International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Speech Enhancement using Convolutional Autoencoder Network

Subhadeep Sengupta
Dept. Electronics and Communication Engineering
National Institute of Technology Karnataka, Surathkal
Roll Number: 181EC147

Pranav Rihal
Dept. Electronics and Communication Engineering
National Institute of Technology Karnataka, Surathkal
Roll Number: 181EC132

Allwin D'Souza
Dept. Electrical and Electronics Engineering
National Institute of Technology Karnataka, Surathkal
Roll Number: 181EE106

***Abstract*—We present an end-to-end deep learning approach to denoising speech signals by processing the raw waveform directly. Given input audio containing speech corrupted by an additive background signal, the system aims to produce a processed signal that contains only the speech content. Recent approaches have shown promising results using various deep network architectures. In this paper, we propose to train a fully-convolutional context aggregation network using a deep feature loss. That loss is based on comparing the internal feature activations in a different network, trained for acoustic environment detection and domestic audio tagging. Our approach outperforms the stateof-the-art in objective speech quality metrics and in large-scale perceptual experiments with human listeners. It also outperforms an identical network trained using traditional regression losses. The advantage of the new approach is particularly pronounced for the hardest data with the most intrusive background noise, for which denoising is most needed and most challenging.**

***Index Terms*—speech enhancement,Fully convolutional denoising autoencoders, single channel audio source separation, stacked convolutional autoencoders, deep convolutional neural networks, deep learning.**

I. Introduction

The project aims at building a speech enhancement system to attenuate environmental noise. Audios have many different ways to be represented, going from raw time series to time-frequency decompositions. The choice of the representation is crucial for the performance of your system. Among time-frequency decompositions, Spectrograms have been proved to be a useful representation for audio processing. They consist in 2D images representing sequences of Short Time Fourier Transform (STFT) with time and frequency as axes, and brightness representing the strength of a frequency component at each time frame. In such they appear a natural domain to apply the CNNS architectures for images directly to sound. Between magnitude and phase spectrograms, magnitude spectrograms contain most the structure of the signal. Phase spectrograms appear to show only little temporal and spectral regularities.

In this project, we will use magnitude spectrograms as a representation of sound in order to predict the noise model to be subtracted to a noisy voice spectrogram.

II. Related Work

Data-driven approaches using regression-based deep neural networks have attracted much interests and demonstrated substantial performance improvements over traditional statistical-based methods for example: Multilayer Perceptrons. Before the popularization of deep networks, denoising systems relied on spectrogram-domain statistical signal processing methods , followed more recently by spectrogram factorization-based methods. Most pipelines still operate in the spectrogram domain. As such, signal artifacts then arise due to time aliasing when using the inverse short-time Fourier transform to produce the time-domain enhanced signal. This particular issue can be somewhat alleviated, but with increased computational cost and system complexity. To capture the temporal nature of speech signals, previous works introduced recurrent neural networks. Recently, there has been growing interest in the design of performant denoising pipelines that are optimized end-to-end and directly operate on the raw waveform. Such approaches aim at fully leveraging the expressive power of deep networks while avoiding expensive time-frequency transformations or loss of phase information. Some of these approaches typically use simple regression loss functions for training the network (e.g., L1 loss on the raw waveform), while ones with more advanced loss functions have shown limited gains in mismatched conditions. Loss function is shown to have also been optimized using Generative Adversarial Networks(GANs).

III. Dataset and Tools

To create the datasets for training, We gathered English speech clean voices and environmental noises from different sources. The clean voices were mainly gathered from the TIMIT Corpus dataset from Kaggle. It consists of 16 speakers from 8 dialect regions, one male and one female from each dialect region.

For this project, We focused on 10 classes of environmental noise: tic clock, foot steps, bells, handsaw, alarm, fireworks, insects, brushing teeth, vacuum cleaner and snoring.

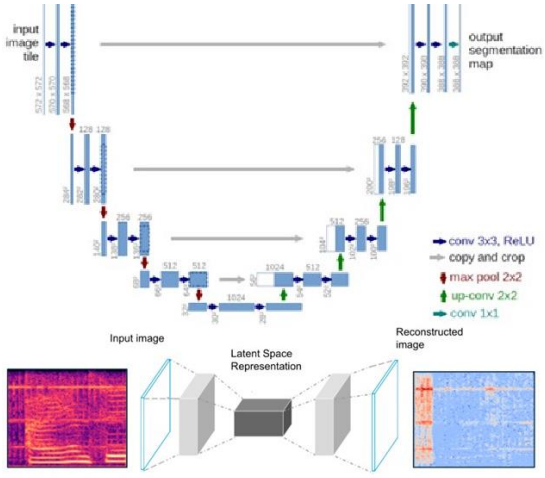

Fig. 1: Model Architecture(U-Net Structure)

IV. Methodology

The model used for the training is a U-Net, a Deep Convolutional Autoencoder with symmetric skip connections. U-Net was initially developed for Bio Medical Image Segmentation. Here the U-Net has been adapted to denoise spectrograms. The architecture looks like a 'U' which justifies its name. This architecture consists of three sections: The contraction, The bottleneck, and the expansion section. The contraction section is made of many contraction blocks. Each block takes an input applies two 3×3 convolution layers followed by a 2×2 max pooling. The number of kernels or feature maps after each block doubles so that architecture can learn the complex structures effectively. The bottommost layer mediates between the contraction layer and the expansion layer. It uses two 3×3 CNN layers followed by 2×2 up convolution layer. But the heart of this architecture lies in the expansion section. Similar to contraction layer, it also consists of several expansion blocks. Each block passes the input to two 3×3 CNN layers followed by a 2×2 upsampling layer. Also after each block number of feature maps used by convolutional layer get half to maintain symmetry. However, every time the input is also get appended by feature maps of the corresponding contraction layer. This action would ensure that the features that are learned while contracting the image will be used to reconstruct it. The number of expansion blocks is as same as the number of contraction block. After that, the

A. Training

- As input to the network, the magnitude spectrograms of the noisy voices.
- As output the Noise to model (noisy voice magnitude spectrogram - clean voice magnitude spectrogram).
- Both input and output matrix are scaled with a global scaling to be mapped into a distribution between -1 and 1.
- Many configurations have been tested during the training. For the preferred configuration the encoder is made of 10 convolutional layers (with LeakyReLU, maxpooling and dropout). The decoder is a symmetric expanding path with skip connections.
- The last activation layer is a hyperbolic tangent (tanh) to have an output distribution between -1 and 1.
- For training from scratch the initial random weights where set with He normal initializer.
- Finally the model is compiled with Adam optimizer and the loss function used is the Huber loss as a compromise between the L1 and L2 loss.

The training curve is illustrated in Figure 2.

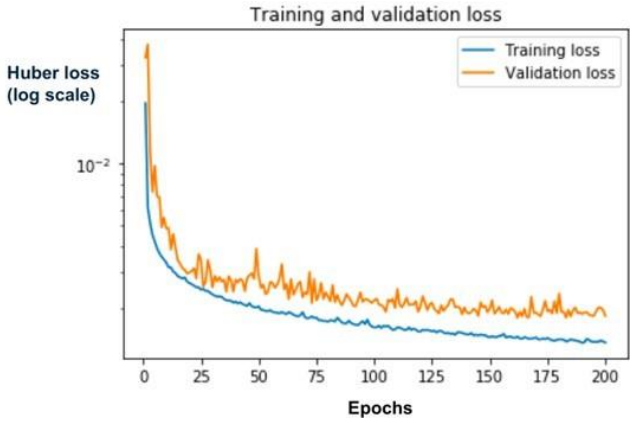

Fig. 2: Training Loss Curve

B. Prediction of Denoised Audio

For prediction, the noisy voice audios are converted into numpy time series of windows slightly above 1 second. Each time serie is converted into a magnitude spectrogram and a phase spectrogram via STFT transforms. Noisy voice spectrograms are passed into the U-Net network that will predict the noise model for each window. Prediction time for one window once converted to magnitude spectrogram is around 80 ms using classical CPU.

The model is then subtracted from the noisy voice spectrogram (here we apply a direct subtraction as it was sufficient for our task, we could imagine to train a second network to adapt the noise model, or applying a matching filter such as performed in signal processing). The "denoised" magnitude spectrogram is combined with the initial phase as input for the inverse Short Time Fourier Transform (ISTFT). Our denoised time series can be then converted to audio.

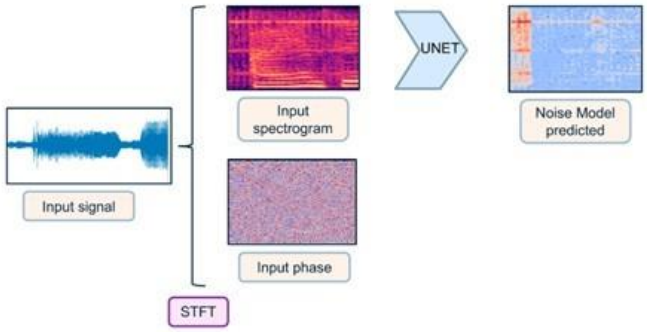

Fig. 3: Audio in Time Series Converted to Spectrogram and Phase

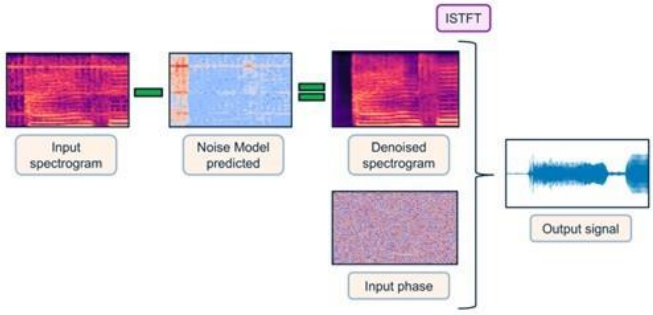

Fig. 4: Audio in Time Series Converted to Spectrogram and Phase

V. Results and Discussion

	Seen Noise					Unseen Noise				
Model	SNR	LSD	MSE	WER	PESQ	SNR	LSD	MSE	WER	PESQ
Noisy Speech	15.18	23.07	0.04399	15.40	2.26	14.78	23.76	0.04786	18.4	2.09
MS	18.82	22.24	0.03985	14.77	2.40	19.73	22.82	0.04201	**15.54**	2.26
DNN-SYMM	44.51	19.89	0.03436	55.38	2.20	40.47	21.07	0.03741	54.77	2.16
DNN-CAUSAL	40.70	20.09	0.03485	54.92	2.17	38.70	21.38	0.03718	54.13	2.13
RNN-NG	41.08	17.49	0.03533	44.93	2.19	**44.60**	18.81	**0.03665**	52.05	2.06
EHNET	**49.79**	**15.17**	**0.03399**	**14.64**	**2.86**	39.70	**17.06**	0.04712	16.71	**2.73**
Clean Speech	57.31	1.01	0.00000	2.19	4.48	58.35	1.15	0.00000	1.83	4.48

Fig. 5: Results according to reference paper

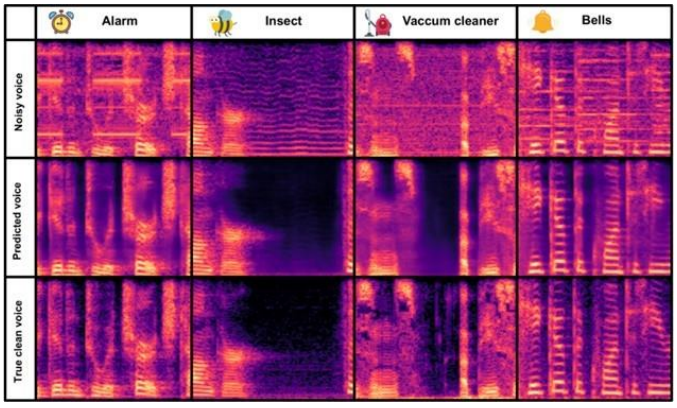

Fig. 6: Spectrograms of the clean, noisy and denoised speech

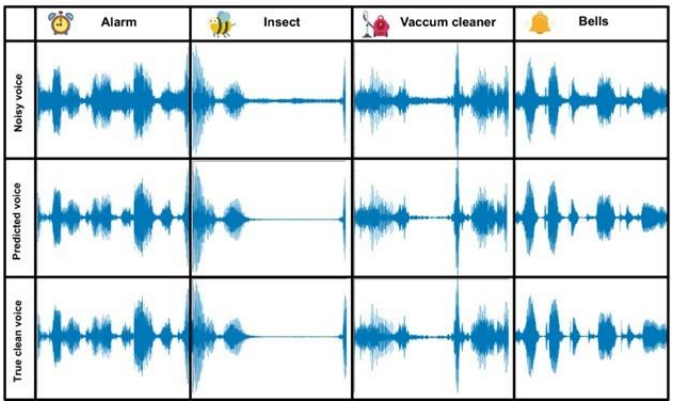

Fig. 7: Time Series representation of the clean, noisy and denoised speech

The model due to computation and time constraints managed to give an SNR of 35 on the test dataset. This was far better than the previous performances of the statistical approaches in related work.

VI. Future Work

We have to work on more datasets to ensure our model actually works properly on unseen data as well. We have to find out architectures and use our current datasets on them and then compare the loss function to see which handles speech denoising better. Another idea might be to use GANs to optimize the loss function and allow for more efficient training.

VII. Conclusion

We presented an end-to-end speech denoising pipeline that uses a fully-convolutional network, using a U-Net architecture pretrained on several relevant audio classification tasks for training. This approach allows the denoising system to capture speech structure at various scales and achieve better denoising performance without added complexity in the system itself or expert knowledge in the loss design. In particular, the presented approach is shown to perform much better in the noisiest conditions where speech denoising is most challenging.

References

[1] Jansson, Andreas, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar and Tillman Weyde.Singing Voice Separation with Deep U-Net Convolutional Networks. ISMIR (2017).
[2] Grais, Emad M. and Plumbley, Mark D., Single Channel Audio Source Separation using Convolutional Denoising Autoencoders (2017). https://arxiv.org/abs/1703.08019.
[3] Ronneberger O., Fischer P., Brox T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://arxiv.org/abs/1505.04597
[4] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia, 2015. DOI: http://dx.doi.org/10.1145/2733373.2806390