic
h(t)
x(t)
HpR
Add
x'(t)
Residual
r(t)
TAE
IFFT

VaPar Synth

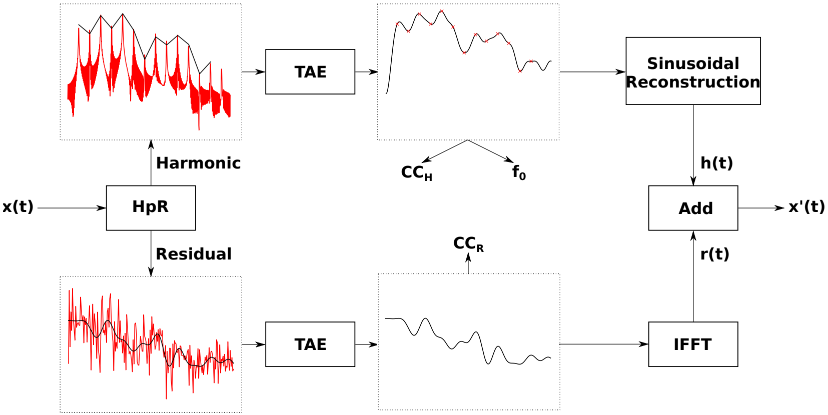


TAE
Sinusoidal Reconstruction
Harmonic
CC_H
f_0
h(t)
x(t)
HpR
Add
x'(t)
CC_R
Residual
r(t)
TAE
IFFT

VaPar Synth

TAE: [Caetano and Rodet, 2012, IMAI, 1979]

Subsampling rates: [Caetano and Rodet, 2013, Serra et al., 1997]



Datasets

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

Datasets

NSynth

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

11 Instruments
MIDI pitch, velocity
Large Number

Datasets

NSynth

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

11 Instruments
MIDI pitch, velocity
Large Number

Datasets

NSynth

Good-sounds

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

11 Instruments
MIDI pitch, velocity
Large Number

Datasets

NSynth

Good-sounds

Individual Note/Scale recordings
Mezzo-forte, MIDI pitch
Initial Experiments

Why **Violin?**

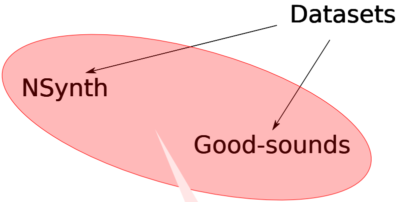
Popular in Indian Music, Human voice-like timbre,
Ability to produce continuous pitch!

Datasets

NSynth

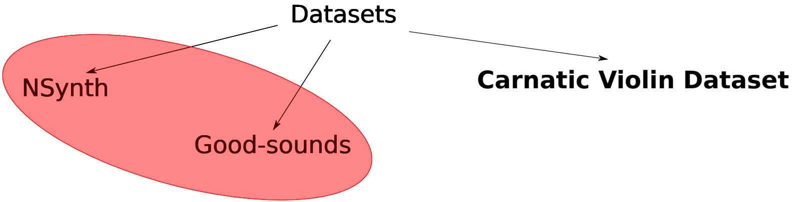
Good-sounds

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



Why insufficient?
MIDI pitches, not Carnatic notes
Not Expressive!

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



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

Datasets

NSynth

Good-sounds

Carnatic Violin Dataset

We record our own dataset!!

Carnatic Violin Dataset

Carnatic Note	Sa	Ri_1	Ri_2	Ga_2	Ga_3	Ma_1
Notation	**Sa**	**Ri1**	**Ri2**	**Ga2**	**Ga3**	**Ma1**
Carnatic Note	Ma_2	Pa	Dha_1	Dha_2	Ni_2	Ni_3
Notation	**Ma2**	**Pa**	**Dha1**	**Dha2**	**Ni2**	**Ni3**

	Description	Notation
Octave	Lower, Middle, Upper	**L, M, U**
Loudness	Soft, Loud	**So, Lo**
Style	Smooth, Attack	**Sm, At**

1. Fixed Notes: *1143* s across *363* instances
2. Raga Recordings: *1075* s with *113* s Gamakas

Carnatic Violin Dataset

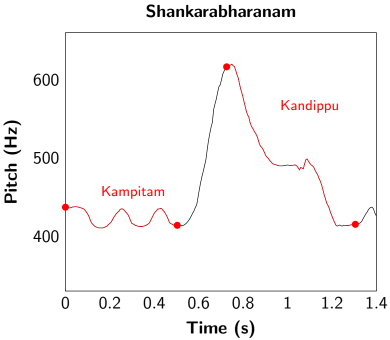
Carnatic Note	Sa	Ri_1	Ri_2	Ga_2	Ga_3	Ma_1
Notation	**Sa**	**Ri1**	**Ri2**	**Ga2**	**Ga3**	**Ma1**
Carnatic Note	Ma_2	Pa	Dha_1	Dha_2	Ni_2	Ni_3
Notation	**Ma2**	**Pa**	**Dha1**	**Dha2**	**Ni2**	**Ni3**

	Description	Notation
Octave	Lower, Middle, Upper	**L, M, U**
Loudness	Soft, Loud	**So, Lo**
Style	Smooth, Attack	**Sm, At**

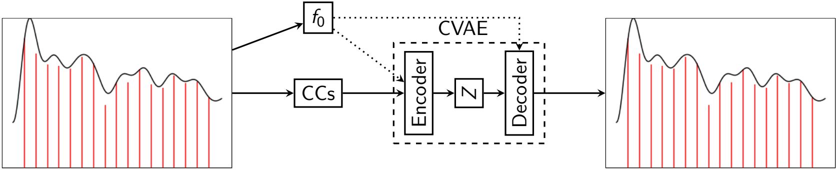
1. **Fixed Notes**: *1143* s across *363* instances
2. Raga Recordings: *1075* s with *113* s Gamakas

Gamakas?

- The subtle shadings of a tone, delicate nuances and inflections around a note that please and inspire the listener [Swift, 1990]
- Ornamentations/Deflections in pitch [SUBRAMANIAN, 2013]



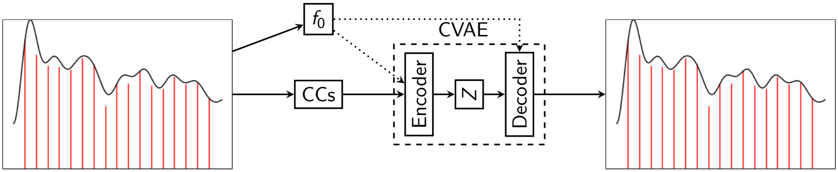
Network Architecture



- Similar network for Residual
- Hyperparameters optimized via MSE plots

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

Network Architecture



- Similar network for Residual
- Hyperparameters optimized via MSE plots

$L \propto MSE + \beta.KLD$

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

Experiments

- 3 questions about our modeling pipeline:

Experiments

- 3 questions about our modeling pipeline:
  1. Compatibility of our parametric model with violin audio

Experiments

- 3 questions about our modeling pipeline:
  1. Compatibility of our parametric model with violin audio
  2. Why CVAE? Why not VAE or AE instead?

Experiments

- 3 questions about our modeling pipeline:
  1. Compatibility of our parametric model with violin audio
  2. Why CVAE? Why not VAE or AE instead?
  3. Coherently modeling harmonic and residual components

Experiments

- 3 questions about our modeling pipeline:
  1. Compatibility of our parametric model with violin audio
  2. Why CVAE? Why not VAE or AE instead?
  3. Coherently modeling harmonic and residual components
- De-mystify these one-by-one ...

Parametric Model for Violin Audio

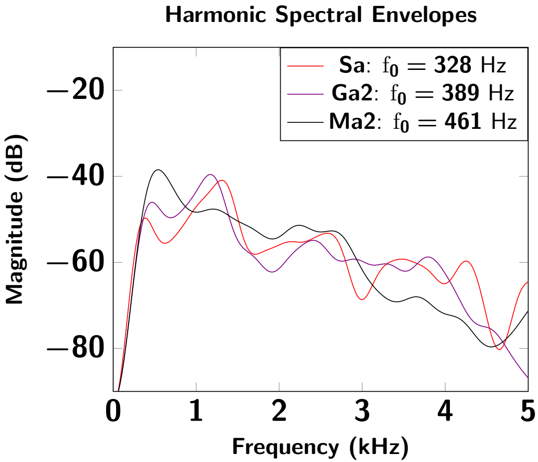
- [Beauchamp, 2017] $\rightarrow$ SF model for Violin

Parametric Model for Violin Audio

- [Beauchamp, 2017] $\rightarrow$ SF model for Violin
- Filter (Spectral Envelope) independent of Source (f_0)?

Parametric Model for Violin Audio

- [Beauchamp, 2017] → SF model for Violin
- Filter (Spectral Envelope) independent of Source (f_0)?



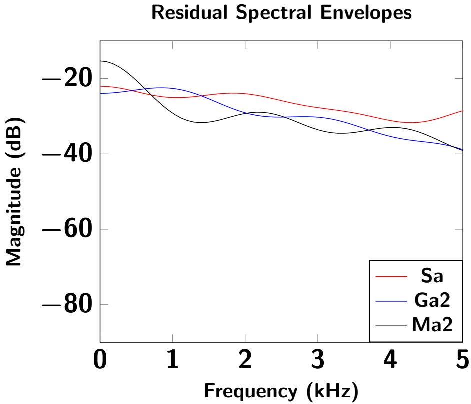
- [Beauchamp, 2017] states violin body (filter) has narrow resonances

- [Beauchamp, 2017] states violin body (filter) has narrow resonances
- f_0 dependent Sub-sampling of harmonic spectral envelope ($K_{CC} < \frac{F_s}{\mathrm{f}_0}$)

- [Beauchamp, 2017] states violin body (filter) has narrow resonances
- f_0 dependent Sub-sampling of harmonic spectral envelope $(K_{CC} < \frac{F_s}{\mathrm{f}_0})$
- Both the above lead to dependence of Harmonic spectral envelop on pitch

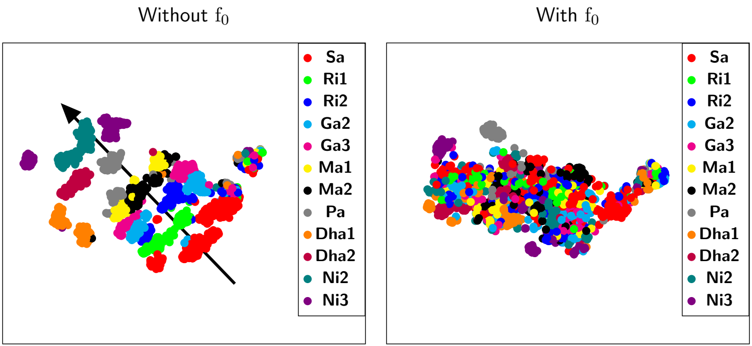
- [Beauchamp, 2017] states violin body (filter) has narrow resonances
- f_0 dependent Sub-sampling of harmonic spectral envelope ($K_{CC} < \frac{F_s}{\mathrm{f}_0}$)
- Both the above lead to dependence of Harmonic spectral envelop on pitch
- What about Residual envelope? Use constant sampling rate as suggested in [Serra et al., 1997]

- [Beauchamp, 2017] states violin body (filter) has narrow resonances
- f_0 dependent Sub-sampling of harmonic spectral envelope ($K_{CC} < \frac{F_s}{f_0}$)
- Both the above lead to dependence of Harmonic spectral envelop on pitch
- What about Residual envelope? Use constant sampling rate as suggested in [Serra et al., 1997]

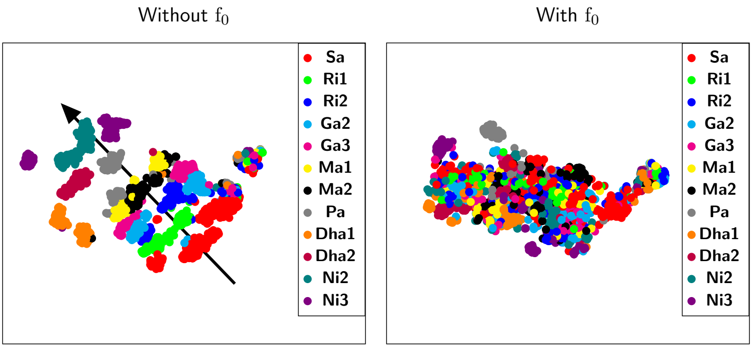


Why CVAE?

- Harmonic spectral envelope depends on pitch
- Conditioning on pitch $\implies$ Network captures dependencies between the timbre and the pitch $\implies$ More accurate envelope generation + Pitch control
- For better understanding, we also visualize the latent space using t-SNE [Maaten and Hinton, 2008]

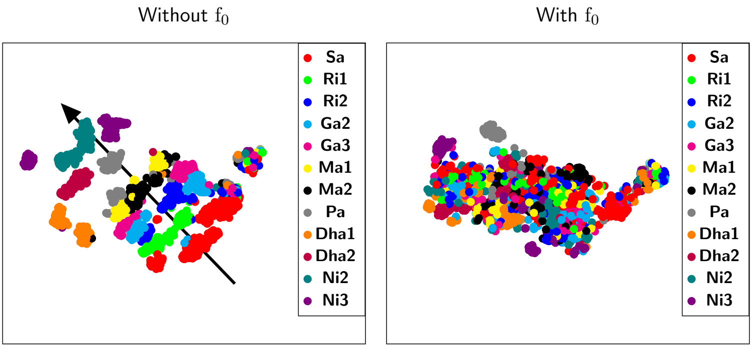


Harmonic CVAE Latent Spaces without and with f_0 conditioning



Harmonic CVAE Latent Spaces without and with f_0 conditioning

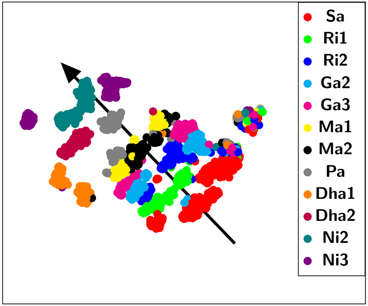
- Clear clustering without pitch conditioning $\implies$ latent space contains pitch information

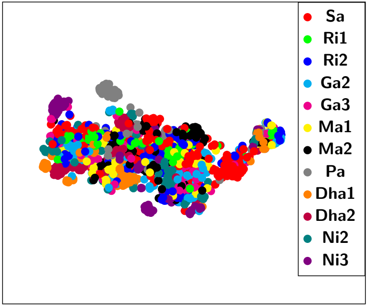


Harmonic CVAE Latent Spaces without and with f_0 conditioning

- Clear clustering without pitch conditioning $\implies$ latent space contains pitch information
- Pitch conditioning $\rightarrow$ optimal spectral envelope for that pitch

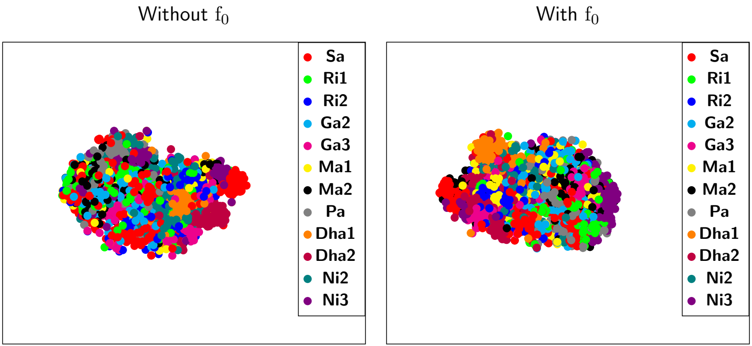
Without f_0 | With f_0



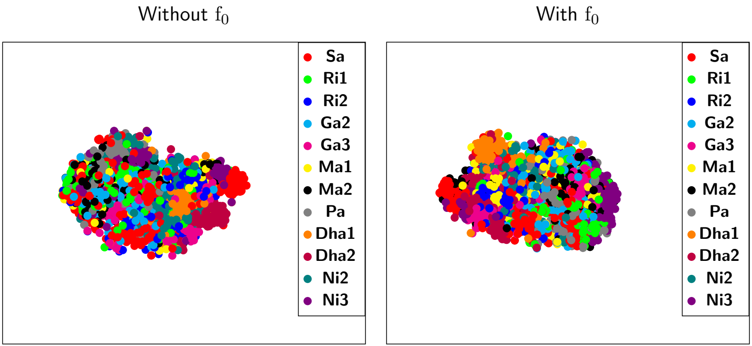


Harmonic CVAE Latent Spaces without and with f_0 conditioning

- Clear clustering without pitch conditioning $\Longrightarrow$ latent space contains pitch information
- Pitch conditioning $\rightarrow$ optimal spectral envelope for that pitch
- What about residual spectral envelope?

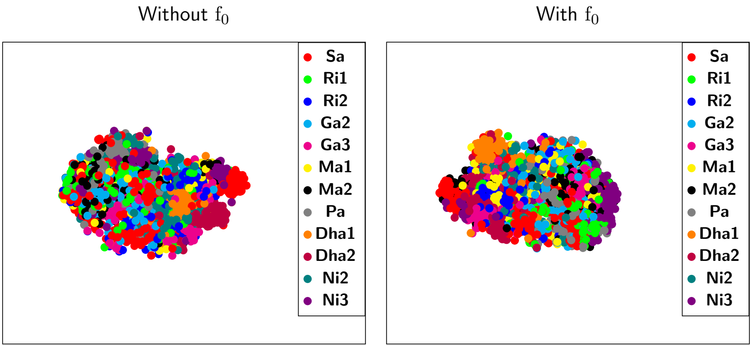


Residual VAE Latent Spaces without and with f_0 conditioning



Residual VAE Latent Spaces without and with f_0 conditioning

- Matches with previous plots of spectral envelope



Residual VAE Latent Spaces without and with f_0 conditioning

- Matches with previous plots of spectral envelope
- No conditioning needed for the residual network!

- Established 2 things so far . . .

- Established 2 things so far ...
  1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies

- Established 2 things so far . . .
  1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies
  2. Residual spectral envelope does not depend on the pitch $\implies$ No conditioning needed

- Established 2 things so far ...
  1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies
  2. Residual spectral envelope does not depend on the pitch $\implies$ No conditioning needed
- Quanititatively verify?

- Established 2 things so far ...
    1. Harmonic spectral envelope depends on the pitch $\Longrightarrow$ CVAE models inter-dependencies
    2. Residual spectral envelope does not depend on the pitch $\Longrightarrow$ No conditioning needed
- Quanititatively verify? **Reconstruction Experiments!!**

- Established 2 things so far ...
    1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies
    2. Residual spectral envelope does not depend on the pitch $\implies$ No conditioning needed
- Quanititatively verify? **Reconstruction Experiments!!**

Reconstruction?

- Established 2 things so far ...
  1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies
  2. Residual spectral envelope does not depend on the pitch $\implies$ No conditioning needed
- Quanititatively verify? **Reconstruction Experiments!!**

Reconstruction?

- Omit pitch instances during training and reconstruct their spectral envelopes

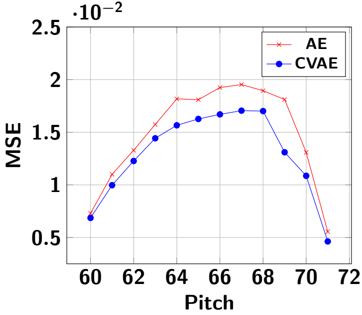
- Established 2 things so far ...
    1. Harmonic spectral envelope depends on the pitch $\implies$ CVAE models inter-dependencies
    2. Residual spectral envelope does not depend on the pitch $\implies$ No conditioning needed
- Quanititatively verify? **Reconstruction Experiments!!**

Reconstruction?

- Omit pitch instances during training and reconstruct their spectral envelopes
- Network's generalization ability to unseen pitches

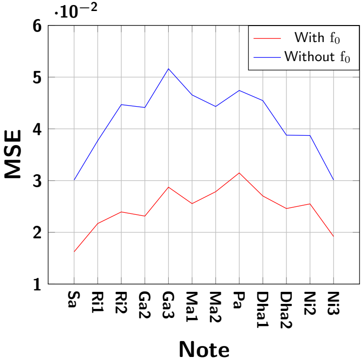
- [Subramani et al., 2020] train on octave endpoints, and reconstruct intermediate harmonic spectral envelopes (Good-sounds)

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

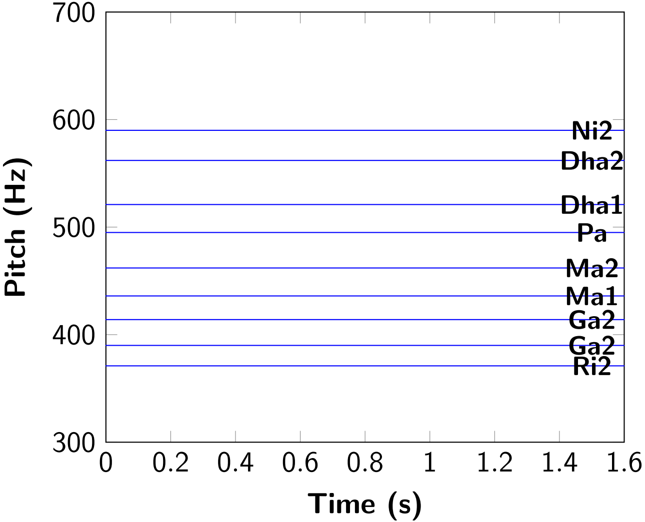


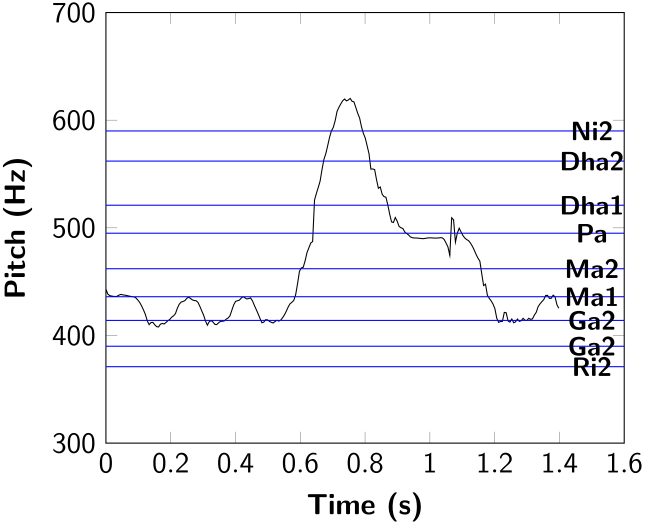
- Conditioning captures the pitch dependency of the spectral envelope more accurately

- Similar experiment with our Carnatic Violin dataset

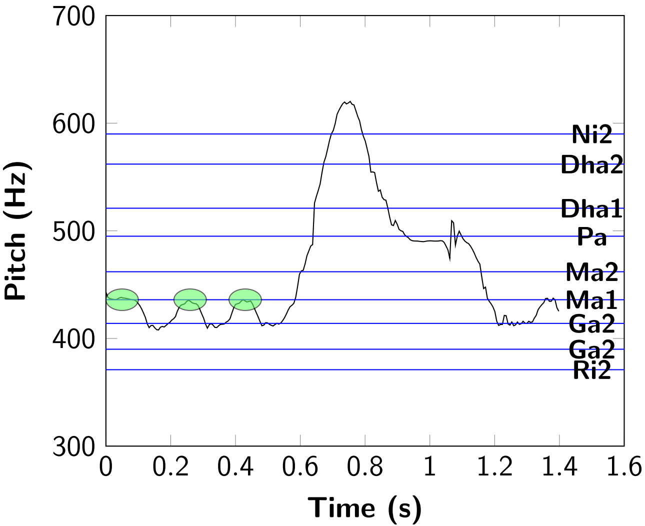


- Pitch conditioning $\rightarrow$ continuous pitch control

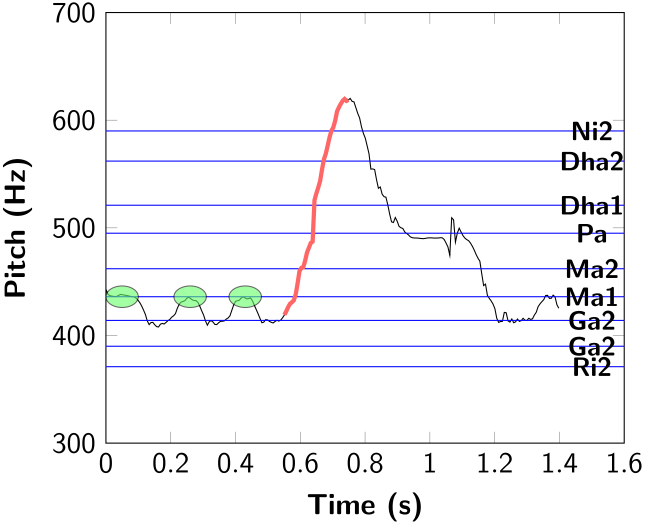

Pitch (Hz)
700
600
500
400
300
Ni2
Dha2
Dha1
Pa
Ma2
Ma1
Ga2
Ga2
Ri2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Time (s)



700
600
500
400
300
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Pitch (Hz)
Time (s)
Ni2
Dha2
Dha1
Pa
Ma2
Ma1
Ga2
Ga2
Ri2



700
600
500
400
300
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Pitch (Hz)
Time (s)
Ni2
Dha2
Dha1
Pa
Ma2
Ma1
Ga2
Ga2
Ri2



700
600
500
400
300
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Pitch (Hz)
Time (s)
Ni2
Dha2
Dha1
Pa
Ma2
Ma1
Ga2
Ga2
Ri2

- Repeat reconstruction with these continuously varying pitch contours, but **only trained on the fixed pitch notes**

$\boxed{1}$ [4] $\boxed{2}$ [5]

Joint Modeling of harmonic,residual

- Why jointly model?

Joint Modeling of harmonic,residual

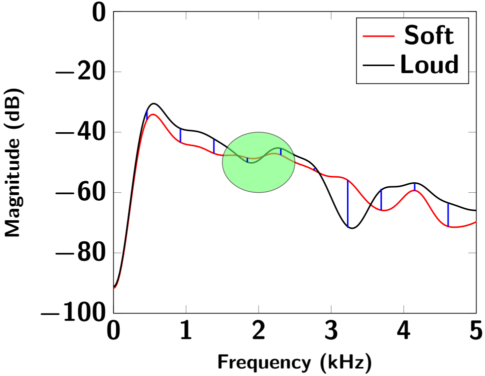
- Why jointly model?
  - [Mathews and Kohut, 1973] ‘Resonant Enhancement’ $\rightarrow$ Violin resonances filter String vibrations

Joint Modeling of harmonic,residual

- Why jointly model?
  - [Mathews and Kohut, 1973] 'Resonant Enhancement' $\rightarrow$ Violin resonances filter String vibrations
  - Harmonic (string vibrations) and residual (bow noise) processed by same resonance $\implies$ not independent

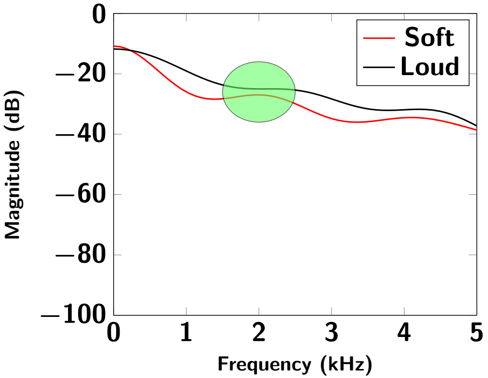
Joint Modeling of harmonic,residual

- Why jointly model?
  - [Mathews and Kohut, 1973] 'Resonant Enhancement' $\rightarrow$ Violin resonances filter String vibrations
  - Harmonic (string vibrations) and residual (bow noise) processed by same resonance $\implies$ not independent

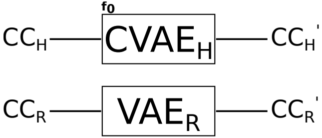


Joint Modeling of harmonic,residual

- Why jointly model?
  - [Mathews and Kohut, 1973] 'Resonant Enhancement' $\rightarrow$ Violin resonances filter String vibrations
  - Harmonic (string vibrations) and residual (bow noise) processed by same resonance $\implies$ not independent

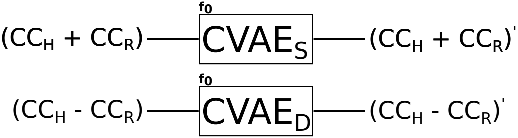


- Harmonic and Residual envelopes could be dependent $\rightarrow$ common underlying origin in the played loudness style of the note

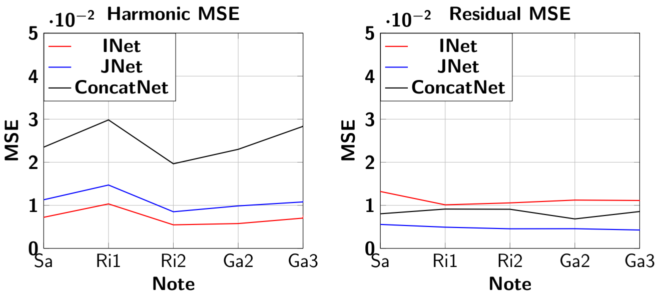


Independent Modeling (INet)

Concatenative Modeling (ConcatNet)



Modeling sum and difference (JNet)

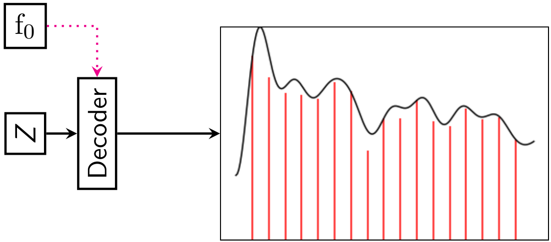


- [Fletcher et al., 1965] mentions that the perceptual impact of the residual is more for higher frequency notes than for lower ones
- Harmonic MSE lower for INet
- Residual MSE lower for joint modeling

1 [6] 2 [7]

Generation

- Interested in using network to 'generate' audio



- How to sample points from Latent Space?
  [Blaauw and Bonada, 2016] performs a random walk with small step size
- Not a good enough emulation of temporal order of frames

1 [8] 2 [9] 3 [10]

Listening Tests

- MSE not perceptually representative $\implies$ Listening tests

Listening Tests

- MSE not perceptually representative $\implies$ Listening tests
- 2 Professionally trained ($\approx$ 15 years) violinists

Listening Tests

- MSE not perceptually representative $\implies$ Listening tests
- 2 Professionally trained ($\approx$ 15 years) violinists
- Present audio examples, take subjective feedback

Listening Tests

- MSE not perceptually representative $\implies$ Listening tests
- 2 Professionally trained ($\approx$ 15 years) violinists
- Present audio examples, take subjective feedback
    1. Reconstruction: Network reconstruction realistic, difficult to differentiate from actual audio

Listening Tests

- MSE not perceptually representative $\implies$ Listening tests
- 2 Professionally trained ($\approx$ 15 years) violinists
- Present audio examples, take subjective feedback
  1. Reconstruction: Network reconstruction realistic, difficult to differentiate from actual audio
  2. Generation: Network generated audio not like a violin, sounds synthetic, even with vibrato

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!)
- ✓ Joint modeling can potentially help in modeling residual better

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!)
- ✓ Joint modeling can potentially help in modeling residual better

But . . .

- × No temporality

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!)
- ✓ Joint modeling can potentially help in modeling residual better

But . . .

- × No temporality
- × Network generated/synthesized audio not realistic

Contributions

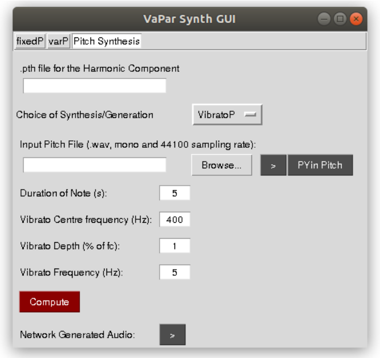
- Published and presented work in ICASSP 2020, ISMIR 2019 and submitted work to ISMIR 2020 [Subramani et al., 2020, Subramani et al., 2019]
- Dataset + Code open source on GitHub[2][3]

[2] https://github.com/SubramaniKrishna/VaPar-Synth

[3] https://github.com/SubramaniKrishna/HpRNet

Contributions

- Published and presented work in ICASSP 2020, ISMIR 2019 and submitted work to ISMIR 2020 [Subramani et al., 2020, Subramani et al., 2019]
- Dataset + Code open source on GitHub[2][3]
- GUI for researchers to better understand our research 1 [11]



[2] https://github.com/SubramaniKrishna/VaPar-Synth
[3] https://github.com/SubramaniKrishna/HpRNet

References I

[Beauchamp, 2017] Beauchamp, J. W. (2017).
Comparison of vocal and violin vibrato with relationship to the source/filter model.
In *Studies in Musical Acoustics and Psychoacoustics*, pages 201–221. Springer.

[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.

[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.

[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.

References II

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

[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.

[Fletcher et al., 1965] Fletcher, H., Blackham, E. D., and Geertsen, O. N. (1965).
Quality of violin, viola,'cello, and bass-viol tones. i.
The Journal of the Acoustical Society of America, 37(5):851–863.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
In *Advances in neural information processing systems*, pages 2672–2680.

References III

[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.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.
Neural computation, 9(8):1735–1780.

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

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

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008).
Visualizing data using t-sne.
Journal of machine learning research, 9(Nov):2579–2605.

[Mathews and Kohut, 1973] Mathews, M. V. and Kohut, J. (1973).
Electronic simulation of violin resonances.
The Journal of the Acoustical Society of America, 53(6):1620–1626.

References IV

[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.

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

[Serra, 1989] Serra, X. (1989).
A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.
Ph.D. Thesis, Stanford University.

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

References V

[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.

[Subramani et al., 2019] Subramani, K., D'Hooge, A., and Rao, P. (2019).
Generative audio synthesis with a parametric model.
arXiv preprint arXiv:1911.08335.

[Subramani et al., 2020] Subramani, K., Rao, P., and D'Hooge, A. (2020).
Vapar synth - a variational parametric model for audio synthesis.
In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 796–800.

[SUBRAMANIAN, 2013] SUBRAMANIAN, S. K. (2013).
Modelling gamakas of carnatic music as a synthesizer for sparse prescriptive notation.

[Swift, 1990] Swift, G. N. (1990).
South indian "gamaka" and the violin.
Asian Music, 21(2):71–89.

[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. Original Sa note
2. Original Sa note harmonic
3. Original Sa note residual
4. Harmonic version of Gamaka
5. Network reconstruction of harmonic version of Gamaka
6. Upper Octave Ri1 recording
7. Upper Octave Ri1 INet reconstruction
8. Network Generated Upper octave Ri2
9. Network Generated Upper octave Ri2 with vibrato
10. Network Generated Gamaka
11. Bohemian Rhapsody Guitar 'Rendered' by our network