Cocktail Party single speaker separation using Convolutional Neural Network based on Denoisng AutoEncoder

Kundan Kumar[1*], *Manikanda Krishnan 1*[1*], *Sanjeev K. Mittal*[1],
Chetan Singh Thakur[1] , *Anirban Chakraborty*[1]

[1]Indian Institute of Science

kundankumar@iisc.ac.in

Abstract

Speech separation has been made possible by the human auditory cortex, which is able to recognize and focus on selective voices in noisy and multi speaker environment. The problem of speech separation has great significance in much research area: neurosciences, computer sciences and auditory modeling.

Traditionally, audio separation is modeled as information processing, which includes; design of filters, hand selected features and computation modeling of human auditory cortex. Recently, deep learning based speech separation models has attracted many attentions, and deep neural network frameworks show their effectiveness in learning the useful representations of the target speaker from a mixture of speakers.

In this paper, we introduce a novel approach for the segregation of monaural sound mixtures based on a Denoising Autoencoders (ADE) network with convolution neural network (CNN) layers. Specially, we explore a training scheme for encoder-decoder network based on a KL divergence cost function proposed by DD Lee [16].

We evaluate our model using the mixture of speakers created from LibriSpeech-ASR [18]. The performance of reconstructed audio is compared by calculation signal-to-noise ratio (SNR) with respect to ground truth speaker audio.

Index Terms: Speech separation, Denoising Autoencoders, Deep Learning, KL Divergence.

1. Introduction

The Cocktail Party problem was introduced in 1953[1] that addresses the human ability to easily listen to one speaker in presence of multiple speakers and background noise. The solution to Cocktail party problem would have wide range of application in defense, industrial plants, mining, voice recording, and human to machine conversation, hearing impairment, and voice based assistance engine. Different techniques has been used to tackle this problem, one such way is by the human cortical modeling to model human auditory cortex.[2][3]

Recently, Deep learning has shown significant improvement in much of the speech processing task. Various types of convolution neural network architecture has been proposed to solve the problem of separating target voice form the mixture of voices. [9][10][11][19][20]. The use of Autoencoders for[1] source separation from mixture of sources is well studied [6][7][8].

Our work is closely related to single channel audio source separation using deep neural network based techniques [4][5]. In this work CNN based Denoising Autoencoders framework is used to solve the problem of target speaker separation from the mixture of two speakers. Denoising Autoencoders (DAE) are a special type of autoencoders [13][14][15] that are trained to reconstruct the source from the noisy representation. The DAE consists of encoder-decoder network, wherein the encoder creates a compressed form of useful representation, which is further used by decoder to generate original input. CNN has robust ability to extract the feature form complex representation. This has shown its effectiveness in image detection task that had been benchmark architecture, winning the prize entry of ImageNet challenge in 2012 [12]. Using CNN along with DAE have become a powerful technique to extract useful features from the speech signals or images.

In this work, we performed the training of CNN based DAE for source separation problem with KL divergence loss function [16]. The DAE is used to generate a spectral mask [9] that is further processed with the original mixture to create a clean representation of the target speaker.

The details of the CNN architecture, DAE model and training framework is given in section 2, followed by experiments section 3 that includes our data preparation and model evaluation strategy is presented in section 4. Experimental results are covered in section 5, followed by discussion in section 6 and conclusion in section 7.

2. Proposed Framework

Figure 1 shows the diagram of proposed framework for speaker separation by using DAE.

A mixture signal is first converted into a time-frequency representation known as spectrogram by applying short-time Fourier transform (STFT). Further, only magnitude is taken into account during the training of DAE. The magnitude spectrogram of the mixture is passed through the layers of encoder-decoder network. The encoder network generates the compressed representation, which is up-sampled by decoder network. The output of DAE is a soft mask Mi of the target

* Both authors contributed equally to this manuscript.

speaker, which is further combined with the original mixture to generate the clean representation of the target speaker Oi.

The KL divergence loss function uses the clean representation with model output, to reduce the reconstruction error between the target clean spectrogram and the reconstructed spectrogram from the model.

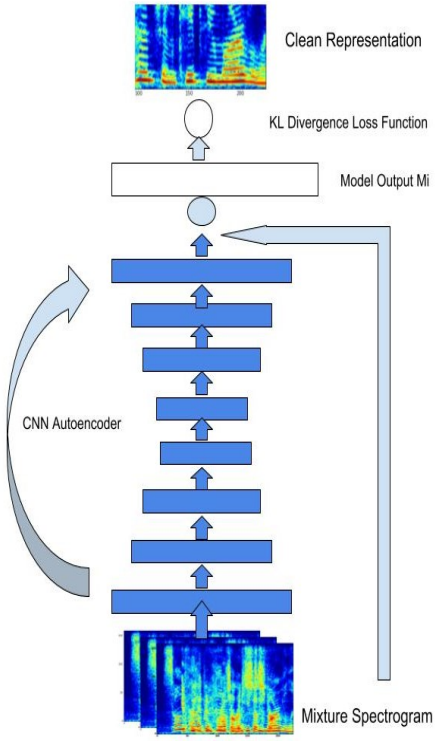

Figure 1: Proposed framework for speaker separation by using DAE

2.1. Model Architecture

The encoder-decoder network contains Convolution neural network layer (CNN). Encoder part consist of four CNN layers. First three layers are each followed by max-pooling, while fourth layer is just a convolutional layer without pooling. The Decoder unit contains similar four CNN layers followed by an up-sampling operation after each layer. The structure of encoder-decoder network along filter sizes and dimension after each layer operation is mentioned in Table 1.

The activation function play important role in training of the neural network, for DAE we chose Relu activation [24], which is an state-of-the-art activation function and has proved its importance in breakthrough models[12][25].

The choice of optimizer algorithm for training of the deep learning models has significant impact on the training time of the model. For our DAE framework we choose The Adam optimization algorithm [21], which has recently been adopted in many computer vision and NLP frameworks.

Table 1 shows the model summary of the DAE, along with the model parameters after each layers and filter sizes.

The filter size of each layers is 3 X 3 from [4], the Max-pooling2D (2, 1) reduce the size of the feature map by 2 time frame, and by 2 in frequency frame. The output shape (128 X 256 X 12) denotes size of 128 time frames, 256 frequency bins and 12 filters.

Model Summary		
Layers	**Filters**	**Output Shape**
Input	-	(256 X 256 X 1)
Convolution2D(3,3) Max-Pooling2D(2,1)	12	(128 X 256 X 12)
Convolution2D(3,3) Max-Pooling2D(2,2)	18	(64 X 128 X 18)
Convolution2D(3,3) Max-Pooling2D(2,2)	36	(32 X 64 X 36)
Convolution2D(3,3)	72	(32 X 64 X 72)
Convolution2D(3,3) Up-sampling2D(2,2)	36	(64 X 128 X 36)
Convolution2D(3,3) Up-sampling2D(2,2)	18	(128 X 256 X 18)
Convolution2D(3,3) Up-sampling2D(2,1)	1	(256 X 256 X 1)

Table 1: The detailed structure of each layer.

2.2. Loss Function

Loss function plays an important role is reducing the training errors. A good choice of loss function in the DAE gives a better reconstruction of the original input signal. We have used the Kullback Leibler divergence loss function for better error reduction from the clean source and the model output which is given equation 1.[16]. Further, L2 loss function is defines by equation 1a [4]

Xij represents the clean source of the target speaker, while Oij is the model output produced by the DAE mask and the mixture spectrogram.

$$J = \sum_{i,j}^{N} \left(Xij \log \frac{Xij}{Oij} \right) - Xij + Oij \qquad (1)$$

$$J = \sum_{i,j}^{N} (Oij - Xij)^2 \qquad (1a)$$

2.3. Spectral Mask

DAE is used to produce the spectral soft mask for the target speaker [4]. In equation 3, Mi is spectral soft mask, which is the output of the DAE. The Mi is further normalized with the sum of other speakers mask. In this case, the sum of other mask is taken to be one, as we have single speaker for separation. Equation (2) is then reduced to equation (3).

$$\sum_i^J Mi(n,f) = 1 \quad (2)$$

$$Li(n,f) = \frac{Mi(n,f)}{\sum_i^J Mi(n,f)} \quad (3)$$

$$Li(n,f) = Mi(n,f) \quad (4)$$

The spectral mask is generated by DAE in equation (4) is multiplied with the original mixture Ci in equation (5) to generate the clean representation Oi of the target speaker.

$$Oi(n,f) = Li(n,f)\, Ci(n,f) \quad (3)$$

3. Experiments

3.1. Dataset

LibriSpeech ASR [18] corpus is used to create the mixture of two speakers, sampled at 50000 points. Each mixture is of 3 seconds time duration. The mixture is created with 0 dB SNR noise and un-normalised magnitude. The mixture data was formed by randomly picking one female speaker and mixed with other male speakers. As these samples were not of same size, they were normalized to create the mixture of equal length. 2000 such samples were utilized for training whereas 200 samples were used to test the network.

3.2. Experimental Setup

Our proposed algorithm separates the target speaker from the mixture of speakers. The mixture was simulated 22 kHz sampling rate. The STFT spectrogram was created based on Librosa framework [23]. The window size of 196 samples along with FFT size of 512 points and 1024 points were used to create the spectrogram.

We used only the magnitude part of spectrogram to train the network. The reconstruction of spectrogram to audio file is done by Single Pass Spectrogram Inversion algorithm proposed in [17] based on magnitude spectrogram and phase information of target speaker. The importance of phase is discussed in discussion part.

The Tensorflow [22] framework is used for implementation of the algorithm.

4. Evaluation

The model is evaluate based on the test set of the mixture created from Librispeech. The SNR value [26] of the clean and mixture is calculated with respect to the contribution of the non-target speaker in the mixture. Equation 4 $SNRi$ is signal to noise ratio of clean target $C(n)$ and the contribution of non-target speaker $(m(n) - C(n))$ in the mixture, where $m(n)$ represents mixture.

$$SNR_i = 10\, log_{10}\left[\frac{c(n)^2}{(\, m(n) - c(n))^2}\right] \quad (4)$$

Similarly, equation 5, $SNRreconstructed$ represents the value of the reconstructed source from the model with respect to the non-target speaker component. Where, $O(n)$ is the output of the model.

$$SNR_{reconstructed} = 10\, log_{10}\left[\frac{O(n)^2}{(m(n) - c(n))^2}\right] \quad (5)$$

Finally, difference of equation 4 and equation 5 is calculated in equation 6, which gives the comparative analysis between the ground truth or clean audio of speaker and the reconstructed audio of the target speaker from the model.

$$diff_{SNR} = SNR_i - SNR_{reconstructed} \quad (6)$$

5. Results

Results of model, as spectrogram of clean, mixture and reconstructed with 512 FFT points, for l2 loss function and KL divergence based loss function is depicted in figure 2.

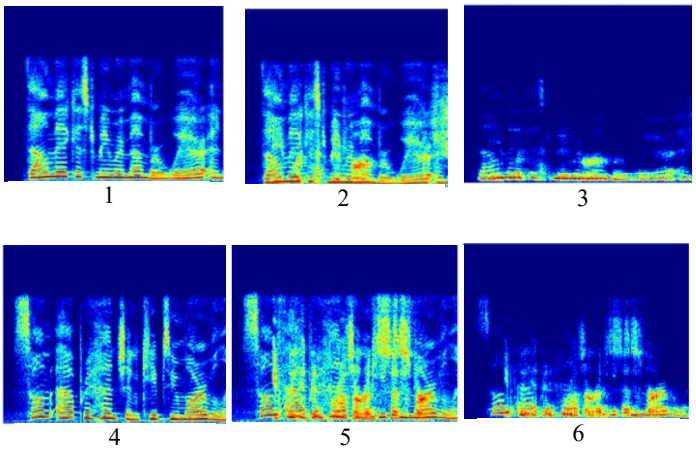

Figure 2: Spectrogram for clean, mixture and reconstructed audio with L2 loss is depicted in 1, 2, 3. Whereas 4, 5, 6 represents spectrogram for clean, mixture and reconstructed with KL divergence based loss function. The FFT size for these spectrograms is 512 points.

Table 2 shows value of $SNRi$, $SNRreconstructed$ and $diff_{SNR}$ with L2 loss and KL divergence loss function and FFT point of 512 and 1024.

Loss Function	SNR_i	SNR_{recon}	$diff_{SNR}$
L2 Loss with FFT 512	-5.597	-2.452	3.146
L2 Loss with FFT 1024	-5.442	-3.860	1.585
KL divergence Loss with FFT 512	-8.651	-8.027	0.628
KL divergence Loss with FFT 1024	-5.216	-3.884	1.330

Table 2: SNR of input, reconstructed and difference of input and output SNR.

6. Discussion

The quality of the source separated from the mixture is measured based on SNR values given in equation 4,5,6. And their values corresponding to different loss function and FFT sizes are given in table 2.
Achieving the less difference in SNR indicates the good separation performance
As it can be seen from Figure 2, the reconstruction of target speaker is better in the case of KL divergence loss as compared to the L2 losses. The KL divergence is logarithmic loss function and provide a stable training framework for this model
We also see from the Table 2, the reconstruction difference with KL divergence loss function is less as compared to the L2 losses, in case of both FFT points of 512 and 1024.

As, most of the literature do not emphasise a use of the mixture phase during the spectrogram to audio generation. In this work audio reconstruction from spectrogram is done by multiplying the clean signal phase with the spectrogram magnitude. Work in [27] discusses importance of original phase information used to reconstruct the audio file from spectrogram.

7. Conclusions and future work

In this work we proposed a scheme of training the denosing autoecode based on KL divergence loss function for single speaker separation from the mixture of speakers. This model is based on convolution neural network based encoder-decoder network along with spectral mask framework. The experimental results with KL divergence loss function is compared with the L2 losses indicates its usefulness in reconstruction for source form mixture.

In our future work we plan to pursue the scaling of the model for both speaker separation. We would also explore the possibility of including phase spectrogram information during training to improve the network performance.

8. Acknowledgements

We express our thanks to Indian Institute of Science, Department of Electronic Systems Engineering for providing us the lab infrastructure and computing facilities. We also thank to thank to our Neuronics lab members for their critical review of our ideas and progress and support throughout this study.

9. References

[1] Cherry E. C., (1953).Some experiments on the recognition of speech with one and two ears, J. Acoust. Soc. Am. 25, 975979.

[2] Krishnan, L., Elhilali, M. & Shamma, S. A. Segregating complex sound sources through temporal coherence. PLoS Comput. Biol. 10, e1003985 (2014).

[3] G. Wolf, S. Mallat and S. Shamma, "Audio source separation with time-frequency velocities," 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Reims, 2014, pp. 1-6. doi: 10.1109/MLSP.2014.6958893

[4] Emad M Grais and Mark D Plumbley. Single channel audio source separation using convolutional denoising autoencoders .arXiv preprint arXiv: 1703.08019, 2017.

[5] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 3734–3738.

[6] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monaural audio source separation using deep convolutional neural networks," in International Conference on Latent Variable Analysis and Signal Separation .Springer, 2017, pp. 258–266

[7] Raj, Bhiksha, Virtanen, Tuomas, Chaudhuri, Sourish, and Singh, Rita. Non-negative matrix factorization based compensation of music for automatic speech recognition. In Proc. INTERSPEECH, pp. 717–720, 2010.

[8] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in International Conference on Latent Variable Analysis and Signal Separation. Springer, 2015, pp. 100–107.

[9] P-S Huang, Kim, M Hasegawa-Johnson, P Smaragdis, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Deep learning for monaural speech separation (IEEE, 2014), pp. 15621566.

[10] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley, Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network, in Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), 2015, pp. 429436.

[11] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multi Channel audio source separation with deep neural networks," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 24,no. 9, pp. 1652–1664, 2016.

[12] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In NIPS, pp. 11061114, 2012.

[13] Hinton, Geoffrey E and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked Denoising Autoencoders: learning useful representations in a deep network with a local denoising criterion" Journal of Machine Learning Research, vol. 11, pp. 3371 –3408, 2010.

[15] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders" in Proc. ICML, 2008, pp. 1096–1103.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing 13 (Proc. NIPS*2000). MIT Press, 2001.

[17] GT Beauregard, M Harish, L Wyse. (2015) Single Pass Spectrogram Inversion. IEEE International Conference Digital Signal Processing (DSP), 427-431, 2015. 12, 2015.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audiobooks. In2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015, pages 52065210, 2015.

[19] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high resolution deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 24, pp. 1424-1437, 2016.M L.

[20] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," arXiv preprint arXiv:1708.07524, 2017..
[21] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. ICLR, 2015.
[22] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015.Software from tensorflow.org, 2015.
[23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th Python in Science Conference, 2015.
[24] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit.Nature, 405(6789):947, 2000.
[25] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. InInterna-tional Conference on Machine Learning, 2010
[26] P. Papadopoulos, A. Tsiartas, J. Gibson, and S. Narayanan, "A supervised signal-to-noise ratio estimation of speech signals," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., May 2014, pp. 8237–8241.
[27] Dubey, G T. Kenyon, N Carlson, A Thresher, in 2017, Does Phase Matter For Monaural Source Separation?, CoRR, 2017