Audio Synthesizer Control using Deep Generative Models and Normalizing Flows

Naotake Masuda, University of Tokyo

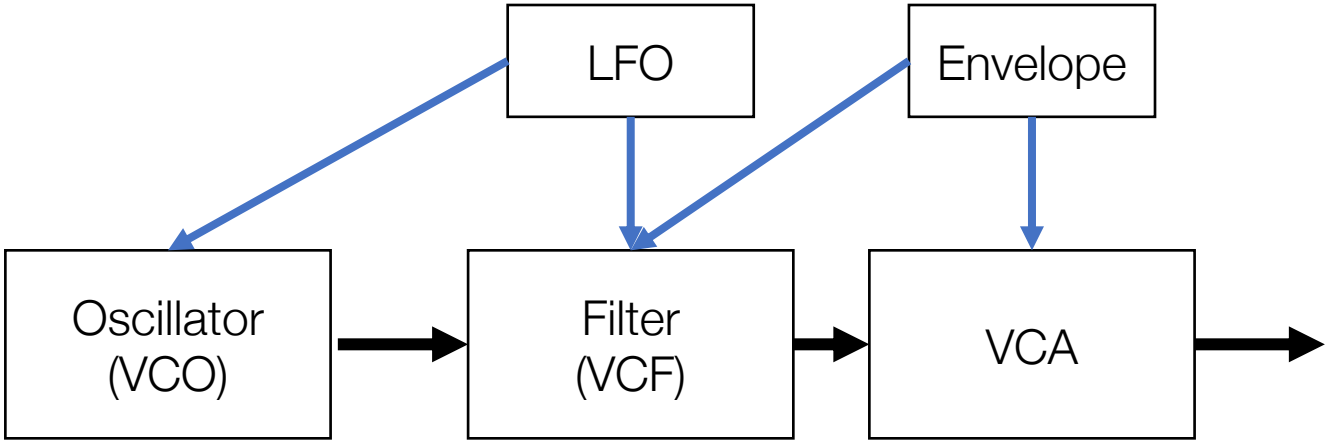
2020/1/15

Synthesizers

- Both analog synthesizers and software synthesizers are often used in music production
- They can produce a wide variety of sounds
- They have many (>100) parameters to control the sound

Subtractive Synthesis

- Most common way of synthesizing audio
- Uses filters to shape the sound



Subtractive Synthesis

Sound we want to create (Hoover Bass)

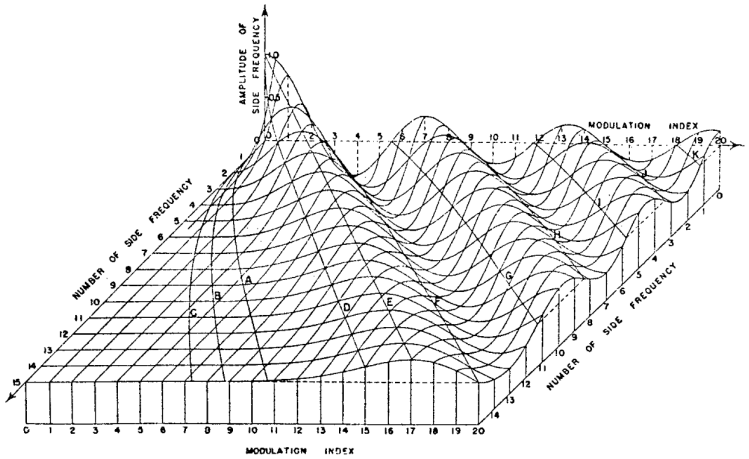
Instructions

- In PartA: Set Oscillator 1 to 4 voices, Retrigerr off, Stereo to 5, Volume to 8, Detune to 3~4, Octave to -3, FINE to -0.07
- In Oscillator 2, set WAVE to SAW, voices to 8, RETRIG OFF, stereo to 6~7, Detune to 4~5, Octave to -1, FINE 0.3
- In amplitude envelope A, set decay to max and release to 1
- In filter A, set input to AB, and filter type to low pass, resonance to 1-2, Cut-off to 20HZ, drive to around 1.8
- In part B, set Oscillator 1 to 8 voices, RETRIG off, detune to 3-4, stereo to 5, fine -0.07
- In Oscillator 2, set WAVE to SAW, 8 voices, RETRIG off, Detune to 4-5, Stereo to 5, FINE to 0.3, Octave to +1
- In amplitude envelope B, set decay to max and release to 1
- Set input to BA and filter type to low pass, resonance to 3~4, cut-off to 30Hz, Drive to 1.8
- In Filter Control, set cut-off to max
- In MOD ENV1 select PitchAB, set xMODEnv1 Dest1 Am to 0.76, Attack to 1.38, Decay to 4.2, Sustain to 1.9
- ……

http://vreap.net/music/haw-to-make-hoover-sound/

FM Synthesis

- Used in Yamaha DX7, etc.
- Small changes in the parameter can change the sound drastically
- Spectrogram is decided by Bessel functions…
- Very difficult

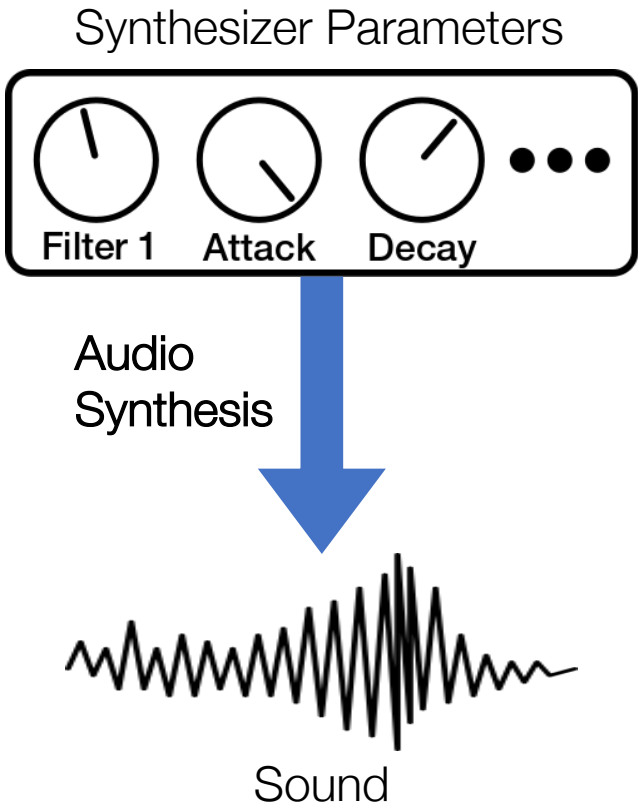


Problems with Synthesizers

- There are thousands of synthesizers but every synthesizer has different controls with different synthesis algorithms

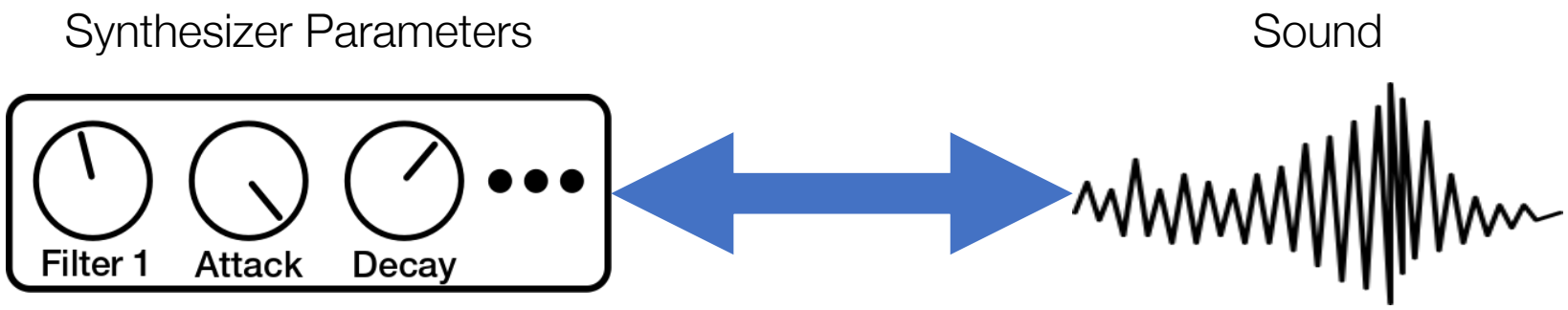
Synthesizers as a function

- Synthesizers can be thought of as a **mapping** from parameters to sounds.
- Parameter's relation to the sound are:
  - **Non-linear**
  - **Sparse** (some knobs are important, while others might do nothing to change the sound)



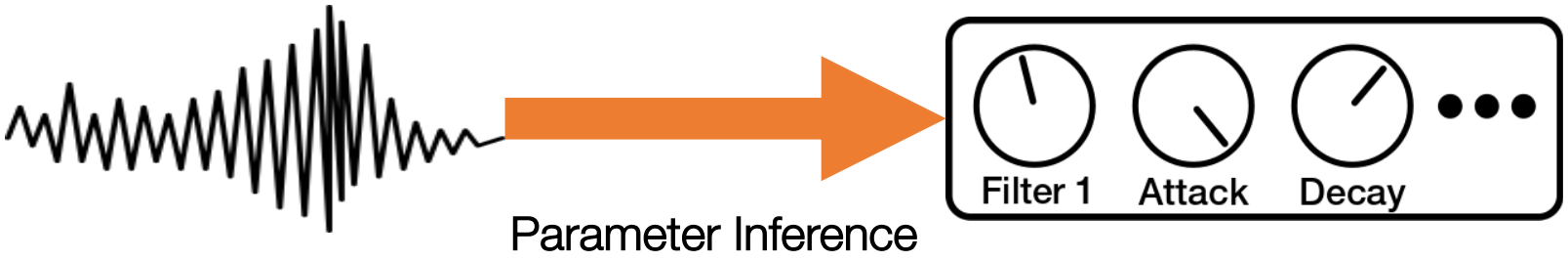
Synthesizer Control Learning

- We aim to learn the relationship between parameters and resulting sound.
- We will refer to learning of this relationship as *Synthesizer Control Learning.*



Related Work

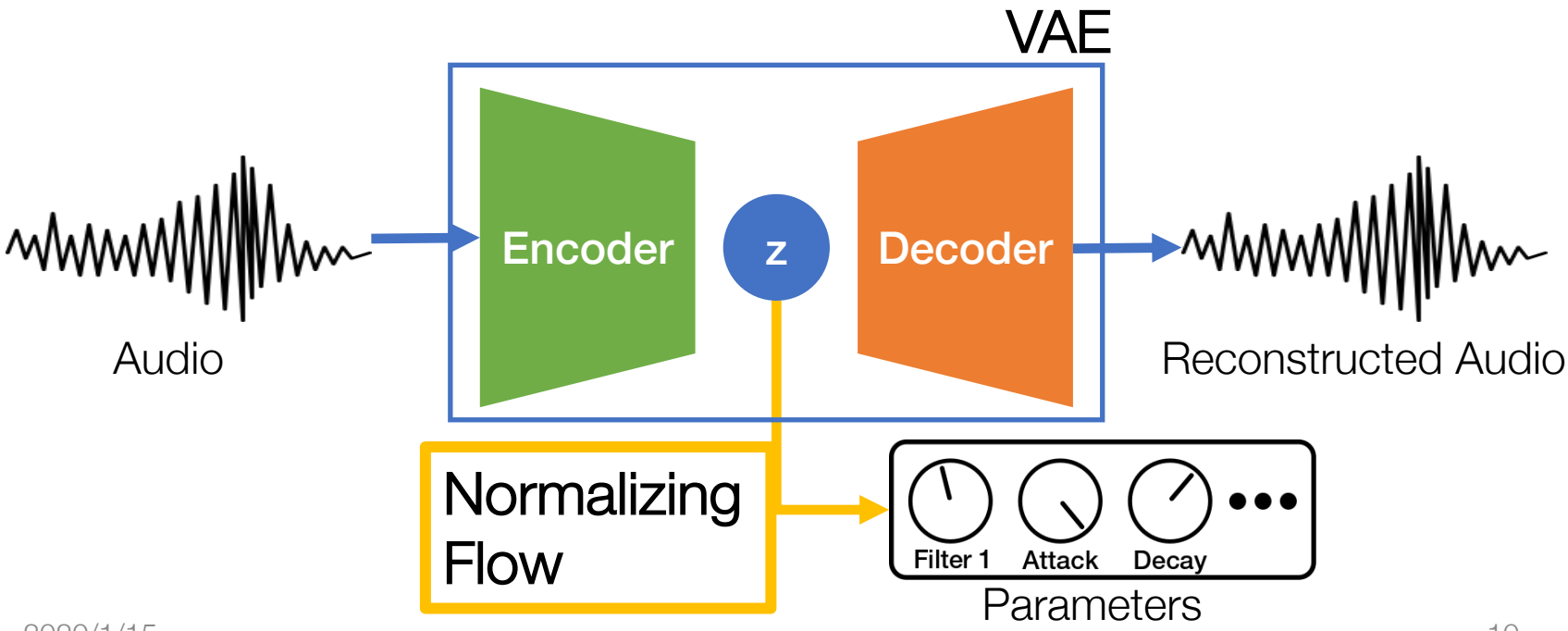
- Yee-king et al. trained an LSTM network to infer the synthesizer parameters from sounds.
- Useful for recreating sounds using synthesizers
- Synthesizer output was only close to the original sound 25% of the time.



M. J. Yee-King, L. Fedden, and M. D'Inverno, "Automatic Programming of VST Sound Synthesizers Using Deep Networks and Other Techniques," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 150–159, Apr. 2018.

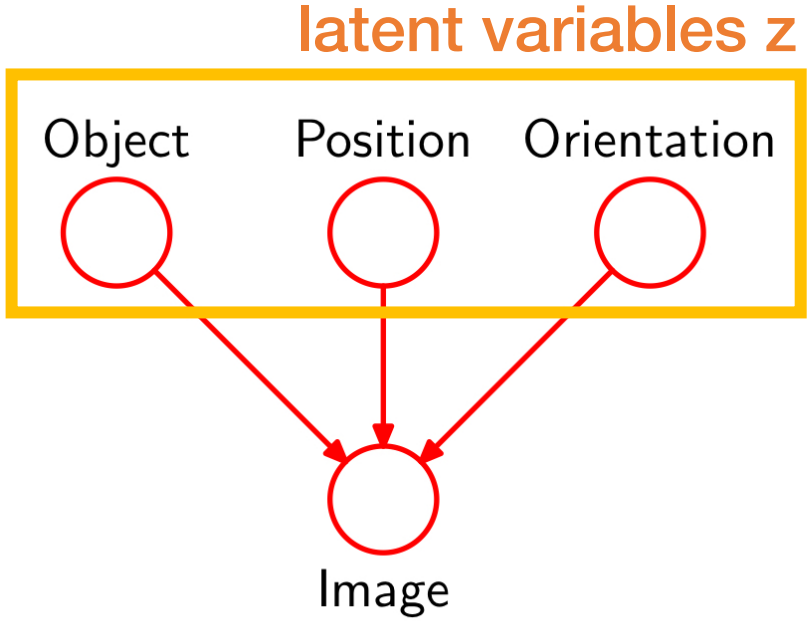
Proposed Method

- Use Variational Auto-Encoder (VAE) to learn an organized auditory latent space
- Use Normalizing Flows (NFs) to learn an invertible mapping from latent space to the parameter space



Background: Generative models

- Generative models try to learn the structure of data: p(x)
- We consider latent variables z that generated the data



Variational Auto-Encoder (VAE)

$$p(x) = \int p(x|z)p(z)$$

- We can use Variational Inference to calculate p(x)
- We consider an approximate distribution $q_\phi(\mathbf{z}|\mathbf{x})$
- KL divergence between p(z|x) and q(z|x) is:

$$D_{KL}[p(z|x)|q(z|x)] = E_{z\sim q}[\log q(x|z)] - \log q(z|x) - \log p(z|x)$$

close to zero if q is powerful enough

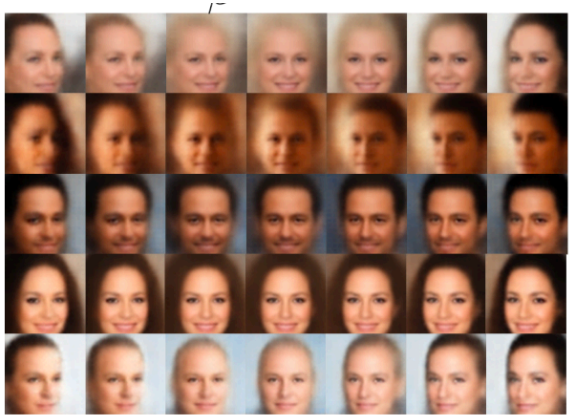
$$\log p(x) - D_{KL}[q(z|x)|p(z|x)] = E_{z\sim q}[\log p(x|z)] - D_{KL}[q(z|x)|p(z)]$$

$$\mathcal{L}(\theta; X) = E_{z\sim q}[\log p(x|z)] - D_{KL}[q(z|x)|p(z)]$$

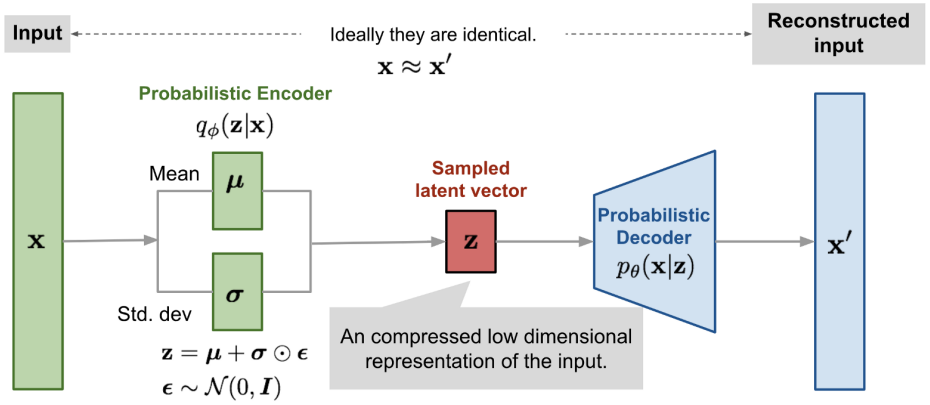
Decoder Encoder

What can VAEs do?

- Generates images (or any data) like GANs
- Encoder can produce latent variables for real data (unlike normal GANs)



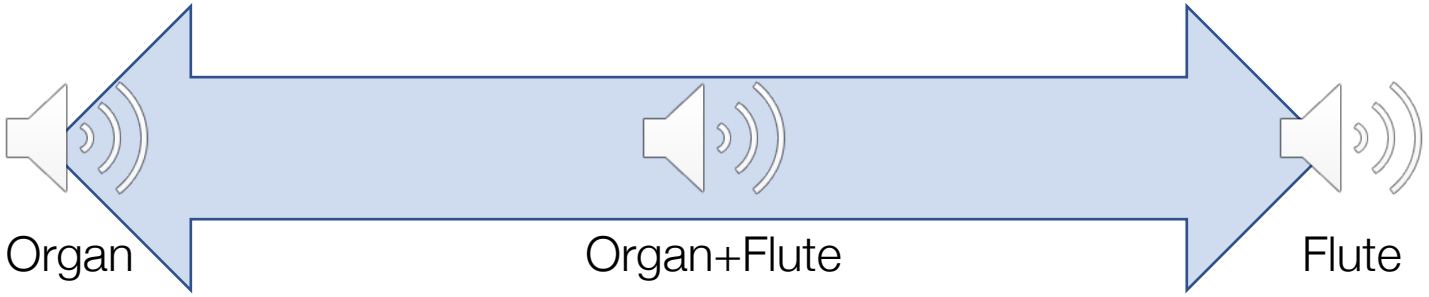
β-VAE latent traversal



https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

VAEs for audio

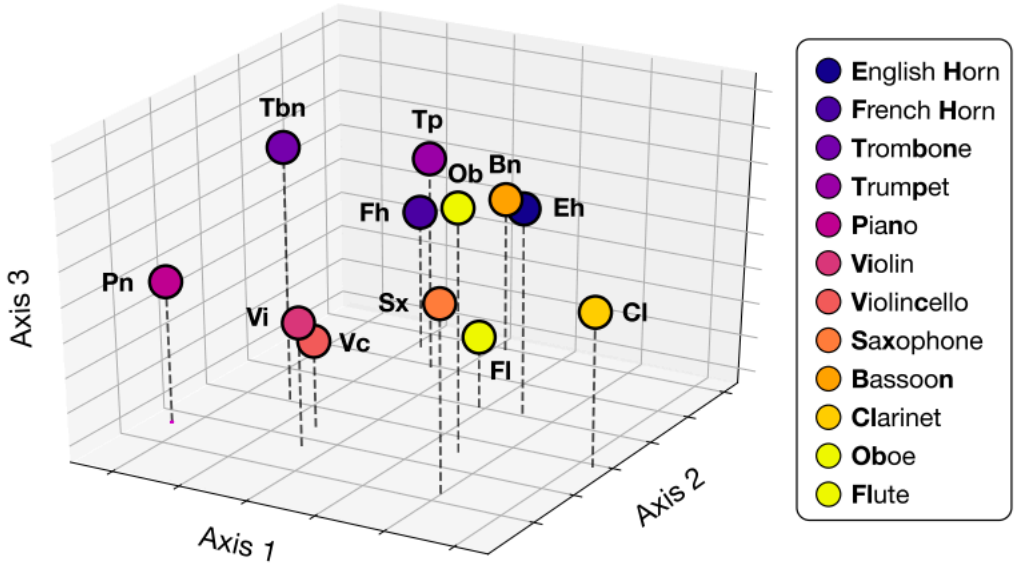
- VAEs can produce a mixture of sounds (e.g. “Clarinet + Tuba”) by interpolating between the latent variables of the data points.



A. Roberts, J. Engel, S. Oore, and D. Eck, “Learning latent representations of music to generate interactive musical palettes,” in *Intelligent Music Interfaces for Listening and Creation*, 2018, vol. 2068.

VAEs for audio

- In the latent space of VAEs, similar sounds are mapped close to each other.



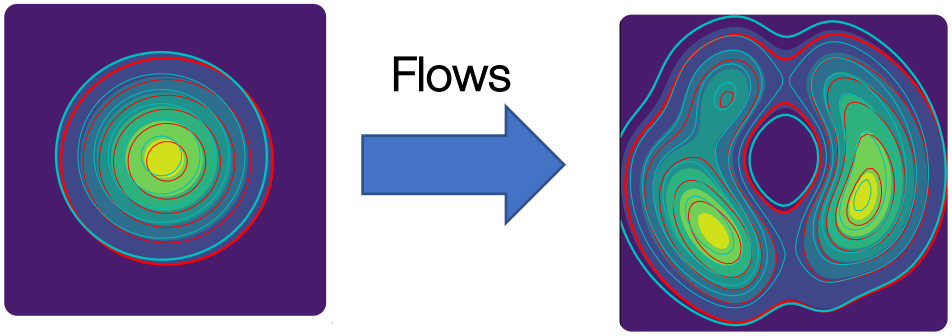
Regularized latent space

P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, “GENERATIVE TIMBRE SPACES: REGULARIZING VARIATIONAL AUTO-ENCODERS WITH PERCEPTUAL METRICS,” in *ISMIR*, 2018.

Normalizing Flows

- Normalizing flows can learn a mapping between simple distributions to more complex ones.
- We apply a transform f to z.
- From *change of variables:*

$$q(\mathbf{z}') = q(\mathbf{z}) \left| \det \frac{\partial f^{-1}}{\partial \mathbf{z}'} \right| = q(\mathbf{z}) \left| \det \frac{\partial f}{\partial \mathbf{z}} \right|^{-1}$$



Normalizing Flows

- We can apply a chain of such transforms:

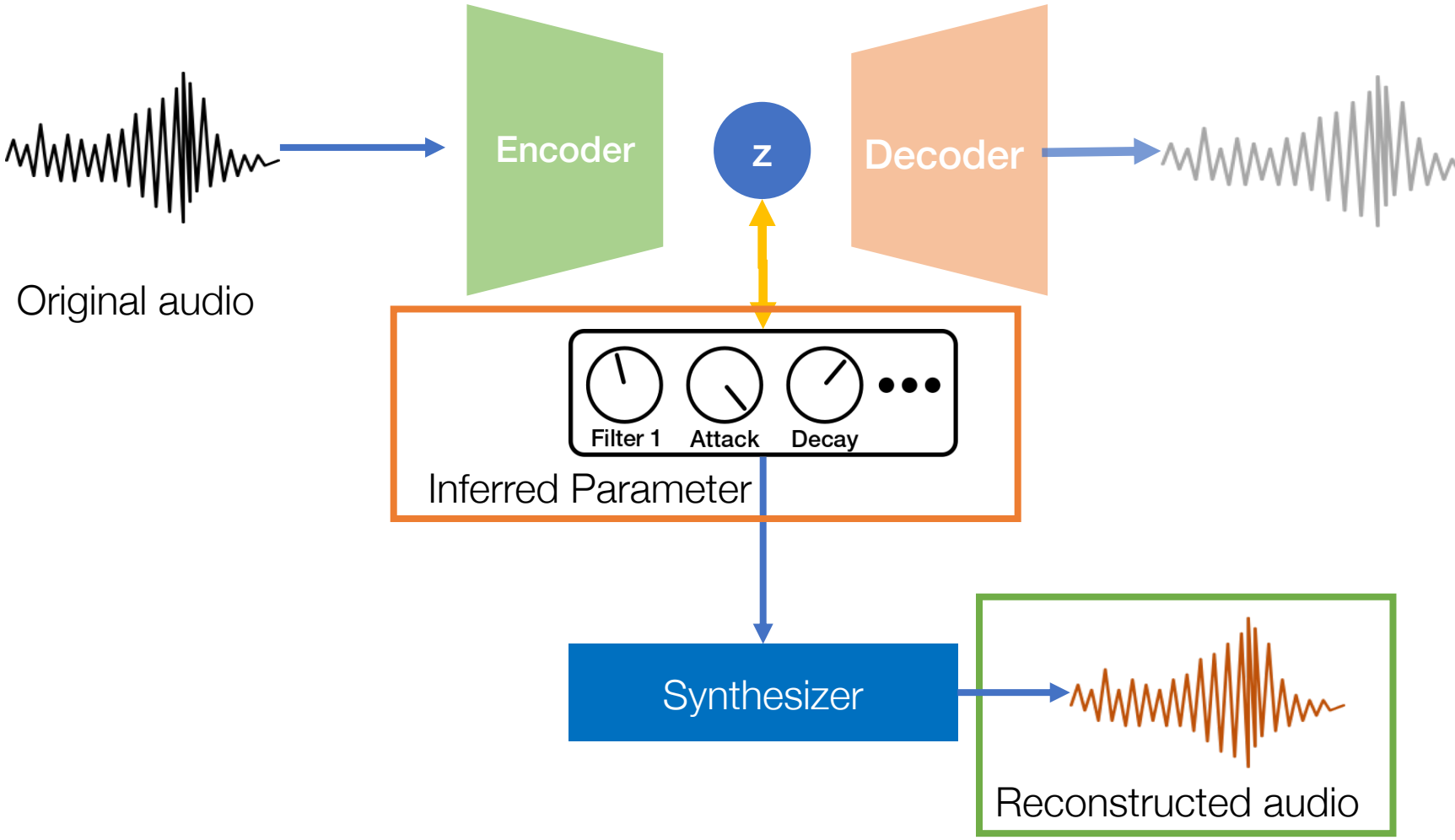
$$\ln q_K(\mathbf{z}_K) = \ln q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \ln \left| \det \frac{\partial f_k}{\partial \mathbf{z}_{k-1}} \right|$$

- Determinant of the Jacobian must be computable.
- Example:
  - Planar flow $f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b)$
  - Inverse auto-regressive flow

What This Model Can Do

- By solving the problem of synthesizer control learning, our system is able to perform:
  1. Parameter Inference / Sound Reconstruction
  2. Macro Control Learning
  3. Semantic Macro Control Learning
  4. Preset neighborhood exploration

Parameter Inference



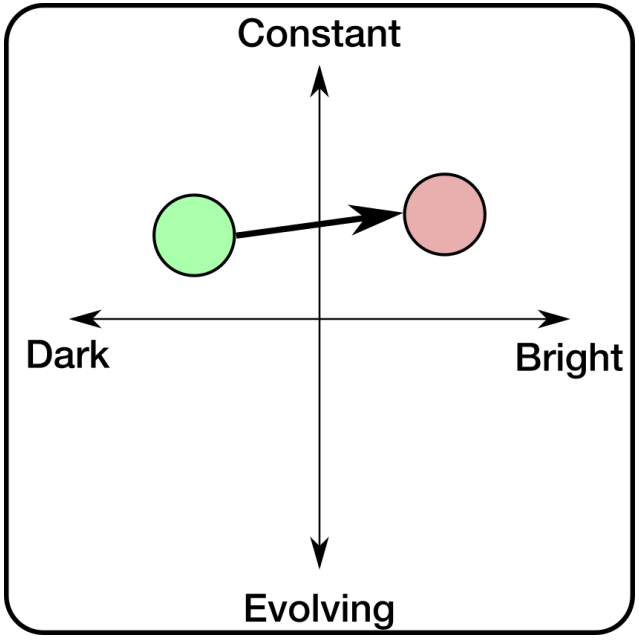
Synthesizer Macro Controls

- Macro controls are mapped to certain parameters to facilitate use of synthesizers.
  - The user can avoid dealing with 100~ parameters
- Macro controls must be designed by hand

Macro Controls (Hissy/Pitched, Spacey/Modulated)

Macro Control Learning

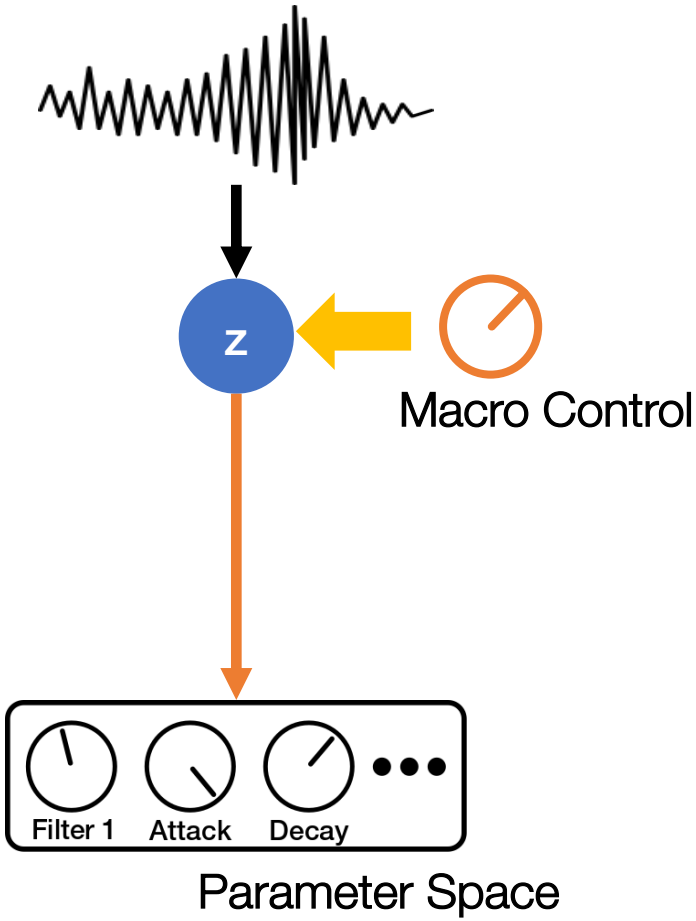
- Not only parameter inference but *macro control learning*
- Learn a more intuitive control for the synthesizer
  - Interactive
  - Allows room for creativity/discovery



Semantic Macro Controls

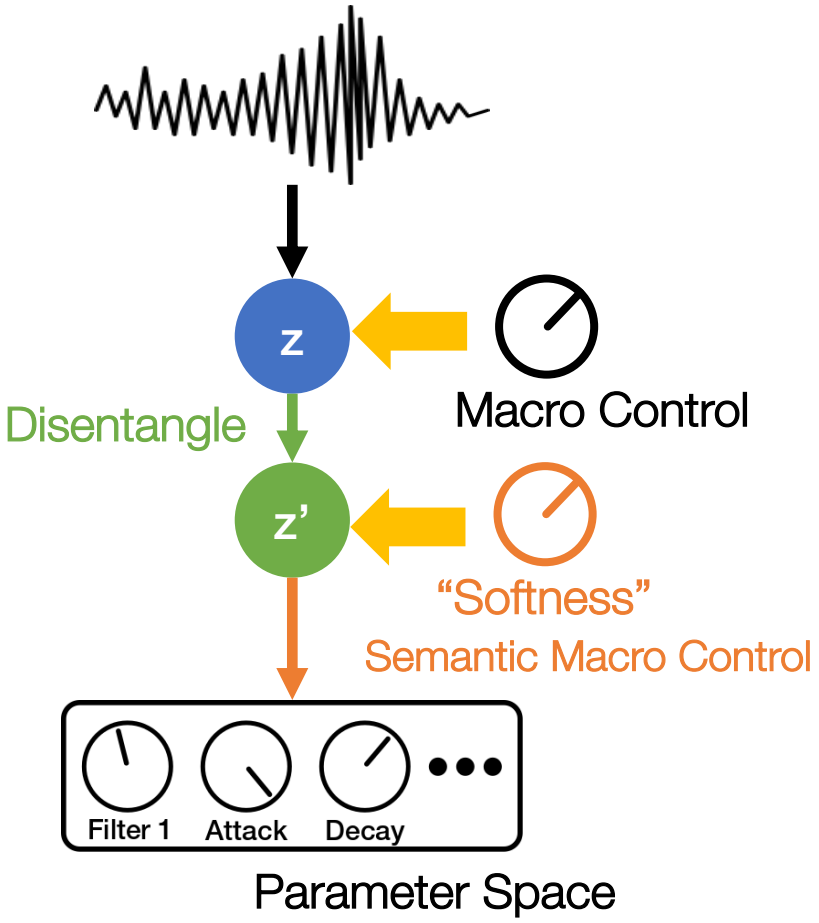
Macro Control Learning

- Traversing the latent variable z changes the sound gradually, so it can be considered as a macro control



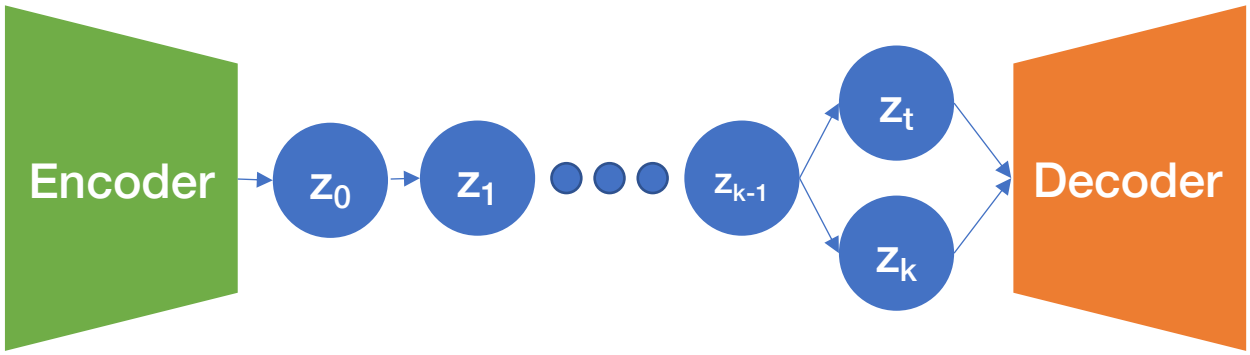
Semantic Macro Control Learning

- Traversing the latent variable z changes the sound gradually, so it can be considered as a macro control
- Use semantic tags to disentangle the latent space, creating a *semantic macro*



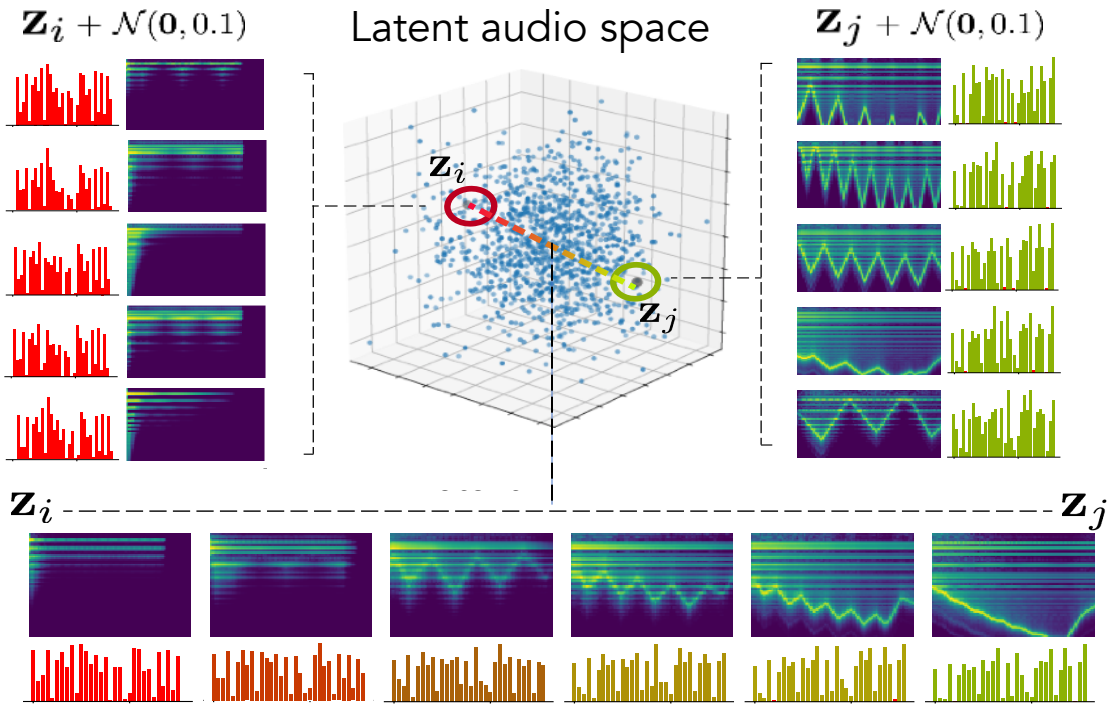
Disentangling Flows

- We use Inverse Auto-regressive Flows (IAF) on the encoder output to model a more complex posterior
- Some dimensions are trained to learn the tag distribution as a type of density estimation problem $p(z_{\mathbf{t}_-}) \sim \mathcal{N}(-\mu_*, \sigma_-)$ and $p(z_{\mathbf{t}_+}) \sim \mathcal{N}(+\mu_*, \sigma_+)$



Preset neighborhood exploration

- Because we learn an invertible mapping, we can map the presets back to the latent audio space
- Interpolating between the latent variables provides a way to mix 2 presets

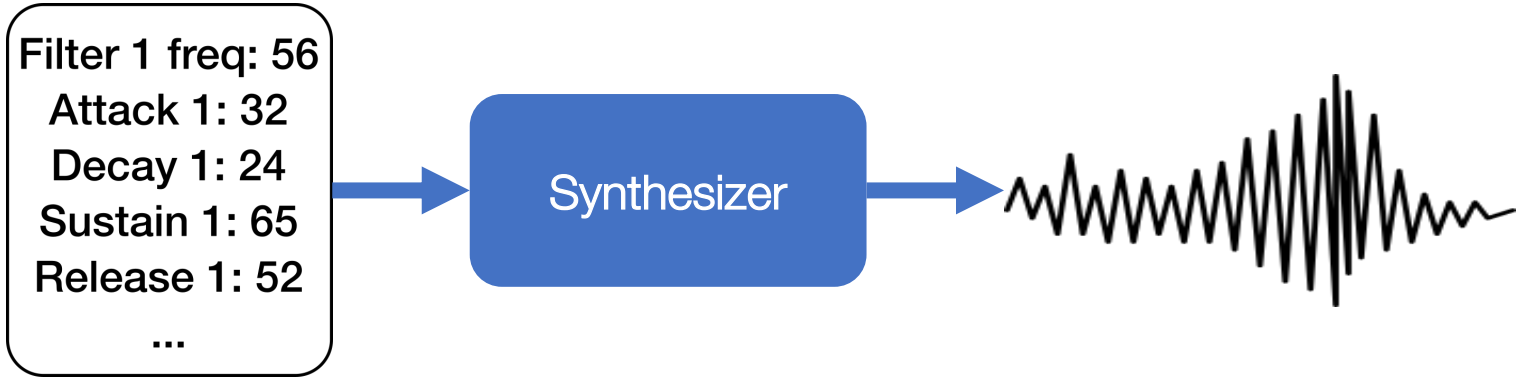


VST Synthesizers

- Diva
  - Subtractive Synthesis + some FM
  - Semantic information available for presets
- Dexed
  - DX7 Clone
  - Complex FM synthesis with 6 operators
  - 32 different algorithms (FM matrixes)

Dataset

- Presets were collected from the internet
- Parameters and tags were extracted from the presets
- Playing of presets to obtain the sound were automated by a program

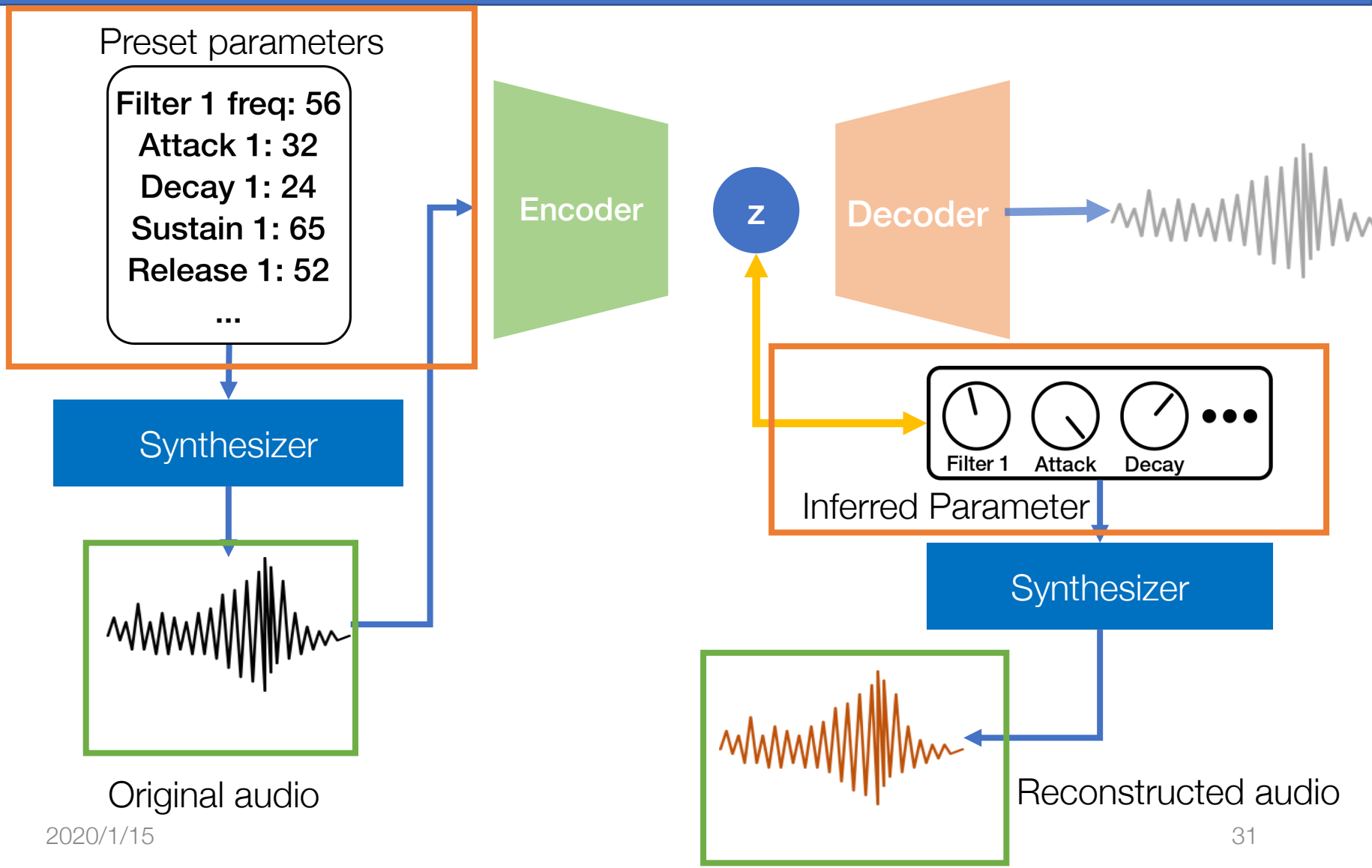


Diva Results

- VAE-based methods outperformed the baselines in audio statistics
- Our model using flows for parameter-audio mapping outperformed MLPs

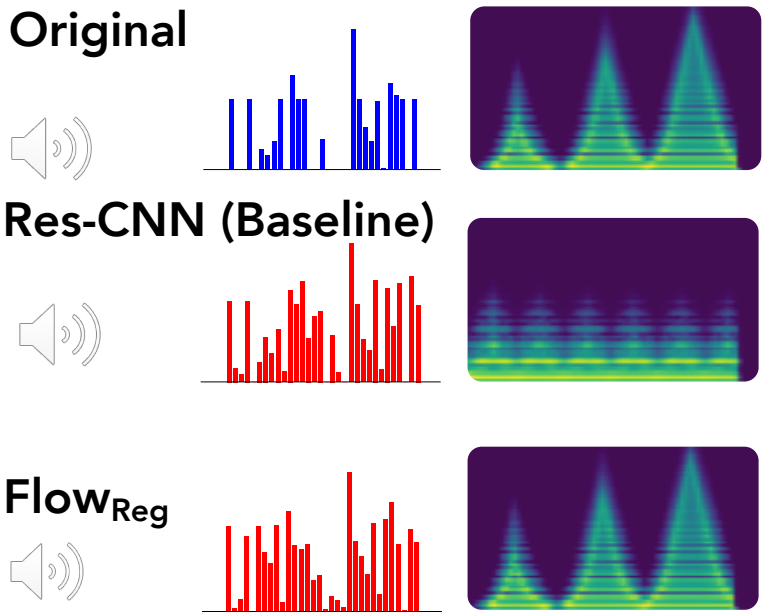
	Test set - 16 parameters			**Test set - 32 parameters**			**Out-of-domain** (32 p.)	
	Params	**Audio**		**Params**	**Audio**		**Audio**	
	MSE_n	SC	MSE	MSE_n	SC	MSE	SC	MSE
MLP	0.236±.44	6.226±.13	9.548±3.1	0.218±.46	13.51±3.1	36.48±11.9	2.348±2.1	37.99±7.8
CNN	0.171±.45	1.372±.29	6.329±1.9	0.159±.46	19.18±4.7	33.40±9.4	2.311±2.2	29.22±8.2
$ResNet$	0.191±.43	1.004±.35	6.422±1.9	0.196±.49	10.37±1.8	31.13±9.8	2.322±1.6	31.07±9.5
AE	0.181±.40	0.893±.13	5.557±1.7	0.169±.40	5.566±1.2	17.71±6.9	1.225±2.2	27.37±7.2
VAE	0.182±.32	0.810±.03	4.901±1.4	0.153±.34	5.519±1.4	16.85±6.1	1.237±1.3	27.06±7.1
WAE	**0.159±.37**	0.787±.05	4.979±1.5	**0.147±.33**	3.967±.88	16.64±6.2	1.194±1.5	26.10±6.4
VAE_{flow}	0.199±.32	0.838±.02	4.975±1.4	0.164±.34	1.418±.23	17.74±6.8	1.193±1.8	27.03±6.4
$Flow_{reg}$	0.197±.31	**0.752±.05**	**4.409±1.6**	0.193±.32	**0.911±1.4**	**16.61±7.4**	**1.101±1.2**	**26.07±7.7**
$Flow_{dis.}$	0.199±.31	0.831±.04	5.103±2.1	0.197±.42	1.481±1.8	17.12±7.9	1.209±1.4	26.77±7.3

Parameter Inference



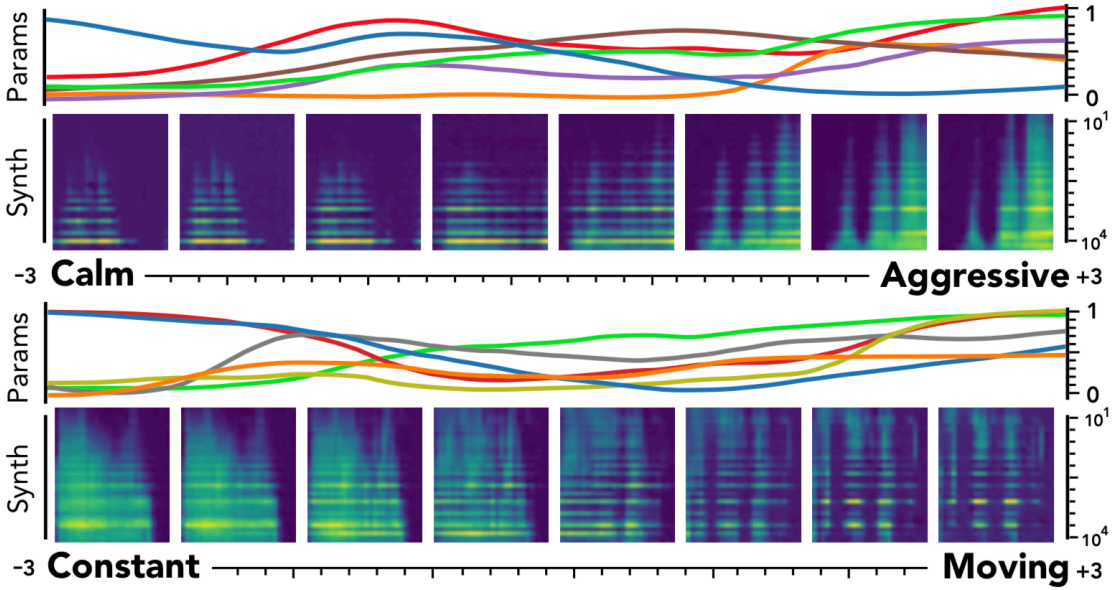
Diva - Parameter inference

- VAE based models showed better reconstruction results
- Baselines perform close inference of the parameters, but results in a different sound



Diva - Semantic Macro Control

- Macro controls are mapped smoothly to parameters
- Shows intuitive results for simple features

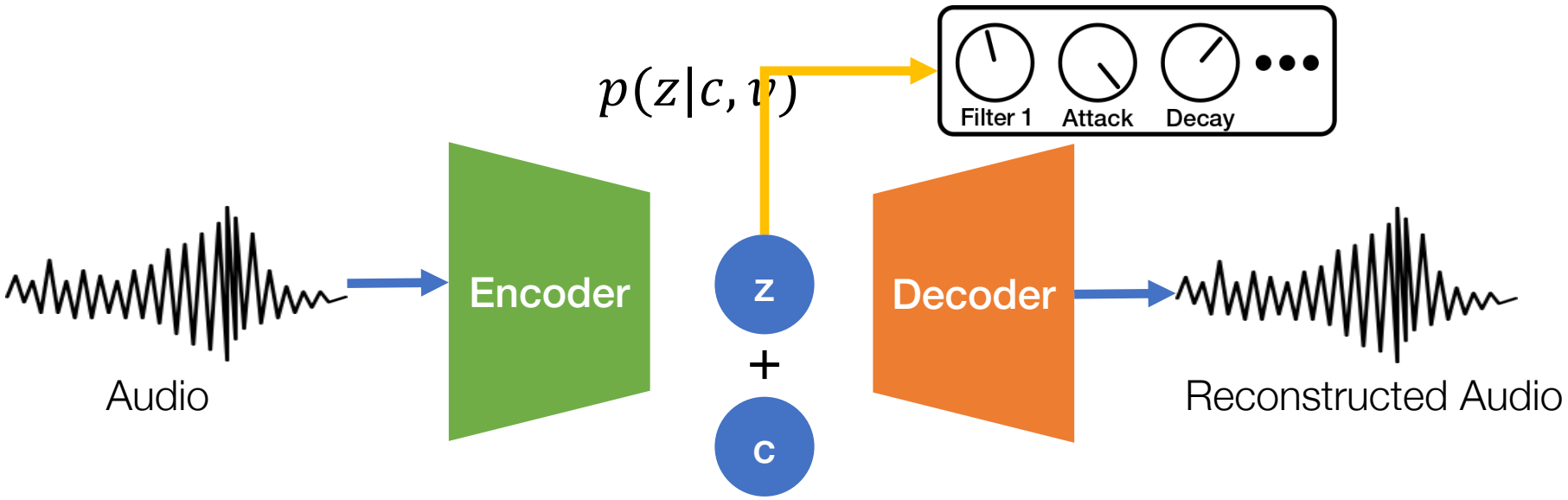


VST Synthesizers

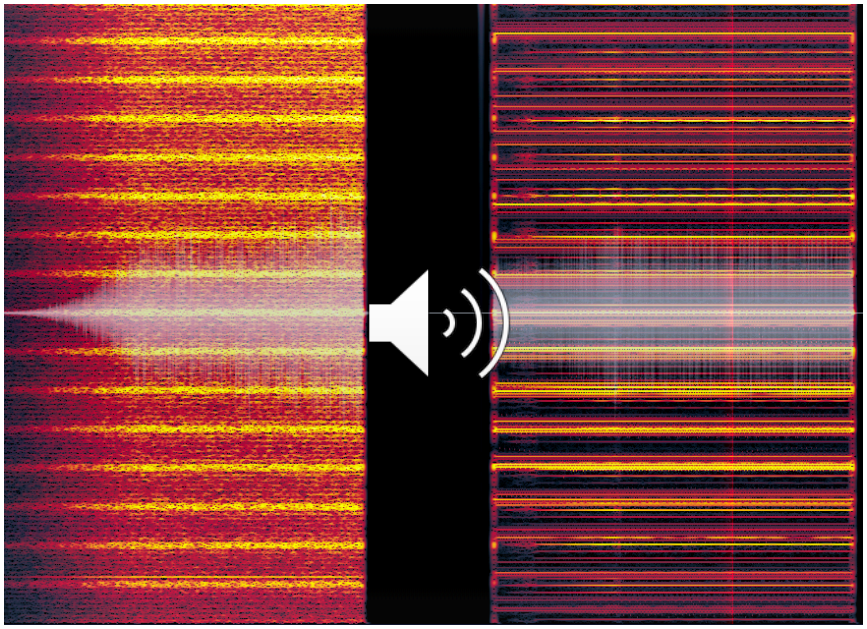
- Diva
  - Subtractive Synthesis + some FM
  - Semantic information available for presets
- Dexed
  - DX7 Clone
  - Complex FM synthesis with 6 operators
  - 32 different algorithms (FM matrixes)

Dexed - Categorical parameter

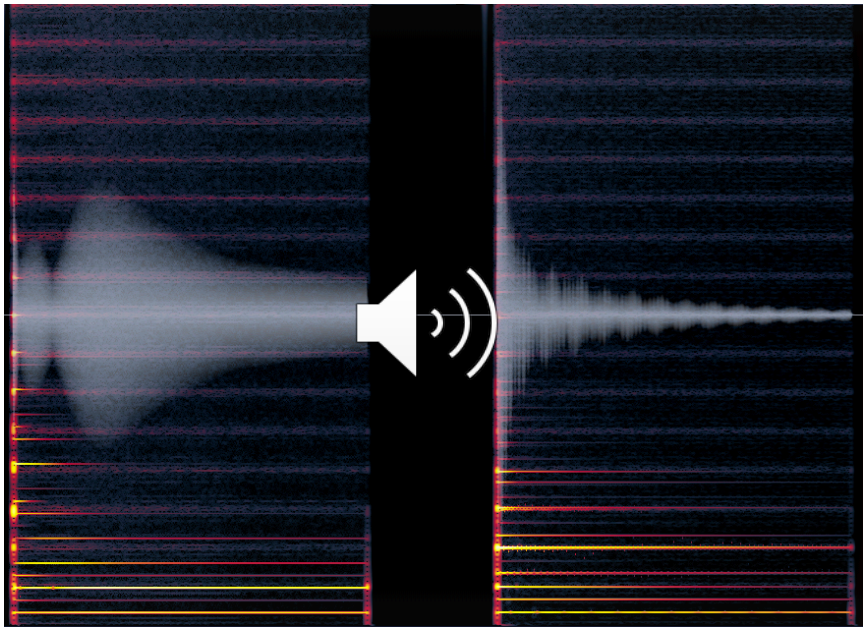
- 32 different configuration for FM synthesis
- We can treat it as a condition for the latent space to parameter model



Dexed – Parameter Inference



Original Reconstruction

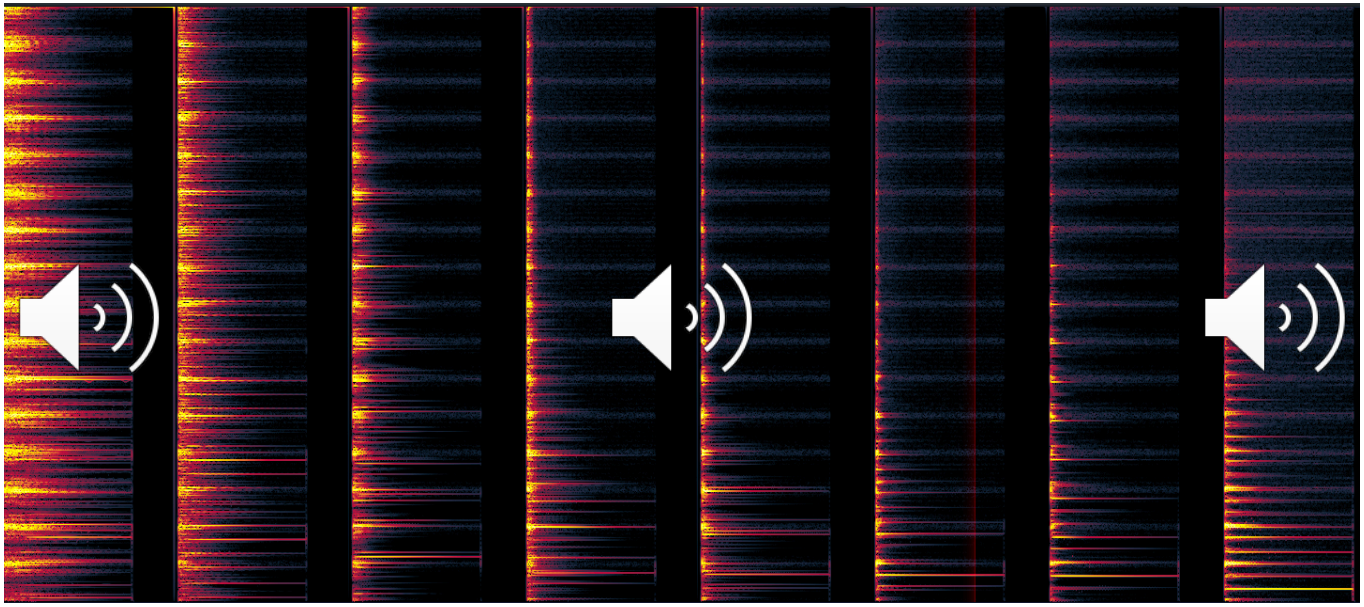


Original Reconstruction

- Envelope of audio is mostly reproduced
- Reconstruction of timbre is often wrong
  - Slight changes in parameters can change the timbre drastically

Dexed - Interpolation

- Interpolation between presets
  - More logical interpolation than in parameter space
- Parameter inference is more difficult



Future Work

- Better formulation for semantic macros
  - Semantic dimensions are not orthogonal
- Presets with tags are hard to obtain
  - Make an interface for labeling audio interactively and integrating user's perception
  - Post-hoc (after the training of VAE) disentanglement based on user feedback