Catch-A-Waveform:
Single-Sample Audio Generation with Multi-Scale GANs

Etienne Maugars Tristan Beruard

1. Problem & Motivation

Goal. Can we train an audio generative model using *a single short waveform* (10–30s), without pretraining?

Challenges specific to audio

- Extremely high temporal resolution (16k–44k samples/s)
- Long-range temporal structure (speech, rhythm)
- Rich frequency content at multiple scales

Limitations of existing approaches

- Autoregressive models require large datasets
- Spectrogram-based GANs struggle with phase
- Patch-based synthesis lacks global coherence

Catch-A-Waveform (CAW) proposes a **single-sample, multi-scale GAN** that learns audio structure progressively from coarse to fine resolutions.

2. Method Overview

CAW represents an audio signal through a **pyramid of temporal resolutions**. Each scale models a distinct frequency band.

Multi-scale formulation

$$x = \sum_{i=0}^{N} x_i, \quad x_i = \text{upsample}(G_i(z_i))$$

- x_0 captures low-frequency, long-range structure
- Higher scales progressively add high-frequency details
- Each generator is trained adversarially on temporal patches

Training proceeds **coarse-to-fine**: once trained, lower scales are frozen.

3. Key Intuition

Frequency = scale.

- Downsampling isolates low-frequency content
- Each scale learns *local stationarity* in its band
- Global structure emerges from the coarsest scale

This design allows CAW to:

- Generate arbitrarily long signals
- Recombine patterns across time and frequency
- Avoid simple copy-paste synthesis

4. Toy Experiments: Understanding the Pyramid

Experiment A — Band isolation. We generate signals while activating only one scale at a time.

- Coarse scale: global rhythm and envelope
- Fine scales: transients, noise, articulation

Takeaway. Each scale contributes a *distinct temporal pattern*, validating the frequency-based decomposition.

Spectrogram placeholders

5. Original Diagnostic Experiments

Nearest-neighbor patch analysis. For each generated window, we compute its closest match in the training waveform.

- CAW produces novel combinations of familiar patches
- Significantly less memorization than naive cut-and-paste

Cross-scale temporal alignment. Low- and high-frequency bands often correspond to *different temporal locations* in the training signal.

Patch similarity / alignment plots

6. Data Efficiency Study

How many seconds are enough?

We train CAW with varying input durations:

$$\{2s,\ 5s,\ 10s,\ 20s,\ 40s\}$$

- Short signals → overfitting, noise artifacts
- Intermediate regime (∼20s) yields best trade-off
- Longer inputs provide diminishing returns

Quality vs duration curve

7. Applications

Bandwidth Extension. Low-frequency content is preserved while high frequencies are synthesized conditionally.

Inpainting. Missing temporal segments are reconstructed using surrounding context.

Style-preserving variation. CAW generates infinite variations consistent with a single audio texture.

Application examples

8. Limitations

- Assumes local stationarity
- Fails on highly non-stationary signals (chirps, strong reverb)
- No semantic understanding (phonemes, harmony)
- Sensitive to first-scale bandwidth selection

Observed failure mode. Reverberation is often transformed into high-pitched noise at fine scales.

9. Discussion & Conclusion

What CAW does well

- Extreme data efficiency
- Interpretable multi-scale structure
- High-quality texture synthesis

What it does not

- Learn symbolic or semantic structure
- Replace large-scale pretrained audio models

Perspective. CAW highlights the power of *architectural inductive bias* and offers a compelling baseline for single-sample generative modeling.

References

[1] Catch-A-Waveform: Learning to Generate Audio from a Single Short Example.
SinGAN: Learning a Generative Model from a Single Natural Image.