האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Deep Audio Waveform Prior

Arnon Turetzky, Tzvi Michelson, Yossi Adi, Shmuel Peleg
The Hebrew University of Jerusalem

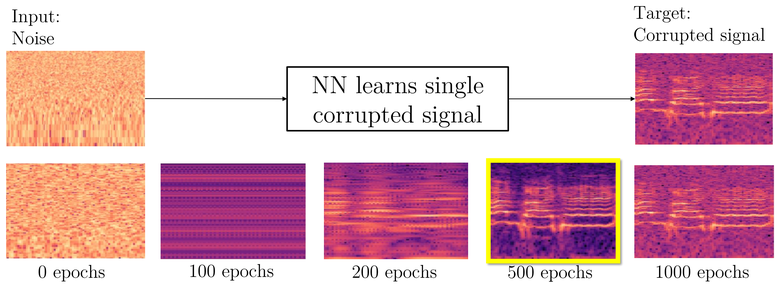
Full Paper

INTERSPEECH 2022
Sep 18 - 22 • Incheon Korea

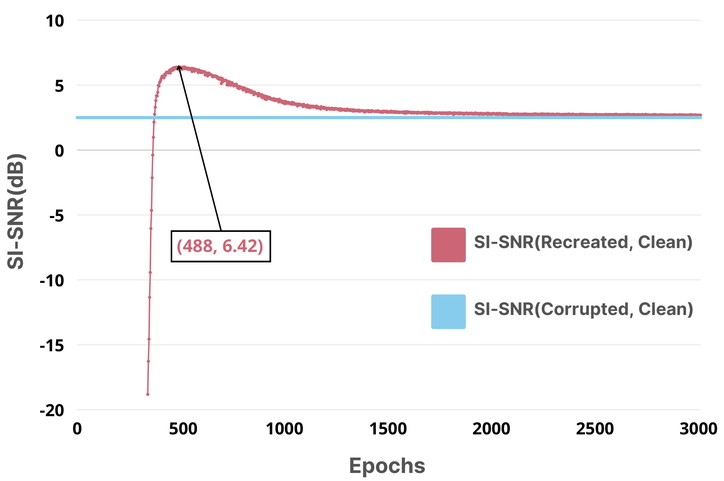
Introduction

- "Deep prior" describes the bias of neural networks (NN) structure to generate "natural" results regardless of training data.
- Existence of deep prior can be demonstrated by training a NN to generate a corrupted signal from pure noise (one sample).
- We demonstrate deep prior for audio waveform using SOTA architecture (Demucs architecture) on several corruptions: reverberations, a gap in the signal (audio inpainting) and background noise.

Deep prior process example for denoising

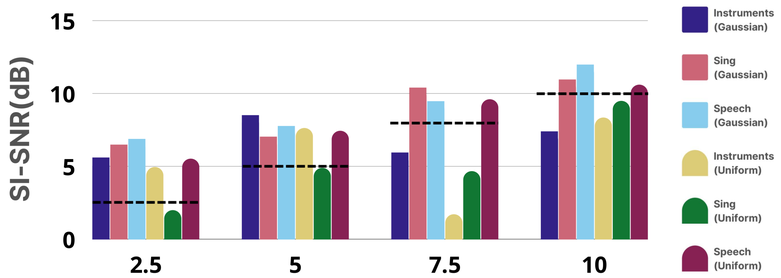


- To discern existence of deep prior we evaluate SI-SNR while training a network to create a single corrupted signal from pure noise.
- In the example the network first recreates a clean version of the signal (after 500 epochs) before recreating the corrupted version of the signal (1000 epochs).
- In the graph, SI-SNR(recreated, clean) above SI-SNR(corrupted, clean) means that the recreated signal becomes more like the clean signal then the corrupted one.



Denoising

- When the bar is higher than the dashed line, the prior denoised the signal.
- When there is more noise (and SI-SNR is lower) the improvement in SI-SNR is greater.



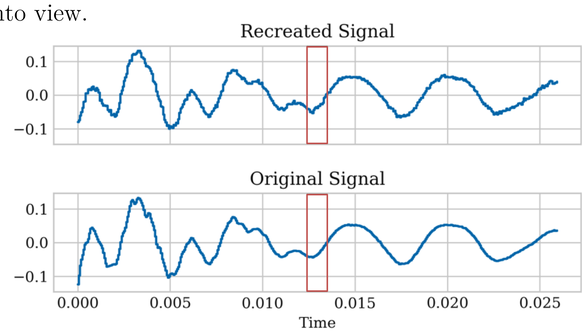
Dereverberation

- Very weak reverberations (Rt60=0.1). The audio sounds very similar to the original. Network improvements are minor.
- Medium strength reverberations (Rt60=0.2). There are more reverberations to clean. The network has a stronger effect.
- High strength reverberations (Rt60=0.5). Reverberations are too strong for the prior. The network has minimal effect.

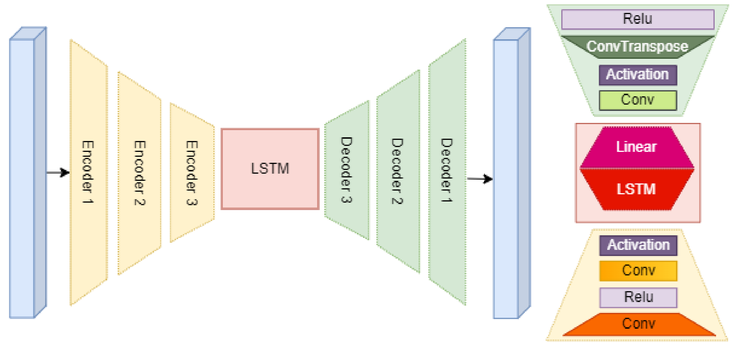
RT60(sec)		SI-SNRi	PSNRi
0.1	Speech	2.22	13.45
	Singing	1.15	14.40
	Instruments	1.15	14.14
0.2	Speech	4.76	16.30
	Singing	5.48	18.45
	Instruments	5.79	18.45
0.5	Speech	-0.11	5.66
	Singing	-0.08	6.65
	Instruments	-0.03	6.45

Audio inpainting

- We masked a 1 to 5ms non-silent segment (red frame) in a 2 second clip. The network recreates this part unsupervised, and its deep priors can come into view.



Ablation



Architecture	SI-SNR	PSNR	PESQ
Conv	3.05	39.12	1.38
Conv + LSTM	5.42	40.56	2.31

- Convolutional layers, aside from containing deep priors themselves, are necessary to prepare the waveform for the LSTM.
- Adding LSTM layers to the convolutional layers improves the network's prior.

Conclusion

- In this work we show for the first time (to the best of our knowledge) the power of deep audio waveform priors in a SOTA architecture.
- We demonstrate the strength of these priors on 3 separate tasks: audio denoising, audio dereverberation, and audio inpainting. We achieve results which would be impossible without deep priors.
- We believe, the findings presented in this study shed light on the recent success of the Demucs architecture on source separation and denoising tasks.