Source Coding of Audio Signals with a Generative Model

Roy Fejgin, Janusz Klejsa, Lars Villemoes, Cong Zhou

Dolby Laboratories

ICASSP 2020

Background

- Significant advances in audio and speech synthesis using Generative Models: WaveNet, SampleRNN, WaveRNN
- Successfully applied to speech coding ([1],[2],[3],[4])
  - a low-bitrate description $\boldsymbol{h}$ is used to condition an autoregressive model: $p_{data}(\boldsymbol{X}|\boldsymbol{h})$
- Application to audio coding an open problem

[1] WaveNet based low rate speech coding (Kleijn et al, 2018)

[2] High-quality speech coding with SampleRNN (Klejsa et al, 2019)

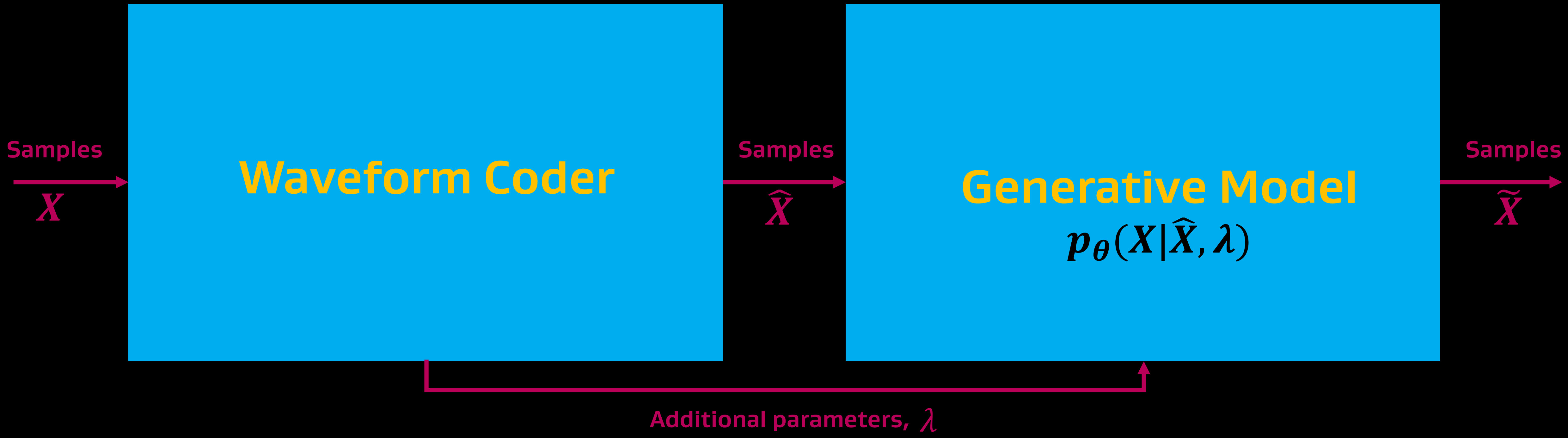
[3] A Real-Time Wideband Neural Vocoder at 1.6kb/s Using LPCNet (Valin et al, 2019)

[4] Low Bit-rate Speech Coding with VQ-VAE and WaveNet (Garbacea et al, 2019)

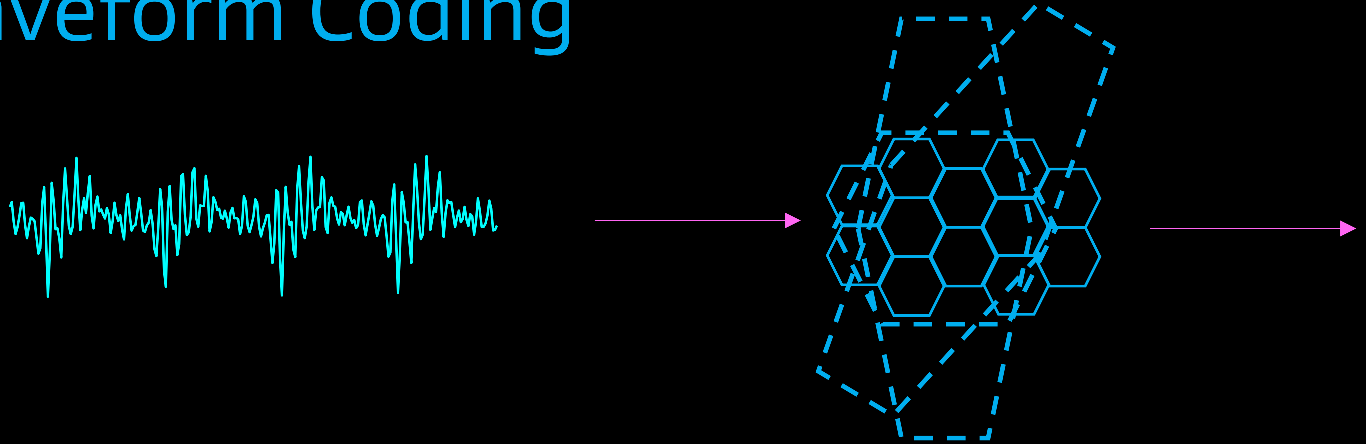
Motivation and Idea

- Generalize to additional signal categories
- What should the conditioning be?
- Waveforms!
- So the conditioning is general, but that does not mean we can *model* all audio out there

System Diagram



Waveform Coding



lossy reconstruction

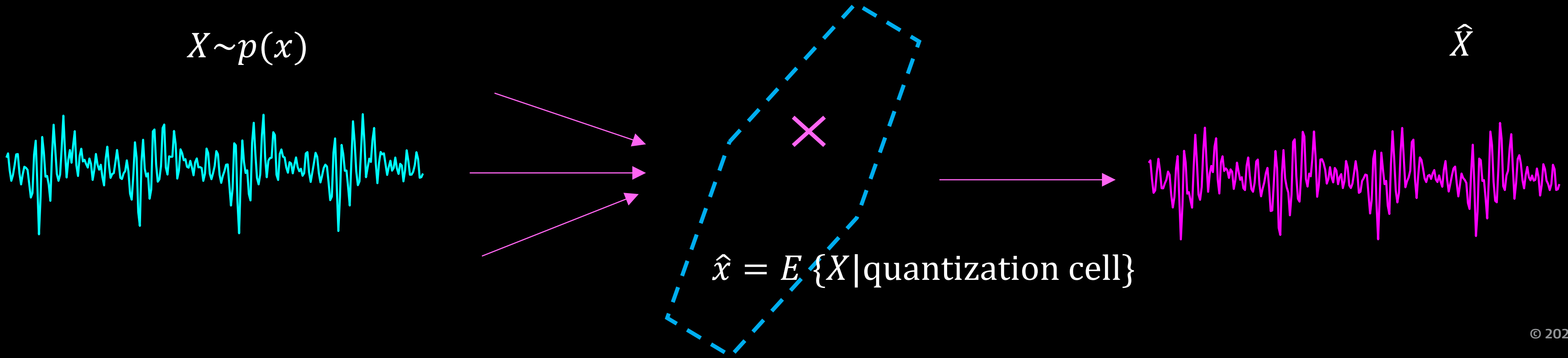
What does a classic waveform coder do?

- Quantizes to provide representation at finite bitrate (e.g. VQ, transform and scalar quantizes)
- Shapes quantization error in a perceptually optimal way (e.g. weighted squared error)
- Exploits statistical dependencies in the signal (e.g., prediction, transform, VQ)

How does it reconstruct the signal?

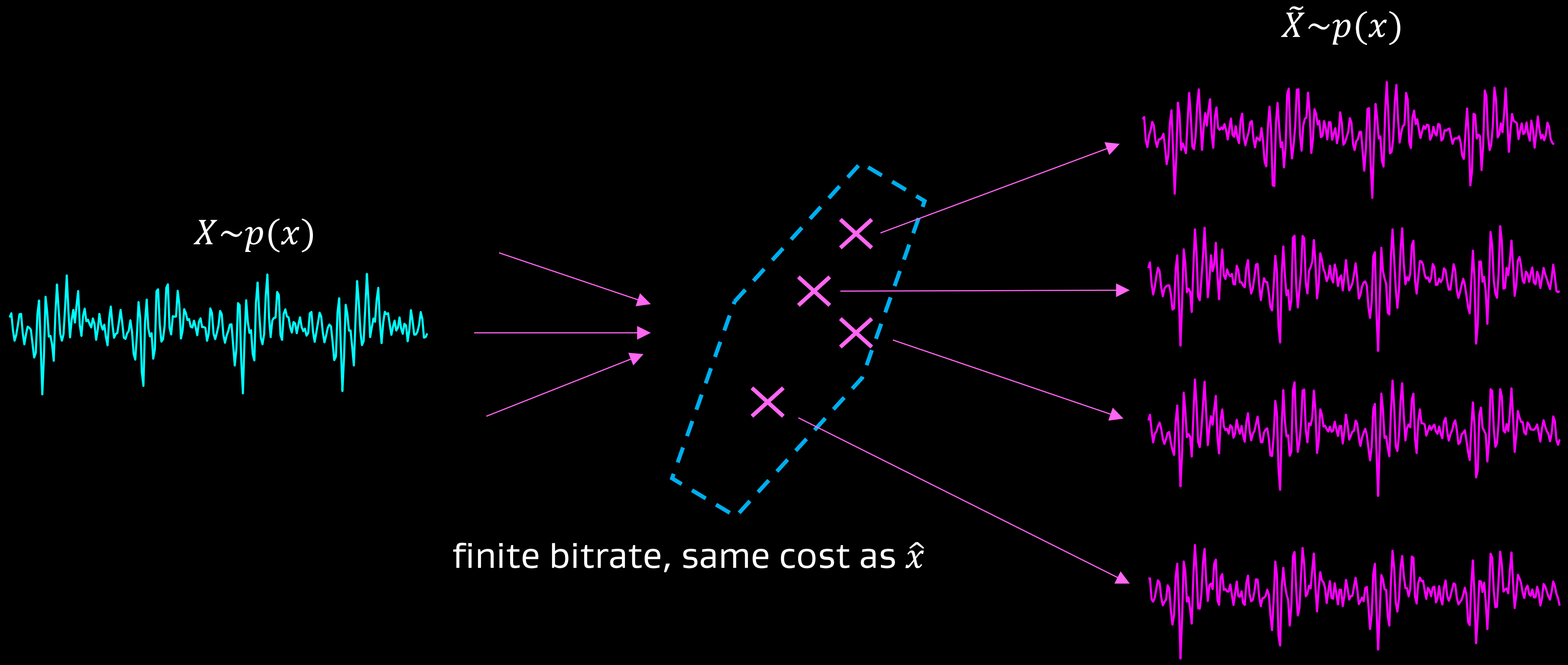
$X \sim p(x)$

$\hat{X}$



$\hat{x} = E\{X|\text{quantization cell}\}$

Source Coding with a Generative Model



reconstruction by random sampling

Theoretical Analysis

Paper includes theoretical analysis resulting in two predictions:

1. Derivation that an NLL-optimal model has

$$\mathbb{E}\{\|\widetilde{\mathbf{X}} - \mathbf{X}\|^2\} = 2\,\mathbb{E}\{\|\mathbf{X} - \mu(\mathbf{X})\|^2\}.$$

Legend:
X original signal
$\tilde{X}$ reconstruction by sampling from generative model

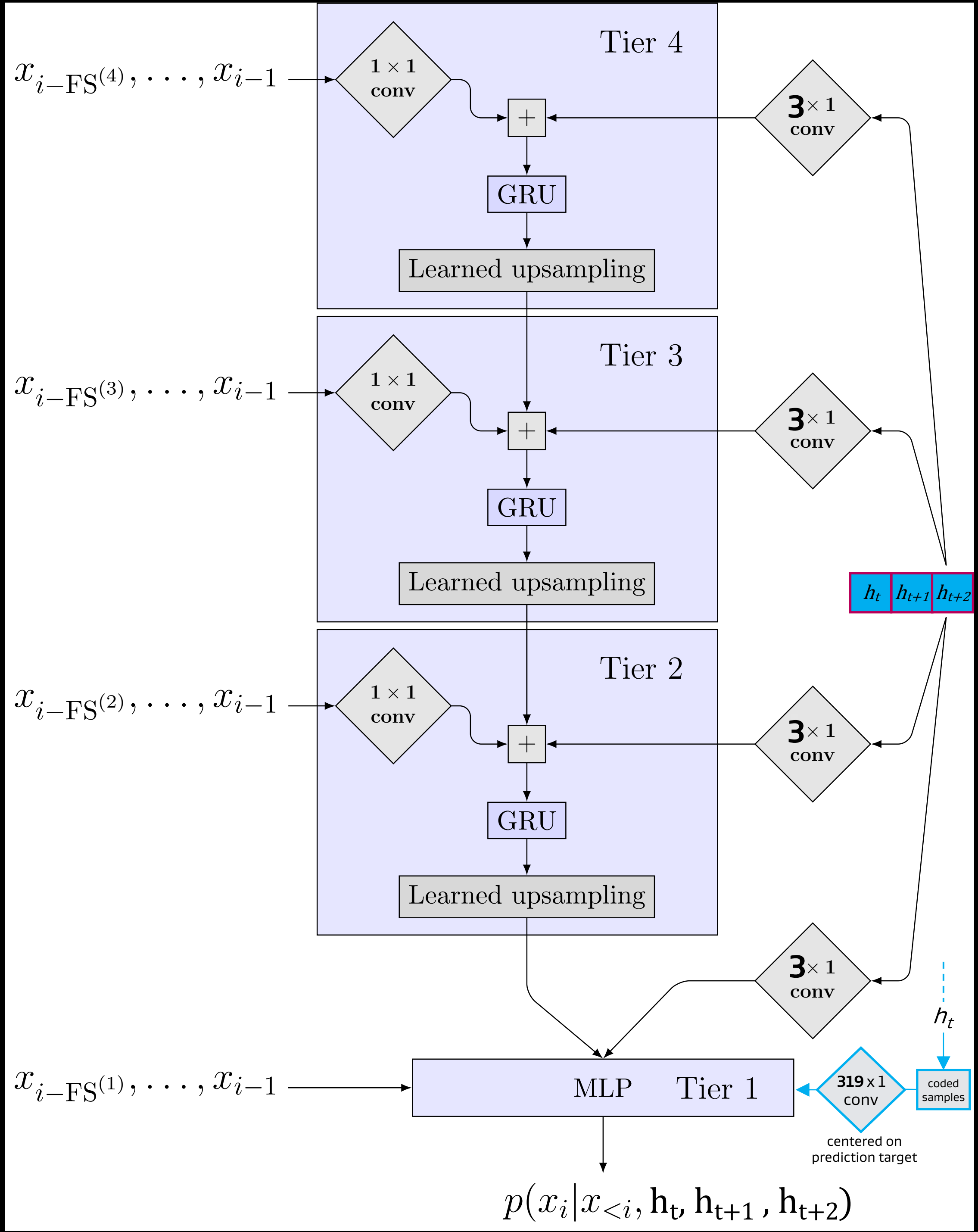
The generative scheme incurs a **3dB penalty** vs reconstruction with the mean

2. The noise shaping properties of the waveform coder will be inherited by the generative model

Conditional SampleRNN

- Similar structure as in [2]
- But conditioning is waveform, envelope
- Frame sizes tuned to signal category
- Architectural enhancements:
    - Two-frame lookahead using 3x1 conv
    - Time-aligned coded samples provided to MLP
        - Referred to as *local context*

[2] High-quality speech coding with SampleRNN (Klejsa et al, 2019)



MDCT-based Waveform Coder

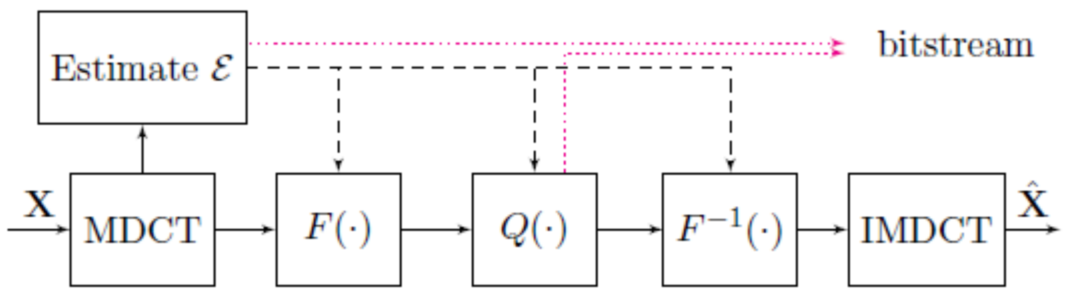


Fig. 2. Diagram of the waveform coder providing conditioning.

- MDCT with a stride of 320 samples (20 ms frames)
- Spectral envelope ε is computed across non-uniform, non-overlapping bands
- SNR allocated proportionally to square root of spectral envelope
- Transmits quantized flattened MDCT coefficients, spectral envelope i_{env}, and bit allocation parameter i_{offset}

Experiment 1: Piano

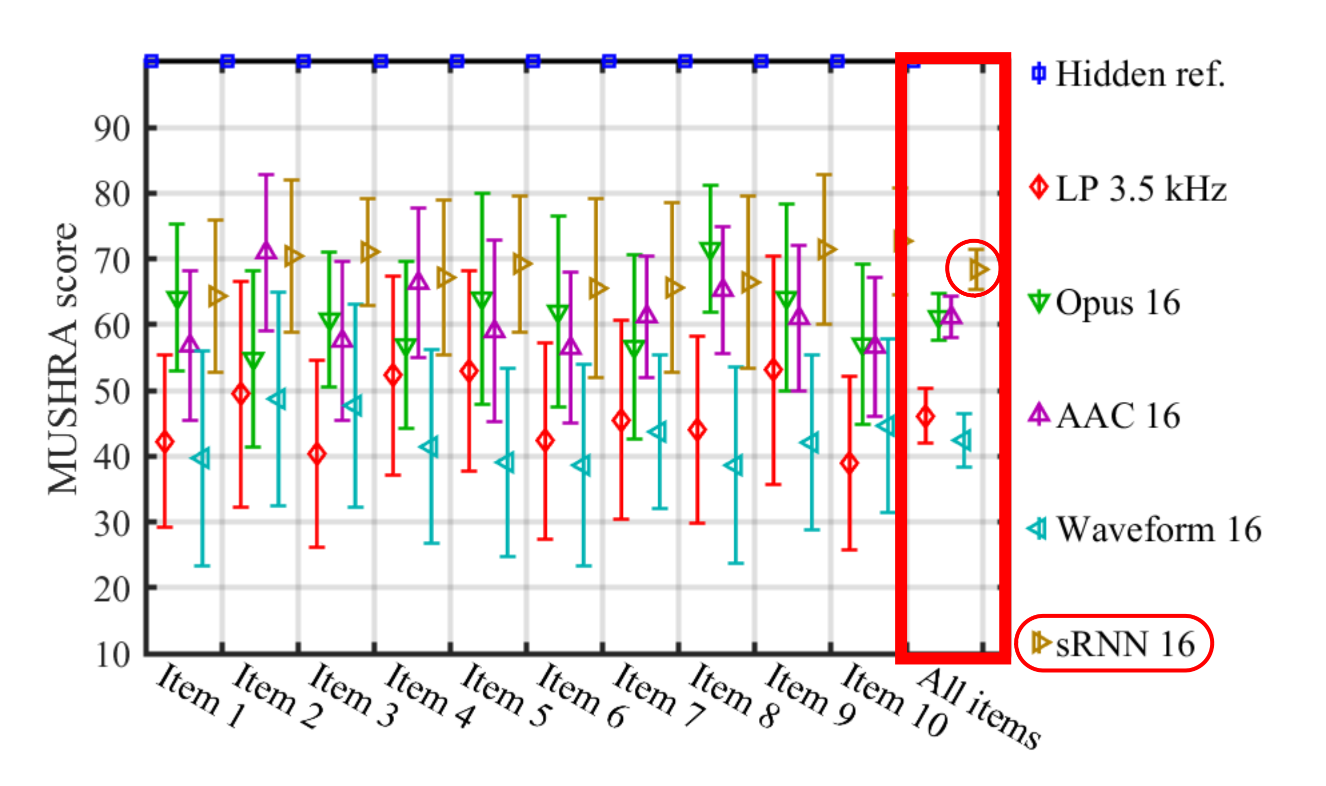
Dataset: Maestro (waveforms only)

- ~200 hours of virtuosic classical piano

SampleRNN configuration:

- frame sizes: 8, 8, 64, 320 (samples)
- 1 logistic component
- No local context

Listening Test



Experiment 2: Speech

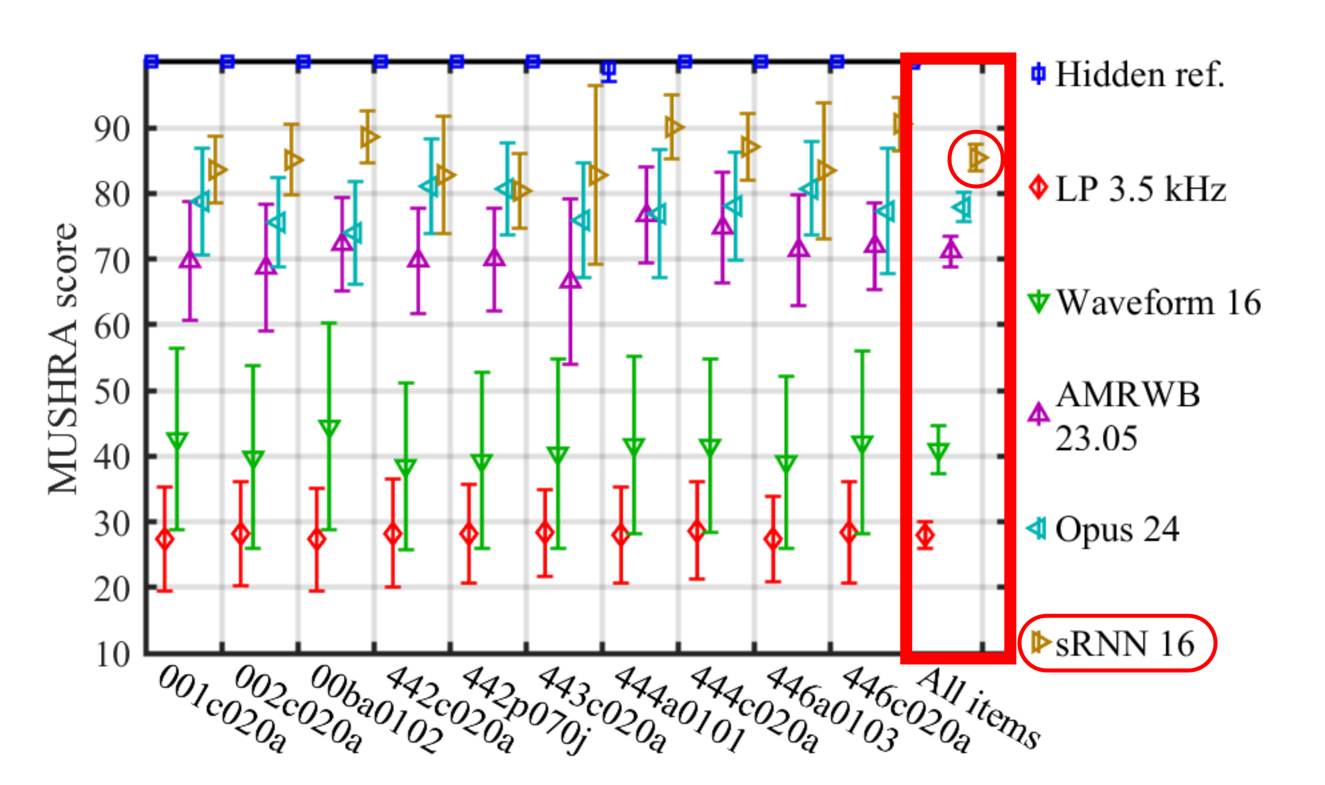
Dataset: WSJ0

- ~70 hours, 16 kHz, multiple speakers

SampleRNN configuration:

- frame sizes: 2, 2, 16, 160 (samples)
- 10 logistic components
- With local context

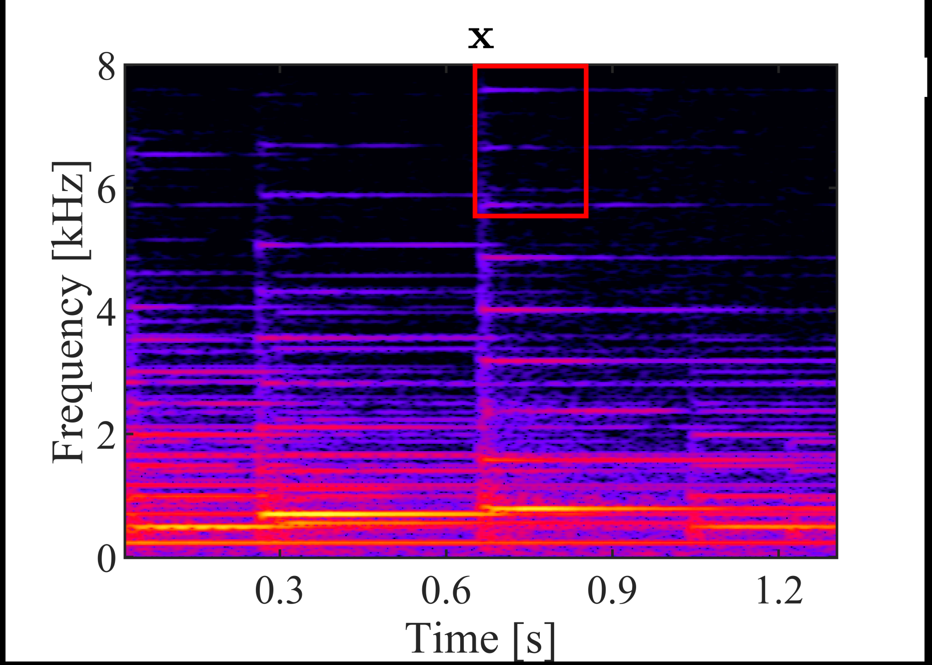
Listening Test

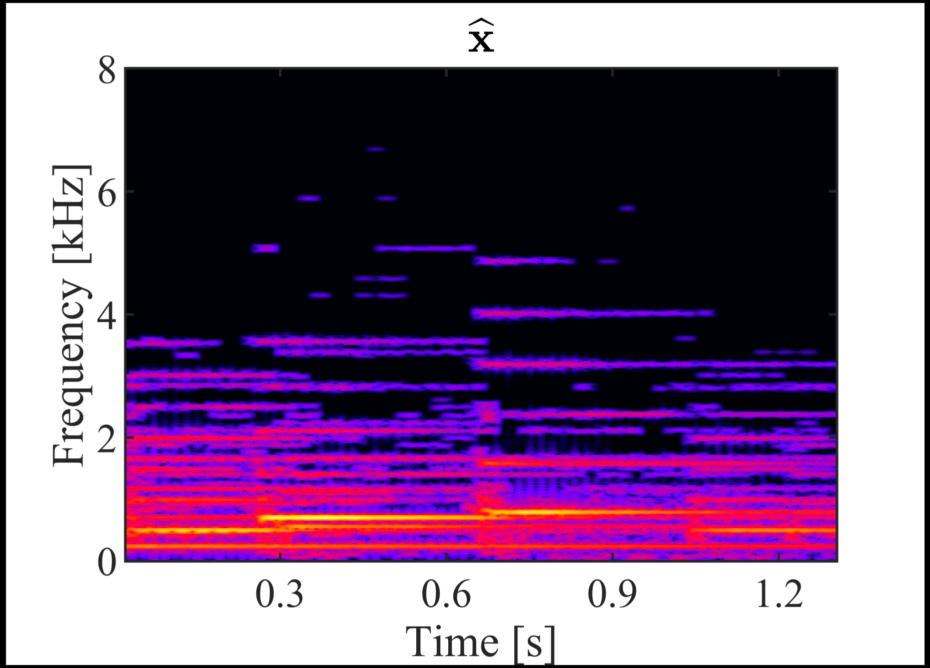


Perceptual advantage

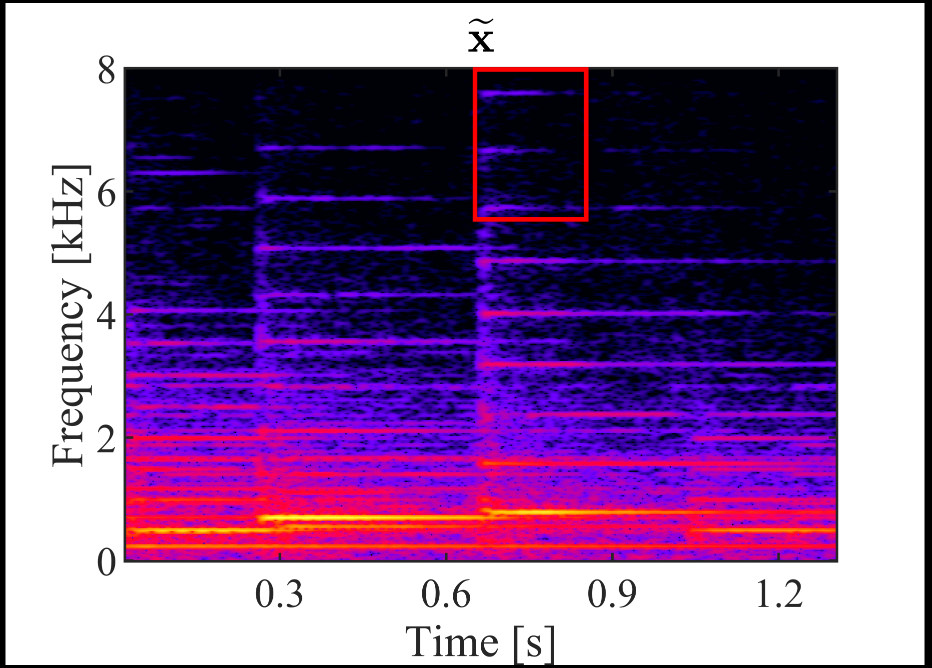
Plausible structures are generated

Original



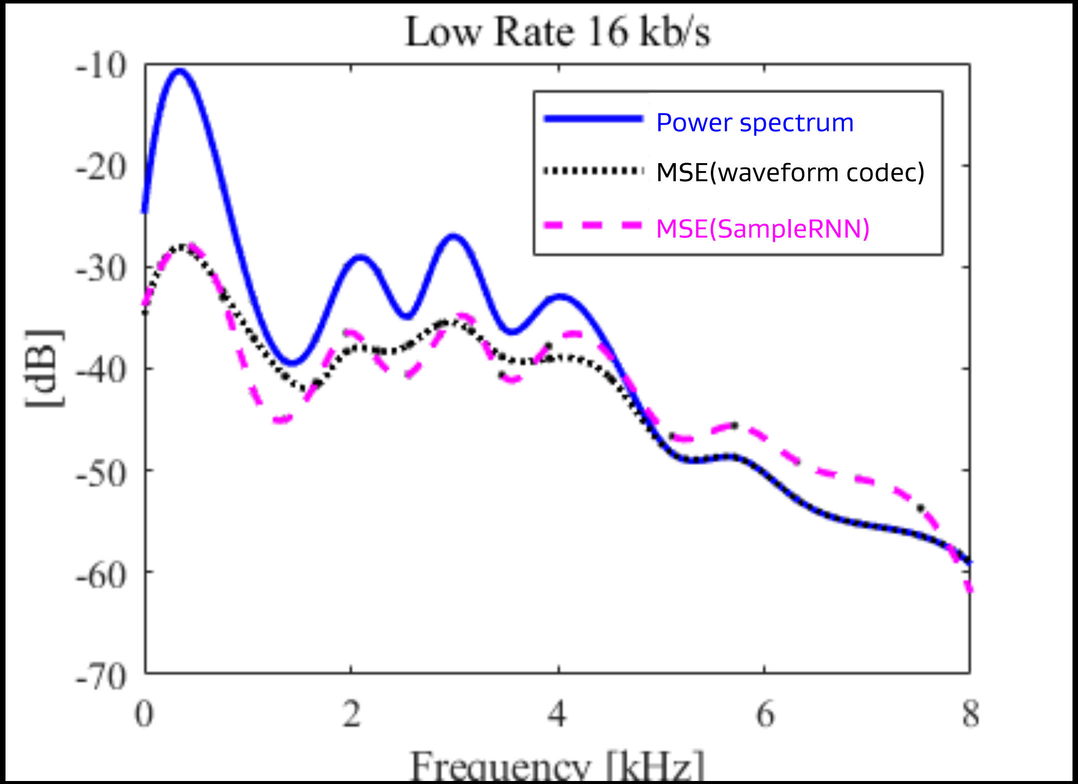


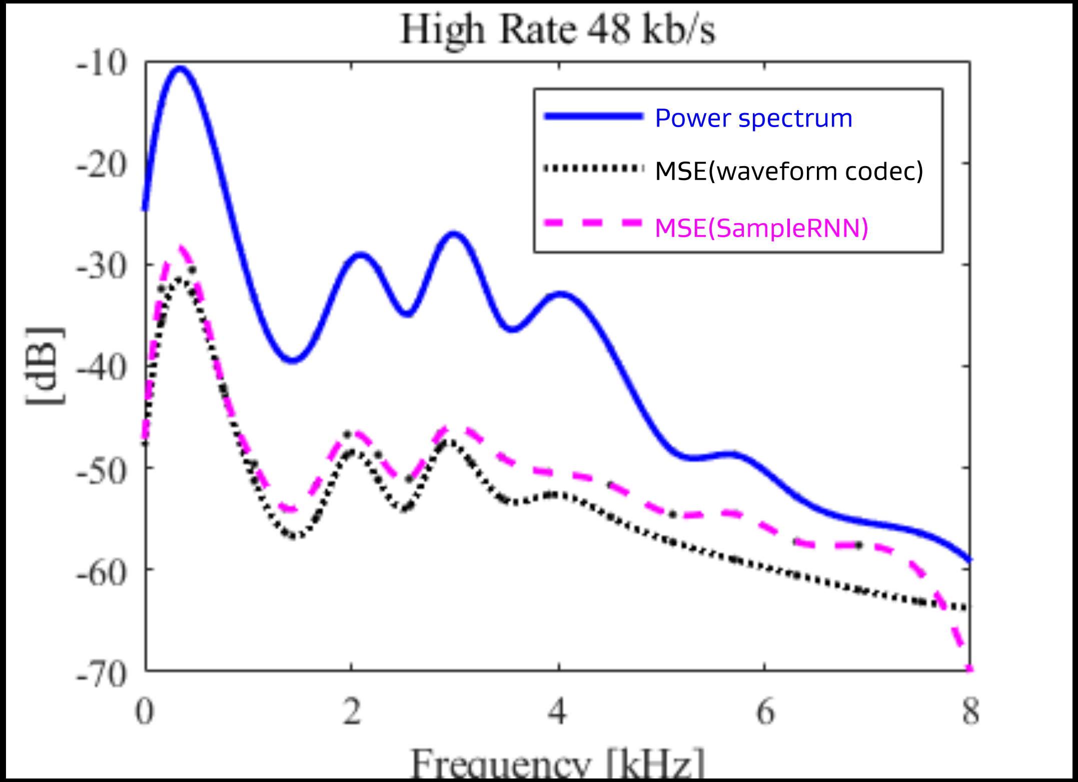
Waveform coder



SampleRNN

Objective Analysis

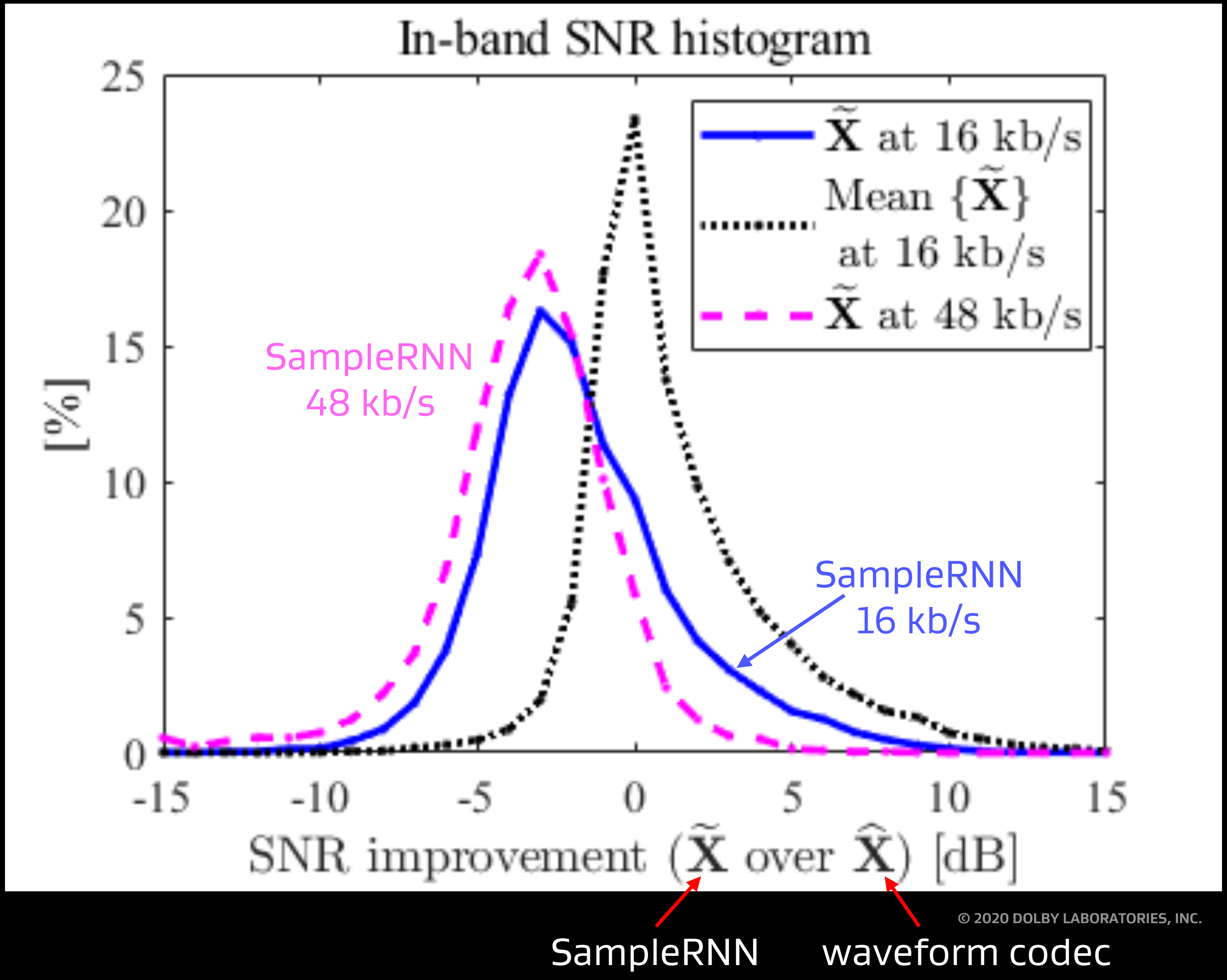




- Noise shaping of waveform codec is preserved
- At high rates, 3dB gap as predicted by theory
- At low rates, in mid frequencies, SampleRNN has lower error

Objective analysis (2)

- Histograms for low and high rate case centered around -3dB
- For low-rate case, positively skewed
- Test theory by estimating μ(X) by averaging 10 realizations
  - Indeed, 3dB gap closed and SNR improvement over baseline



Demo

Piano and speech samples can be found here:

https://sigport.org/documents/source-coding-audio-signals-generative-model

Summary

- A general set-up
- Combines advantages of waveform and parametric coding
- Operation can be described and predicted analytically
- Generalization to general audio is an open problem