Learning to Denoise Historical Music

Yunpeng Li, Beat Gfeller, Marco Tagliasacchi, Dominik Roblek {yunpeng,beatg,mtagliasacchi,droblek}@google.com

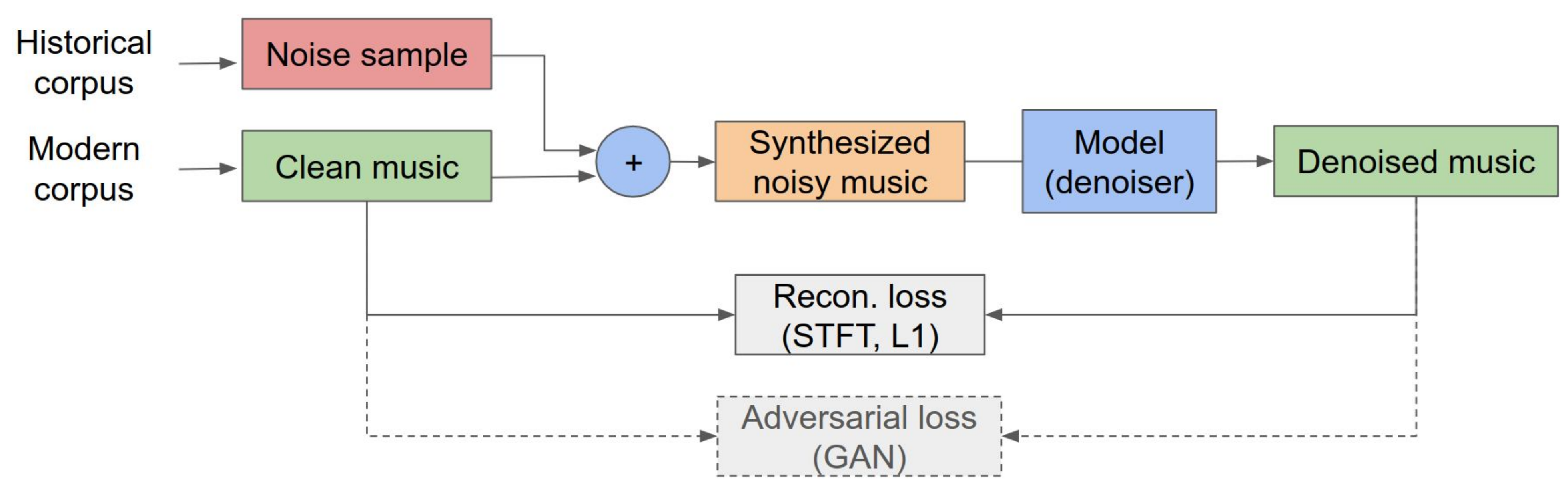
Introduction

Motivation

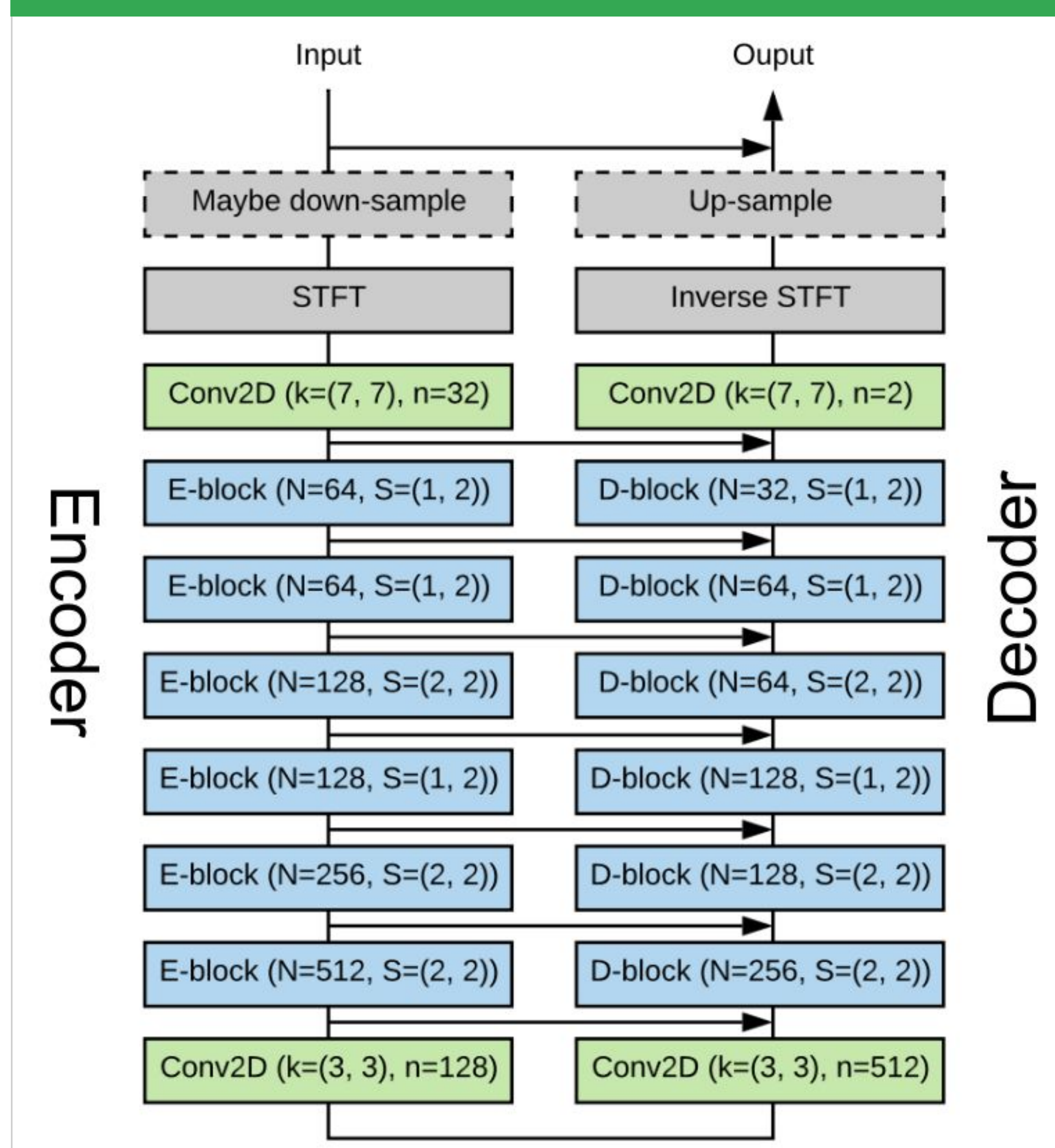
- Historical music: old equipment, analogue media → Noisy!
  - Popping, crackling, hissing, etc.
- Manual "remastering" is labor-intensive
- Objective:
  - Automated method
  - Direct audio-to-audio
  - Handles ***real*** recording with ***real*** noise

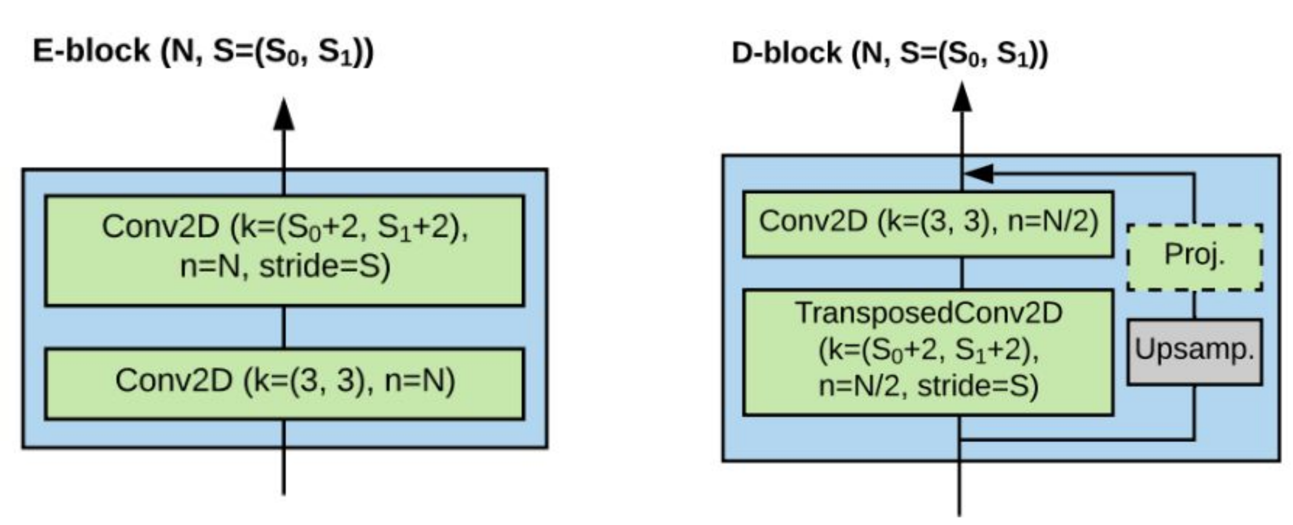
Approach

- Neural nets + supervised learning
- Obstacle: Historic music has no ground truth!
- But: We have "ground truth" noise samples -- in "silence"
- Solution: Synthesize
  - Clean target = clean modern recording (plenty)
  - Noisy input = clean target + real noise samples
    - Also simulate frequency loss by bandpass filtering
- Training objective:
  - STFT reconstruction loss + optional adversarial loss



Model Architecture





Fast: real-time speed on CPU!

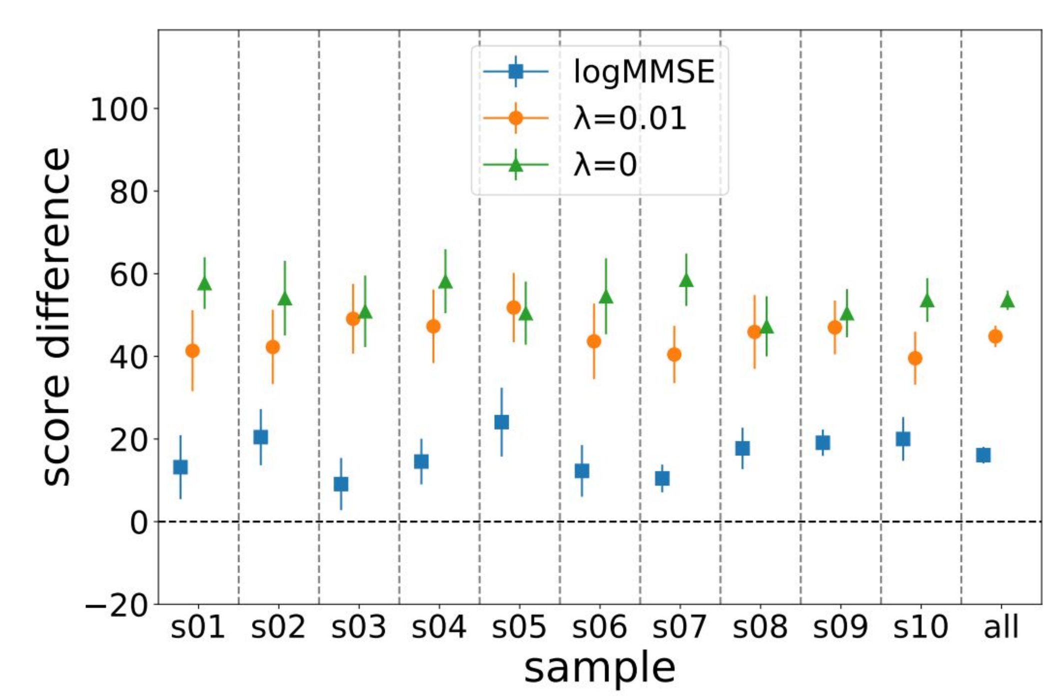
More Details

- U-Net: Encoder/decoder + skip connections and residual blocks
- 2D convolutional on complex STFT spectrograms
  - Real/imaginary component = 2 input/output channels
- Can be made multi-scale
  - Together with multi-scale reconstruction/adversarial losses
- Discriminator (if used):
  - Same architecture as encoder (w/o U-Net skip connections)

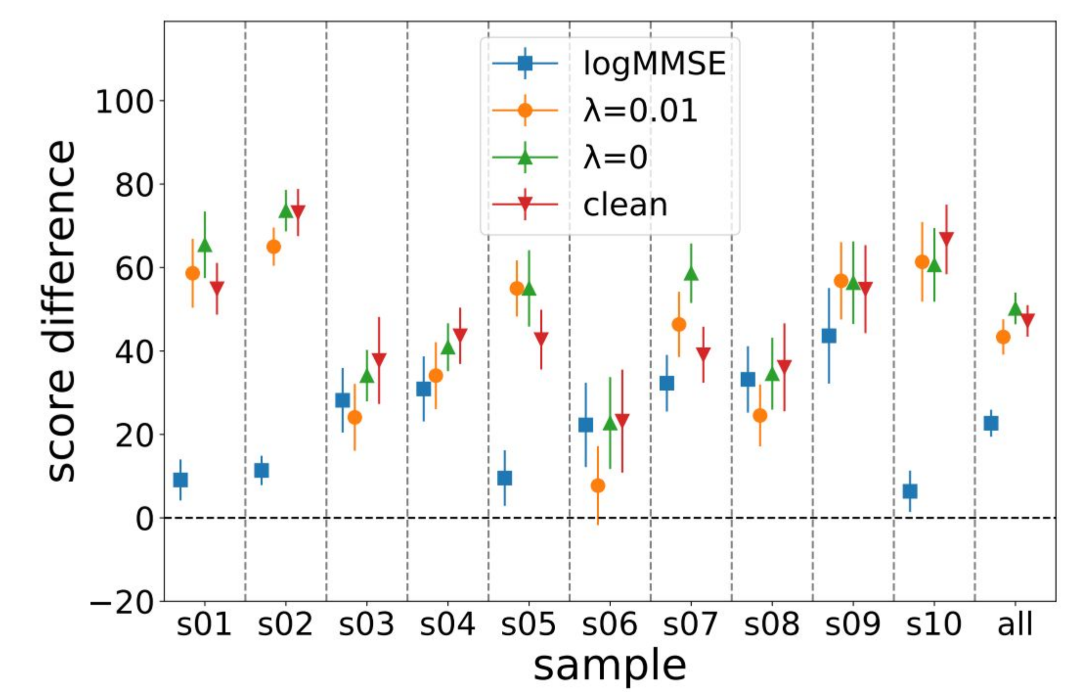
Evaluation

Methodology

- MUSHRA, which measures human-perceived quality
  - Score scale: 0--100
- 11 rates, 10 audio samples of 5 seconds each
  - Sampling rate: 44.1KHz
- On both real historical music and synthesized noisy music
  - **λ=0.01**: Trained with adversarial loss (weight=0.01)
  - **λ=0**: No adversarial loss, i.e., reconstruction loss only
  - **LogMMSE**: A well-established signal processing baseline
- Scores shown: Difference between a method's output and noisy input.



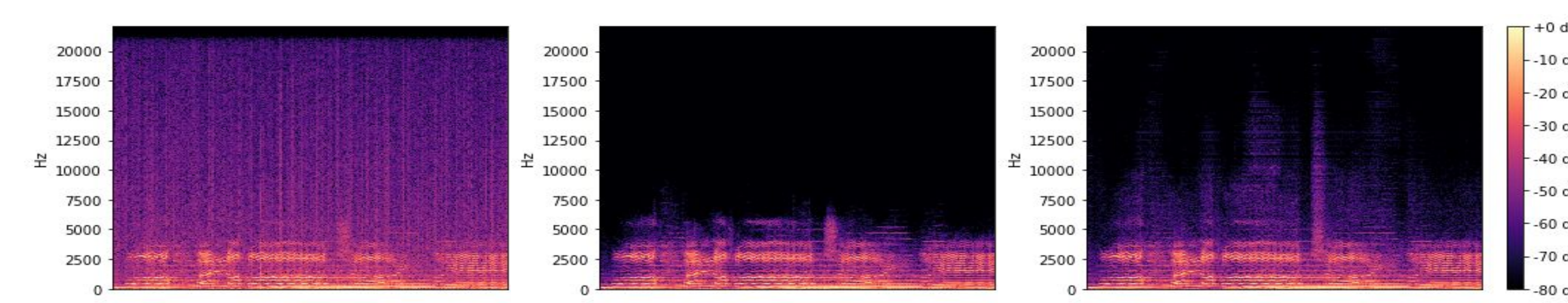
Real historic music recording samples



Synthesized noisy music samples

Observations

- Our method much better than LogMMSE
- Reconstruction-loss-only model (λ=0) scored better than adversarially trained model (λ=0.01)
  - Difference is statistically significant
  - Reason:
    - Adversarially-trained model enhances more aggressively. Possibly wider frequency range, but also more artifacts
      - See STFT spectrograms above
    - Human are more sensitive to artifacts
- On synthesized test samples, the average score of our better model (λ=0) is statistically indistinguishable from that of ground truth



Input Denoised, λ=0 Denoised, λ=0.01

Web Links

Audio Samples

- Github page: https://google-research.github.io/seanet/music-denoising
  - YouTube playlist of interleaved noisy and denoised audios (direct link)
  - Live-switchable demo between noisy and denoised audios (direct link)

Paper

- arXiv: https://arxiv.org/abs/2008.02027