Hearing Aid Speech Enhancement Using U-Net Convolutional Neural Networks

Paul Kendrick

kenders2000@gmail.com

Music Tribe

Overview

- Motivation
- Approach
- U-what-Net?
- Window processing
- Training
- Hearing aid model
- Results and Conclusions

Motivation for entering competition

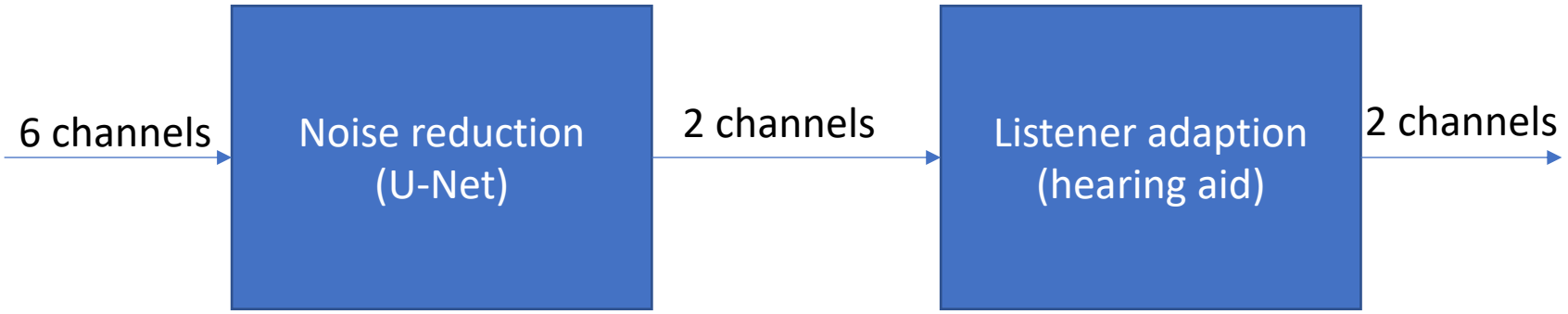
- Learn more about deep learning in audio processing
- Interest in hearing loss
- Build a deep learning workstation

Overview of the solution

Down-sample to 16 kHz

Stage 1) Noise reduction (U-Net)

Stage 2) Listener adaptation (simple hearing aid)



U-Nets

- Ronneberger et al. 2015: *'Convolutional Networks for Biomedical Image Segmentation'*[1]
  - Convolutional Neural Network
  - Semantic segmentation of images

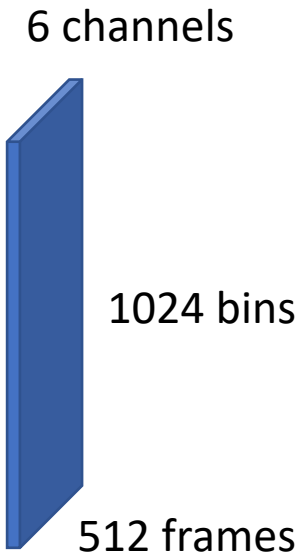
[1] https://arxiv.org/pdf/1505.04597.pdf

U-Nets

- Ronneberger et al. 2015: *'Convolutional Networks for Biomedical Image Segmentation'*[1]
  - Semantic segmentation of images
- Jansson et al. 2017: *'Singing voice separation with deep u-net convolutional networks'* [2]
  - Operates on magnitude Spectrograms (128 x 512, 11s)
  - A mask is predicted from magnitude spectrograms
  - Original mixture phase used in reconstruction

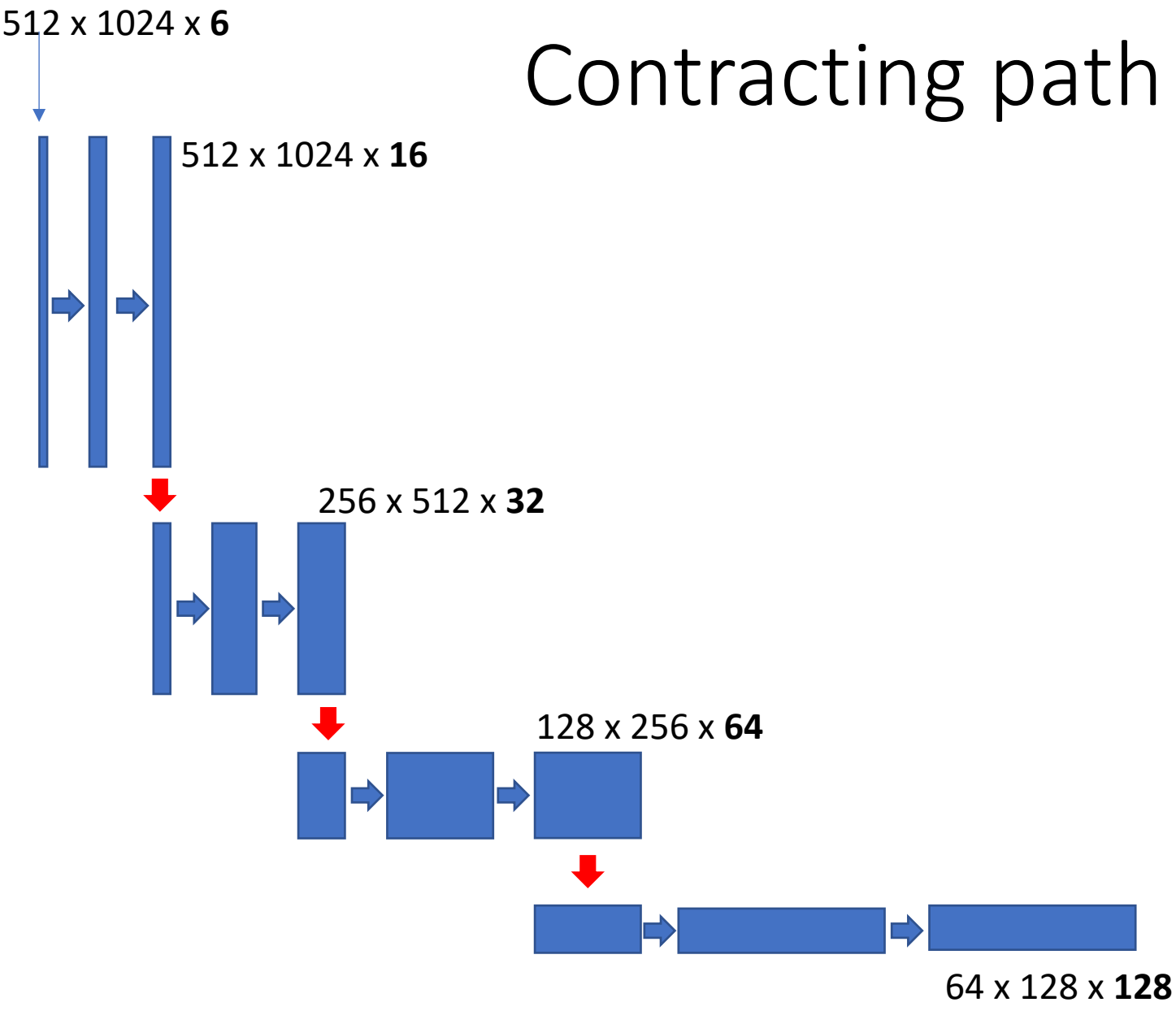
[1] https://arxiv.org/pdf/1505.04597.pdf
[2] https://ejhumphrey.com/assets/pdf/jansson2017singing.pdf

Proposed: U-Net Input Shape



- Input : 6s audio
- Frame/fft size 1024, hop 256
- Spectrogram : 376 x 513 x **6**
- Zero padded to 512 x 1024 x **6**

Contracting path (encoder)



Conv 3x3 ReLU

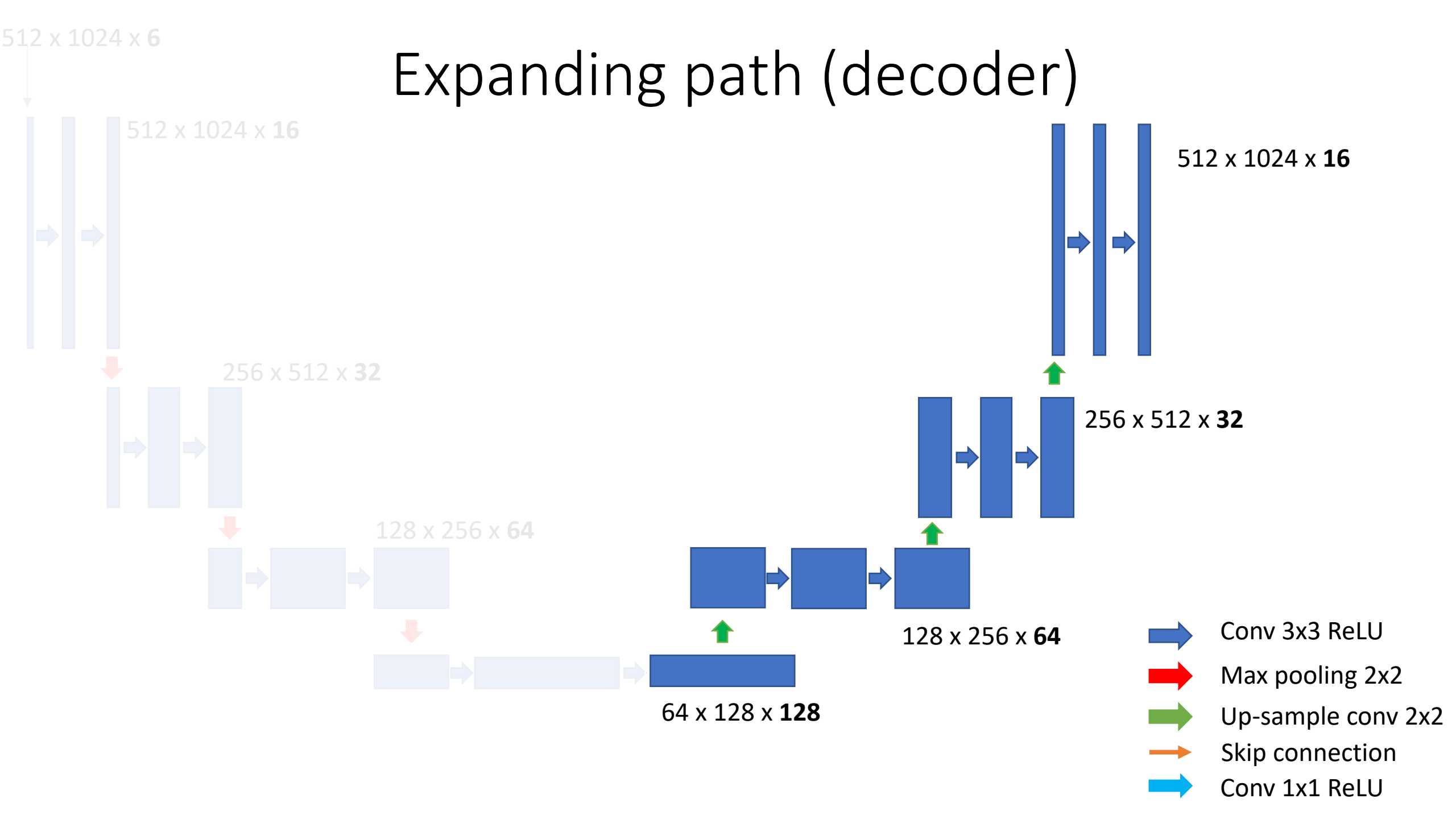
Max pooling 2x2

Up-sample conv 2x2

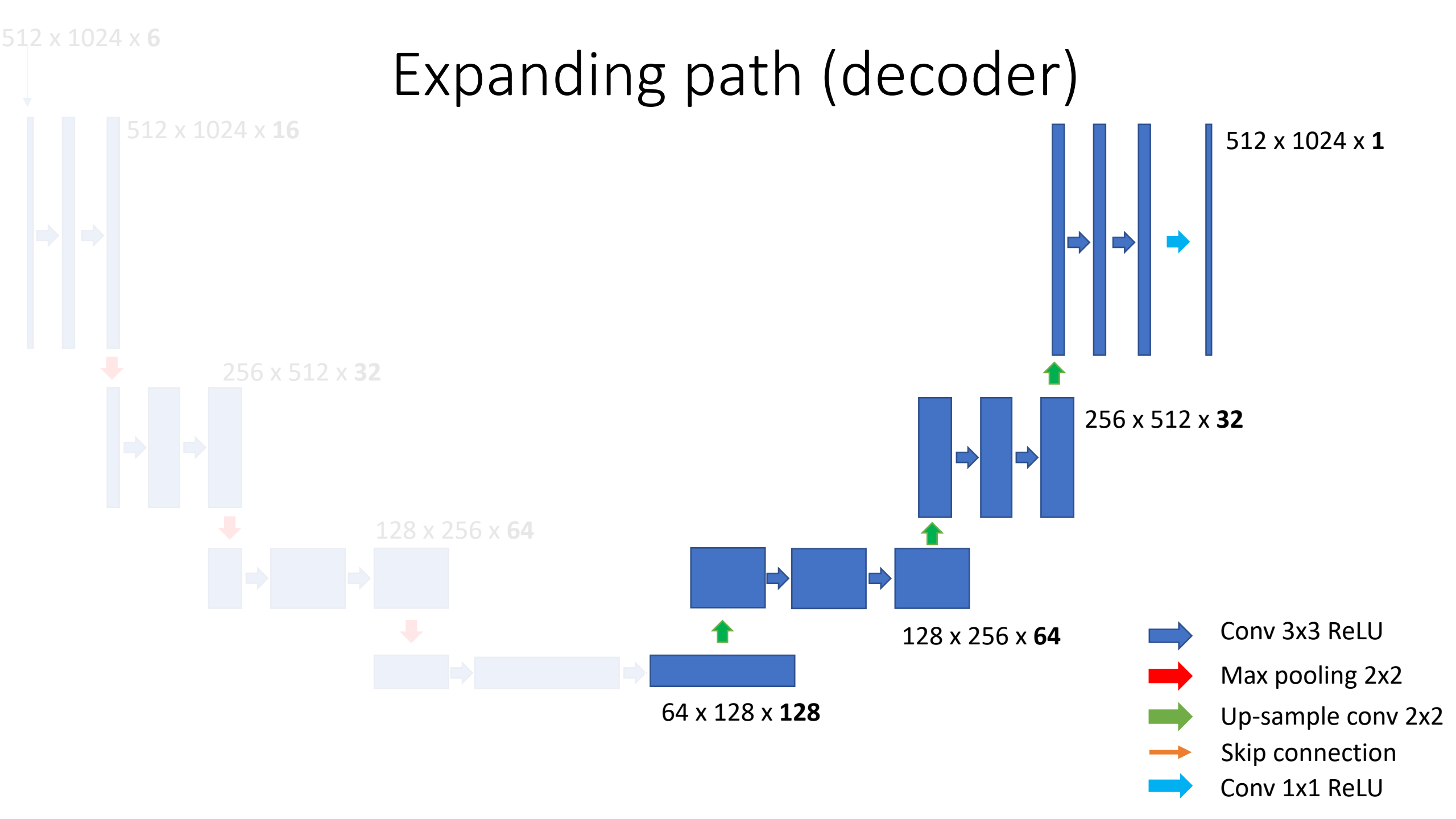
Skip connection

Conv 1x1 ReLU

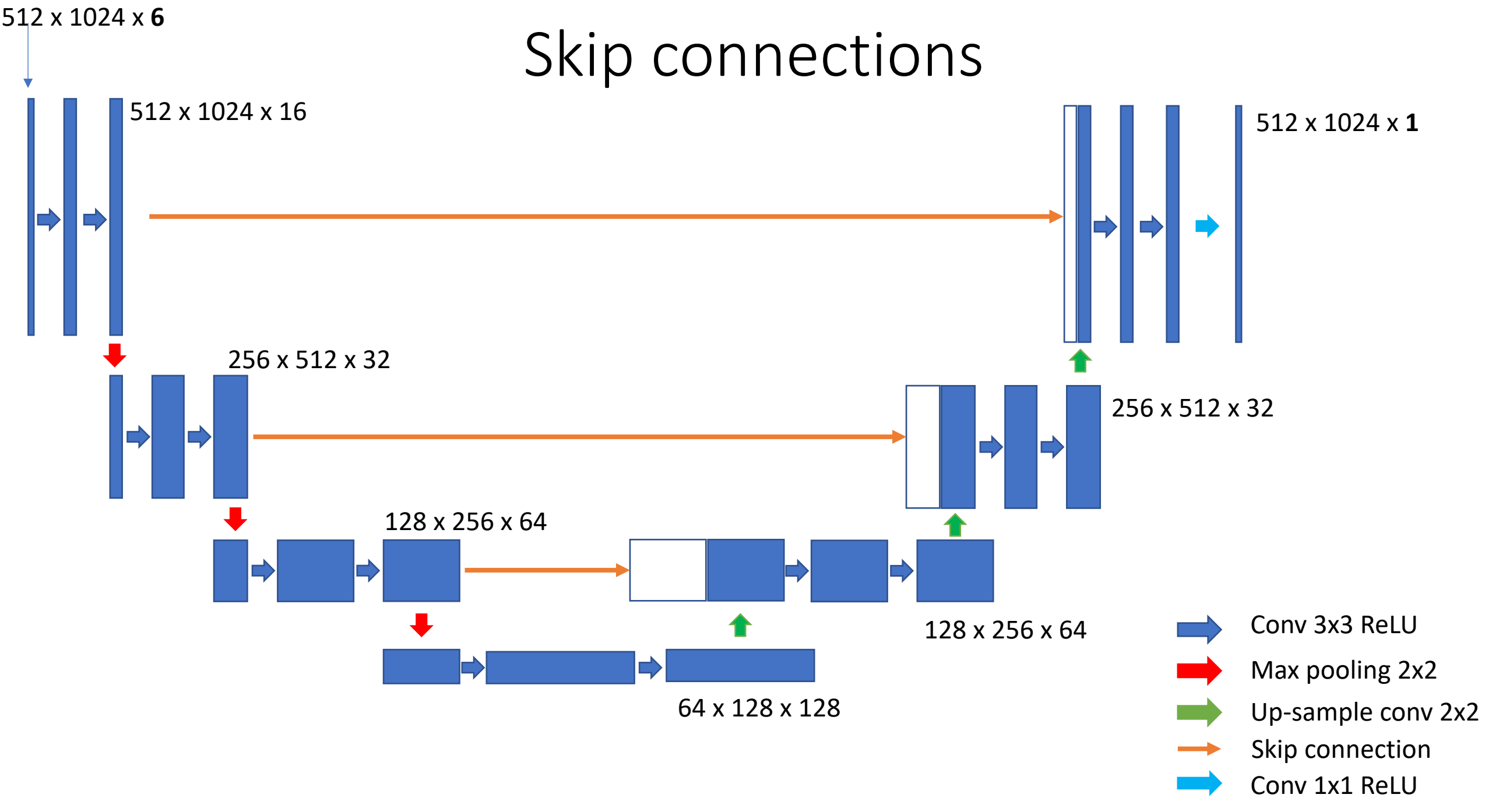
Expanding path (decoder)



Expanding path (decoder)

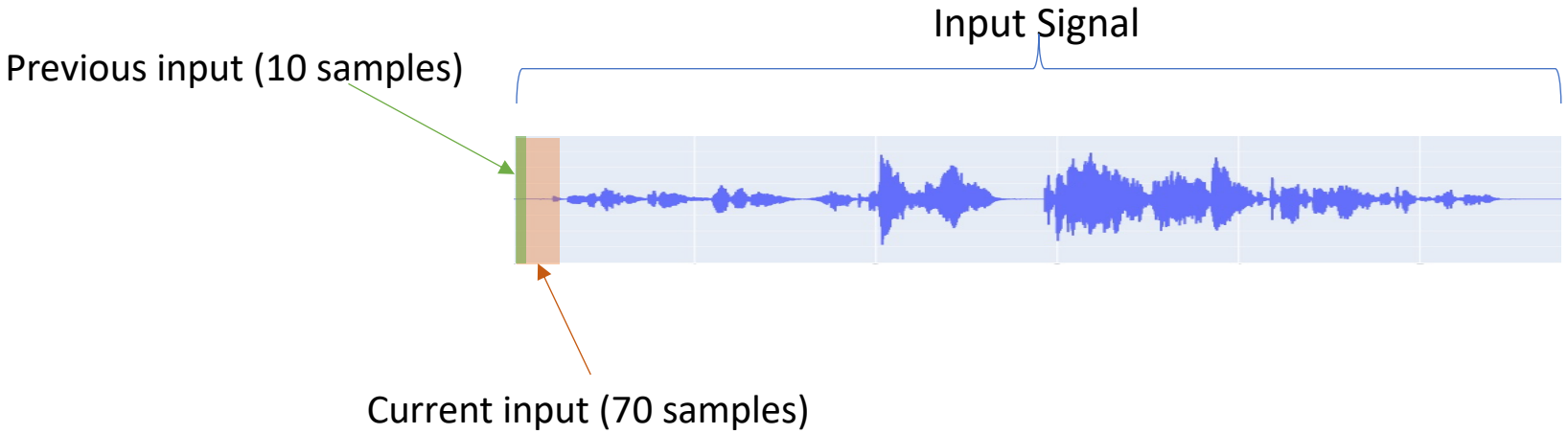


Skip connections

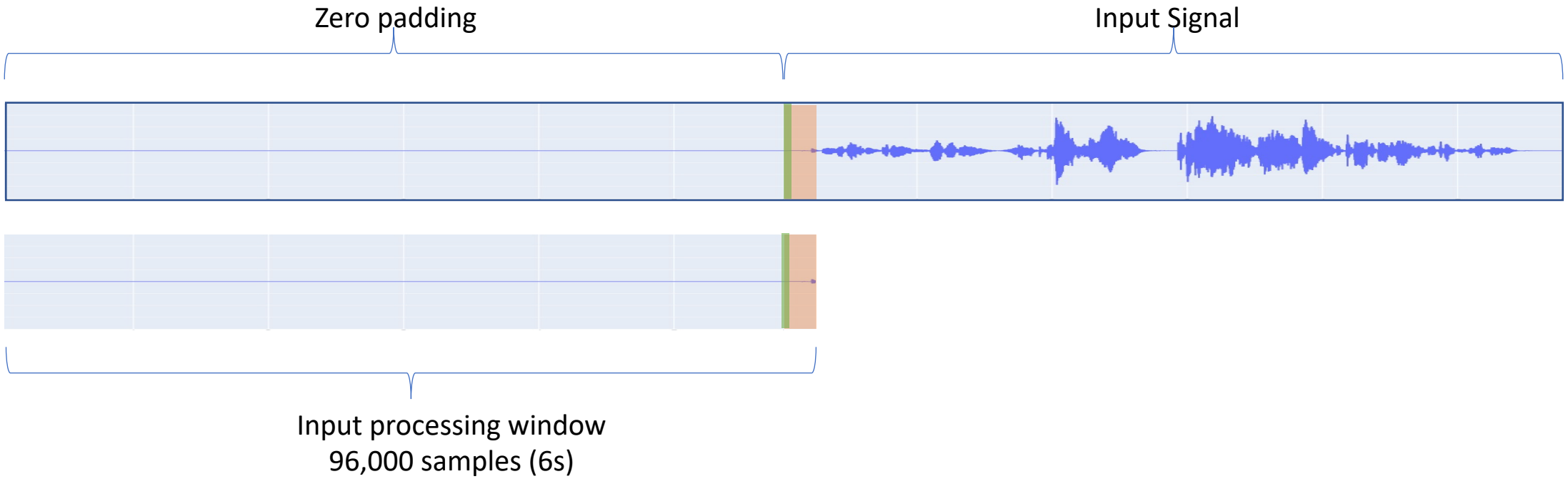


Window processing

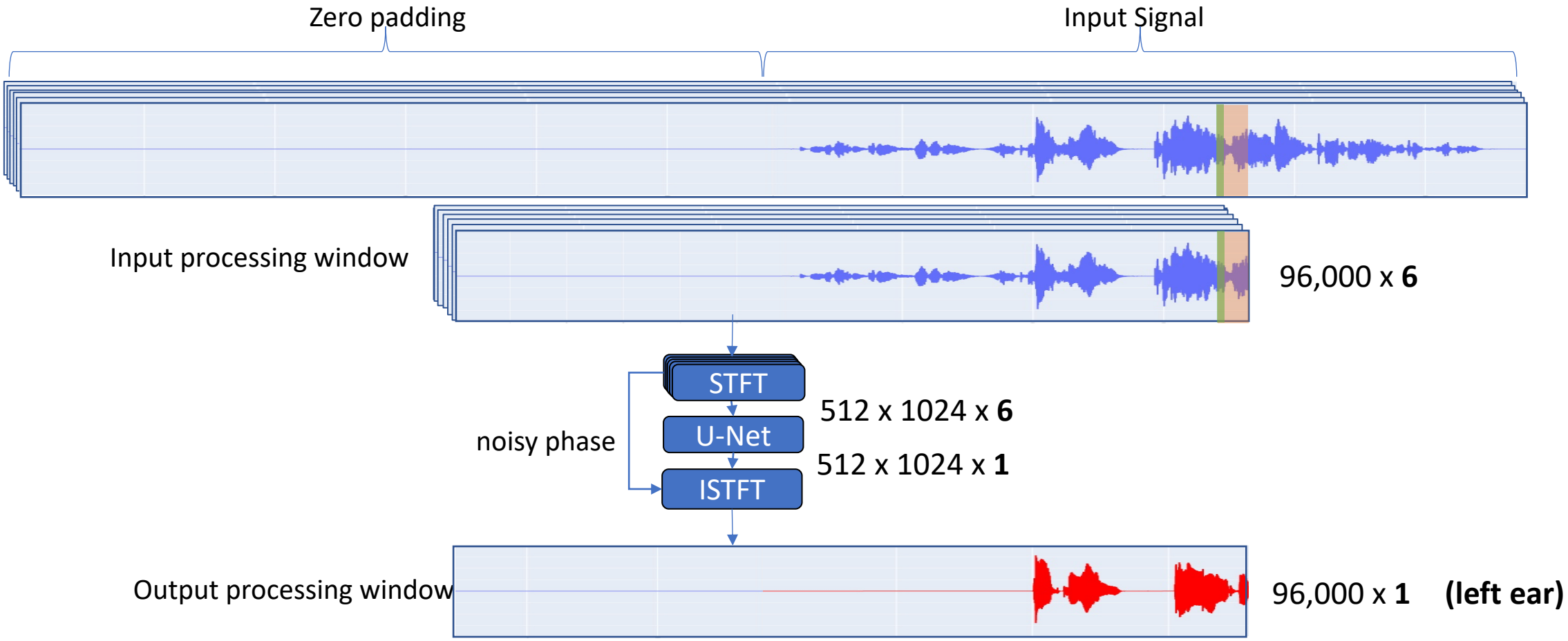
- 80 sample input
- 70 sample hop



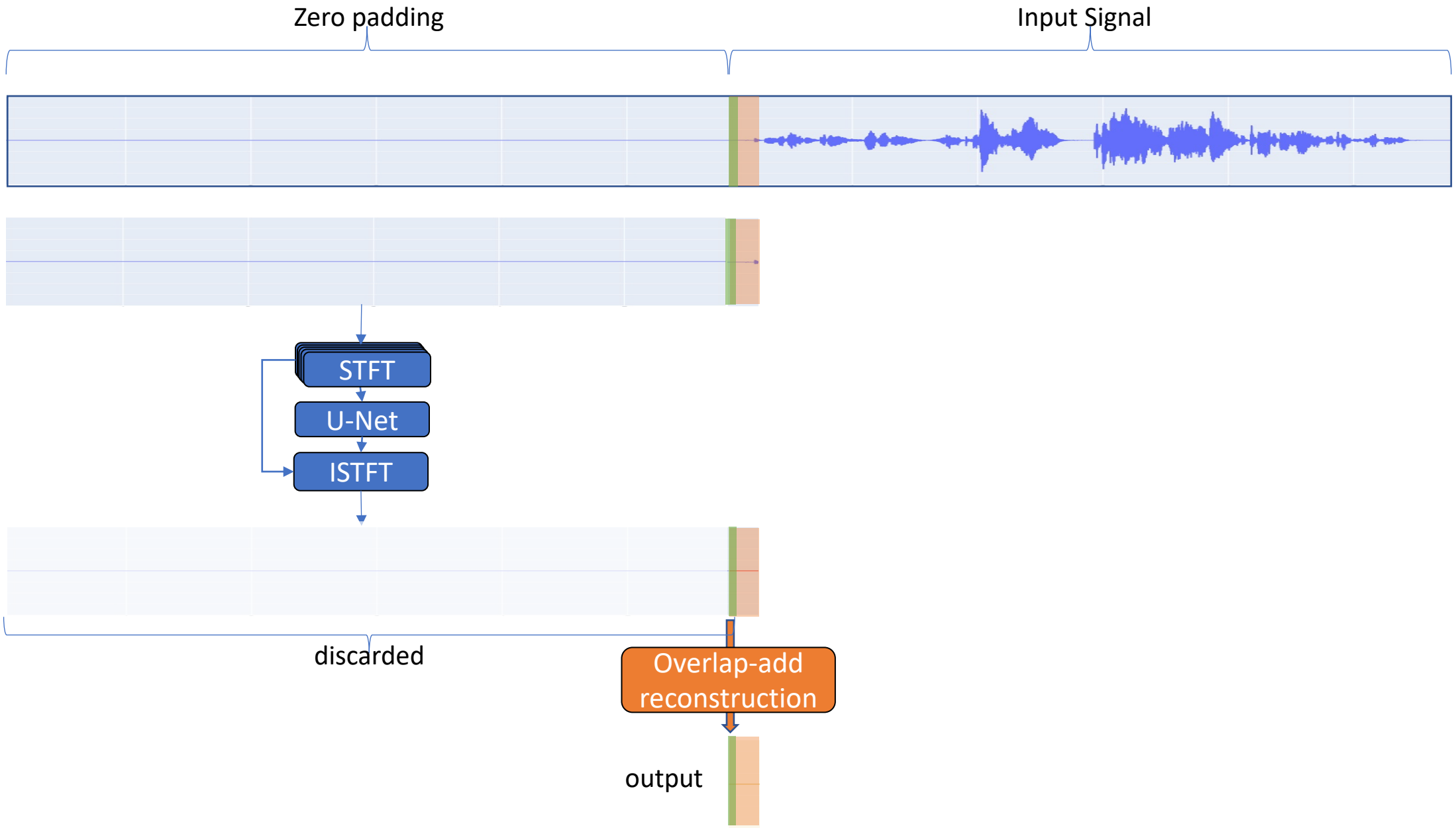
Window processing



Window processing



Window processing



Training the U-Net

- Synchronize clean target with the noisy signal (1st channel, left ear)
- After window processing then under-sampling (10%)
  - Inputs: **1372**x376x513x6 -> **378**x376x513x6
  - Targets: **1372**x376x513x1 -> **378**x376x513x1
- Trained using TF 2.4.1, Adam, LR 0.001, mean absolute error
  - Nvidia GTX 1080ti, 11Gb
  - 10 epochs – 22 Days!

Inference

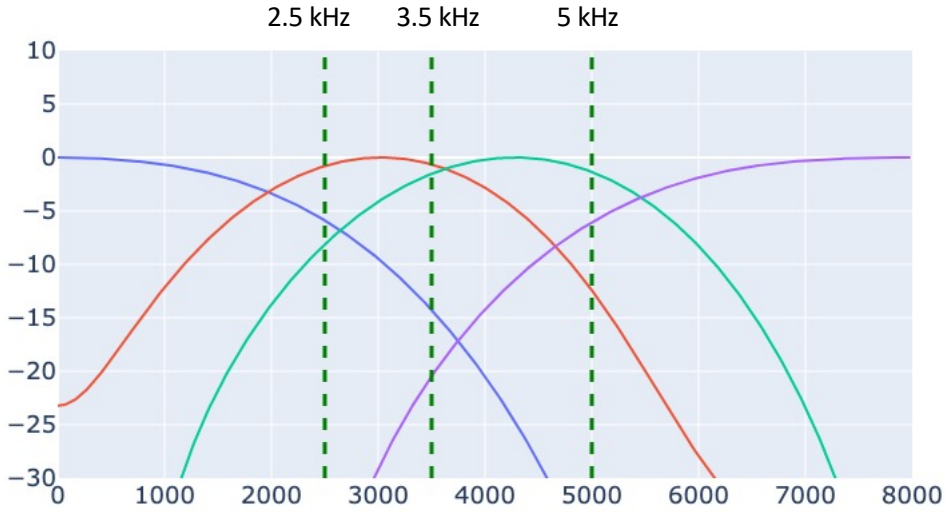
- Left ear:
  - Inference on all 1372 input STFTs
  - ISTFT of all 1372 U-Net output STFTs
  - Overlap add reconstruction
- Right ear
  - Mirror the head by swapping ears channels around in input STFTs

Hearing aid model

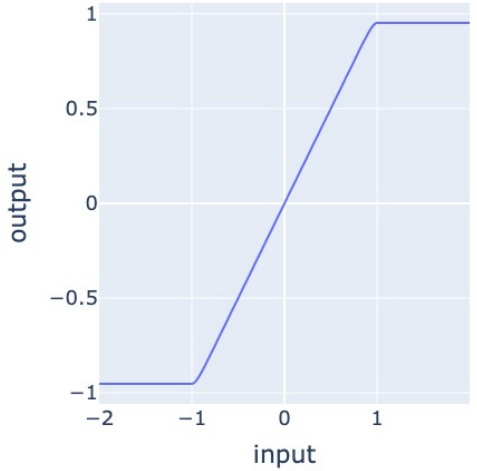
Filter bank
- Zero phase FIR filters
- 11 taps
- 5 sample lookahead

→ **2-channel compressor** → **Soft clipping (d = 21)**

Cut-off frequencies



-6 dB threshold
ratio of 5:1
attack time of 4ms
75ms release time.

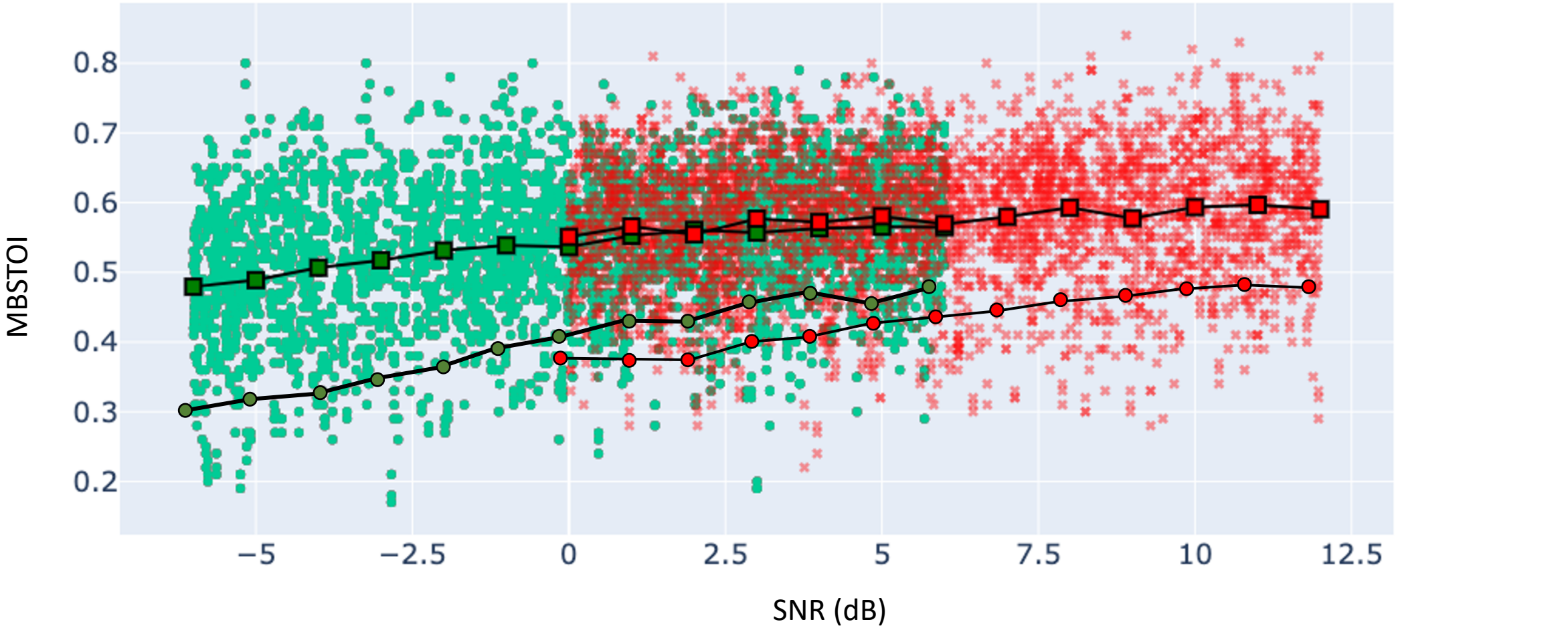


Results

MBSTOI correlation with SNR

Proposed : speech: $\tau = 0.12$, $p < 0.001$; noise: $\tau = 0.27$, $p < 0.001$)

Baseline : speech: $\tau = 0.35$, $p < 0.001$; noise: $\tau = 0.49$, $p < 0.001$)



Overall performance

Evaluated over full dev set

Method	**Mean MBSTOI**
Baseline (dev)	0.41
Proposed (dev)	0.56
Baseline (eval)	0.31
Proposed (eval)	0.66

Evaluated over subset of dev set (first 10 scenes)

Method	**Mean MBSTOI**
Proposed (no hearing aid)	0.54
Proposed (with hearing aid)	0.57

Conclusions

- Improvement in the MBSTOI measure compared with the baseline
- hearing aid model only provided marginal improvement
- https://github.com/kenders2000/u_net_speech_enhancement

Further work

- Include frequency equalization in network
- Large scope for optimisation
- Alternative loss functions
- Used STFT transforms as layers
- Wave-U-Net

Thanks for listening!

An inefficient, expensive, but effective way to heat up your garage!

Choosing gains for the hearing aid

$$G_{ij} = \min(G_{max}, T_{ij} - T_{best})$$

- A low-pass filter with a cut-off of 2500 Hz (average of 250 Hz, 500, 1 kHz and 2 kHz bands)
- A band-pass filter centered at 3 kHz with cut-off frequencies of 2.5 kHz and 3.5 kHz (3kHz band)
- A bandpass filter centered at 4 kHz with cut-off frequencies of 3.5 kHz and 5 kHz (5 kHz band)
- A high-pass filter with a cut-off of 5 kHz (average of 6 & 8 kHz bands)

Examples

Noisy

Cleaned (no hearing aid)