DDSP

Differentiable Digital Signal Processing

T1 - PRESENTATION

CentraleSupélec - 2A

Peizhou ZHANG - 1900291
Molin LIU - 1900262
Xinjian OUYANG - 1900273
Haoyu YU - 1900287

CentraleSupélec

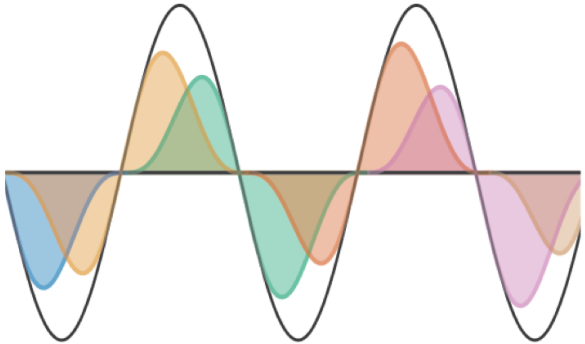
CONTENTS

PART 01

State of Art

A brief introduction to the challenges of neural audio synthesis and related work.

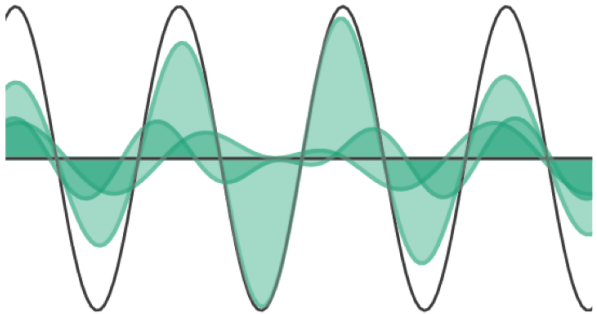
Part 1.1 | Models generate waveforms in the t/f domain



Strided Convolution
Phase Alignment
(WaveGAN, SING)

◆ **Models that apply a prior over generating audio with aligned wave packets rather than oscillations**

- Generate waveforms directly with overlapping frames.
- The model must precisely align waveforms between different frames and learn filters to cover all possible phase variations.

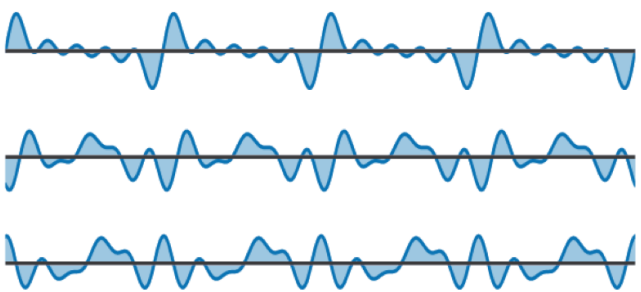


Fourier Representation
Spectral Leakage
(Tacotron, GANSynth)

◆ Fourier-based models

- Suffer from the phase-alignment problem, as the Short-time Fourier Transform (STFT) is a representation over windowed wave packets.
- Must contend with spectral leakage, where sinusoids at multiple neighboring frequencies and phases must be combined to represent a single sinusoid when Fourier basis frequencies do not perfectly match the audio.

Part 1.1 | Models generate waveforms in the t/f domain



Autoregressive
Waveform != Perception
(WaveNet, SampleRNN)

◆ Autoregressive waveform models

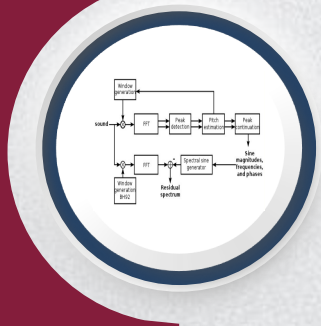
- Generate the waveform a single sample at a time, not constrained by the bias over generating wave packets and can express arbitrary waveforms.
- Require larger and more data-hungry networks.
- The use of teacher-forcing during training leads to exposure bias during generation, where errors with feedback can compound.
- Incompatible with perceptual losses such as spectral features , pretrained models and discriminators (inefficient).

Oscillator Models (vocoders or synthesizers)

- Directly generate audio with oscillators.
- Physically and perceptually motivated.
- Use expert knowledge and hand-tuned heuristics to extract synthesis parameters (analysis) that are interpretable (loudness and frequencies) and can be used by the generative algorithm (synthesis).
- Fall short of end- to-end learning.

◆ Neural Source Filter (NSF)

- The NSF can be seen as a specific DDSP model, that uses convolutional wave-shaping of a sinusoidal oscillator to create harmonic content, rather than additive synthesis explored in this work.
- Generate audio in the time domain and impose multi-scale spectrograms losses in the frequency domain.

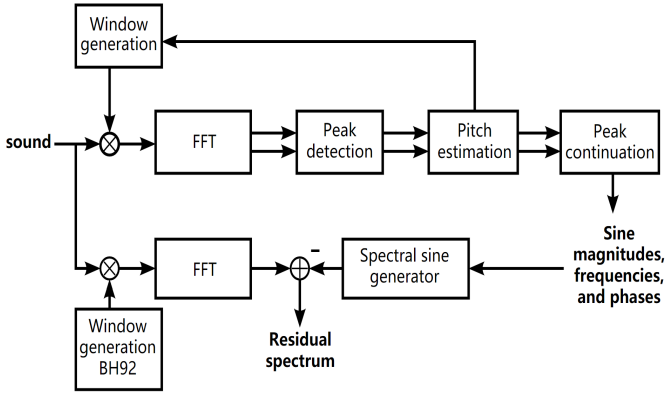


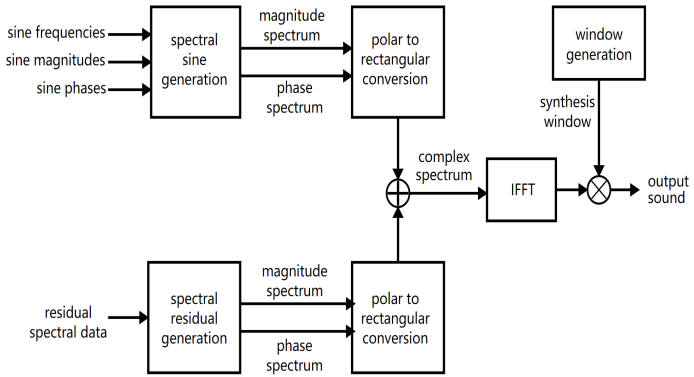
PART 02

DDSP Components

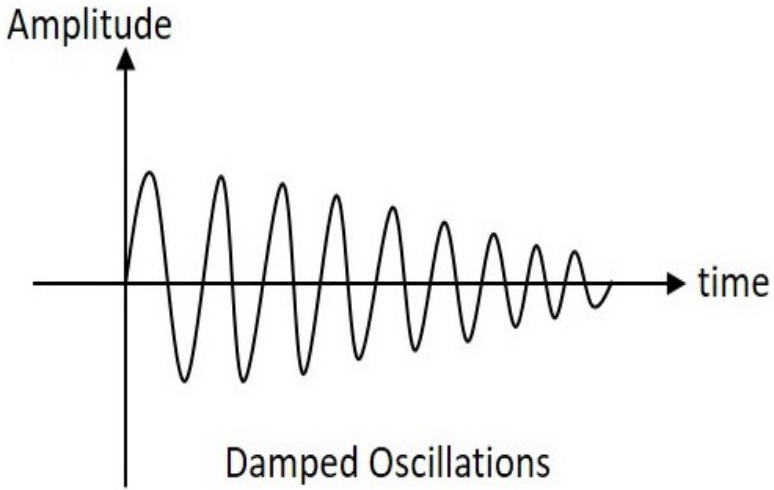
Core components (feedforward functions) of DDSP: oscillators, envelopes, and filters

Part 2.1 | Spectral Modeling Synthesis





- SMS is an acoustic modeling approach which generates sound by combing an additive synthesizer (**harmonic content**) with a subtractive synthesizer (**noise content**);
- It's widely used in many types of audio signals;
- One of its advantages is that it has many more parameters than other models, which makes it amenable to control by neural network.



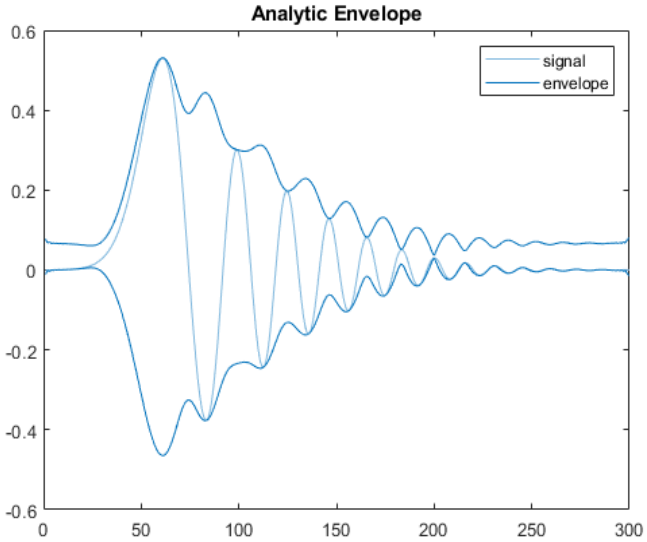
- The sinusoidal oscillator is the basis of the synthesis techniques in this paper;
- The output is entirely determined by the time-varying fundamental frequency $f_0(n)$ and harmonic amplitude $A_k(\mathrm{n})$.

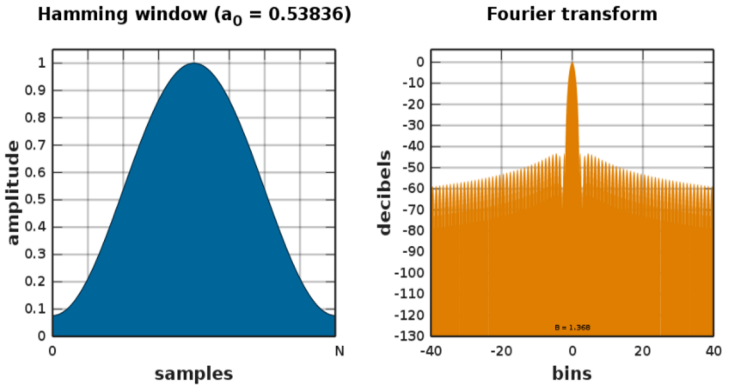
$$x(n) = \sum_{k=1}^{K} A_k(n) \sin(\phi_k(n)),$$

$$\phi_k(n) = 2\pi \sum_{m=0}^{n} f_k(m) + \phi_{0,k},$$

$$A_k(n) = A(n)c_k(n).$$

Part 2.3 | Envelopes





Hamming window, a_0 = 0.53836 and a_1 = 0.46164. The original Hamming window would have a_0 = 0.54 and a_1 = 0.46.

$$w[n] = a_0 - \underbrace{(1 - a_0)}_{a_1} \cdot \cos\left(\frac{2\pi n}{N}\right), \quad 0 \leq n \leq N,$$

- A smoothed amplitude envelope helps to prevent artifacts for the amplitude and harmonic distribution of the additive synthesizer;
- Adding overlapping Hamming windows helps to design an ideal envelope.

◆ Frequency Sampling Method

- Method: convert the networks outputs into impulse responses of linear-phase filters;
- Advantage: ensure interpretability and prevent phase distortion;

In the Fourier domain:

$$Y_L = H_l X_l$$

$$y_l = IDFT(Y_l); h_l = IDFT(H_l)$$

With:

- h_l Impulse responses
- x_l Non-overlapping frames
- y_l The frame-wise filtered audio

◆ FIR Filters

- FIR means “finite impulse response”, it’s a filter whose impulse response is of limit duration;
- Basic properties:
  - It requires no feedback;
  - It is inherently stable;
  - It can easily be designed to be linear phase;

- With the help of **Harmonic plus Noise model**, the natural sounds could be captured by combining the output of an additive synthesizer with a stream of filtered noise.

  - Harmonic part:

$$h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j\, k\omega_0(t)\, t}$$

  - Noise part:

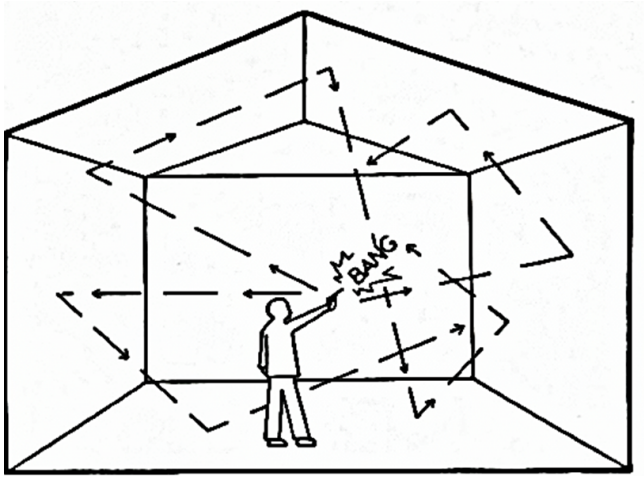
$$n(t) = e(t)\,[v(\tau, t) \star b(t)]$$

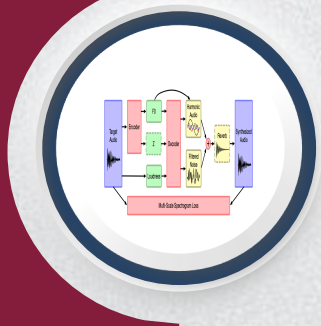
  - Speech:

$$s(t) = h(t) + n(t)$$

Part 2.6 | Room Reverberation(reverb)

- It is created when a sound or signal is reflected causing numerous reflections to build up and then decay as the sound is absorbed by the surfaces of objects in the space;
- A realistic room impulse response can be as long as several seconds;
- We implement reverb by **performing convolution** as multiplication in the frequency domain.



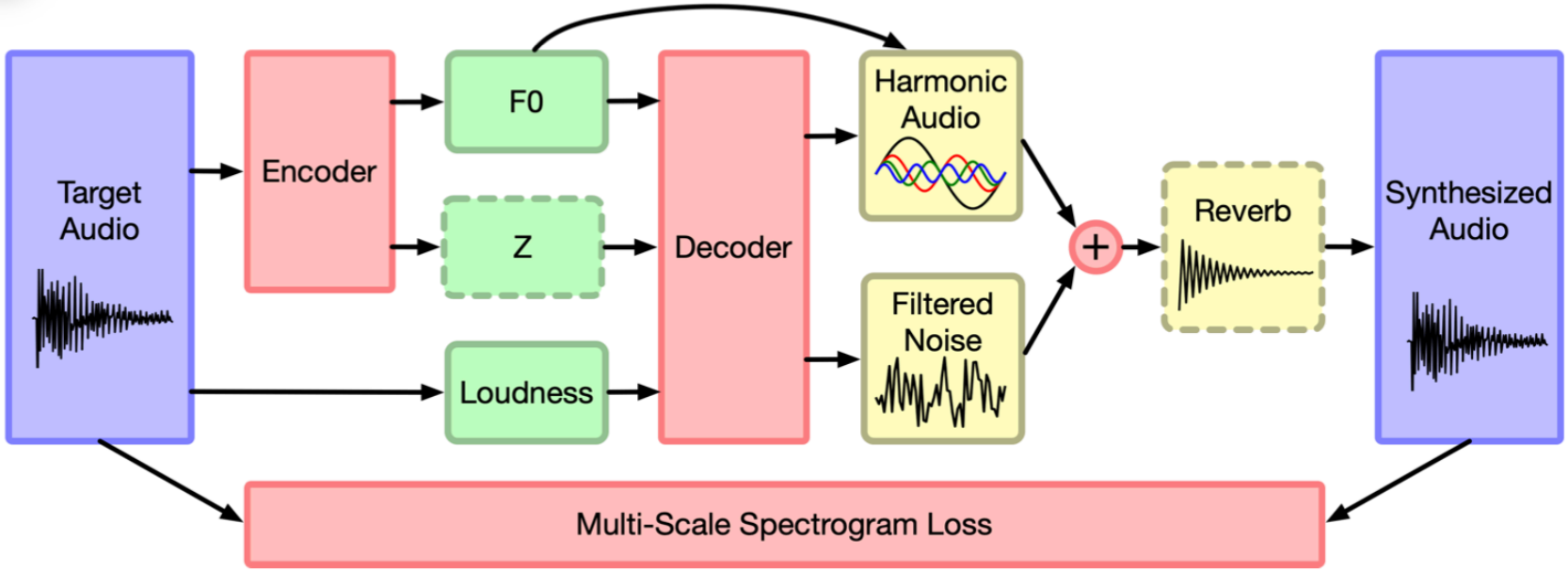


PART 03

Model Details

Detailed model of encoder & decoder and an introduction of MLP.

Part 3.1 | General Structure (Autoencoder)



- Red components are part of the neural network architecture;
- Green components are the latent representation;
- Yellow components are deterministic synthesizers and effects;
- Z is what we want to achieve.

Part 3.2 | Encoder

- f - encoder:

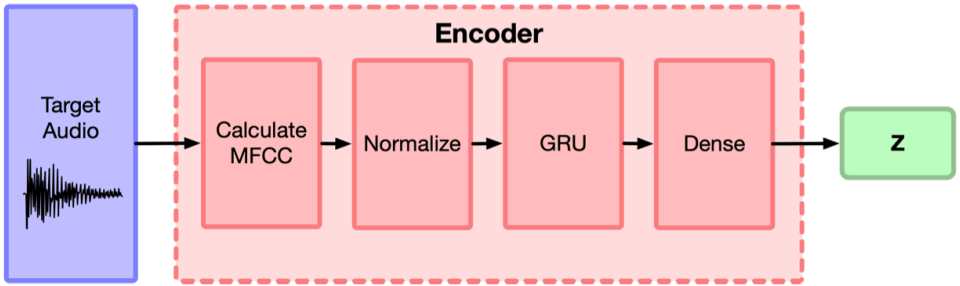
We use a pretrained CREPE pitch detector as the f -encoder to extract ground truth fundamental frequencies (F0) from the audio.

- ℓ - encoder:

We use identical computational steps to extract loudness. Namely, an A-weighting of the power spectrum, which puts greater emphasis on higher frequencies, followed by log scaling.

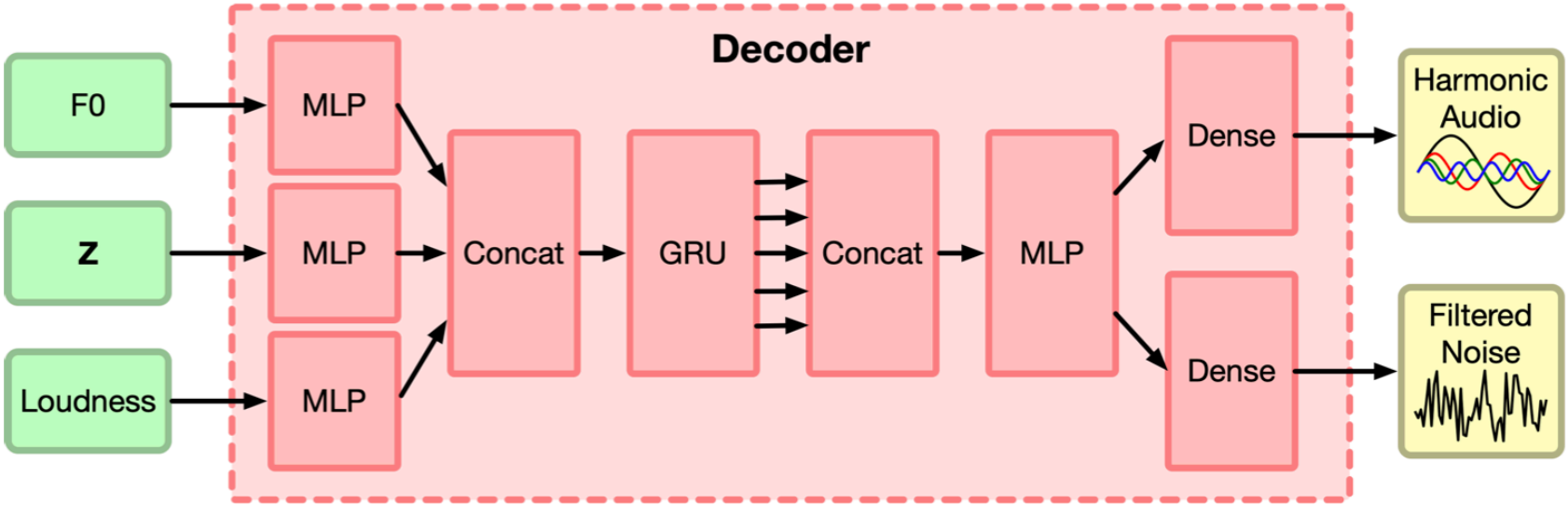
- Z - encoder:

As shown in the figure below, the encoder first calculates MFCC's (Mel Frequency Cepstrum Coefficients) from the audio.



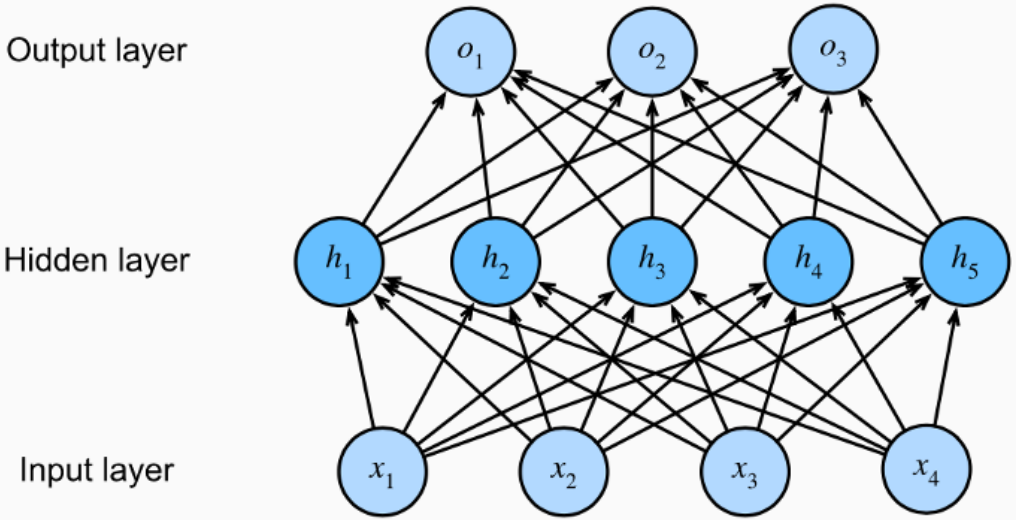
MFCC is computed from the log-mel-spectrogram of the audio with a FFT size of 1024, 128 bins of frequency range between 20Hz to 8000Hz, overlap of 75%

Part 3.3 | Decoder



- The decoder's input is the latent tuple (f(t), l(t), Z(t)) (250 timesteps). Its outputs are the parameters required by the synthesizers;

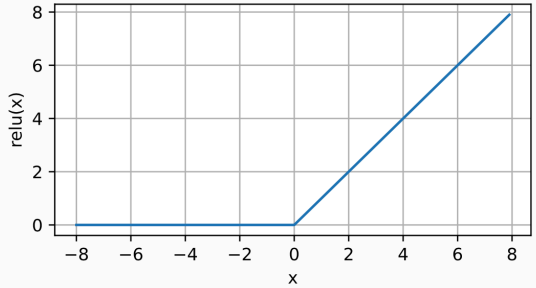
- A "shared-bottom" architecture, which computes a shared embedding from the latent tuple, and then have one head for each of the (a(t), H) outputs.



- With activation functions decide whether a neuron should be activated or not by calculating the weighted sum and further adding bias with it.

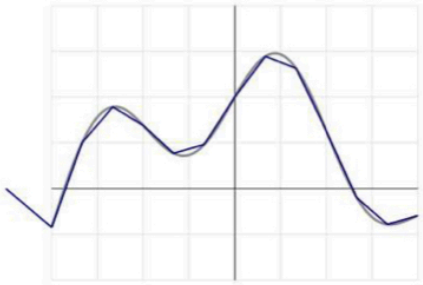
- The most popular choice is the *rectified linear unit* (*ReLU*).

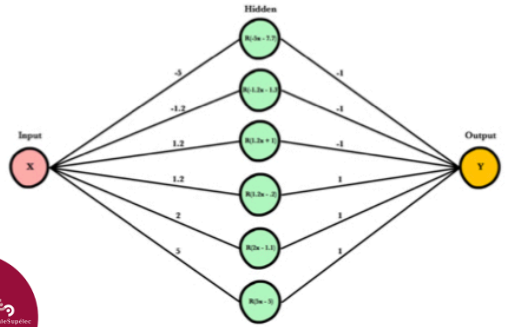
$$ReLU(x) = max(x,0)$$



Universal approximation

We can approximate any $f \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions

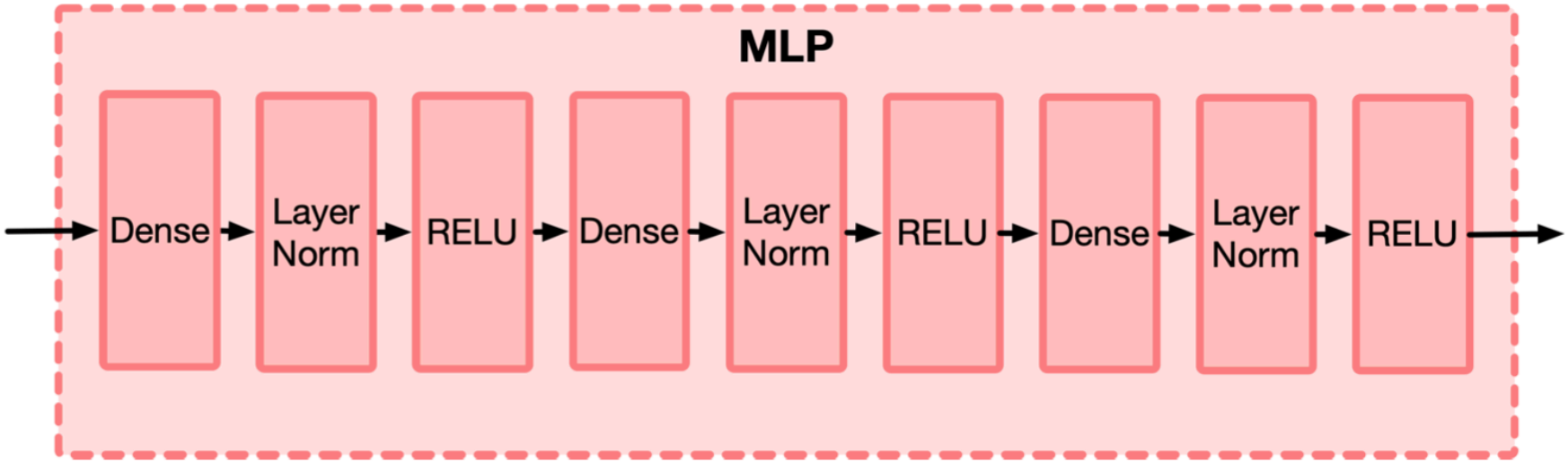




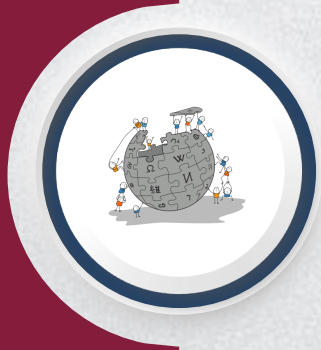
$$y = \sum_i \mathrm{Relu}(\mathrm{a_i} \times \mathrm{x} + \mathrm{b_i})$$

$$\mathbb{R} \to \mathbb{R} \text{ generalised to } \mathbb{R}^n \to \mathbb{R}^p$$

Part 3.4 | MLP



- A standard MLP with a layer normalization before the RELU nonlinearity.
- All the MLPs have 3 layers and each layer has 512 units.



PART 04

Conclusion

Contributions of DDSP compared with previous work.

DDSP models benefits from the inductive bias of using oscillators, while retaining the expressive power of neural networks and end-to-end training.

Models employing DDSP components are capable of generating high-fidelity audio without autoregressive or adversarial losses.

Further, the interpretability and modularity of these models enable:

- Independent control over pitch and loudness during synthesis.
- Realistic extrapolation to pitches not seen during training.
- Blind dereverberation of audio through separate modelling of room acoustics.
- Transfer of extracted room acoustics to new environments.
- Timbre transfer between disparate sources, converting a singing voice into a violin.
- Smaller network sizes than comparable neural synthesizers.

REFERENCE

MAJOR REFERENCE

1. ***DDSP: DIFFERENTIABLE DIGITAL SIGNAL PROCESSING*** - Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, & Adam Roberts (Google Research, Brain Team) – *arXiv: 2001.04643v1 [cs.LG] 14 Jan 2020*

2. ***Dive into Deep Learning*** - *https://d2l.ai/index.html*

3. ***Slide of Machine Learning Course 3*** - David Rousseau (IJCLab-Orsay) - *CentraleSupélec, ST4 PNT, Spring 2020*

Q. & A.

CENTRALESUPÉLEC

2A – Pôle Projet

DDSP - Differentiable Digital Signal Processing

T1 – PRESENTATION

Peizhou ZHANG Molin LIU

Xinjian OUYANG Haoyu YU