International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-6, November 2022
https://doi.org/10.55524/ijircst.2022.10.6.7
Article ID IRPV1065, Pages 39-42
www.ijircst.org

A New Residual Convolutional Neural Network-Based Speech Improvement

M. Balasubrahmanyam[1], B. Haribabu[2], and P. Uday Kumar[3]

[1,2,3]Assistant Professor, Department of Electronics and Communication Engineering,
PACE Institute of Technology and Sciences, Ongole, India

Correspondence should be addressed to M. Balasubrahmanyam; subramanyam_m@pace.ac.in

ABSTRACT- Among the most crucial methods for denoising a noisy voice signal and enhancing its quality is speech enhancement. This study makes use of Adaptive Residual Neural Network technique to reduces maximum off background noise. This method continuously monitors the background noise depends upon the environmental changes using SNR parameter. It has two functions first one is non linear functions followed by convolutional neural networks and second one is linearity followed due to Residual neural networks. By using these factors remove background noise even SNR is low conditions. Compared to other techniques this technique is new,fastest, requires less training and also reduces size.

KEYWORDS- Convolutional Neural Networks, Liver, Segmentation, Tumor, ResNet, Deep Learning.

I. INTRODUCTION

For a very long time, speech enhancement was utilised as preprocessing. step in a variety of applications that include speech, like hearing Aids [1, 2], speech coding [3, 4], cochlea implants [4, 5], and recognising speech automatically [6]. As time has passed quite a bit of research has been done. in this area beginning Weiner filtering's fundamental methods [7], Spectral Subtraction [8], factorization of a non-negative matrix [9]. These methods don't extract anything valuable from the data, but they do present a blind estimate using pre-built noise and speech models. Their Model assumptions are complex and incongruent with reality stationary noise, which results in these approaches failing in practical applications

In this, numerous the use of deep neural networks investigated. Despite a domain's promising performance, practically everyone suffers the drawbacks of complexity and resource consumption. Traditionally, utilising deep neural networks fully connected were investigated for the task of improving voice, but neural network size normally scales up to hundreds of megabytes, then usage in embedded systems is challenging. Convolutional Neural Networks, however,due to their smaller number of factors, the hand sharing property with weight. The models are more accurate and faster less resource-intensive and can be integrated into portable devices devices. Historically significant models like Fully In-depth neural networks with connections (DNN) take 20–30 times as long. greater parameters for similar amount of layers and greater based on convolutional layer reduction parameters and a measure of the inputs reduce training period for achieve the same results, and CNN performs more effectively when identifying local dependencies and resulting in much better outcomes than dense networks.

The goal of the research is to identify a memory-efficient denoising method that eliminates background noise from test samples with hidden speakers and hidden noise environments. Unseen samples refer to those that were not utilised in the network's training to confirm that the model generalises well to conditions in the real world and does not cling to the training-specific conditions. Two convolutional neural network designs are used in the research to analyse the findings. The suggested architecture, a skip convolutional neural network connections, and even a model with no skip connections are both neural networks. The suggested model performs better than depending on the examination, various models metrics when both of to assess the performance, these architectures are trained both with and without dropouts. measures.

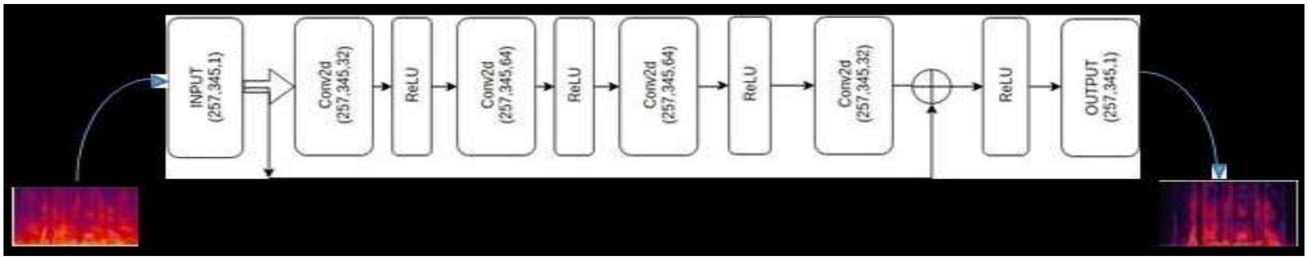

Figure 1: Architecture model four convolutional layers with identical padding and ReLU activation make up the RCNN. Clean speech spectra are produced at the output after input of noisy speech spectra

II. CONCEPT OF THE PROBLEM

Let x ϵ X fxt and y ϵ X fxt be the noisy parallel and its associated clear profound and significant, with t denoting the spectrogram's length and f denoting the extent of its frequency bins. The issue might be Finding a mapping between loud and clean has been formalised signal, i.e. Z: X fxt --> X fxt, which maps raucous utterances to a clean one, incidentally minimising L2 norm, officially known as Using the mean-square-error of the input and output signals, the optimal weights for the neural model are determined. assuming that the training set Train = {(Xm, Ym)}m= n pairs of spectrograms of clean and noisy signals, ranging from 1 to n.

$$\Delta w = \frac{1}{n}\sum_{m=1}^{n} ||Z(x_m) - y_m||^2$$

*Δ*w is our model's loss, which indicates how far our model's predictions are from the predicted outputs.

III. CONVOLUTIONAL NEURAL NETWORK, RESIDUAL

The spectrogram can be examined for the frequency and temporal dimensions of regional trends using neural networks using convolutions(CNN). Using convolutional layers the sole kinds of layers used in CNN as shown in figure.1. Short-term Fourier Transform (stft) is used as the source for convolutional layers, and each layer receives inputs from the preceding layer in the form of a small rectangular grid, which are then multiplied by the weights matrix (w). To find particular kinds of regional patterns, the Localized filter, also known as the Wrights matrix, is duplicated throughout the whole input space. The component neurons of an outline map will all have the identical weights. Our method uses a 2-D plane regularly and over time axes as with the entry point for voice enhancement. same padding without a pooling layer.

$$Padding = \frac{(Kernal - 1)}{2} \quad Output = \frac{(Width - Kernal + 2 * Padding)}{stride} + 1$$

Here, kernal-based padding is added, and the output is according to the width of the preceding window, the kernal size, the stride, and the the output's padding window, layer of convolution. Additionally, network features bypass connections, linking the output and input layers. This was done owing to the structure of output and input created is similar, and many characteristics perhaps lost when recreating last year's work. since the gradient's diminishing. Skip connections assist in maintaining the knowledge to prevent the deterioration speech for us signal, as well as we keep the positive qualities. Skip connections enabling gradients to penetrate the network more deeply will ultimately producing better outcomes.

$$Y^n = (w^{n-1,n} * a^{n-1} + b^n) + (w^{0,1} * x^0 + b^1)$$

$$Y^{out} = ReLU(Y^n)$$

in which Yn is the result of the final convolution layer, in which wn-1, n denotes its weight, The stimulation of an-1 layers before it, also bn implies its bias, these are combined with w0, 1, x0, and the first layer's bias b1. The final layer's output, Yout, is created by feeding Yn through the Unified Rectified Unit (ReLU) non-linearity.

We suggest implementing this using the Convolutional Residual Neural Network depicted in Figure 1. Our model's architecture of a neural network exclusively employs convolutional layers. As a result, the model simply aims to understand the chronological tight relationships between input data. Using Convolutional layers alone lowers quantity of layers used.

IV. INTERVENTIONS

A. Data Set

The Edinburgh dataset [12], which is available to the public, was used to test the model. Data volume and speakers match the description of the issue. At 16kHz, the entire dataset is resampled. The training set has 28 speakers, each of whom has 400 sentences. There are 10 different types of noise and 4 different SNR levels (15 dB, 10 dB, 5 dB, 0dB). Therefore there are 10 phrases per noise and SNR condition. speaker. speakers not seen and noise settings that weren't present in the model's training set are present in the test set. It has 2 speakers, 5 different noise kinds, and 4 different SNR conditions that are each 18.5, 13.5, 6.5, and 3.5 dB. As a result, each speaker has twenty sentences each condition.

B. Evaluation Matrixs

As most often used estimators of voice quality, distortion of speech and noise reduction are employed to assess our model's performance.

C. Preprocessing

We must first extract characteristics our unprocessed audio wave that will be used to train the model before we begin the real training process. We spoke loudly for two seconds. phrases used in our 512-point Hamming window and model training 75% overlap is used. The model is given the impure speech as input. seeing such that fewest features contain the most The signal's potential information is displayed. Switching The task is accomplished from time in terms of frequency. domain. due to this Term Fourier Short Transform, or STFT, is used.

At this stage, the amplitude phasing values of the intricate FFT sequence that currently serves as our signal's representation can be fed through the neural network. Although, it appears that we have so far only the magnitude of our neural network's input was expanded.

Signal phase is not taken into account. The phase, which is eventually utilised to create the noise-free audio file by combining the phase from the noisy file with the output from the neural network signal, is not taken into consideration by the neural model, which is just optimised for the audio's size. signal. But the effectiveness of the outcomes that may be acquired using this strategy is constrained.

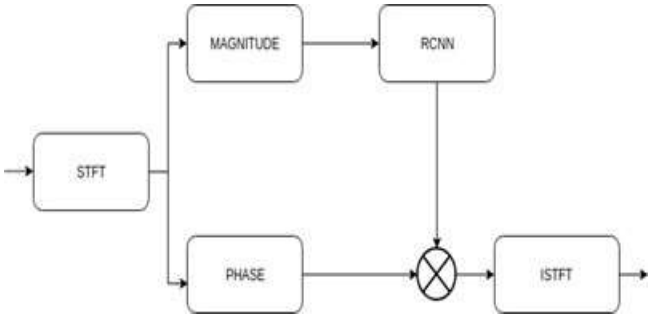

Figure 2: ISTFT model

Fig.2: steps applied to the raw audio waveform before processing. In order to create the final output audio file, the ISTFT is used after the waveform is provided as input to the STFT and the expected results have been multiplied by the phase from the noisy signal.

D. Instruction of the Model

Since The simulation is prepared via the Google Collaboratory, it seems to be essentially given training for just two hundred eras and produces respectable results. A smaller multilayer the network is employed since the model's lack of computing requirements. Since the skip link is beneficial to conserve low level characteristics and prevents their disappearance, It's been shown that including the skip connection the design results in a better signal than the model without skip connections. It has been observed that adding a dropout layer to the model during training reduces the quality of signals since it discards both random and valuable characteristics, which are difficult to recover and generate distorted signals. It also requires an extension of time to the specified set of features.

Best model is honed using a variety of nonlinearities, including elu, selu, prelu, and relu. Relu produced the most attractive outcomes as a result of less likely to have issues like disappearing gradients. Relu significantly lowers the net's training's computational cost. greater net sizes and more parameters can be trained using this method similar computational burden, increasing capacity and frequently Additionally, test set accuracy. On a learning rate of 0.0005, various optimizations including sgd, adagrad, and adam are used, with adam emerging as the victor.

V. RESULTS AND DISCUSSION

The study investigates the application of speech-generating convolutional neural networks augmentation. Two fundamental network designs that use skip connections and the other without are used for the studies. We also conduct analyses to verify the outcomes of employing dropouts when training our models for novelty. As seen in table 1, overall dropouts do not improve the network's efficiency in converting the noisy spectra to clean. While taking into account the equipment and resource.

Constraints that are currently in place as well as the requirement for creating a lightweight, convergent model more quickly utilises less data, the suggested skip links network trained with no dropouts outperforms other architectures and yields reasonable results as shown in figure.3.

Table 1: Models' NR and SD Performance

Model	Noise in the background Reduction	Speech Disruption
CNN's drop-out	0.854	0.142
CNN without a break	0.834	0.057
With dropout RCNN	0.892	0.054
Without dropout, RCNN	0.969	0.052

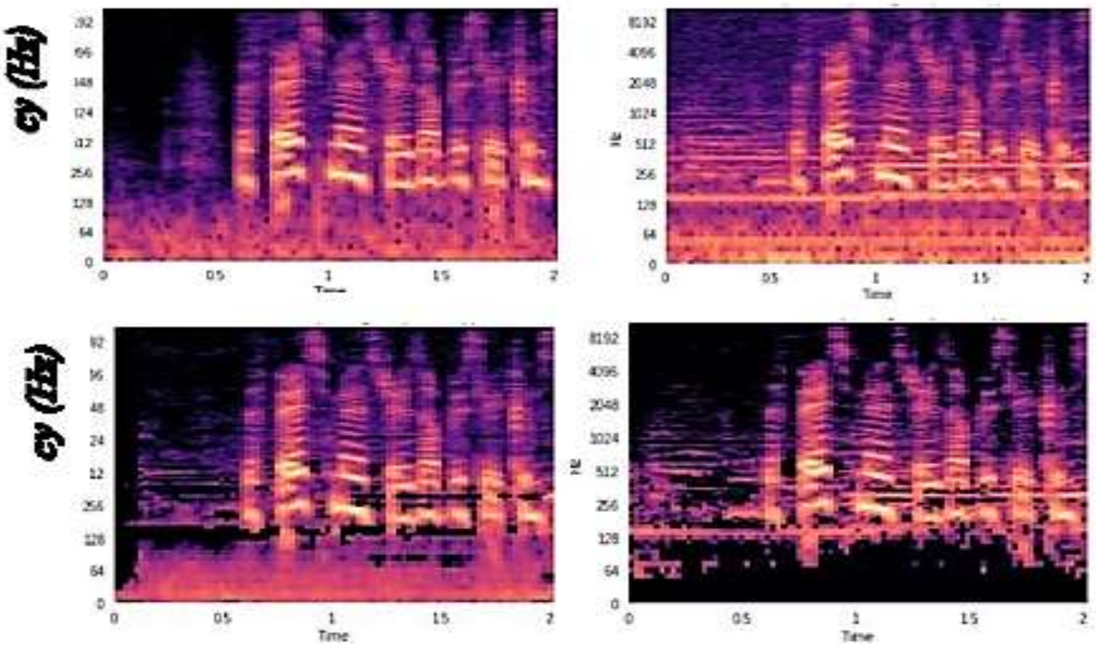

Figure 3: [a] The Clean Speech Spectrogram and the [b] Noisy Speech Spectrogram
[c] RCNN's improved speech without interruptions, [d] CNN's improved speech without interruptions

The table 2 shows the comparison of different deep learning models for storage.

Table 2: Comparison of Model Size

Pattern	Capacity (Megabytes)
Using Deep Neural Networks to Improve Speech	100-160
based on reinforcement learning Enhancement	60-70
Speech Enhancement using Generative Adversarial Network(SEGAN)	50-60
Convolutional Auto Encoder for Speech Enhancement	11-20
Residual Convolutional Neural Network For Speech Enhancement	0.54-0.63

Our model's size is far less than that of the other networks that have been suggested, yet performance levels haven't been greatly affected. This represents a significant advancement in the integration of neural network-based speech enhancement models in wearable technology. In future work for security of speech signals we design a wireless model by using the methods of [13-17]

VI. CONCLUSION

This research focuses on improving speech, a problem that is tackled by using neural networks. Here, the emphasis is on developing a neural network design that is simple to use, train, and deploy in the real world, thereby removing the major shortcomings of neural networks. Speech noise is reduced in the proposed model by 96.97%, and 4.1% less speech distortion is produced. Only a 10% performance decline and a 99.45% model size reduction from the top performing model are observed. This research is a ground-breaking effort to fully utilise neural networks' capabilities without concentrating enlarging them for better performance. More clamour kinds SNR values may also taken into account in further work to provide a more reliable

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] H. Levitt, "Noise reduct ion in hearing aids: A review," Journal of rehabilitation research and development.

[2] D. Wang, "Deep learning reinvents the hearing aid," IEEE Spectrum.

[3] J. Li, M. Akagi, S. Sakamoto, S. Hongo, & Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," Speech Communication, 2011.

[4] Y. Hu, F. Chen, & M. Yuan, "Evaluat ion of noise reduct ion methods for sentence recognit ion by mandarinspeaking cochlear implant listeners," Ear & hearing, 2015.

[5] Y.-H. Lai, F. Chen, X. Lu, Y. Tsao, S.-S. Wang , & C.-H. Lee, "A Deep Denoising Autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," IEEE Trans on Biomedical Engineering, 2016.

[6] L. Deng, J. Li, Y. Gong, and R. Haeb- Umbach, "An overview of noise robust automat ic speech recognition," IEEE/ACM Trans on Audio, Speech, and Language Processing, 2014.

[7] P. Scalart , "Speech enhancement based on a priori signal to noise estimation," in International Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1996.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtract ion," IEEE Trans on Acoustics, Speech and Signal Processing, 1979.

[9] K. W. Wilson, P. Smaragdis, B. Raj & A. Divakaran, "Speech denoising using nonnegat ive matrix factorizat ion with priors," in Internat ional Conf. on Acoust ics, Speech and Signal Processing (ICASSP), 2008

[10] Szu-Wei Fu, Yu Tsao, Xugang Lu & Hisashi Kawai, “Raw Waveformbased Speech Enhancement by Fully Convolut ional Networks”, P roc. of APSIPA Annual Summit and Conference 2017.

[11] C. Valent ini-Bot inhao, X. Wang, S. akaki and J. Yamagishi, "Speech Enhancement for a Noise-Robust Text -to-Speech Synthesis System using Deep Recurrent Neural Networks", In Proceedings, Interspeech 2016.

[12] Wang, DeLiang, and Jitong Chen. supervised speech separat ion based on deep learning: An overview." IEEE/ACM Trans. on Audio, Speech, and Language Processing (2018).

[13] S. Patibandla, M. Archana, and R. C. Tanguturi, “Object Tracking using Multi Adaptive Feature Extraction Technique,” International Journal of Engineering Trends and Technology, vol. 70, no. 6, pp. 279–286, Jun. 2022, doi: 10.14445/22315381/ijett-v70i6p229.

[14] G. Sadineni, A. M, and R. C. Tanguturi, “Optimized Detector Generation Procedure for Wireless Sensor Networks based Intrusion Detection System,” International Journal of Engineering Trends and Technology, vol. 70, no. 6, pp. 63–72, Jun. 2022, doi: 10.14445/22315381/ijett-v70i6p208.

[15] S. Patibandla, Dr. M. Archana, and Dr. R. C. Tanguturi, “DATA AGGREGATION BASED HYBRID DEEP LEARNING TECHNIQUE FOR IDENTIFYING THE UNCERTAINTIES AND ACCURATE OBJECT DETECTION,” Indian Journal of Computer Science and Engineering, vol. 13, no. 3, pp. 697–708, Jun. 2022, doi: 10.21817/indjcse/2022/v13i3/221303049.

[16] Dr. S. R. Anand, Dr. R. C. Tanguturi, and S. D S, “Blockchain Based Packet Delivery Mechanism for WSN,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2, pp. 1112–1117, Jul. 2019, doi: 10.35940/ijrte.b1627.078219.

[17] M. V. Bharathi, R. C. Tanguturi, C. Jayakumar, and K. Selvamani, “Node capture attack in Wireless Sensor Network: A survey,” 2012 IEEE International Conference on Computational Intelligence and Computing Research, Dec. 2012, doi: 10.1109/iccic.2012.6510237.