Audio Denoising

Ido & Yifu

Introduction

Removing unwanted noise from an audio signal

Closely related to source separation

Various applications： speech recognition, music production, and hearing aid devices

Traditional methods： spectral subtraction, Wiener filtering, and wavelet-based methods.
struggle with complex noise patterns and can result in distortions in the audio signal.

Deep learning methods for denoising

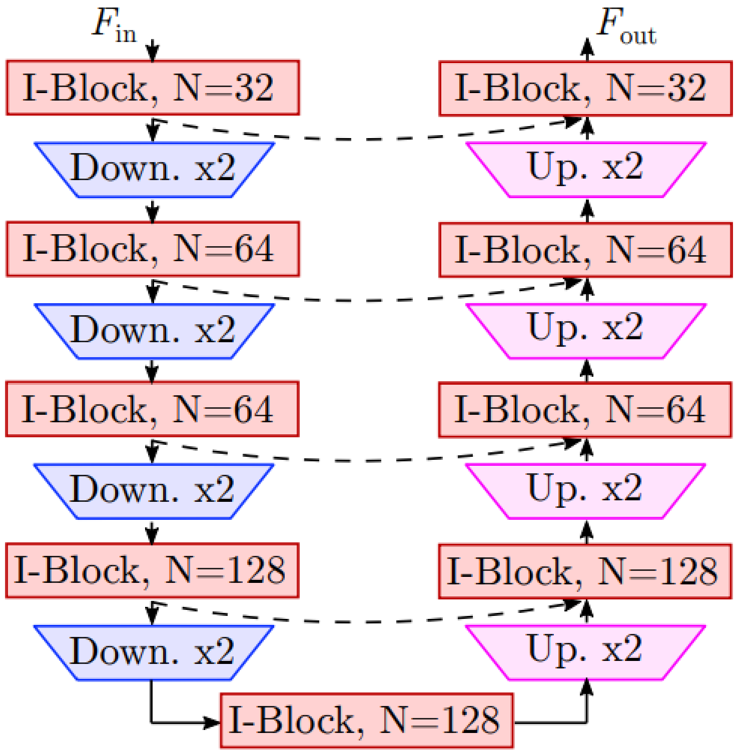
higher accuracy and robustness
learn from large and diverse datasets.

- Variational Autoencoders (VAEs)
- Generative Adversarial Network (GAN)
- **Supervised learning with a U-Net**

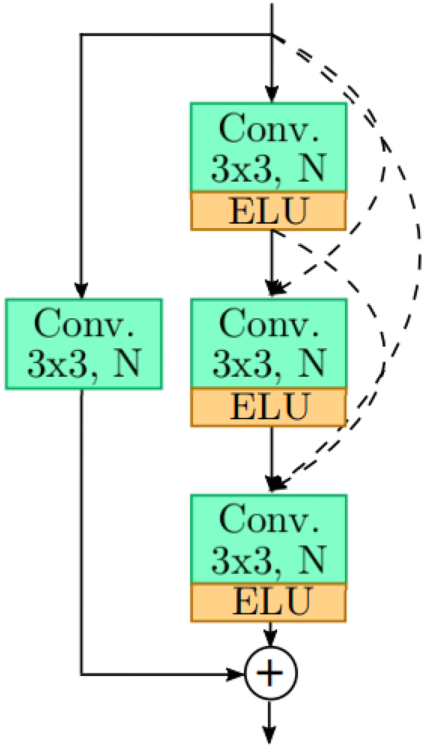
U-Net- Convolutional Neural Network for Image Segmentation

- First introduced by Ronneberger, Fischer and Brox for Biomedical purposes [1]
- Useful for semantical segmentation- individual pixel classification for multi-channel input
- How is this relevant to denoising **audio signals**? STFT representation is processed as a 2-channel (magnitude, phase) image-like representation of the signal

U-Net Structure



U-Net Structure



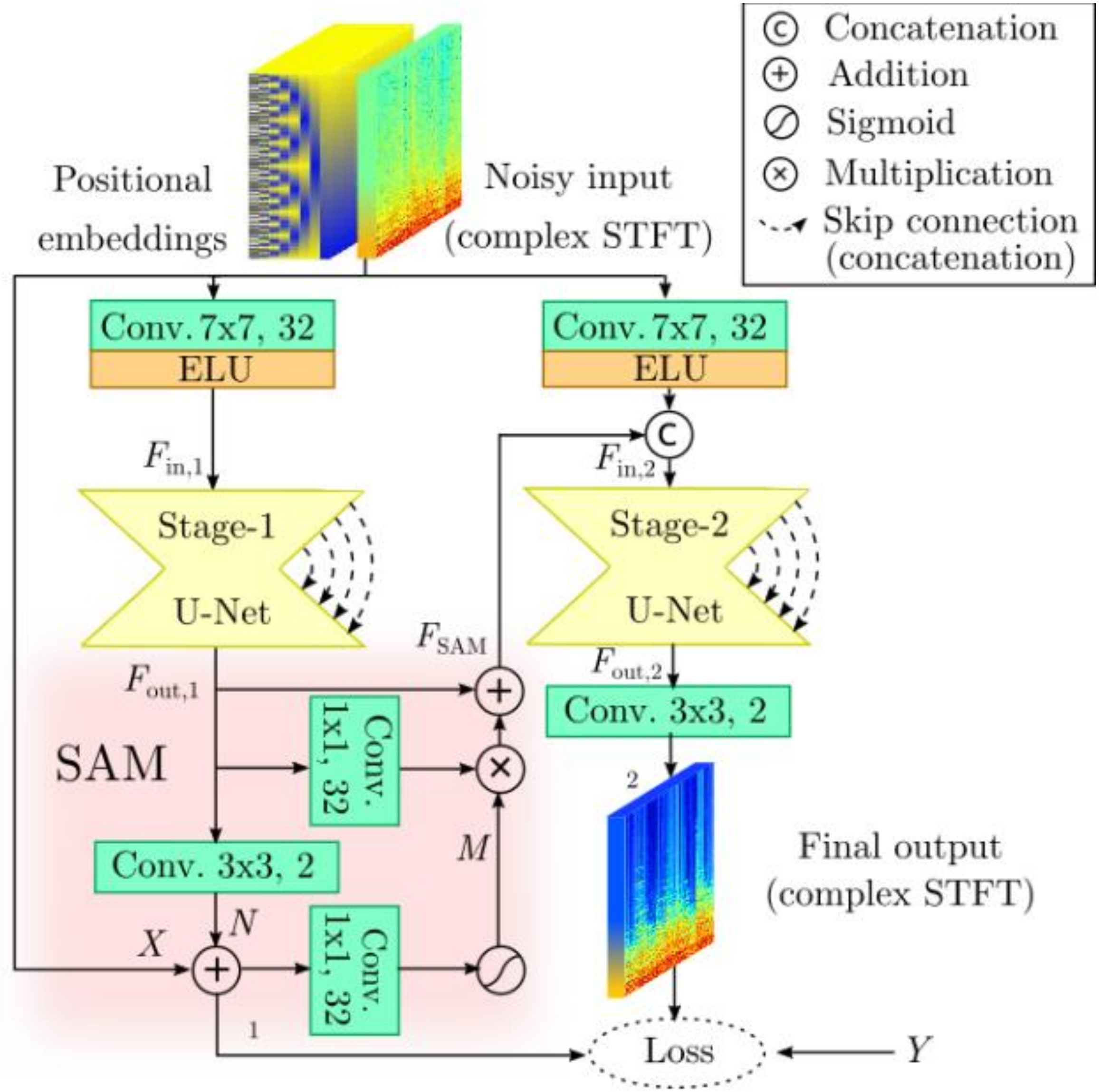
IBlock Structure

One part of the network acts as an encoder, with alternating intermediate convolution blocks and downsampling layers.

Each block sends skip connections to its counterpart in the decoder part- which alternates the same blocks with upsampling layers (Convolution transpose)

Structure

- Transform input sample to STFT-representation, add positional embeddings (10 additional channels)
- Process input through Stage-1 U-Net, SAM (supervised attention mechanism)
- Creates higher-level features (concatenated to input) for second U-Net to learn from
- Training loss is sum of L1 (mean of absolute difference) losses of both SAM and Stage-2 UNet outputs (we try to minimize error on SAM output to give the Stage-2 U-Net meaningful features)
- Smaller Model- only Stage-1 U-Net, preceded and followed by convolution layers, added 2-channel input
- **Smaller model has 6877826 trainable parameters, larger model has 13793668 (twice as many)**



Our task - denoise wildlife recordings (get rid of environmental noises)!

Dataset & Preprocessing 1

- The **ESC-50** dataset is a labeled collection of 2000 environmental audio recordings
- 5s-clips, 40 clips each of 50 different classes across natural, human and domestic sounds
- Choose animals class samples as "clean" audio, human created sound as noise.

$$X = \beta(Y + \alpha N)$$

X: Mix audio
Y: Clean audio (animal)
N: Noise

β： level scaling factor [-6 dB, 4 dB]
α： signal-to-noise-ratio [2 dB,20 dB]

Baby crying 🔊

Wind 🔊

Footstep 🔊

Dataset & Preprocessing 2

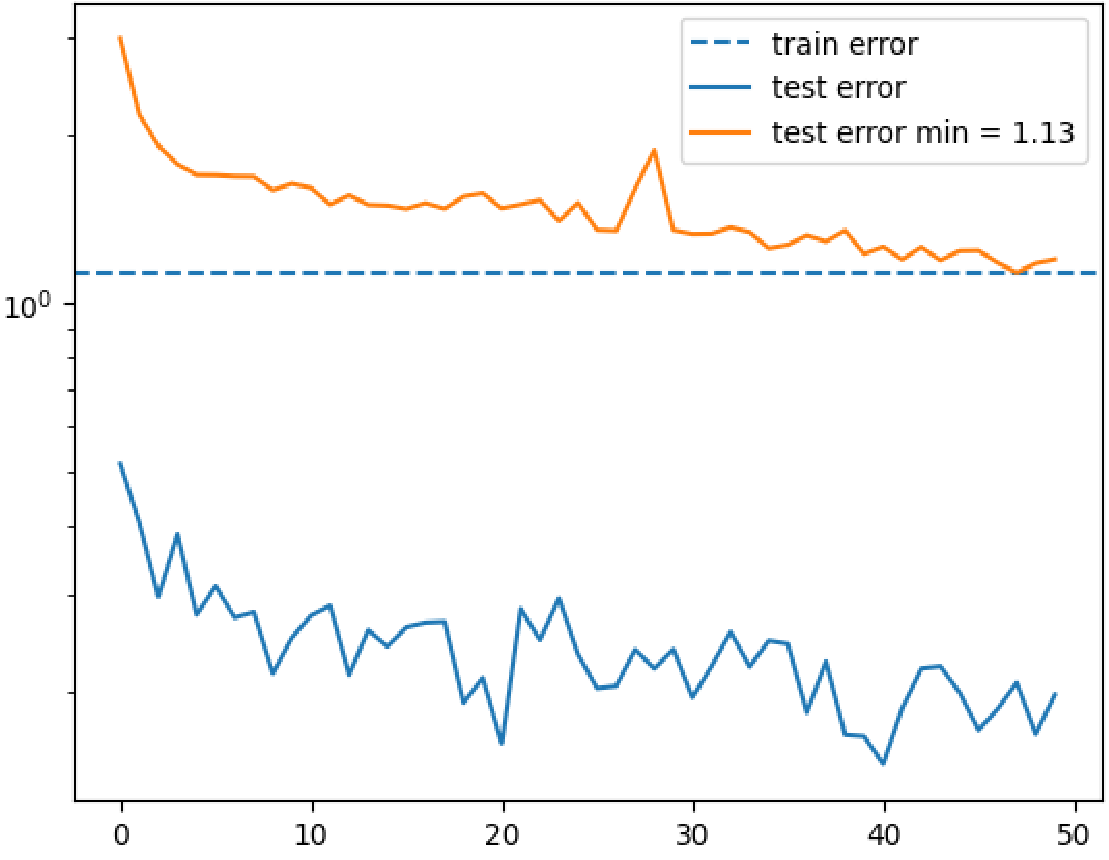
- Noise injection: truncate the noise-classes into small segments(2048 samples), window them (with Hamming window), and then combine them randomly with clean sample to get noisy audio.
- Prevention against: the model "over-fitting" to the noisy samples

Plane

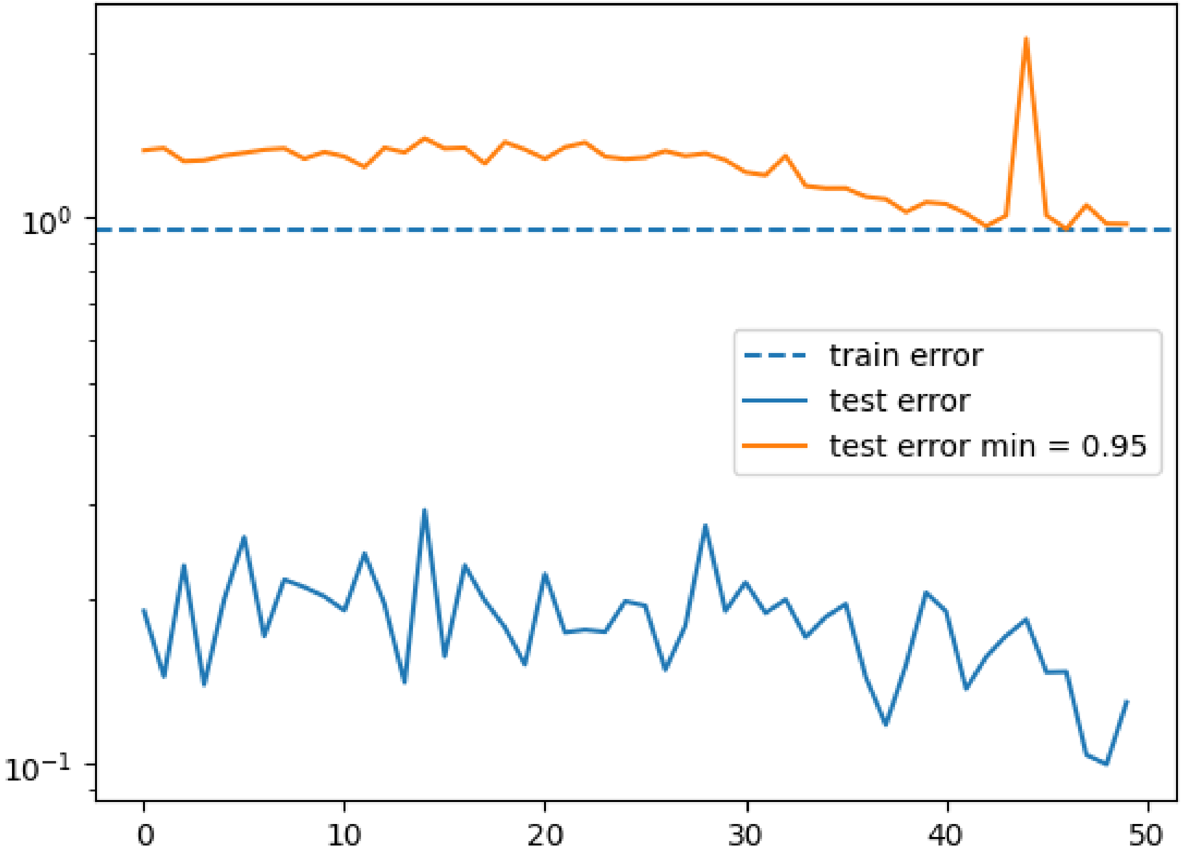
- Training set: 75% of samples(90 audios), Validation 25%(30 audios), make sure both sets have **different** clips from the dataset
- Toy dataset: One class of animal (Rooster), one class of noise (Plane)
- Train across 50 epochs, use adam optimizer with parameters lr=0.0005, betas = (0.7,0.9)
- Use L1 Loss:

Results: loss comparison on training/test sets

Simple model



Full-model (better)



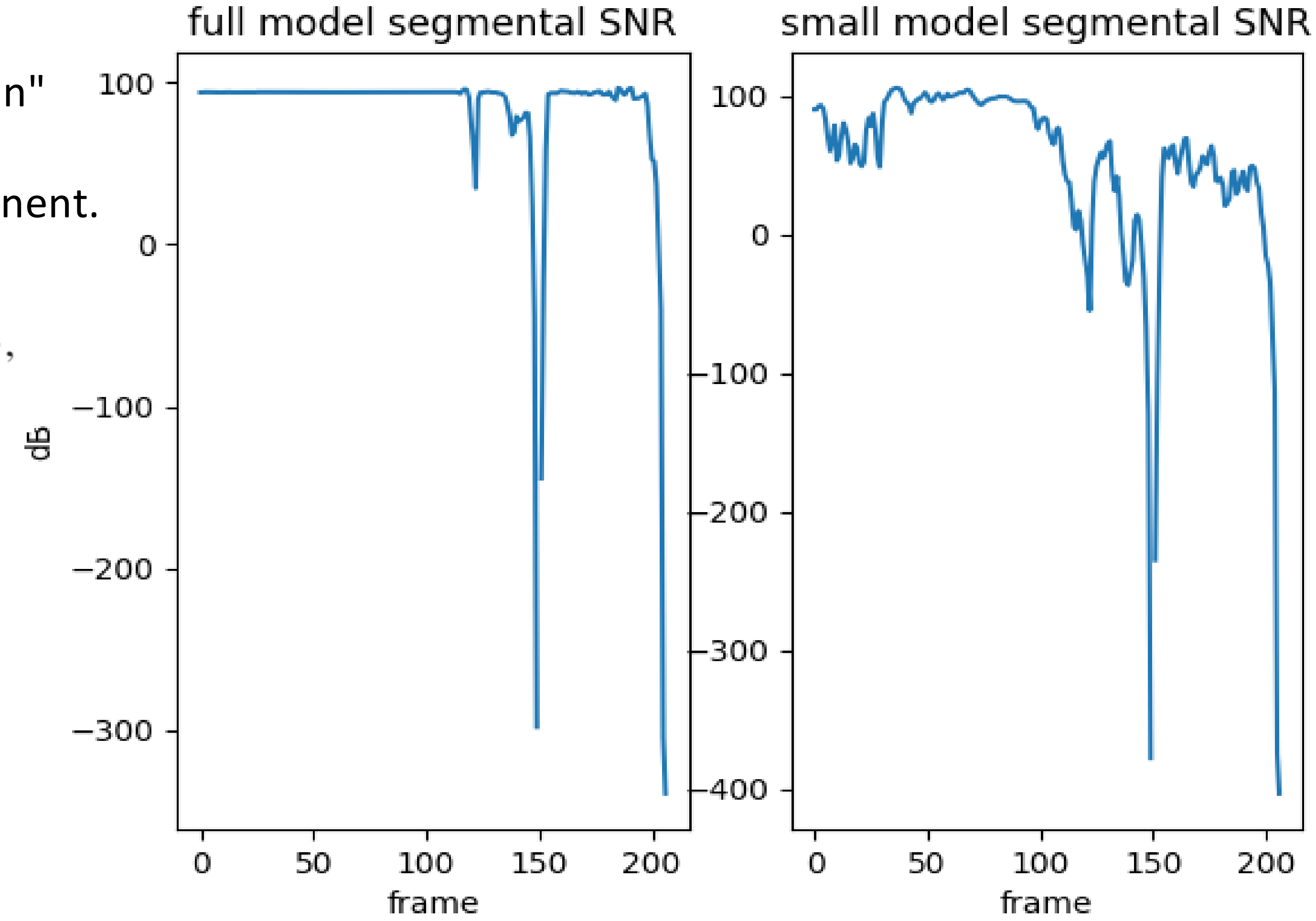
Sound results(subjective)

- **Random Test sample (injected airplane noise):**
- Denoised by simple model:
- Denoised by full model:
- **Sample found on the internet:**
- Denoised by simple model:
- Denoised by full model:

Results- Segmental SNR

Compare denoised test samples to original "clean" sample- treat denoised as an estimation of the clean, subtract the two to evaluate noise component.

$$\mathrm{SNR}(k) = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |E(f,k)|^2} = \frac{\sum_{f=0}^{F} |S(f,k)|^2}{\sum_{f=0}^{F} |S(f,k) - \hat{S}(f,k)|^2},$$



Reference

- Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
- Piczak K J. ESC: Dataset for environmental sound classification[C]//Proceedings of the 23rd ACM international conference on Multimedia. 2015: 1015-1018.
- Moliner E, Välimäki V. A two-stage u-net for high-fidelity denoising of historical recordings[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 841-845.