International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

CNN only Single Source Speech Denoising, for inference at edge

Atul Yadav[1], Vishesh Jain[2], Rishi Kumar Verma[3]

[1,2,3]UG Student, dept. of Computer Science Engineering, Government Engineering College, Raipur, India

---

Abstract - *Speech Denoising refers to the removal of any environmental noise/artifacts from audio signals. We propose a method of using CNN, trained with appropriate loss functions; MSE can perform much better than recurrent neural networks with the size ~670kb. The loss functions are motivated from BSS_EVAL like source to distortion ratio (SDR), source to interference ratio (SIR) and source to artifacts ratio (SAR).*

Key Words*: *Deep Learning, Speech Denoising, Autoencoder-decoder, convolutional neural network

1. INTRODUCTION

Speech denoising is an outstanding problem; oldest works include S.Boll's spectral subtraction method[1] from 1979. With deep learning many new methods and models have been proposed, John R. et al.[2] has proposed use of deep unfolding neural networks which is very suitable for real time applications.

Some amazing results are observed when visual features are also considered with audio signal, Ariel Ephrat et al.[3] they are computing a 1024-D vector from the face of the speaker using [4], and feeding it as input to their model along with STFT spectrogram. A similar approach is seen in [5], [6], [7]. Rationale behind using visual clues is to mimic how humans leverage lips reading to compensate for noisy or less auditory audio. Although this performs really great visual features are not readily available in all scenarios and use of BLSTM layers make these models considerably slower. Google coral TPU is a deep learning runtime accelerator for edge computing, it lacks the support for any RNNs as of writing this; similar is the case with other edge hardwares, thus making solutions with recurrent layers very less useful at the edge.

Convolutional Neural Networks usually have a much lesser number of parameters compared to FNNs and RNNs. CNNs have been used for style transfer on speech[8], speech recognition [9][10].

We inspire our work from Se Rim Park et. al. [11], they have proposed a fully convolutional neural network called R-CED as shown in Fig 1.

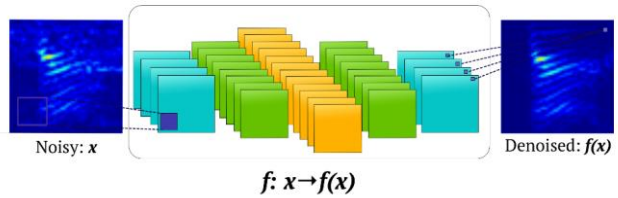

Fig-1: FCNN for Speech Enhancement

While the R-CED model works well and even efficiently given it's small size, the MSE loss function used by Se Rim Park et. al. makes it harder to converge and degrades speech quality when higher magnitude of noise is introduced, therefore we propose 3 metrics for loss functions inspired from BSS_EVAL[11]

Paper is organized as follows. Section 2 describes the datasets used and synthetic training set preparation, section 3 describes the architecture of neural network used, section 4 presents our proposed loss function, section 5 contains the results which were observed and in section 6 we conclude our work with further ideas to improve the model's performance.

2. Dataset

We have used 2 datasets that are publicly available, The Mozilla Common Voice and The Urban Sound 8K, the former is used for speech sound and the latter is used to derive noise from. UrbanSound8K has 10 classes of noise in it.
For training we removed any silent frames from the speech audio and added noise such that a SNR of 0dB is maintained, a 256-point Short Time Fourier Transform (STFT) is computed from the resultant signal to be fed to the network. The STFT window has a length of 256 and hop size of 64, this ensures 75% overlap. Finally the input vector is composed to have a shape (129, 8) i.e. current STFT noise vector plus 7 previous noise STFT vectors. Output of the network would have a dimension of (129, 1).

This is sometimes referred to as autoregressive modeling where the model predicts current output based on previous observations, here previous observations are fixed to be 7.

The data was finally divided into 3 parts; train, test and validate with respective proportions 40:40:20.

3. Convolutional Network Architecture

The Redundant Convolutional Encoder-Decoder model used by us was proposed first by Se Rim Park et. al. [11], it's architecture is very similar to a conventional Convolutional Encoder-Decoder model. Unlike an usual Encoder-Decoder model with R-CED we don't compress the features in encoder part and upsample in decoder; rather we it encodes feature into higher dimension on the encoder part and compresses along the decoder part, this completely eliminated the need of upsampling layers.

Rationale behind this is that auto encoder-decoder are usually interested in compressed representation as an embedding or just compressed data, but here we are only concerned with the input and output from the network.

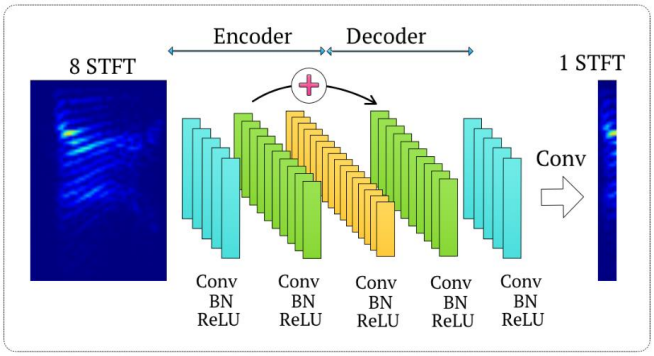

Fig-2: R-CEDNetwork

A concept of 1-D convolution is also used, it basically means using a convolution filter of size 1xnumber of rows in input, as illustrated below in Fig-3, while Fig-4 is a normal filter used in image recognition projects.

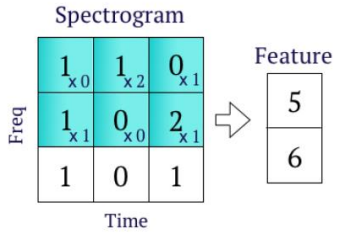

Fig-3: 1-D convolution

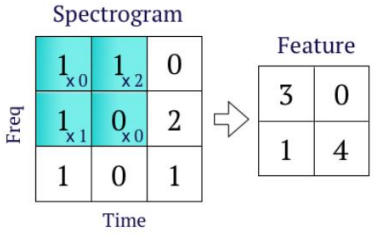

Fig-4: 2-D convolution

4. Loss Function

We propose 3 new loss functions along with MSE inspired from BSS_Eval, we found that MSE highly benefits from suppressing the whole audio signal to achieve optimal results. BSS_Eval is used a lot in source separation problems, where the sound of a singler speaker has to be separated from multiple speakers; speech denoising is also a very similar problem.

We aim to remove interference and artifacts from the signal. Interference can be defined as the additional sound except the speaker's voice, source-to-interference ratio(SIR) is used to measure and remove any interferences. Artifacts refer to sounds additionally introduced to the signal which were originally not present; this is due to the algorithm and several transformations, it can be measured and removed with source-to-artifacts ratio (SAR). Meanwhile source-to-distortion ratio(SDR) is a combination of the above two which captures how well both noise and speech signals are separated.

We are using the following notations, **s** refers to speakers sound, **d** refers to noise and **s'** refers to speech + noise signal. These refer to the time domain signal of the speech and **s'** and **d** are constant when the network is optimized for **s.** Notation [xy] refers to $x^T.y$

For maximizing SDR,

$$\begin{aligned} maxSDR(s', s) &= max \frac{[ss']^2}{[ss][s's'] - [s's]^2} \\ &\equiv min \frac{[ss][s's'] - [ss']^2}{[s's]^2} \\ &= min \frac{[ss][s's']}{[s's]^2} - \frac{[s's]^2}{[s's]^2} \\ &\equiv min \frac{[s's']}{[s's]^2} \end{aligned}$$

For maximizing SIR,

$$maxSIR(s', s, d) = max \frac{[dd]^2[s's]^2}{[ss]^2[s'd]^2} \equiv min \frac{[s'd]^2}{[s's]^2}$$

For maximizing SAR we are assuming **s** and **d** to have a phase difference 90degrees, i.e. **|s.d|** = 0; this assumption helps with following simplification

$$maxSAR(s', s, d) = max \frac{\| \frac{[s's]}{[ss]} s + \frac{[s'd]}{[dd]} d \|^2}{\| s' - \frac{[s's]}{[ss]} s - \frac{[s'd]}{[dd]} d \|^2}$$

$$\equiv min \frac{\| s' - \frac{[s's]}{[ss]} s - \frac{[s'd]}{[dd]} d \|^2}{\| \frac{[s's]}{[ss]} s + \frac{[s'd]}{[dd]} d \|^2}$$

$$= min \frac{[s's'] - \frac{[s's]^2}{[ss]} - \frac{[sd]^2}{[dd]}}{\frac{[s's]^2}{[ss]} + \frac{[s'd]^2}{dd}}$$

$$= \frac{[s's']}{\frac{[s's]^2}{[ss]} + \frac{[s'd]^2}{[dd]}}$$

5. Results

Performance of the system was evaluated based on subjective listening by the authors. As opposed to looking for a model with least loss we are evaluating it on the perpetual quality of the output. Following combination of loss functions were compared -

1. MSE
2. SDR
3. 0.7 x SDR + 0.15 x SIR + 0.15 x SAR

As expected MSE performed the worst with suppressing the speech along with noise signal, SDR performed marginally better but still had some artifacts left. Last function performed significantly better than the baseline MSE model.

6. Conclusion and Future Scope

We have demonstrated the superiority of loss function derived from the problem domain over generic functions like MSE, further research needs to be done in engineering better loss functions that are differentiable, thus can be used for backpropagation. In our knowledge STOI and deep features can further improve the performance of this neural network.

Also the saved model instance is just ~760kB in size, due to weight sharing capabilities of a CNN it is very efficient for edge applications. Further study is required on runtime metrics of this model in readily available hardwares, CNNs can be accelerated using special vector instruction set on the CPU; alternatively microcontrollers such as MAX78000 or Kendryte K210 have built-in convolution neural network computation capabilities.

REFERENCES

[1] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction", 1979.

[2] John R. et. al., "Deep Unfolding:Model-Based Inspiration of Novel Deep Architectures", 2014.

[3] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. "Synthesizing normalized faces from facial identity features", 2017.

[4] Ahsan Adeel, Mandar Gogate, Amir Hussain, William M. Whitmer, "Lip-Reading Driven Deep Learning Approach for Speech Enhancement", 2018.

[5] Jen-Cheng Hou; Syu-Siang Wang; Ying-Hui Lai; Jen-Chun Lin; Yu Tsao; Hsiu-Wen Chang; Hsin-Min Wang, "Audio-visual speech enhancement using deep neural networks", 2017

[6] Aviv Gabbay, Asaph Shamir, Shmuel Peleg, "Visual Speech Enhancement" 2017

[7] Prateek Verma, Julius O. Smith "Neural Style Transfer for Audio Spectrograms", NIPS 2017

[8] Ossama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang,Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," 2014

[9] Dario Amodei, Rishita Anubhai, Eric Battenberg, CarlCase, Jared Casper, Bryan Catanzaro, Jingdong Chen, MikeChrzanowski, Adam Coates, Greg Diamos, et al.,"Deepspeech 2: End-to-end speech recognition in english and man-darin," 2015

[10] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005

[11] Se Rim Park and Jinwon Lee, "A Fully Convolutional Neural Network for Speech Enhancement" , 2016