Neural granular sound synthesis for raw waveform generation

2nd Seminar on Artificial Intelligence applied to Sound and Music Composition, Tokyo University of the Arts

Adrien Bitton | bitton@ircam.fr

PhD in the *Artificial Creative Intelligence and Data Science* (ACIDS) team
Under the supervision of Pr. Agon and Pr. Esling at IRCAM / Sorbonne Université

Japanese Society for Promotion of Science (JSPS) fellowship (10.2019 - 04.2020)
@ Machine Intelligence Laboratory (MIL) - The University of Tokyo

Presentation overview

1) Granular sound synthesis

2) Motivations for **neural granular sound synthesis**

3) Generative neural networks (state of the art)

4) Neural granular sound synthesis

5) Results

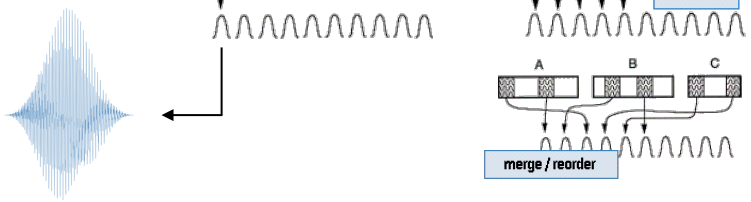
Granular sound synthesis

https://ccrma.stanford.edu/CCRMA/Courses/SoundSynthesis97/Lectures/4/Slides.html

A popular **sound synthesis** technique

by re-arranging signal windows

= **grains** of short and fixed duration (10-100ms)



Corpus-based = extracting grains from an audio library

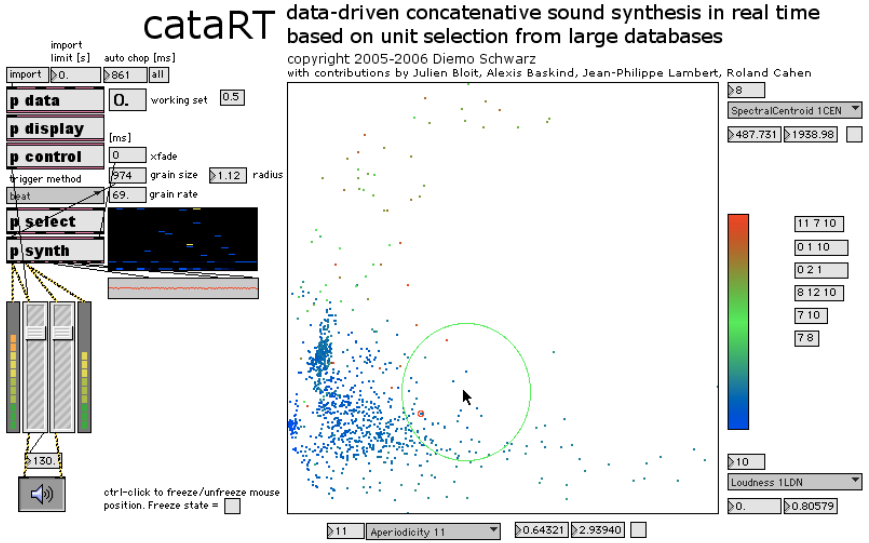
Analysis with acoustic descriptors

⇨ to structure their audio similarities

⇨ to control the synthesis

traversals of the grain space

criterion for data-driven synthesis



A grain space visualization with Aperiodicity, Loudness and Spectral Centroid features.

Motivations for Neural granular sound synthesis

Such grain space is **not invertible**

☹ cannot synthesize audio besides the scattered grains

☹ needs to store the library for generating

The quality of the analysis dimensions is bound to the **choice of the descriptors**

Visualization of grain-level audio similarities but **no temporal structure**

⇨ Use of **generative neural networks**

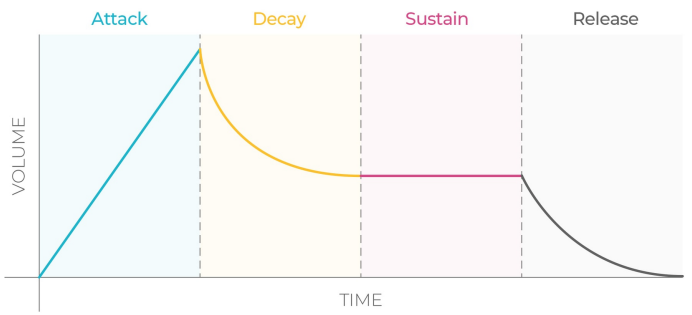
☺ to **learn analysis dimensions** from the data

☺ to **continuously** invert to signal domain

☺ to model **higher-level temporal relationships**

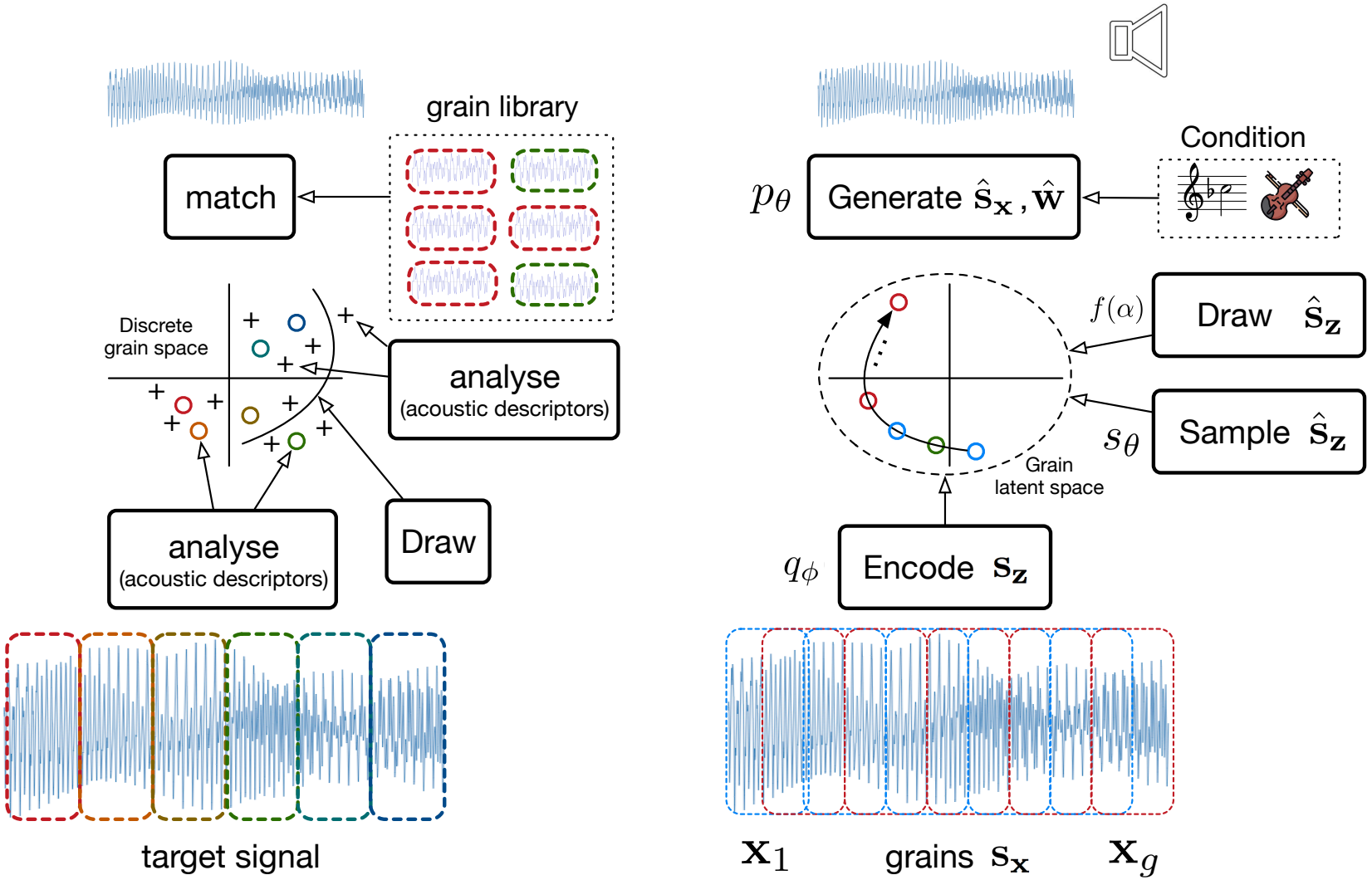
☺ to embed the dataset in the model parameters

☺ for **creative applications** of neural sound synthesis



temporal dynamics of a musical note

Motivations for Neural granular sound synthesis



Generative neural networks

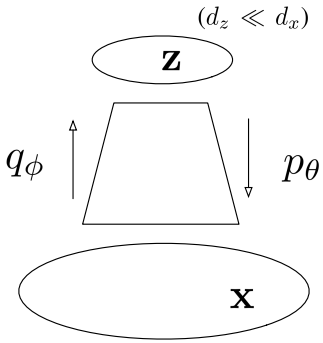
Use a dataset $\mathbf{x} \in \mathbb{R}^{d_x}$ to estimate a **probability distribution** $p(\mathbf{x})$

The model approximates this data distribution $p_\theta \sim p(\mathbf{x})$ to generate new samples

Representation and control using **continuous latent variables** $\mathbf{z} \in \mathbb{R}^{d_z}$

⇨ extract salient and unsupervised features of the data $(d_z \ll d_x)$

⇨ higher-level space generating any given example



Relate to granular sound synthesis with an **Auto-Encoder** architecture

encoder (analysis) $\mathbf{z} = q_\phi(\mathbf{x})$

decoder (generation) so that $p_\theta(q_\phi(\mathbf{x})) \sim \mathbf{x}$

Replacing the acoustic descriptor space with the learned latent variables

Probabilistic latent space

Reconstruction objective over every data examples $p_\theta(q_\phi(\mathbf{x})) \sim \mathbf{x}$

but no guarantee of the **smooth and continuous invertibility** of the latent dimensions

⇨ Learn a **latent probability distribution** $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$

Optimization of this joint distribution with a **Variational Auto-Encoder (VAE)**

and a **Gaussian prior** distribution over the latent space $p_\theta(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

probabilistic encoding $q_\phi : \mathbf{x} \rightarrow \{\mu(\mathbf{x}), \sigma(\mathbf{x})\}$

~ parameterize a « volume » inside the prior

sampling $\mathbf{z} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x}))$

Encoder distribution matched to the latent prior with the Kullback-Leibler (KL) divergence

$$\mathcal{L}_{\theta,\phi} = \underbrace{-\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction}} + \beta * \underbrace{\mathcal{D}_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})\right]}_{\text{regularization}}$$

Raw waveform generation

Time series at **high-sampling** rate
eg. 16kHz, 44.1kHz

Challenge of modeling:

- **multi-scale** temporal relationships

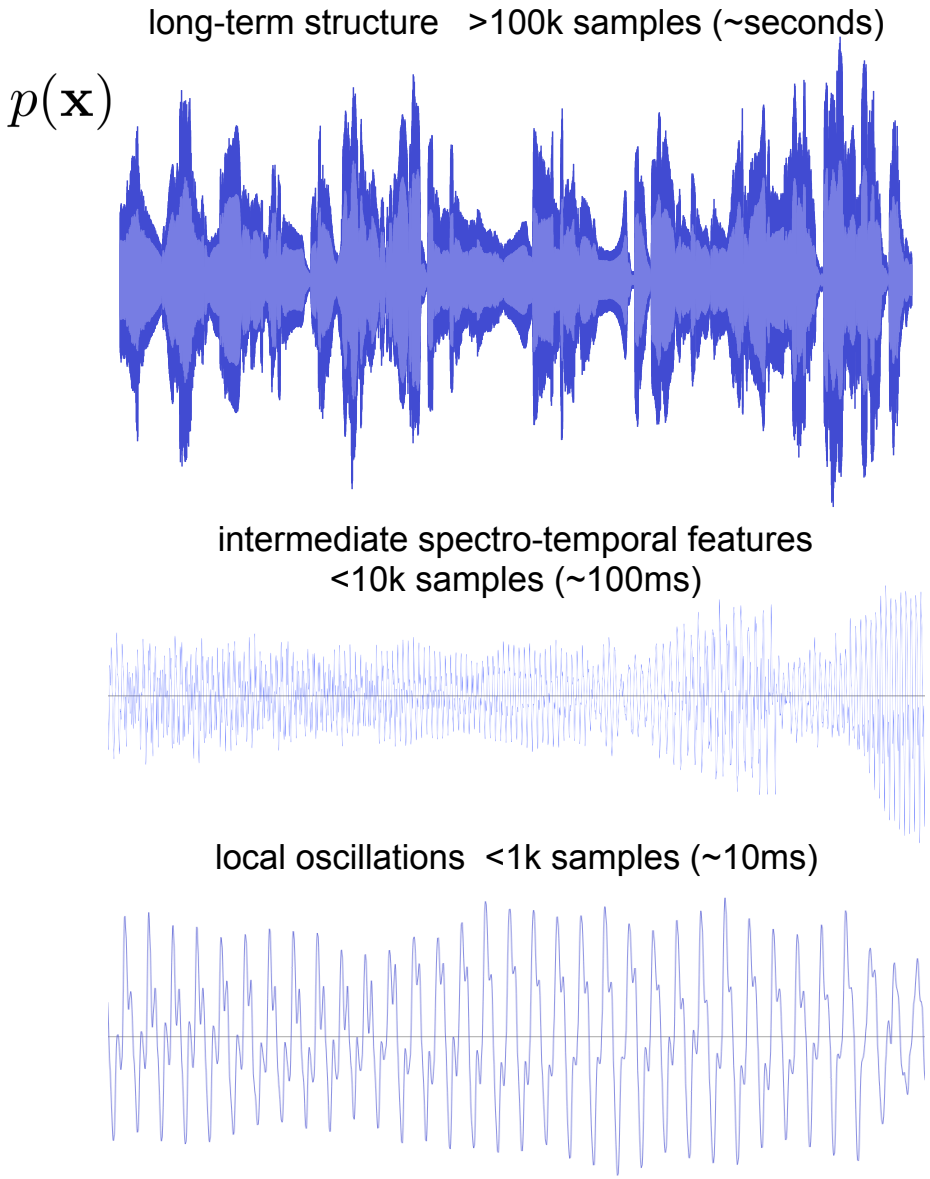
local features (audio quality)

longer-term dependencies (structures)

- interesting generative controls
  eg. music composition

Benefits:

- no loss of information
- direct synthesis
- can target any sound



Neural waveform models

- Can use the **causality** of audio signals $\mathbf{x} = \overrightarrow{\{x_1 \dots x_T\}}$

WaveNet autoregressive estimate of sample likelihood $p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1 \dots x_{t-1})$

- Can use a **frame-level** modeling

Symbol-to-Instrument Neural Generator (SING)

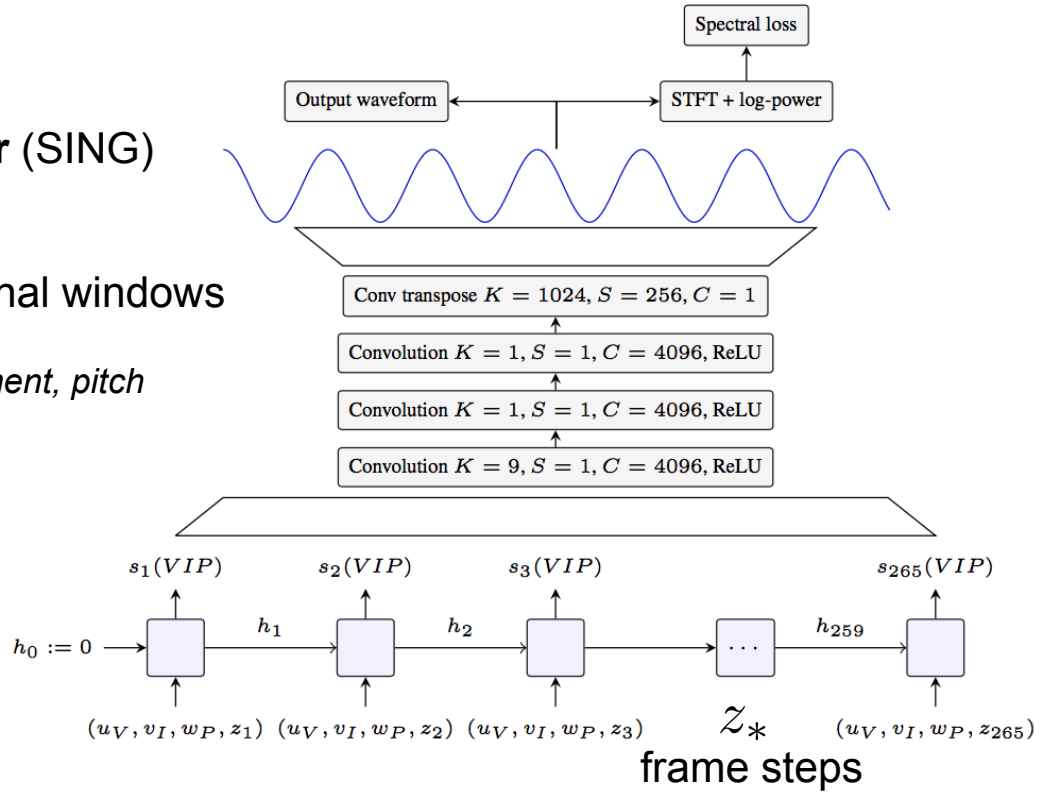
recurrent decoder and **overlap-add** of signal windows

$\hat{\mathbf{x}} = d(s(V, I, P))$ *velocity, instrument, pitch*

spectrogram reconstruction

$$l(\mathbf{x}) = \log(\epsilon + |\mathrm{STFT}[\mathbf{x}]|^2)$$

$$\underset{D,S}{\operatorname{argmin}} ||l(\mathbf{x}), l(\hat{\mathbf{x}})||_1$$



Neural DSP implementations

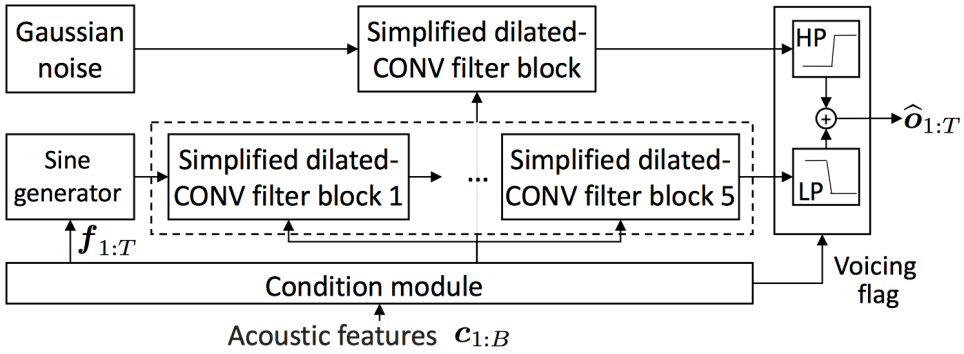
Explicitly learning **Digital Signal Processing** (DSP) operations

eg. **harmonic + noise** sound synthesis

Neural Source-Filter (NSF) model

temporal filtering for sines and noise

multi-scale reconstruction $\mathrm{STFT}_{1\ldots N}$

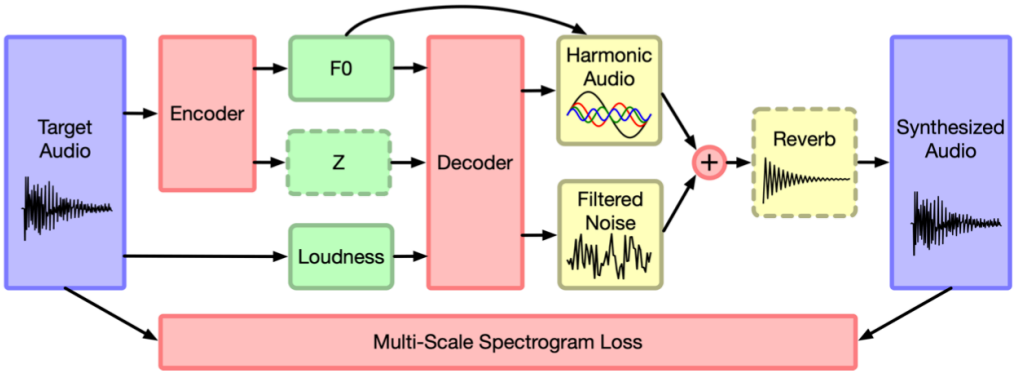


Differentiable Digital Signal Processing (DDSP) model

additive harmonic synthesizer

subtractive noise synthesizer
frequency domain filtering

control with a recurrent decoder



Neural granular sound synthesis

Substituting acoustic descriptors with **latent dimensions of a VAE**

⇨ to address limitations of granular sound synthesis

⇨ to efficiently model waveform at **multiple time scales**

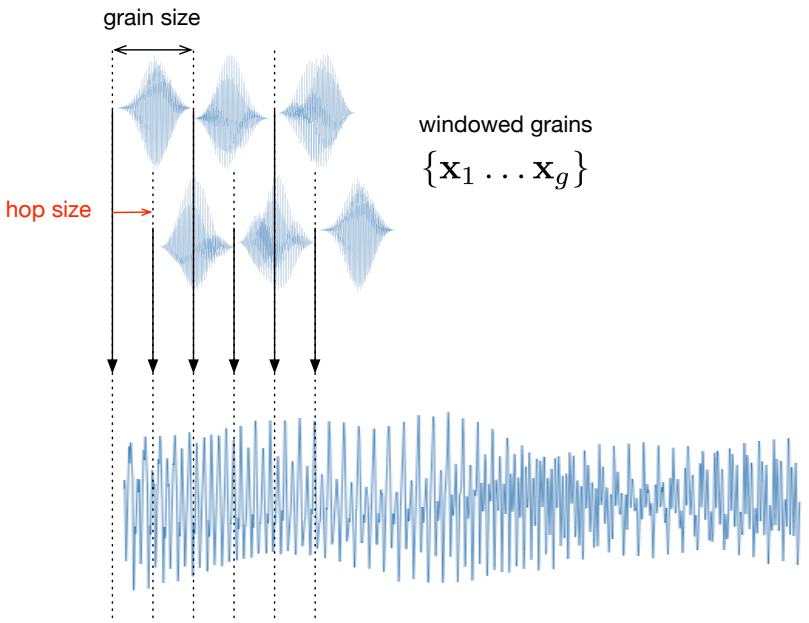
Dataset $\mathbf{x} \in \mathbb{R}^{d_x}$ of grains with size d_x

modeled with a **granular latent space** $\mathbf{z} \in \mathbb{R}^{d_z}$

Waveform $\mathbf{w}_{1..T}$ reconstruction with **overlap-add**

over contiguous grain series $\mathbf{s}_\mathbf{x} = \{\mathbf{x}_1 \dots \mathbf{x}_g\}$

down-sampling with the hop size



Model architecture

Grain sequences are encoded into **latent paths**

$$\mathbf{s_z} = \{\mathbf{z}_1 \dots \mathbf{z}_g\}$$

Each latent point is decoded into a grain

by **subtractive synthesis** on grains of noise

The grain latent space is probabilistic

⇨ model of the **grain similarities**

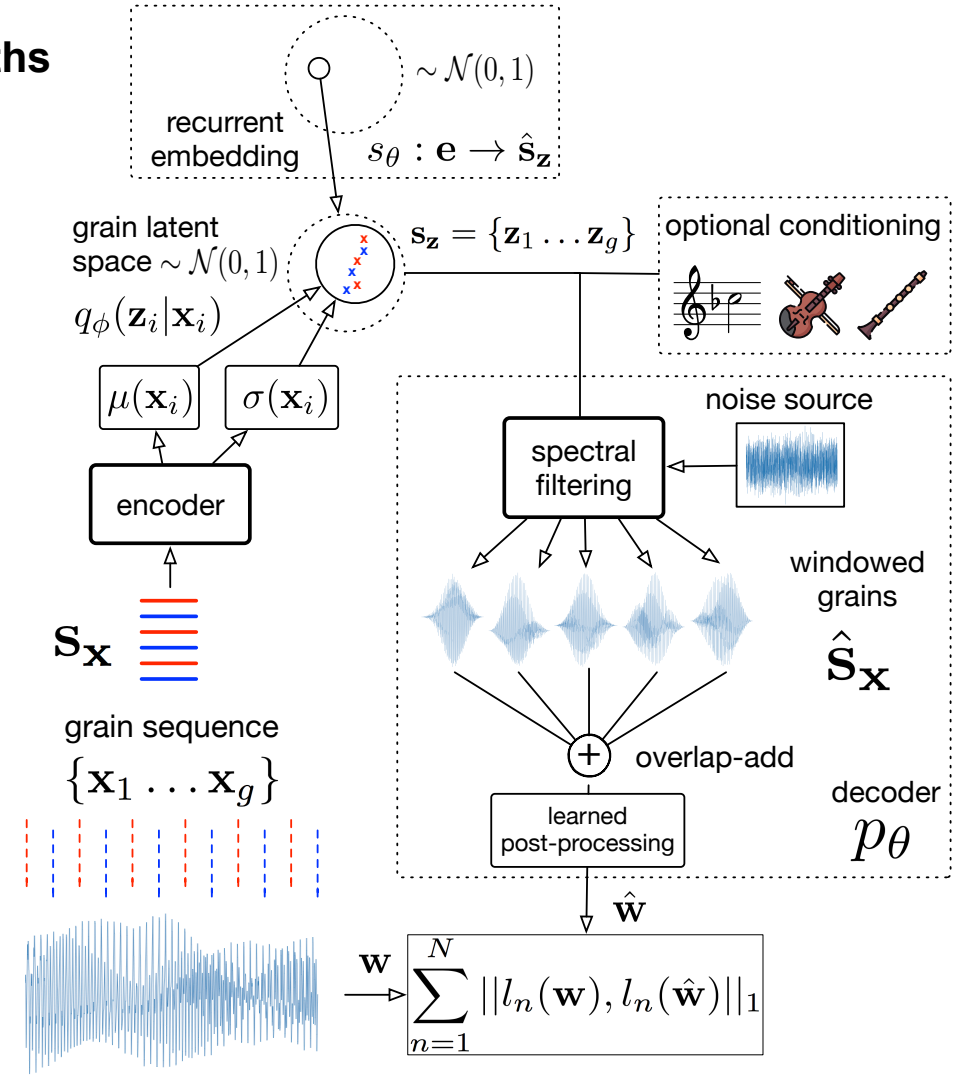
⇨ **continuously invertible**

A recurrent embedding can be trained

⇨ to **learn structured latent paths**

⇨ to generate consistent dynamics
eg. musical notes, drum hits

Generation can be conditioned with additional **user controls** *eg. target pitch or instrument class*



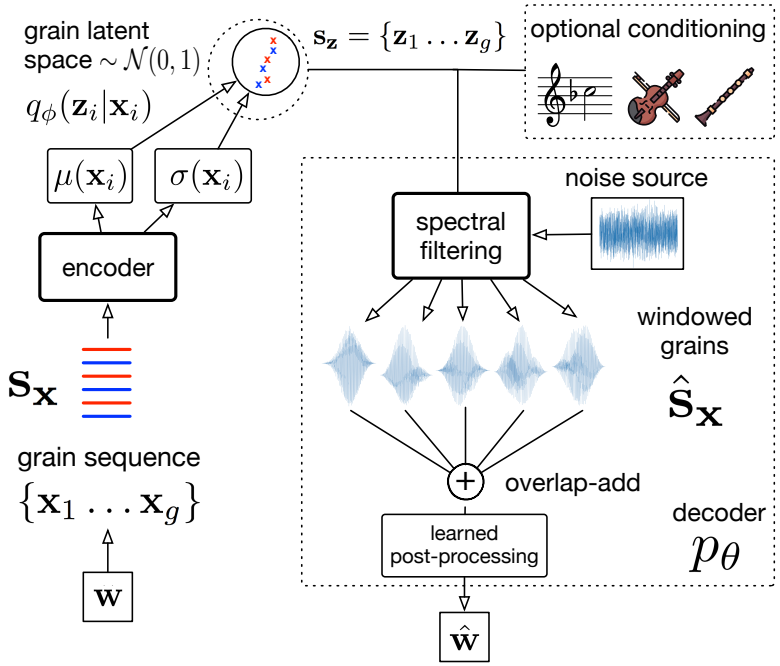
Analysis and synthesis

Encoder with down-sampling convolutions

grain analysis $q_\phi : \mathbf{x}_i \rightarrow \{\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i)\}$

Gaussian sampling $\mathbf{z}_i = \mu(\mathbf{x}_i) + \eta * \sigma(\mathbf{x}_i)$

from the prior $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$



Decoder predicts frequency coefficients $\mathbf{H}_i \in \mathbb{R}^{d_h}$

conditioned on $\mathbf{Z}_i$ to **filter grains of uniform noise**

<u>Filter design derived from DDSP</u> $\mathbf{n}_i \sim \mathcal{U}^{d_x}_{[-1,1]}$ $\hat{\mathbf{X}}_i = \mathbf{H}_i * \mathrm{DFT}(\mathbf{n}_i)$

Synthesized waveform grains $\hat{\mathbf{x}}_i = \mathrm{iDFT}(\hat{\mathbf{X}}_i)$ are overlap-add

Refined waveform quality with a **learnable post-processing**

Generation can be additionally conditioned by concatenating one-hot class labels to $\mathbf{Z}_i$

$$p_\theta : (\hat{\mathbf{s}}_\mathbf{z}, \mathbf{oh}_{\text{class}}) \rightarrow \hat{\mathbf{s}}^{\text{cond.}}_\mathbf{x} \rightarrow \hat{\mathbf{w}}^{\text{cond.}}$$

Model objective

Given $\hat{\mathbf{X}}_i = \mathbf{H}_i * \mathrm{DFT}(\mathbf{n}_i)$ and an even grain size

the decoder predicts a **filter size** $d_h = d_x/2 + 1$

⇨ grain and hop sizes directly relate to the **resolution for spectrogram synthesis**

Accuracy is evaluated with multi-scale spectrogram reconstruction

several settings $\mathrm{STFT}_{1\ldots N}$ for $l(\mathbf{x}) = \log(\epsilon + |\mathrm{STFT}[\mathbf{x}]|^2)$

Grain-level latent distribution by **individual regularization and sampling**

$$\mathcal{L}_{\theta,\phi} = \underbrace{\sum_{n=1}^{N} ||l_n(\mathbf{w}), l_n(\hat{\mathbf{w}})||_1}_{\text{reconstructions}} + \beta * \underbrace{\sum_{i=1}^{g} \mathcal{D}_{KL}\left[q_\phi(\mathbf{z}_i|\mathbf{x}_i) \parallel p_\theta(\mathbf{z})\right]}_{\text{regularizations}}$$

Sequence embedding

Frame-level waveform modeling over contiguous grain sequences

⇨ Longer-term temporal relationships

⇨ Learning **structured paths inside the latent grain space**

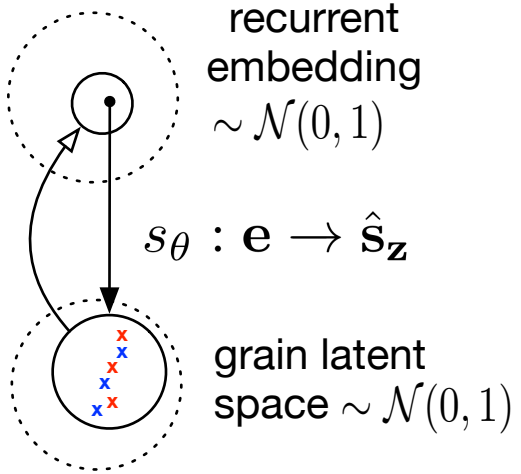
Recurrent encoder to « summarize » $\mathbf{s_z} = \{\mathbf{z}_1 \dots \mathbf{z}_g\}$

into a probabilistic embedding $\mathbf{e} \in \mathbb{R}^{\mathbf{d_e}}$

Given $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I})$

a **recurrent decoder** generates a path $s_\theta : \mathbf{e} \rightarrow \hat{\mathbf{s}}_\mathbf{z}$

Sufficient for modeling audio events such as:
- a musical note
- a drum hit
- a short audio loop
- an animal sound



$\mathbf{s_z} = \{\mathbf{z}_1 \dots \mathbf{z}_g\}$

Results

The model is **interpretable** and efficient

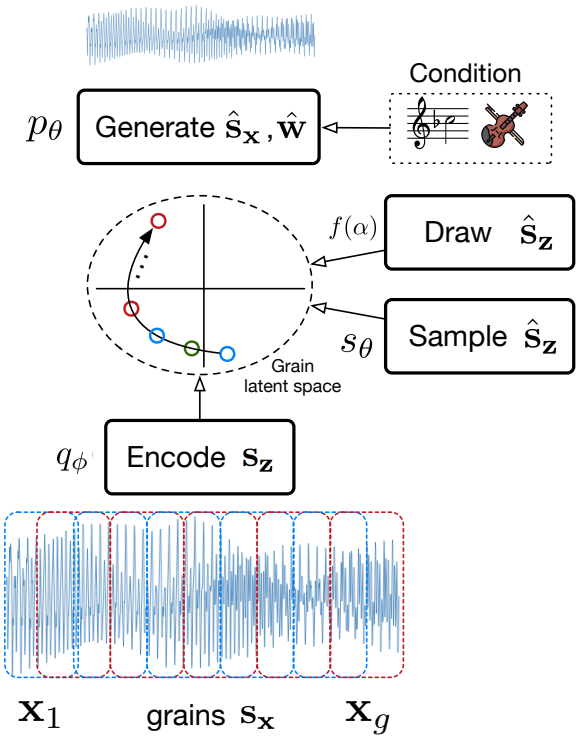
It can be adapted to **diverse sound domains** and interactions

We train **unconditional** and **conditional** models on datasets of

- individual notes for different orchestral instruments and playing techniques
- drum sounds
- animal sounds

It refines granular sound synthesis

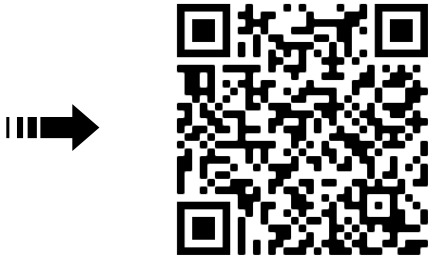
- continuous free-synthesis
- data-driven re-synthesis
- sampling paths with temporal structure
- control over some target styles



Thank you

Let’s listen and improvise something !

find more on this demonstration page



References

D. Schwarz et al., "**Real-Time Corpus-Based Concatenative Synthesis with CataRT**" in Proceedings of the International Conference on Digital Audio Effects, 2006.

D. P. Kingma et al., "**Auto-encoding variational bayes**", International Conference on Learning Representations, 2014.

A. van den Oord et al., "**Wavenet: A generative model for raw audio**", arXiv 1609.03499, 2016.

A. Defossez et al., "**Sing: Symbol-to-instrument neural generator**" in Advances in Neural Information Processing Systems, vol. 31, p. 9041–9051, 2018.

X. Wang et al., "**Neural source-filter waveform models for statistical parametric speech synthesis**", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 402–415, 2019.

J. Engel et al., "**DDSP: Differentiable Digital Signal Processing**", International Conference on Learning Representations, 2020.