Speech Enhancement

Peter Lukač, Jakub Zárybnický

May 8, 2020

1 Introduction

Speech enhancement is one of the tasks in the domain of audio processing, others being source separation, speech recognition, or source localization. In particular we focus on the task of single source speech enhancement. In this report, we present our research into existing methods of speech enhancement, our approach, and compare the results we obtained with results of other attempts.

2 Task

Given a noisy sound signal, the processing pipeline aims to enhance the contrast of the speech signal against noise, with the ideal outcome being a clean signal without any noise.

There is a set of standard metrics for evaluating the quality of de-noising:

- Signal-to-Noise Ratio (SNR) is the ratio of the clean signal to noise in the estimated signal in the signal, quantified by the average power.
- Signal-to-Distortion Ratio (SDR) enhances SNR to account for distortion as well.
- Short-term Objective Intelligibility (STOI) measures uses time-frequency band analysis over short windows (˜400ms) comparing the clean and noisy signals.
- Perceptual Evaluation of Speech Quality (PESQ) is the standard method of evaluating speech enhancement techniques. On a scale of 1 (terrible) to 5 (great), this is a composite measure of quality.

3 Existing solutions

Speech enhancement is a rather traditional audio processing task, commonly performed in ways other than neural networks. Usual approaches include various filtering techniques (Wiener, subspace filtering, spectral subtraction), or spectral restoration using various objective estimators.

However, these approaches often improve only speech quality and not intelligibility, and often also introduce new distortions into the synthesized signal. This is also reflected in commonly metrics (the above-presented ones but also others), where it is the time-frequency spectrum that is compared between clean and noisy signals.

Neural network-based approaches usually rely on a base of convolutional layers, in combination with a fully-connected or recurrent layers as well. Most recent approaches use recurrent layers - bi-directional LSTM in particular, which is what we have used in the end.

Unfortunately, there have been no comprehensive review articles focusing on speech enhancement using neural networks since 1999 [5], despite the number of review articles focusing on speech recognition using a variety of approaches (the most recent one being a review [1] from 2019).

In evaluating out network's performance, we've used a small-scale review presented as part of a paper presenting a combination of MSE estimators with a convolutional LSTM network (DeepXi [2]), and in particular the PESQ measure.

4 Solution

Our solution uses TensorFlow as its basis (the Keras interface in particular), as well as a number of sound processing libraries for preprocessing the raw audio files (namely Librosa and SoundFile).

The neural network we use is composed of four convolutional layers, two LSTM layers, and finally four deconvolutional layers. While our experiments have included fully-connected, bidirectional LSTM, or other layers, this combination is one that has produced the best results.

The dataset we have used is a combination of the OpenSLR LibriSpeech sample library [3], and a number of noise sample sets (cars, cafés, noisy street, heavy machinery, . . .). The network has been trained using the Adam optimiser and MSE loss, in 64 epochs of minibatch.

5 Evaluation

In obtaining the PESQ measure of our cleaned samples (rescaling it so that it corresponds to the results of DeepXi), we can see that our work does not measure up to the current state-of-the-art.

Method	PESQ
Original	1.97
Apriori SNR	2.22
GAN	2.16
MSE-GAN	2.53
DeepXi	2.95
ConvLSTM (our)	**2.09**

The row marked *Apriori SNR* [4] is a paper from 1996 that does not use neural networks at all, whereas both of the other approaches use GANs and custom loss measures [**?**, **?**] - as opposed to our approach, where we've used a convolutional network without an adversary, and the common MSE loss.

Unfortunately, while the review part of DeepXi contained more than these works, they did not include the results of the PESQ measure.

6 Conclusion

Overall, we can conclude that while this approach is not untenable, it does not reach the standards of the current state-of-the-art. There are the obvious hyper-parameters to optimize (number or size of layers, changing the type of the LSTM used), or using epochs or larger datasets, but perhaps using a different loss criterium - as demonstrated the papers our work has been compared against - might also be a useful approach to try.

References

[1] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[2] Aaron Nicolson and Kuldip K. Paliwal. Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, 111:44 – 55, 2019.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[4] P. Scalart and J. V. Filho. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632 vol. 2, 1996.

[5] Eric A Wan and Alex T Nelson. Networks for speech enhancement. *Handbook of neural networks for speech processing. Artech House, Boston, USA*, 139:1, 1999.