Speech Enhancement using Convolution-Recurrent Networks & Wavelet Pooling

Parthasarathi Kumar
ECE Department
UC San Diego
pakumar@ucsd.edu

Saqib Azim
ECE Department
UC San Diego
sazim@ucsd.edu

Abstract

In this report, we present an end-to-end data-driven system for enhancing the quality of speech signals using a convolutional-recurrent neural network. We present a quantitative and qualitative analysis of our speech enhancement system on a real-world noisy speech dataset and evaluate our proposed system's performance using several metrics such as SNR, PESQ, STOI, etc. We have employed wavelet pooling mechanism instead of max-pooling layer in the convolutional layer of our proposed model and compared the performances of these variants. Based on our experiments, we demonstrate that our model's performance on noisy speech signals using haar wavelet is better than when using max-pooling. In addition, wavelet based approach results in faster convergence during training as compared to other variants.

Index Terms: *Convolutional neural networks, Recurrent neural networks, Wavelet pooling, Speech enhancement.*

Code: *Our implementation can be found here on github.*

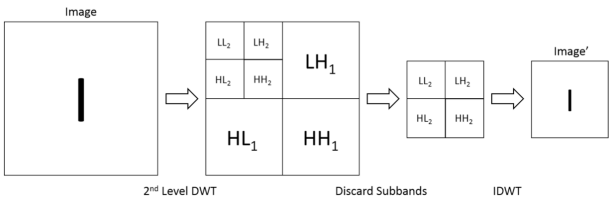
1. Introduction

Speech Enhancement is one of the most important tools in modern speech recognition and communication systems and has numerous applications in these areas. Currently, most machine learning systems rely on large amounts of data to train large-scale deep neural networks. One such system is an automatic speech recognition system. Due to the distribution mismatch between clean data used to train the system and noisy test data encountered during deployment, there is often a degradation in recognition accuracy. Thus, speech enhancement algorithms can act as a preprocessing module helping to reduce the noise in speech signals before it is fed into these systems. The goal of speech enhancement is typically to recover clean speech from noisy, reverberant, and often bandlimited signals in order to yield improved intelligibility, clarity, or automatic speech recognition performance. However, the acoustic goal for a great deal of speech content such as podcasts, demo videos, lecture videos, voice calls, etc. is often not merely clean speech, but speech that is aesthetically pleasing. In addition, a growing amount of speech content nowadays is being recorded on common consumer devices such as tablets, smartphones, iPads, and laptops in common but non-acoustically treated environments such as homes, offices, public cafes, etc. The goal of enhancing such recordings should not only be to make it sound cleaner as would be done using traditional speech enhancement techniques but to make it sound like it was recorded and produced in a professional recording studio.

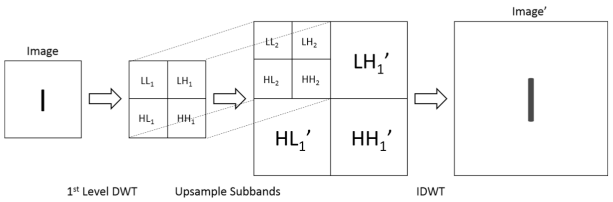
2. Related Work

Speech enhancement has attracted a lot of research efforts over the past decades from the research community. But with modern advancements in science and technology, there are new upcoming challenges every day and the threshold for better quality speech keeps rising. In the past years, data-driven approaches have gained popularity due to availability of large amounts of data as well as computational resources. Past works using multi-layer perceptrons (MLPs) for speech enhancement revolves around applying MLP as a non-linear function approximator to learn the mapping between the noisy speech and its corresponding clean speech [1, 2] using regression based training. However, the fully-connected network structure of MLPs usually cannot exploit the rich spatial and temporal patterns in spectrograms. Work done by [3] introduced the use of RNNs to automatically capture the temporal nature of speech signals, thereby removing the need of explicitly feeding context windows in MLPs.

A popular approach to solve this problem is to treat it as an image-image translation problem where we use the STFT spectrogram of the noisy speech signal to product clean spectrograms. EHNet by Zhao et. al. [4] is one such work that proposes to use a combination of convolution and recurrent neural networks to enhance the quality of speech signals. The idea here is to use CNN to exploit the local



(a) Wavelet pooling - Forward propagation algorithm



(b) Wavelet pooling - Backward propagation algorithm

Figure 1. Wavelet pooling

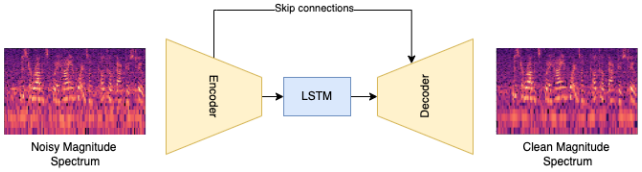
structures in the frequency and time domain, and the RNNs help in modeling the dynamic correlations between adjacent frames. It is purely a data-driven approach and does not make any underlying assumption about the noise in the speech. It also demonstrates substantial performance improvements over traditional statistical-based methods. One drawback of this approach is that the authors in [4] only clean up the magnitude spectrum and use the noisy phase as it is to recover the clean speech signal. Recent work by Chhetri et. al. [5] goes one step ahead and proposes to split the STFT spectrogram into their real and imaginary channels and use separate decoder branches to denoise both channels, and finally combine them together to produce output clean speech signal.

2.1. Wavelet Pooling for CNNs

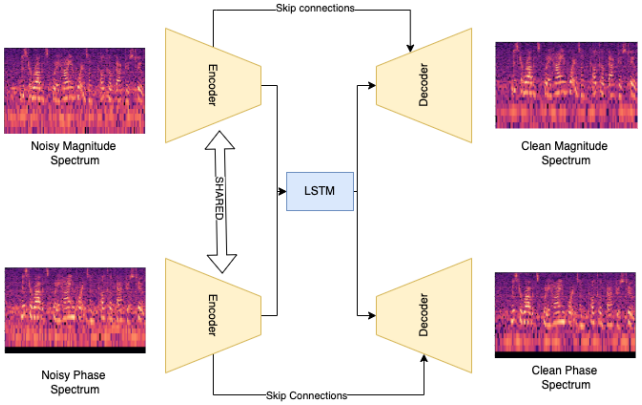
In the past decade, CNNs have become very popular and common in various visual applications or tasks such as image and object classification, object detection, pose estimation, etc. Due to the specific spatial structure in CNNs, they are most apt for image and video data as compared to linear vector-based simple multi-layer perceptron. Pooling layers are very commonly used across most CNN-based architectures such as, max pooling, average pooling, mixed and stochastic pooling, etc. But they often suffer from overfitting which hinders the potential for optimal learning. Wavelet pooling [6], on the other hand, addresses this by proposing to decompose features using second-level wavelet decomposition and discards the first-level subbands to subsample features. This algorithm more accurately represents the feature

3. Problem Formulation

In this work, we assume $X \in \mathbb{R}^{d \times t}$ to be the noisy spectrogram and $Y \in \mathbb{R}^{d \times t}$ to be its corresponding clean



(a) Model 1 architecture overview



(b) Model 2 architecture overview

Figure 2. Architecture overview across model 1(a) and model 2(b)

spectrogram, where d is the number of frequency bins in the spectrogram and t is the time-length of the spectrogram. Given a training set $D = \{(X_i, Y_i)\}_{i=1}^{n}$ of n pairs of noisy and clean spectrograms, the problem of speech enhancement can be formalized as finding a mapping $g_\theta : \mathbb{R}^{d \times t} \rightarrow \mathbb{R}^{d \times t}$ that maps a noisy version to a clean one, where g is parametrized by θ. We use the MSE loss as defined in Eq 1 between the predicted spectrogram and the ground-truth spectrogram of the clean speech as our training objective and try to optimize the model parameters to learn the mapping.

$$\min_{\theta} \sum_{i=1}^{n} \|g_\theta(x_i) - y_i\|_F^2 \tag{1}$$

4. Proposed Approach

We have implemented a U-Net-like convolutional-recurrent network with an encoder-decoder architecture and an intermediate bidirectional recurrent neural network module with input as STFT spectrograms of noisy speech signals as shown in Fig 2. The idea here is similar to that in [4] and [7] where the convolutional layers capture the local structure in the time and frequency domain while the RNN module captures the temporal dependencies. Further, we have also added another decoder after the RNN which is responsible for predicting the clean phase spectrogram.

Based on [6], we have implemented wavelet pooling and compared the performance with traditional max-pooling to see the variation in performance across different wavelets.

Layer	Specifications	Layer	Specifications
Encoder Block	[Convolution layers Pooling Batch Normalization Non-Linear Layer]	Decoder Block	[De-convolution layers Batch Normalization Non-Linear Layer]

Table 1. Specifications of the module in the encoder and decoder blocks

4.1. Network Architecture

The proposed network architecture has a common encoder module that has 5 layers of the block described in Table 1. This block consists of convolution layers that is configured such that the input and the output spatial dimensions remain the same. This is followed by a pooling module which varies across wavelet pooling and max-pooling. Finally, we apply batch normalization and element-wise non-linear ReLU activation on the output downsampled feature map as it helps to alleviate the gradient vanishing problem [8]. The output of the convolutional encoder block is fed to the bidirectional LSTM module that have recurrent connections in forward and backward directions and it helps to model long-term interactions in the speech spectrogram.

The decoder module has 5 layers of the block as shown in Table 1. This block uses transposed convolution layers to up-sample the input feature map to twice the input dimension. The input to this block is the concatenation of the output of the previous block in the decoder along with the corresponding block in the encoder module using skip-connections. The encoder feature module helps in up-sampling the feature maps conditioned on the input data. This is followed by the standard batch normalization and non-linear activation. Further details of the model architecture can be found in Fig 3.

We have performed our experiments using 2 model variants - model 1 and model 2 - whose architecture overview is shown in Fig 2. A detailed architecture specifying the skip connections can be found in Fig 3.

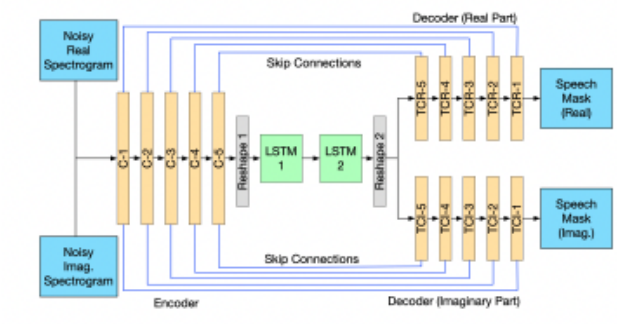


Figure 3. Model architecture for denoising magnitude and phase spectrogram of the noisy speech signal

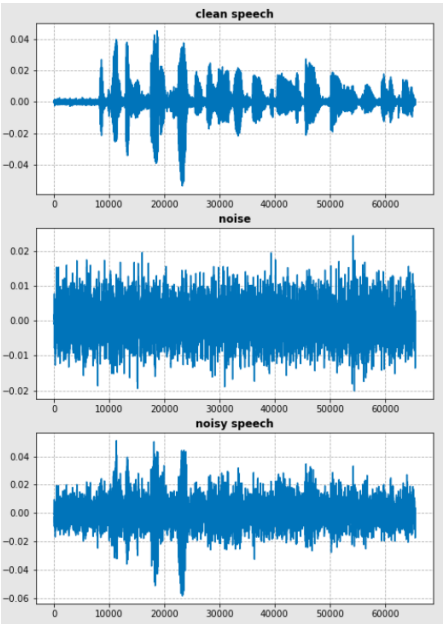


Figure 4. Figure showing a sample time-domain clean speech signal, random noise, and corresponding noisy speech signal obtained using Eq 2.

5. Experiments and Results

5.1. Dataset and Setup

We have used the CSR-I (WSJ0) dataset [9] which consists of $\sim$ 42000 clean speech recordings of different speakers. We synthetically generate noisy speech signals by adding randomly sampled noise from a noise corpus obtained from [10] to the clean speech signals as per Eq 2. A sample instance of a clean signal and its corresponding noisy speech signal is shown in Fig 4.

$$\mathcal{X}_{\text{noisy}}[n] = \mathcal{X}_{\text{clean}}[n] + \alpha * \mathcal{X}_{\text{noise}}[n] \tag{2}$$

As a preprocessing step, we use Short-Time Fourier Transform (STFT) to extract the spectrogram for each signal utterance. While generating data for our experiments, we ensure that the length of each data point (or noisy/clean speech

signal) in the time domain is the same for all samples, so that the resultant STFT spectrogram is of a fixed dimension $(d \times t)$. This has several advantages both during training as well as for overall objective learning. Firstly, it helps in batching the data during training which results in faster convergence and decreases the overall training time. Secondly, it ensures that each data point has equal weightage during training. We trimmed each speech signal upto a length N in time domain and used STFT window size of 1024 (or 256) resulting in a spectrogram size of 1024 (or 256) frequency bins and t = time frames. Finally, we sampled the noise weight factor α in Eq 2 randomly in the range $[0.6, 0.9]$. We create and use a train set consisting of 6000 noisy speech samples, a validation set of 1000 noisy samples, and a test set of 100 noisy samples. To train our network, we fix the number of epochs to be 25, and used SGD algorithm with a learning rate of 0.001.

5.2. Evaluation Metrics

To thoroughly measure the enhancement quality of speech signals, we use the following three metrics to evaluate different models and variants learned for the task of speech enhancement:

- **Signal-to-Noise Ratio (SNR):** SNR is a measure that compares the level of a desired signal to the level of background noise and is defined as the ratio of signal power to the noise power, often expressed in decibels.

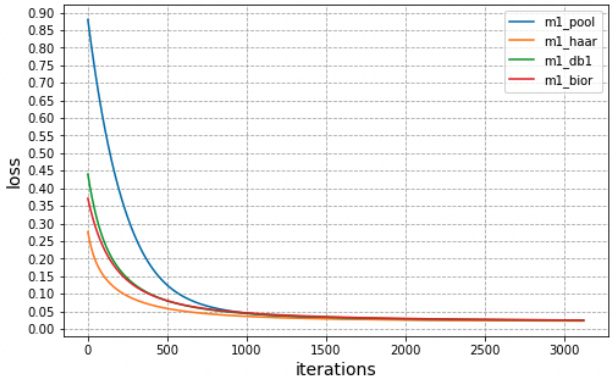
$$\text{SNR}_{dB} = 10 \log_{10} \frac{P_{\text{signal}}}{P_{\text{noise}}} = 10 \log_{10} \left(\frac{A_{\text{signal}}}{A_{\text{noise}}} \right)^2 \quad (3)$$

- **Perceptual Evaluation of Speech Quality (PESQ):** It comprises of a test methodology used by phone manufacturers, telecom operators, etc. for automated assessment of speech quality as experienced by a user. It is designed to predict subjective opinion scores of a degraded audio sample where a larger PESQ score is desirable.

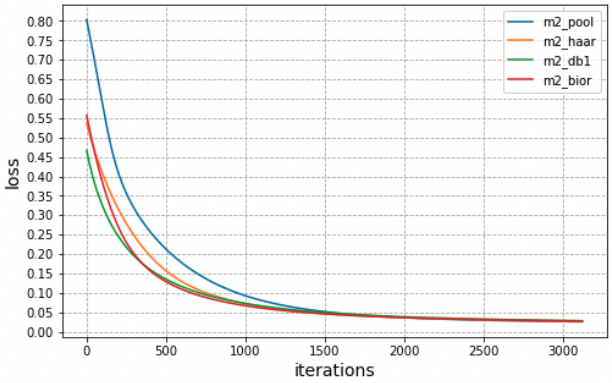
- **Short-time Objective Intelligibility (STOI) [11]:** Intelligibility measure is highly correlated with the intelligibility of degraded speech signals, e.g., due to additive noise, single-/multi-channel noise reduction and is a function of the clean and degraded speech signals. As with PESQ and SNR, we aim for a higher STOI value.

5.3. Results and Analysis

Analyzing the training curves in Fig 5, we observe faster convergence using wavelet pooling across the 2 models. This is consistent with the expected behavior as per [6]. However, there was no clear trend across the different wavelets used for wavelet pooling. In fact, we observed similar trends using wavelet pooling and max-pooling for various experiments conducted using the 2 models such as,



(a) Training loss of Model 1



(b) Training loss of Model 2

Figure 5. Training Loss curves for model 1 and model 2 comparing the variation of loss across different pooling variants including max-pooling, haar wavelet, daubechies I and biorthogonal wavelets.

SNR performance in Fig 6. This might be due to the lack of spatial structure in the spectrogram images, unlike typical RGB natural images which enable wavelet-based pooling to perform better than standard pooling methods such as max-pooling, stochastic pooling. In addition, both max-pooling and wavelet pooling are non-learnable layers without any learnable parameters and hence show similar trends in these experiments.

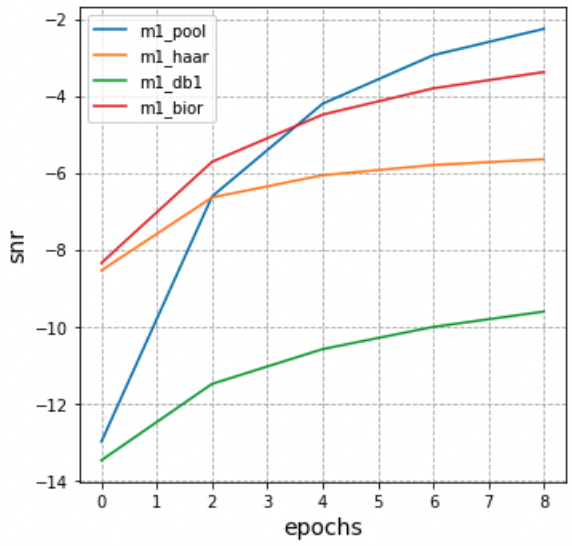
Comparing performances across model 1 and model 2, we see an increase in the SNR for the latter confirming our hypothesis that cleaning up only the magnitude spectrogram is not sufficient. By addressing both the magnitude and phase spectrogram via the proposed approach, model 2 outperforms model 1 by achieving a better SNR on validation data. We plot other metrics mentioned in section 5.2 in Fig 6, Fig 7, and Fig 8.

6. Conclusion

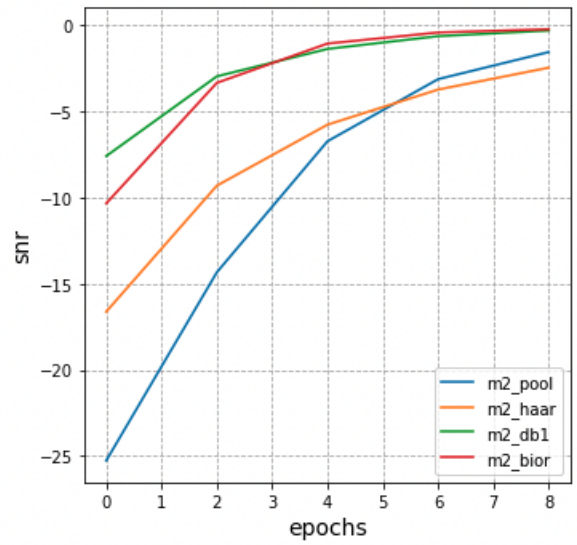
In this project, we present a system that combines both convolutional and recurrent neural networks for speech en-

Metrics	Max Pooling		Haar		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

Table 2. Test set performance across model 1 and 2 using different pooling mechanisms



(a) Validation SNR of model 1



(b) Validation SNR of model 2

Figure 6. SNR curves evaluated on validation data during training stage after every few epochs.

hancement. The convolution layers exploit the local structures in the time and frequency domain, whereas bidirectional RNNs model the dynamic correlations between adjacent frames. Due to sparse nature of convolution layers, our model requires less computation as compared to MLPs. We experimented across 2 model variants for the given task of speech enhancement. We also tried different pooling mechanisms such as max-pooling, stochastic pooling, wavelet pooling with different wavelet types. Finally, we compared our results for different models using SNR, PESQ and STOI evaluation metrics. We observed improvement in SNR values obtained on validation set during training for both the models across different variants.

References

[1] S. Tamura, "An analysis of a noise reduction neural network," in *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 2001–2004 vol.3, 1989. 1

[2] F. Xie and D. Van Compernolle, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. ii, pp. II/53–II/56 vol.2, 1994. 1

[3] A. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *INTERSPEECH*, 2012. 1

[4] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," 2018. 1, 2

[5] A. S. Chhetri, P. Hilmes, M. Athi, and N. Shankar, "On the robustness of deep learning-based speech enhancement," in *IEEE ICMLA 2022*, 2022. 2

[6] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *International Conference on Learning Representations*, 2018. 2, 4

[7] K. Tan and D. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proc. Interspeech 2018*, pp. 3229–3233, 2018. 2

[8] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," 2013. 3

[9] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete." `https://catalog.ldc.upenn.edu/LDC93s6a`. 3

[10] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016. 3

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. R. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, 2010. 4

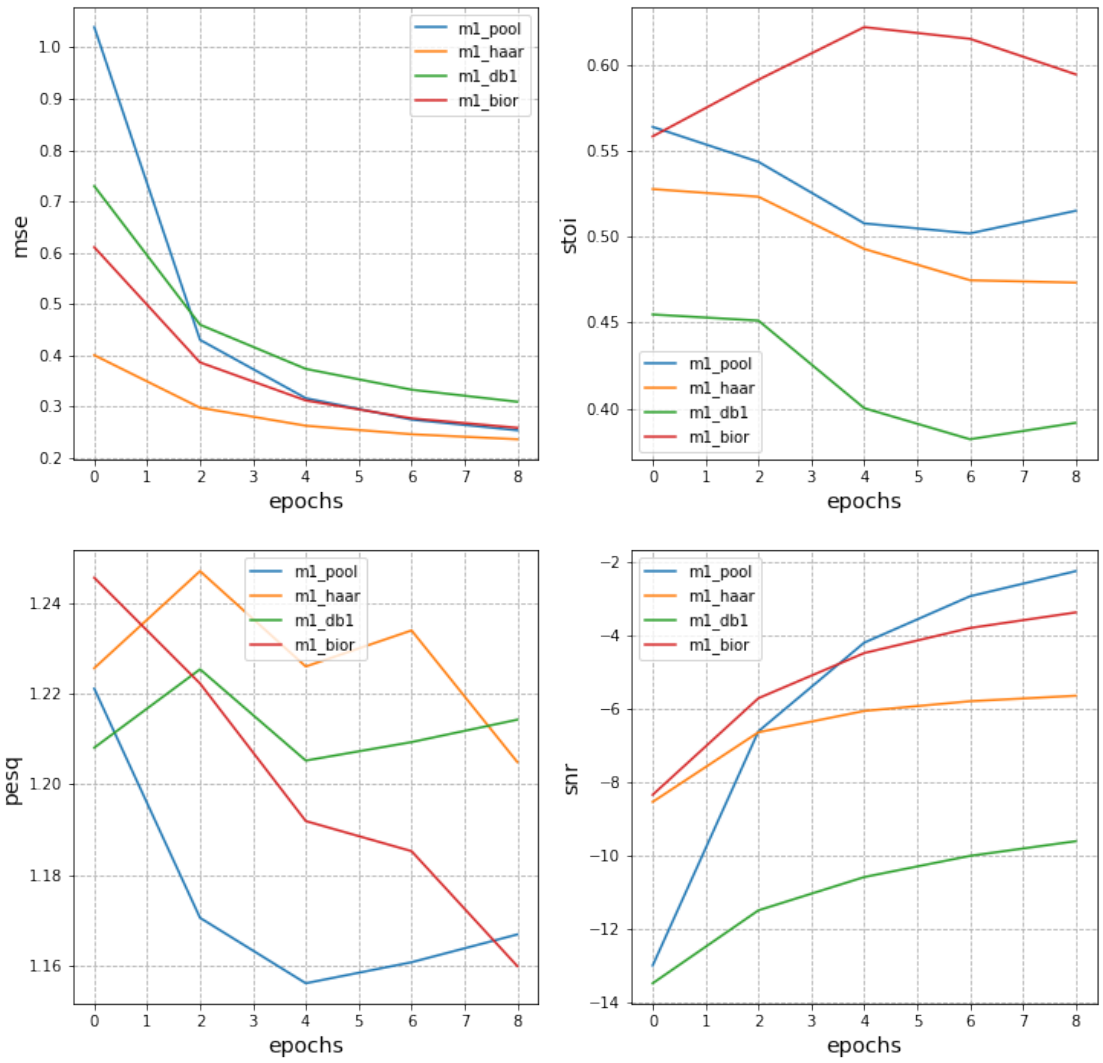


Figure 7. Evaluation curves of model 1 obtained on validation data during training

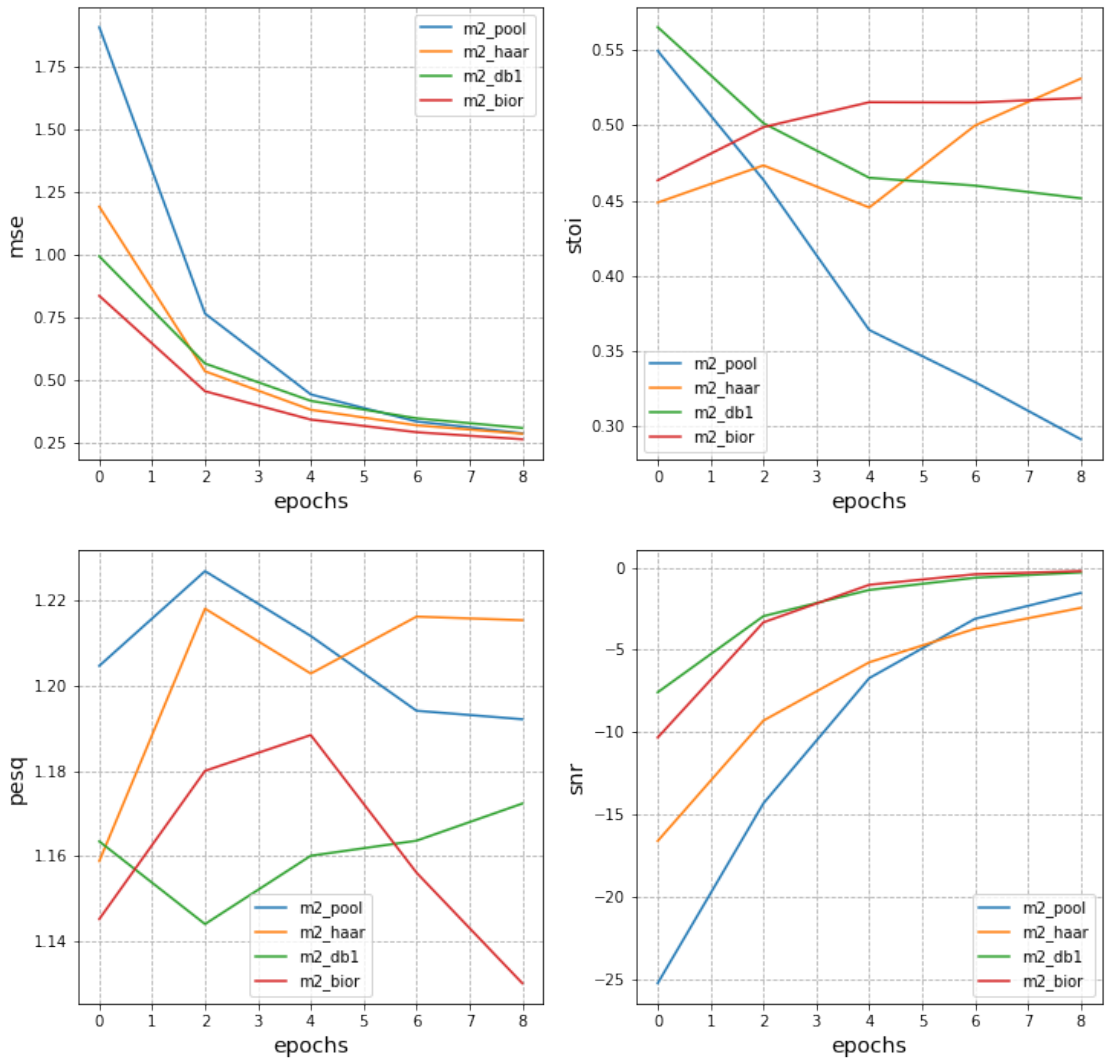


Figure 8. Evaluation curves of model 2 obtained on validation data during training