Speech Enhancement using Convolutional-Recurrent NNs and Wavelet Pooling

ECE 251C Project Final Presentation

Team-11: Saqib Azim, Parthasarathi Kumar

Date: Dec 8, 2022

Problem Definition

- To enhance the quality of speech signals for improved intelligibility, clarity of audio, music, recorded videos

- Goal is to recover clean speech from noisy signals

- Applications:
    - In preprocessing noise reduction modules for speech recognition systems

    - To improve audio quality on the receiver side in a noisy communication system

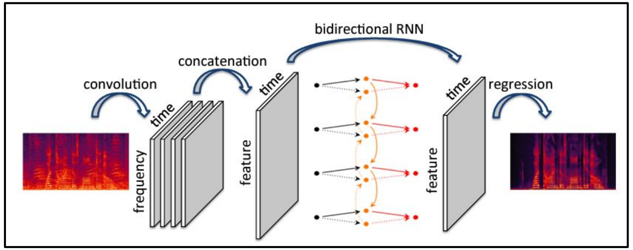
    - Noise reduction in videos/audios recorded on common consumer devices (smartphones, laptops, etc.) in noisy environments

Existing Approaches

Paper 1 [EHNet]:

- EHNet proposes a convolutional-recurrent network based approach to denoise the noisy magnitude spectrogram

- Feedforward noisy spectrogram generated from noisy speech signals to a U-Net based encoder-decoder architecture with a bidirectional LSTM layer in between.



Paper 2 [GCRN]:

- Cleaning magnitude spectrum is not sufficient → Authors propose similar model to denoise the phase spectrum

- Proposed two separate decoders for the real and imaginary parts of magnitude spectrum

Dataset

- **CSR-I (WSJ0) dataset [1]**
  - contains clean speech recording of different speakers.

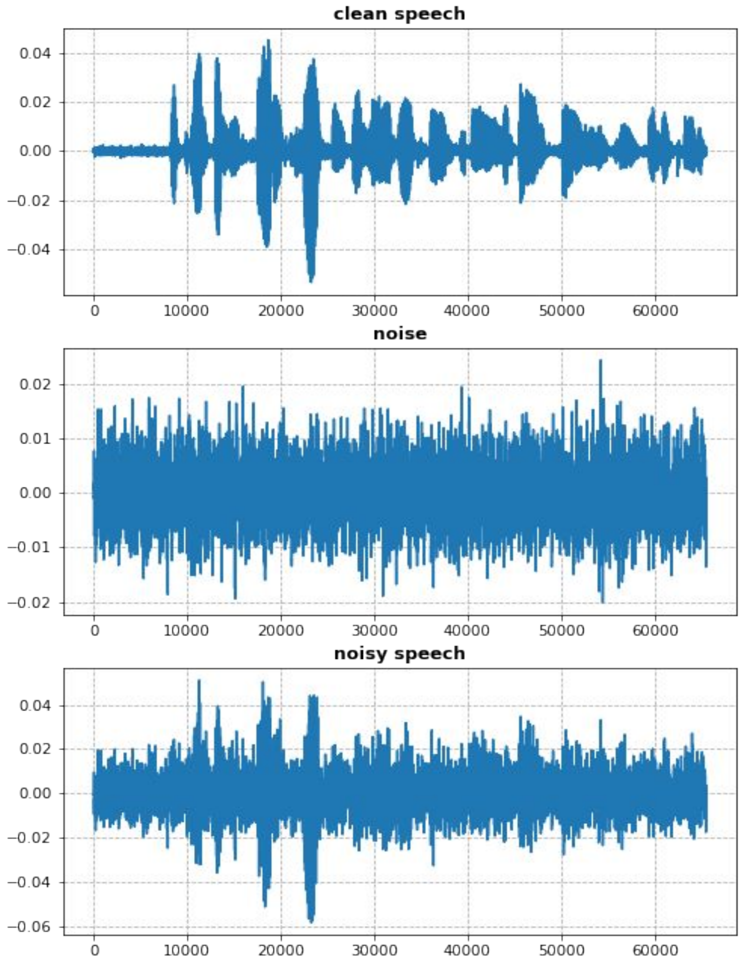
- Noise randomly sampled from a corpus of noise recordings [2] and added to the clean speech

- Sampled noise randomly added to clean speech to create data

$$\mathcal{X}_{noisy}[n] = \mathcal{X}_{clean}[n] + \alpha * \mathcal{X}_{noise}[n]$$

CLEAN SPEECH

NOISY SPEECH



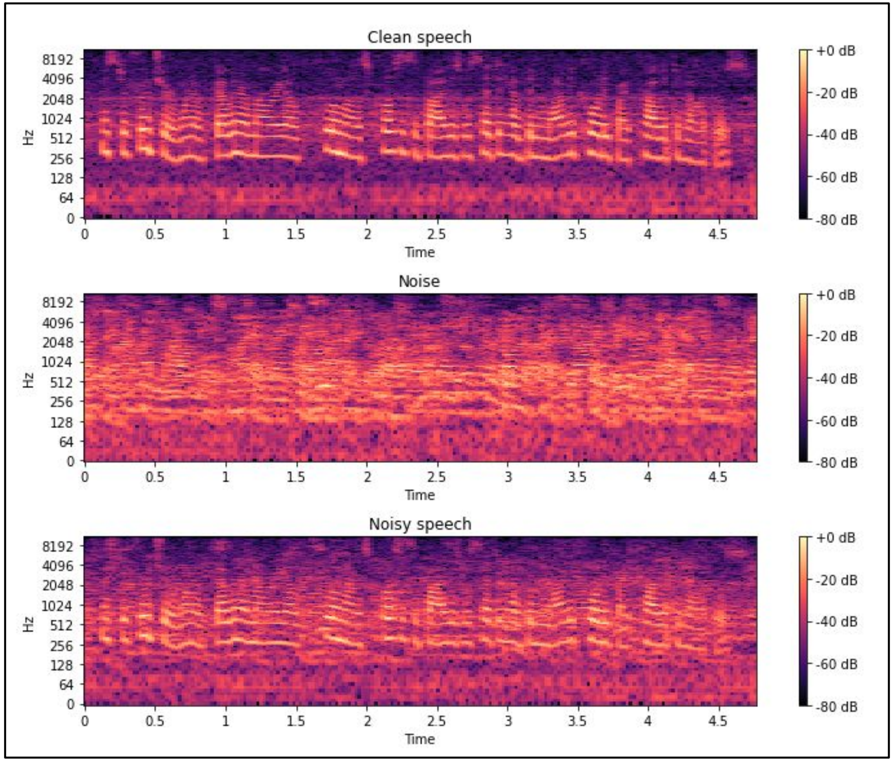
[1] https://catalog.ldc.upenn.edu/LDC93S6A
[2] http://www.ee.ic.ac.uk/naylor/ACEweb/index.html

Dataset Preparation

- Considered speech signals of constant time duration.

- Spectrogram calculated using STFT with varying window size (512, 256, 1024)

- Dataset size -
  - Train : 6000 samples
  - Val : 1000 samples
  - Test : 100 samples

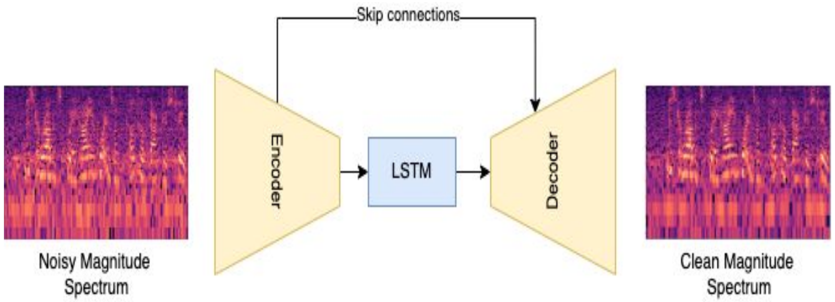


[1] https://catalog.ldc.upenn.edu/LDC93S6A
[2] http://www.ee.ic.ac.uk/naylor/ACEweb/index.html

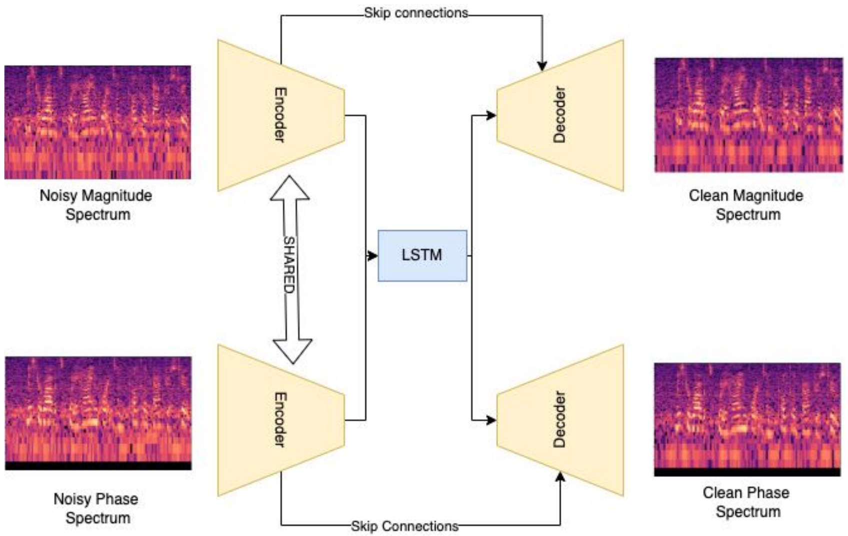
Model Architecture

- U-net like encoder-decoder architecture with a bidirectional RNN in-between to exploit local structures in frequency and temporal domains.

- Bidirectional RNNs model the dynamic correlations between adjacent frames



Model - 1

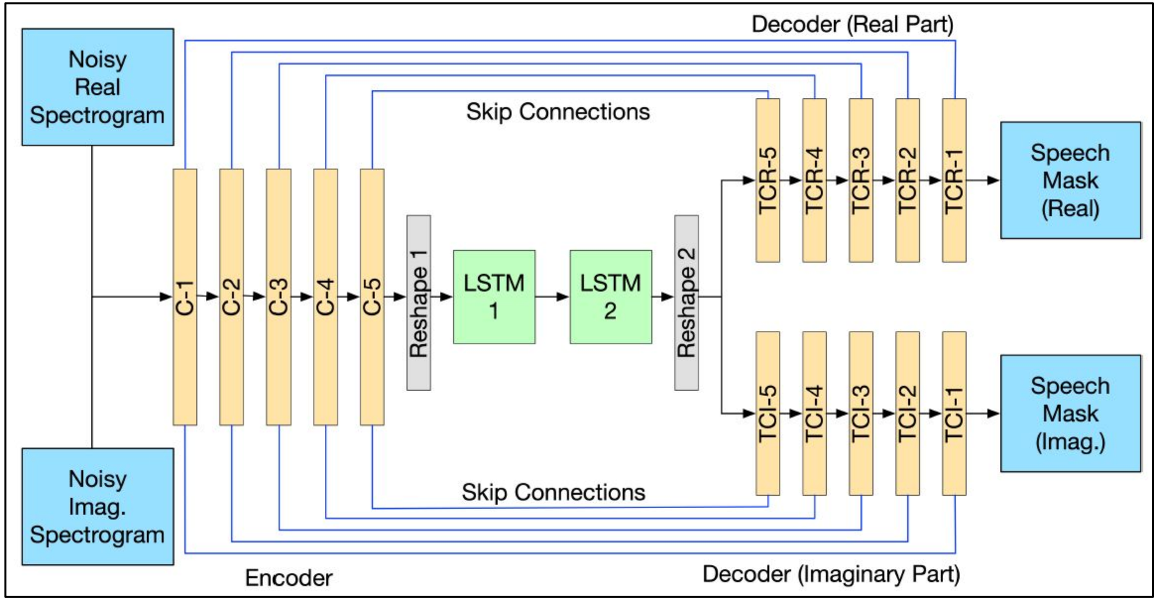


Model - 2

Proposed Method - Model

- Clean up noisy phase by having different decoders for real and imaginary spectrograms

- Weights shared across encoders.



Architecture block diagram for the proposed baseline network

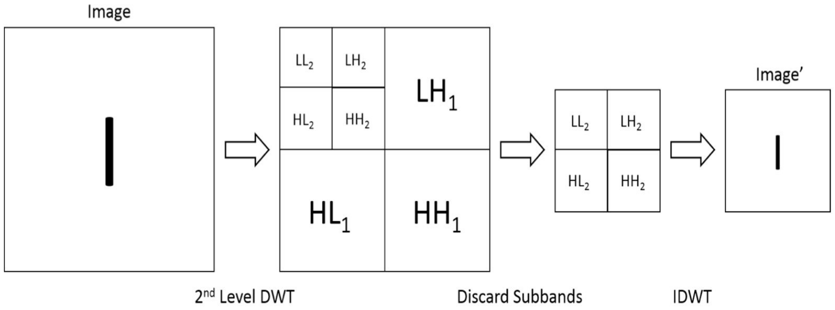
https://www.amazon.science/publications/on-the-robustness-of-deep-learning-based-speech-enhancement

Wavelet Pooling

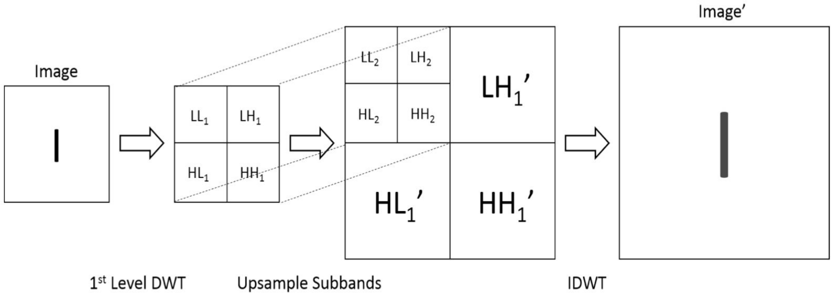
- Alternative to traditional pooling mechanisms that does a better job at compressing the features.
- Uses the 2nd order wavelet subbands to reconstruct the compressed feature
- Process reversed for backpropagation

$$W_\varphi[j+1,k] = h_\varphi[-n] * W_\varphi[j,n]|_{n=2k,k\leq 0}$$

$$W_\psi[j+1,k] = h_\psi[-n] * W_\psi[j,n]|_{n=2k,k\leq 0}$$



Forward Propagation



Back Propagation

https://openreview.net/forum?id=rkhlb8lCZ

Implementation & Training details

- Wavelet pooling - Used torch library for wavelet toolbox - [ptwt]
    - Used backward hook functionality in pytorch modules to realise backpropagation mentioned in paper
    - 4 variants - max-pooling, wavelet pooling using haar, db1 and biorthogonal wavelets

- Considered fixed length input signal (equivalent to spectrogram of size 256 x 64) → helped in batching data

- Training Details -
    - Batch size : 8
    - ~25 epochs
    - SGD optimizer
    - MSE Loss between the predicted and ground-truth spectrogram

$$\boxed{\min_{\theta} \quad \frac{1}{2}\sum_{i=1}^{n} ||g_\theta(\mathbf{x}_i) - \mathbf{y}_i||_F^2}$$

Evaluation metrics

- **SNR** - Signal to Noise Ratio

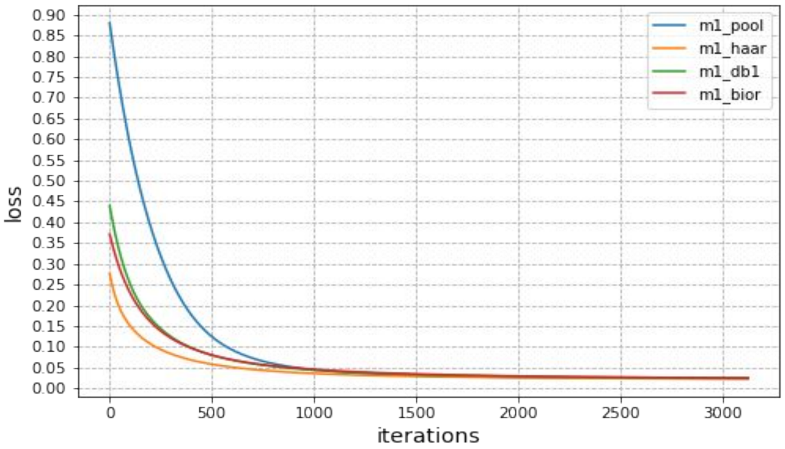
$$\mathrm{SNR}_{dB} = 10\log_{10}\frac{P_{\text{signal}}}{P_{\text{noise}}} = 10\log_{10}\left(\frac{A_{\text{signal}}}{A_{\text{noise}}}\right)^2$$

- **PESQ** - Perceptual Evaluation of Speech Quality
  - Designed to predict subjective opinion scores of a degraded audio sample.
  - PESQ returns a score from 4.5 to -0.5, with higher scores indicating better quality.
- **STOI** - Short time Objective Intelligibility
  - Highly correlated with the intelligibility of noisy speech signals, e.g., due to additive noise

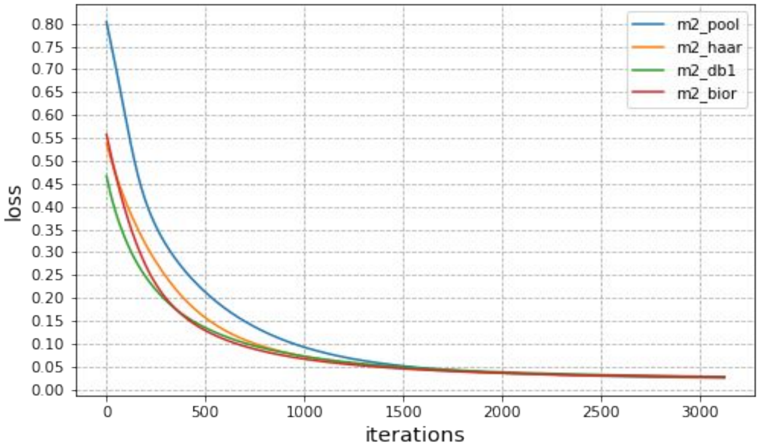
Results - Model 1 vs Model 2

- The loss curves show faster rate of convergence for wavelet pooling.
- However, we could not observe any clear pattern across the different wavelet types.

Model 1 - Magnitude only

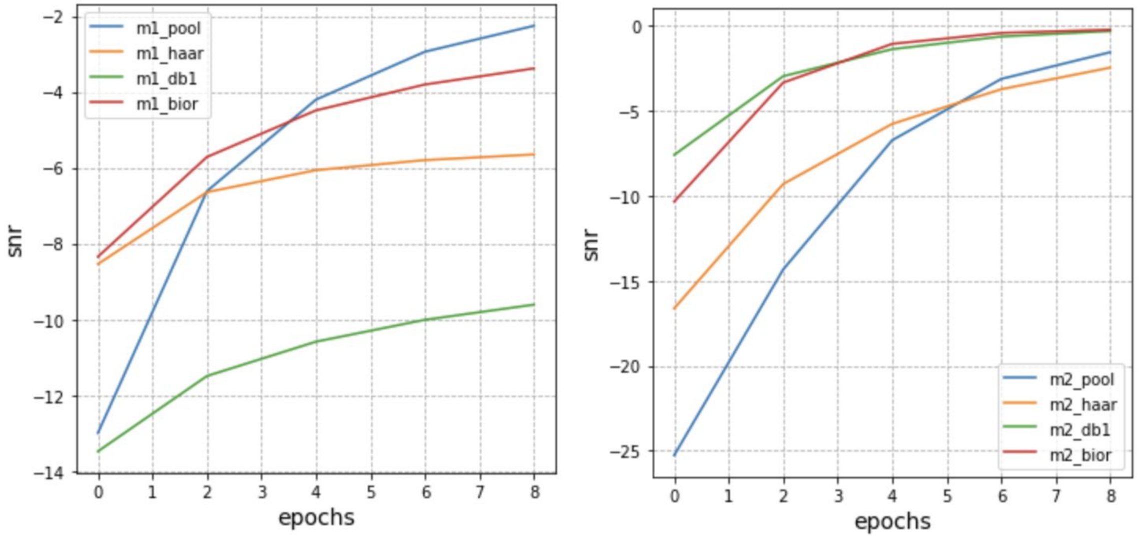


Model 2 - Magnitude and phase



Evaluation Results - SNR

- Cleaning phase and magnitude (model 2) results in better SNRs than model 1.
- Using wavelet pooling performs similar to max pooling and we observe similar

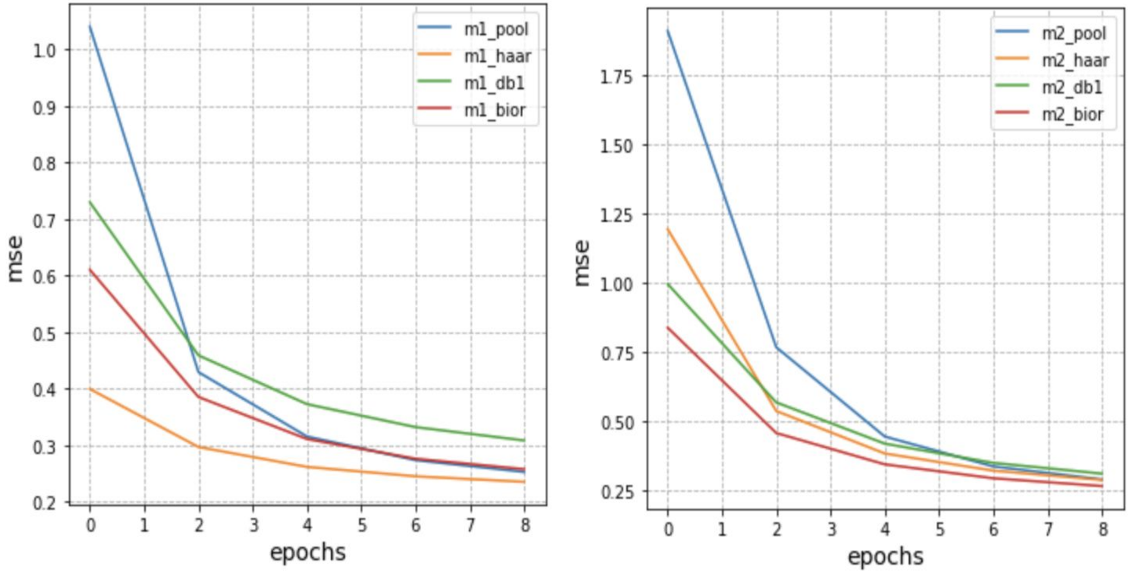


SNR comparison across Model - I & II using different pooling mechanisms

Evaluation Results - MSE

- Not enough structure in spectrogram image data (unlike typical RGB images) for wavelets to significantly outperform max-pooling.

- Just like max-pooling, we observe that wavelet pooling does not have any learnable parameters that can boost performance over max-pooling

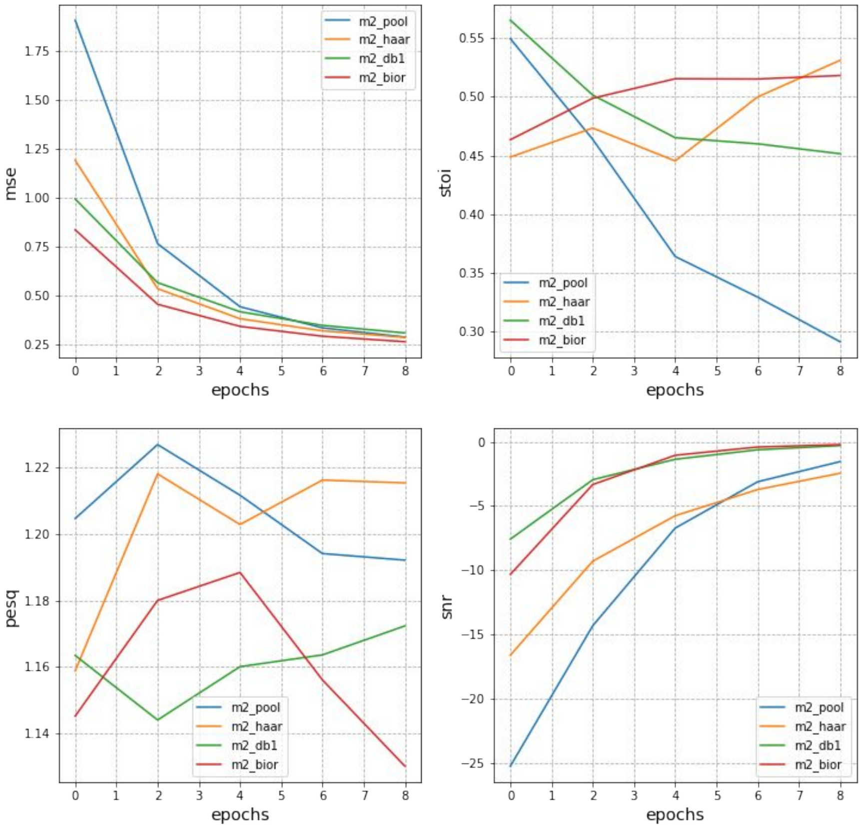


MSE comparison across Model - I & II using different pooling techniques.

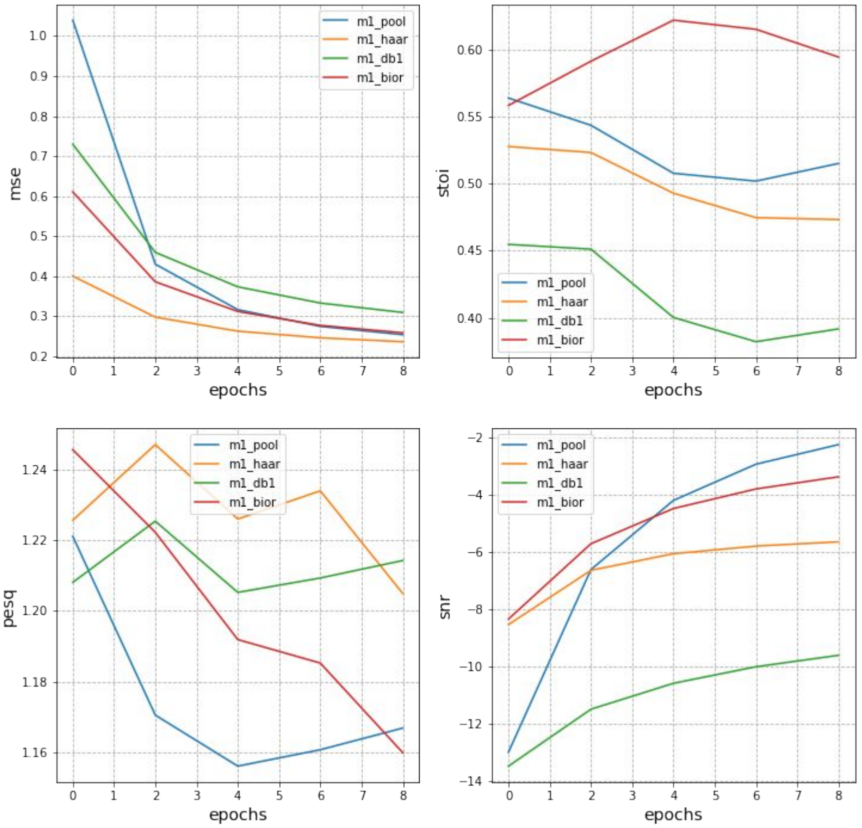
Evaluation Results

Metrics	Max Pooling		Haar Wavelet		Daubechies 1		Biorthogonal	
	M1	M2	M1	M2	M1	M2	M1	M2
SNR	-0.3933	**-0.1119**	-0.8252	-0.7809	-2.0521	-0.1893	-0.830	-0.2712
MSE	**0.0132**	0.0230	0.0145	0.0253	0.0163	0.0227	0.0162	0.0227
STOI	0.4405	0.4066	0.4484	0.4569	**0.5293**	0.5123	0.4452	0.4945
PESQ	1.1482	1.1843	1.242	1.1273	1.193	**1.3258**	1.171	1.1561

Additional Results and Conclusion



Model 2 - Magnitude and Phase Denoising



Model 1 - Magnitude only Denoising

Thank You