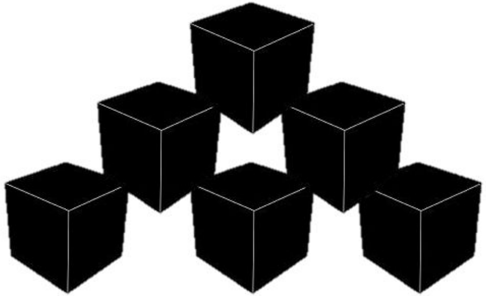
DDSP: DIFFERENTIABLE DIGITAL SIGNAL PROCESSING

Research by: Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, & Adam Roberts

Presented by: Tobias Oberkofler

Problem Description (What's it all about?!)

- Most existing generative audio models **do not utilize existing knowledge of how sound is generated and perceived**
- Can't be **interpreted**
- They require **huge amounts of data** (hours of play)
- Are **expensive to compute**(SOTA generative models often take several GPU days to weeks to train from scratch)



Problem Description (What's it all about?!)

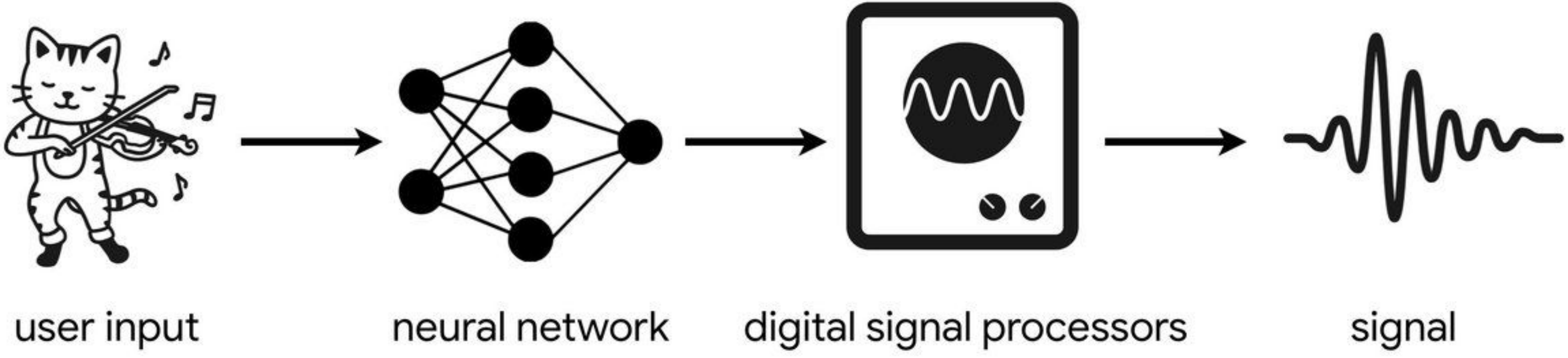
- Existing models **are not scalable**
- **Badly generalize** to unseen examples
- Can't be easily adjusted. **No way of live integration**
- No way to **control** parts separately

Proposed solution

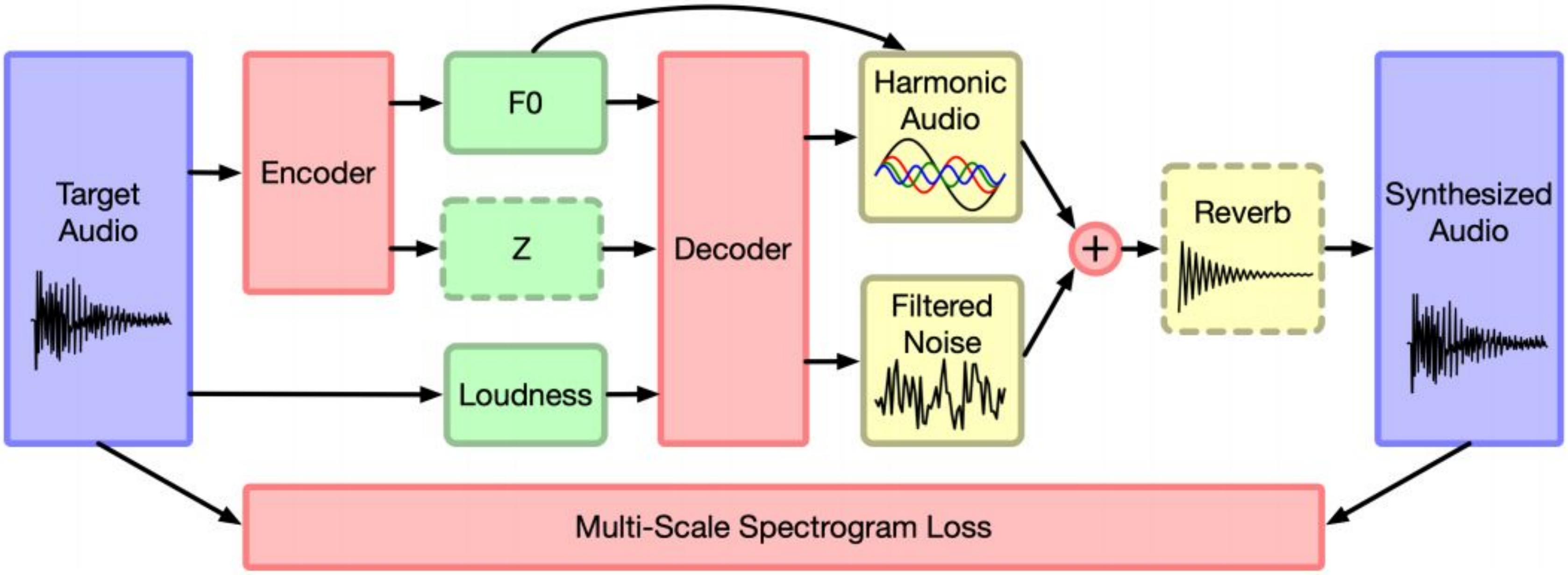
- **Utilize extensive knowledge on dsp and human perception**
  - No need for autoregressive models or adversarial losses

- Make it **modular**

- Keep it **simple** (and be aware of **limitations**)

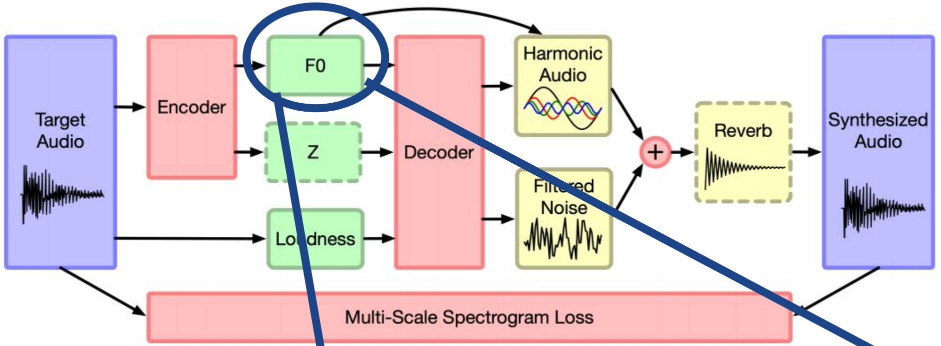


Methods



Target Audio
Encoder
F0
Z
Loudness
Decoder
Harmonic Audio
Filtered Noise
+
Reverb
Synthesized Audio
Multi-Scale Spectrogram Loss

Methods



- F0:
  - “Pitch detector”
  - Fundamental frequencies
  - Pretrained CREPE (Kim et al., 2018) network

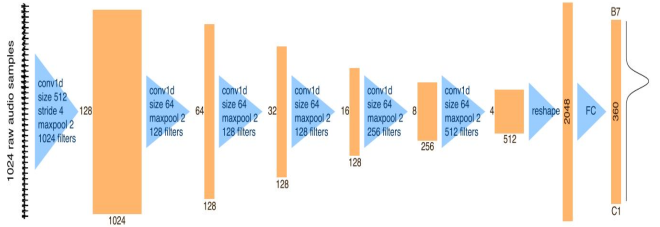
    Fixed weights for supervised task

    A convolutional network for pitch estimation (SOTA)
  - Unsupervised task: use a Resnet architecture (He et al., 2016)

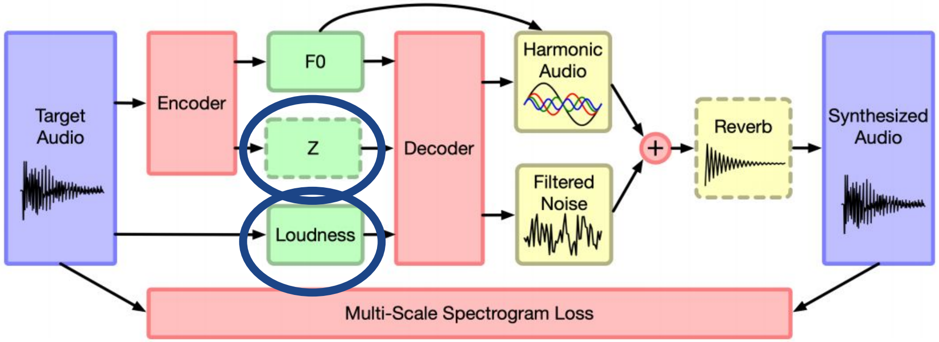
    Extremely deep

    Smart use of skip connections

    Cleverly linked conv blocks, normalizations,… fight vanishing gradient



Methods

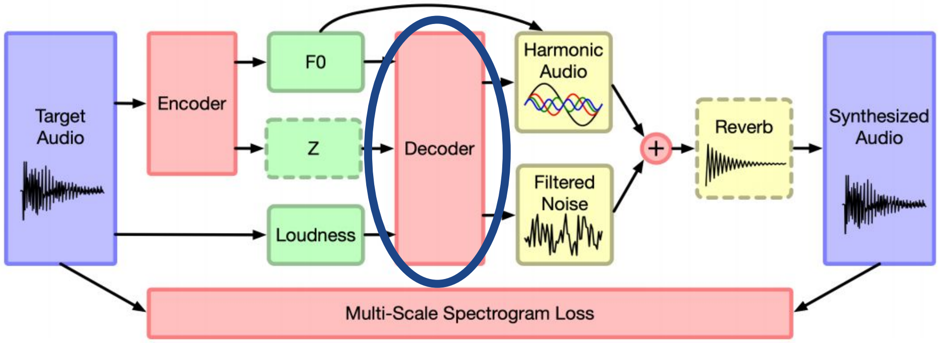


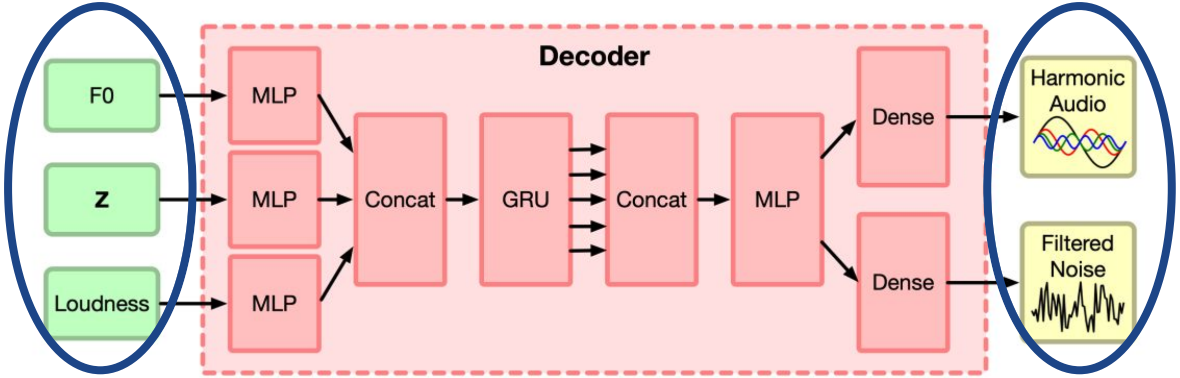
- Loudness:
  - Extracted directly from audio
  - Detect note segmentation?

- Z:
  - “residual information”
  - Use Mel-frequency cepstral coefficients(MFCC) coefficients (30 per a frame)
  - transformed by a single GRU layer

Methods

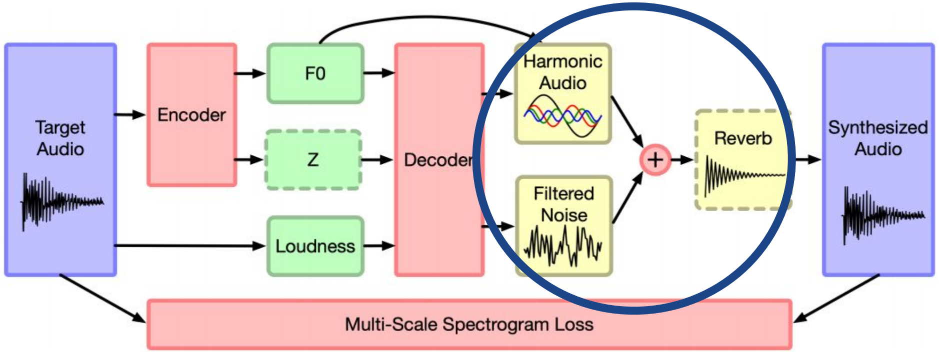
- **Map f(t), l(t) and z(t)to control parameters for the additive and filtered noise synthesizers**
- 3 simple multi layer perceptron networks (ANNs)
- Concatenate latent space
- GRU yet again
- Dense network to obtain estimated parameters

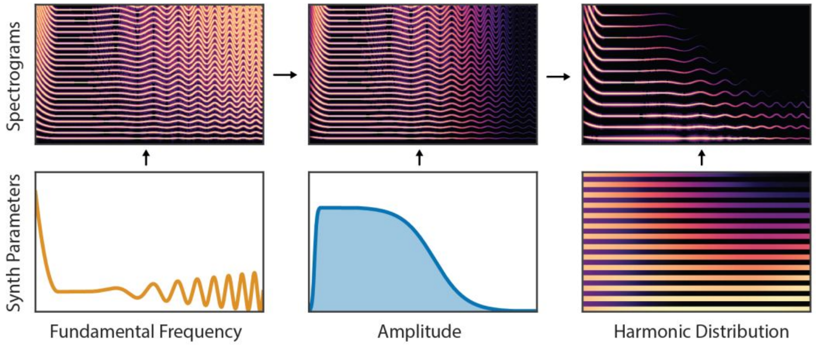




Methods

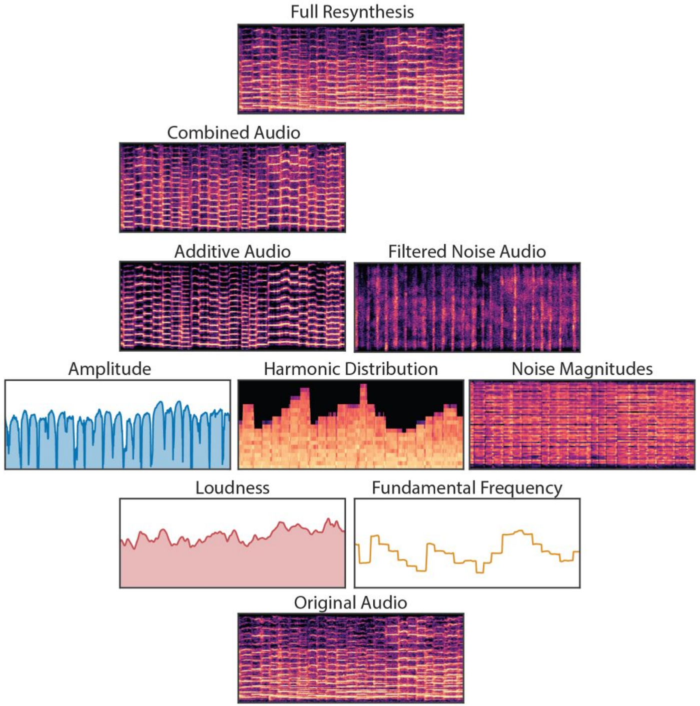
- Use **Additive Synthesizer** to generate sound out of the **Fundamental Frequency, Amplitude and Harmonic Distribution component**
- The synthesizer generates audio as a sum of sinusoids at multiples of fundamental frequency
- Allow parameters to be controlled externally





Results

- Comparable to **SOTA** model for **NSynth dataset**
- Outperforms SOTA despite more general loss function
- Even unsupervised version outperforms supervised WaveRNN



- Qualitatively show **good interpolation** (independent control c generative factors: for example loudness adjustment) and **extrapolation quality**(generalize to unseen data)
- Qualitatively demonstrated abilities in deverbaration and acoust transfer as well as timber transfer.

Demonstration Results

Conclusion

DDSP is:

- A way to **utilize** our extensive existing **dsp knowledge**

- A very **light weight network** that can be trained within hours without large amounts of data(violin model based on 13 minutes of data)

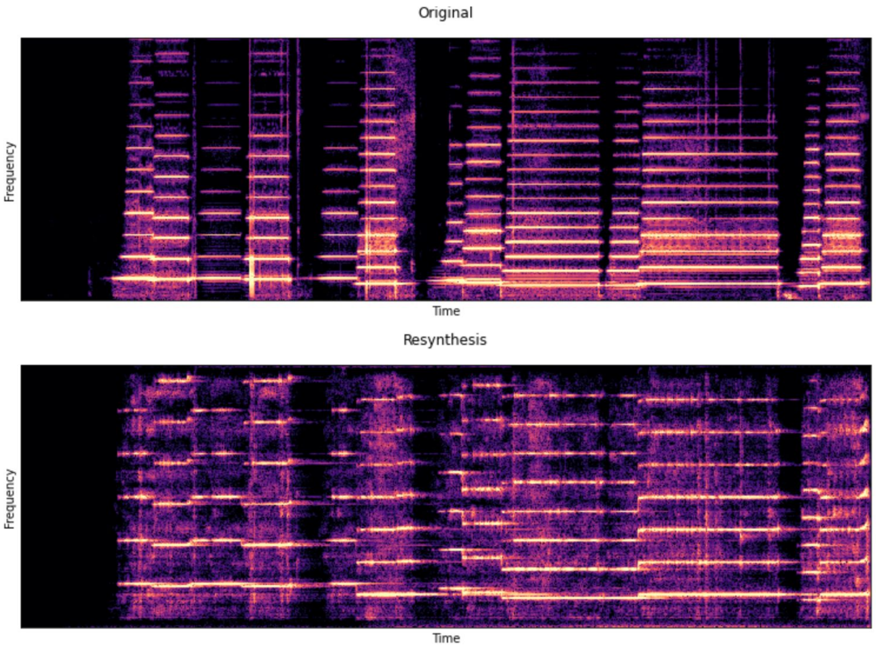
- Enables **live interaction** with DL output. From passive to active role in the process

- Astonishing **timber transfer** and reverberation ability

- Limited to **monophonic audio**. Can only handle single instrument data (extension in progress).

  Samples should share a **consistent room environment**.

Demonstration Results



Thank you for your attention!

Stay curious

Universiteit Leiden
The Netherlands

Additional interesting details on the Methods: What is a GRU?