Sound Source Separation Using Neural Network

Shreya Sose [1], Swapnil Mali [2] , S.P. Mahajan [3]
[1] College of Engineering, Pune, SPPU University, India
sosesss17.extc@coep.ac.in

***Abstract*— We have given an unsupervised separation of the sound for combination of the unknown different sound in the channel based on the DNN i.e. deep neural network. here it can be separate the different voices with the assumption of the dissimilarity measures of the audio signal. It has been shown that the different sound can be classified and separate. Here uses DNN based algorithm to acquire a mapping of mixed signal to the recovered signal. Also, it gives a separation architecture where it constructs an DNN separation module from which the separation of sound is done. After the training and testing of the sound separation it is shows that the performance of the system is better than other sound separation system and its very useful in segregate the multiple mixed signal. sound separation has varied uses in most real-time applications. In this paper, we use the mask for training targets of DNN for speech separation. The experiments are done with adding different noise conditions. The evaluation of this done using STOI evaluation parameter.**

***Index Terms*— speech separation, DNN, neural network, Ideal ratio mask**

I. INTRODUCTION

Automatically detecting the sound coming from the speaker has an varies application in real world. E.g. automatic speech recognition (ASR) In the audio processing there are many processes are scrutinized from which we can separate the audio signals. The independent component analysis (ICA) is a case of blind source separation. The cocktail party problem is referring to listen the particular voice within the different voices which are simultaneously coming in a crowded place [10] [11]. The mixed signal coming in such place is 'A'. which is combination of different voices. A= x+y. where the x, y are the different interfering voices and the speech of the targeted person.

In the neural network there are two method supervised and unsupervised neural network. The main difference between a supervised and unsupervised network is in the training of the neural network. The supervised neural network has the given dataset with the label for it while train the dataset is provided with label which will easily separate the data further.[13] And also train the network according to it. But it has different scenario in the case of unsupervised network, the dataset for the training not has a label generated the input is provided from which features are extracted which help network to train. Where DNN is mostly used for the unsupervised sound separations.

The unsupervised learning has the different methods from which it makes presumption from dataset containing input data without the response.[9] Computational auditory stream analysis (CASA) is method based on the ability of the humans to detect the sound and acknowledge unique sound sources from the combination of the sound source are introduced as an auditory scene analysis.[10]

This network is attentive for monaural speech separation. As measure monaural speech with multi-microphone solutions, it is not much delicate toward the room reverberation.[2] From the signal processing perspective, different methods have been suggested to determine the ideal wiener filter, which is most appropriate filter to retrieve clean speech in the MMSE (i.e. minimum mean squared error). The statistical based models are alternative to this wiener filter. It gives information about the speech spectral coefficients gives the noisy observation under prior distribution assumptions for speech and noise.

For the training target it is more significant that ideal target could obtain superior separation result. The targets are mostly ideal binary mask and the ideal ratio mask. [1] The intelligibility and quality of speech is better in the ideal ratio mask (IRM). There are different learning machines are found by the literature, they are support vector machine (SVM), Gaussian mixture model (GMM), Deep Neural Networks (DNN). The effective modelling of non-linear relation of speech signal and auditory environment also vigorous structure of speech is enabled because of well-built learning capacity.

II. RELATED WORK

There are different methods generally used for source separation. The Beamforming, ICA i.e. Independent component analysis. In beamforming the sound separation done with suppressing the other sound source which is not required. Extract target sound from specific spatial direction with a sensor array. The multiple audio signals cannot be separate out. Where as ICA does the separation of multiple signal at a time. ICA find a demixing matrix from the source signal. The major advantage of using DNN for source separation is it will separate audio signal from the multiple mixture of source signal with intelligence. The disadvantage of these previous system is overcome with the help of DNN.

III. DNN BASED SPEECH SEPARATION

A. DNN structure

DNN speech separation include the pretraining of unsupervised and fine tuning with the supervised method.as the machine learning algorithm for without label input data used is

unsupervised learning. At the pretraining it handles each successive pair of layers as a restricted Boltzmann machine (RBM). Where, RBM is a stochastic artificial neural network. The RBM network well informed with probability distribution of a collection of input data.

Each circle represents the similar structure of neuron called as node. The two nodes are not connected at same layer. But nodes are linked with different nodes of another layer. Beside the unlabeled unsupervised data, the algorithm gives some supervised statistical intelligence to the network. The main intension of using DNN for speech separation that the separation of the desired data should be done from the mixed data provided to network.

As the structure is shown in figure, the DNN structure has the input layer then hidden layers and at last the output layer. As fig1 shows the structure of DNN network. The extracted feature is provided as an input to the network. The where the output provided by DNN is features of a noise and clean speech. So, these features are extracted from feature extraction method and provide to first layer of DNN. After which the weights are calculated according the functions are applied at the node. The mean square error method is used to calculate the error between the desired and the obtained output. According to which the weights are again adjusted.

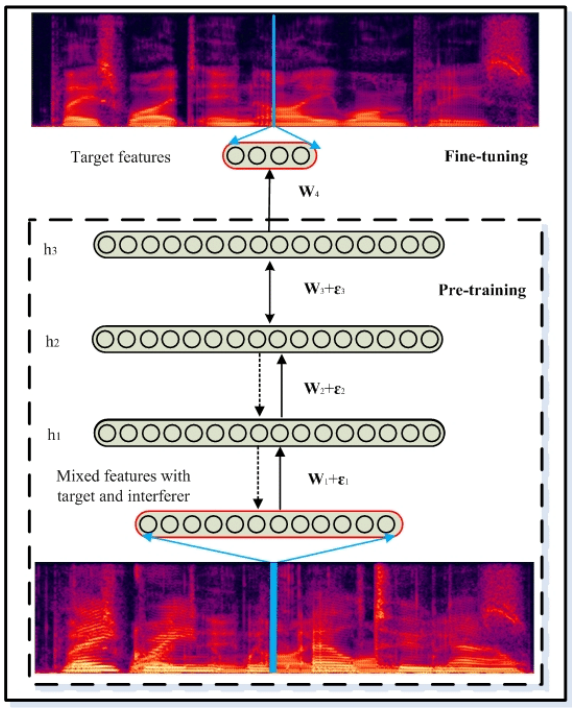

Fig.1. DNN Structure

The flow of the source separation model is as shown in fig.2. the first training data is provided which has the number of sound signal in the .wav form. The next step is a feature extraction of the signal. The important features are extracted from the sound source is crucial part. The DNN module is set according to the configuration. Which has decided the number of hidden layers in network the configurations required for the DNN model as learning rate, activation function, optimization technique, loss function of the model.[2]

Various training targets are their which are most important to get the clear speech separation. Whereas the different masking is available as ideal binary mask (IBM), ideal ratio mask (IRM), phase sensitive mask (PSM).[2] The algorithm evaluates the IRM in the Mel frequency domain with DNN. With the help of the mask the noise from noisy Mel frequency can be eliminated.

As the parameters are decided the training of the separation module is done. For the testing of the model the similar sound sources are provided which are mixed with the noise. Again, the feature extraction is done. Which is compared with the training parameter. The separation module gives the desired audio signal as a separation output. The audio file of the separated audio is reconstructed with the help of the features of a sound source.

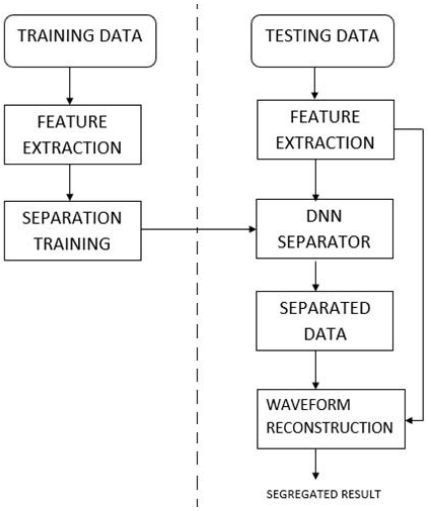

Fig.2. DNN Sound Separation Structure

B. *Feature Extraction*

Feature extraction is an important part where the different characteristics of the spoken utterances are detected. The arithmetic block for this generally include the pre-emphasis, windowing, Fast Fourier transform (FFT)/Discrete Cosine transform (DCT), log. The different set of complementary features are they are in framework. Feature extraction mostly has three stages. The first stage is speech analysis which does the analysis of spectra temporal of the signal. The second stage produce feature vector which combine the dynamic and static features.[3] At the end, extended feature vector into more compress and vigorous vector which are provided to recognizer. The different type of feature extraction methods is Mel frequency cepstrum coefficients (MFCC), Linear predictive code (LPC), Perceptual Linear Prediction (PLP), Relative spectra filtering of log domain coefficients (RASTA).[6][7]

C. *Ideal Ratio Mask*

The ideal ratio mask is defined as follows:

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)}\right)^{\beta} \quad (1)$$

$$= \left(\frac{SNR(t,f)}{SNR(t,f)+1}\right)^{\beta} \qquad (2)$$

Particularly in T-F unit, the speech and noise energy denoted as the $S^2(t,f)$ and $N^2(t,f)$ respectively. b Is the tunable parameter. The wiener filter in time domain and the IRM are closely related. With the different experimental results, it is found that the value of b=0.5 to the best extend. [5] Where it much more similar to the square root wiener filter. Which is nothing but most appropriate determinate of power spectrum. Similar to the other mask IRM also acquire 64-channel gammatone filter bank. The range of IRM is in between 0 to 1.[4]

D. DNN configuration

DNN is made up of four hidden layers. The sigmoid activation function used for the targets in range of [0,1]. The dropout regulation for the back propagation (rate 0.2) is provided in the training of the network. Gradient descent along with momentum term is given as an optimization technique for it. the rate of momentum is 0.5 during the epoch.it used the mean square error as a cost function.

IV. EXPERIMENTAL DETAILS

A. Training Dataset

The utterances are random selected human voices. The 600 audio clips in form of .wav file are used for the training of the network. The noise is the different random audio which are mix with the training audio data. The different combinations of the data are made as the 600 sound are mix with different noise signal. The similarly the 120-audio signal are mix with noise is used as a testing dataset. The training and testing data are in form of audible .wav files. For the different dataset combination, the code is written so that various data can get from basic files which has obtain. The noise signal is approximately 4min long. While creating dataset the training set use random 10 slice which are from the noise signal. The testing utterances at the -2db. At the testing the noise part is used are different to ensure the testing and training data are distinct from each other.

B. Evaluation criteria

There are many different criteria for the evaluation of the audio network. The signal to artifact ratio (SAR), signal to interference ratio (SIR), short time objective intelligibility (STOI) are the main major for performance evaluation.[2] Here we had use the STOI for estimation of objective intelligibility. The represent an association of short-term temporal case into separated and clean speech. It has been validating that STOI is majorly associate with the intelligibility score of human speech. The range for STOI is in between [0,1]. [4] The objective speech quality also measured with perceptual evaluation of speech quality (PESQ) score.[5]

C. Experimental Results

As the output of a separation module is to get recovered the original signal from the mixed signal. The waveform of a mixed signal, original signal and the output i.e. separated audio is generated. The waveforms below show the results generated.

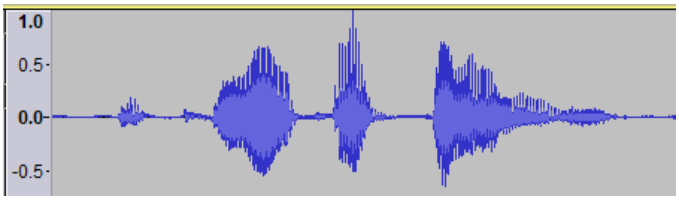

Fig.3. Original audio signal

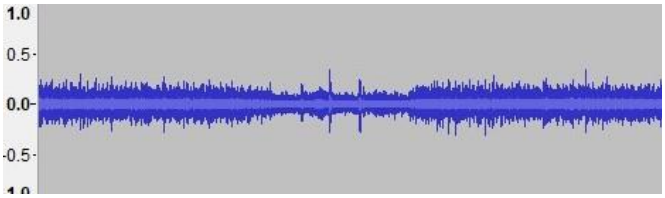

Fig.4. factory noise signal

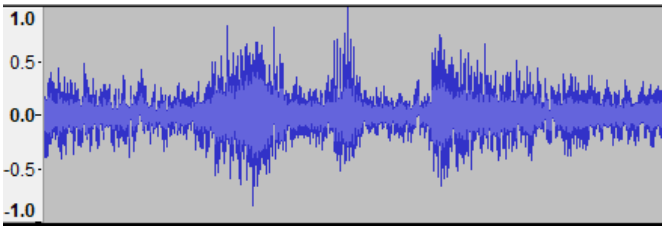

Fig.5. Mixed audio signal

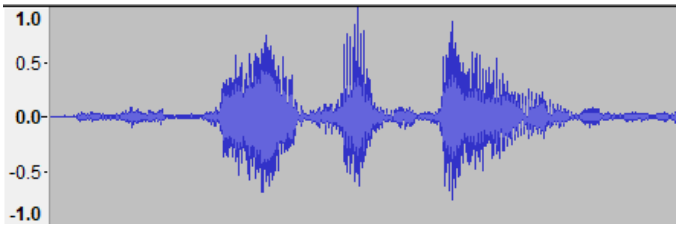

Fig.6. Recovered signal/ separated audio

The separation results of the testing data are evaluated using STOI is measured as follow,

The ideal value for the specified signal is given in column similarly the after the separation the recovered signal i.e. estimated value of the separated signal is calculated are given below. After the testing of an all testing samples the average value of an STOI is calculated.

TABLE I
CALCULATED STOI PARAMETER FOR DIFFERENT DATSET

Targets	Test 1	Test2	Average
Unprocessed STOI	0.6349	0.5447	0.6235
Ideal STOI	0.9023	0.8595	0.9044
Estimated STOI	0.7543	0.6667	0.7577

TABLE II
CALCULATED PERFORMANCE PARAMETER OF THE AUDIO SIGNAL

	Recovered signal	Estimated value	Average value signal
SAR	Signal 1	5.41	5.71
	Signal 2	5.19	
	Signal 3	6.24	
	Signal 4	5.61	
	Signal 5	6.13	
SDR	Signal 1	5.37	5.72
	Signal 2	5.16	
	Signal 3	6.28	
	Signal 4	6.37	
	Signal 5	5.42	
SIR	Signal 1	26.80	26.38
	Signal 2	27.71	
	Signal 3	25.34	
	Signal 4	26.42	
	Signal 5	25.64	
STOI	Signal 1	0. 5629	0.5975
	Signal 2	0.6264	
	Signal 3	0.5528	
	Signal 4	0.6524	
	Signal 5	0.5930	

Fig.7. performance parameter

The images below show the spectrogram of a signal which is provide for the testing with the mixing of noise. And the result of a sound separation. Which shows the maximum similarities of the recovered signal.

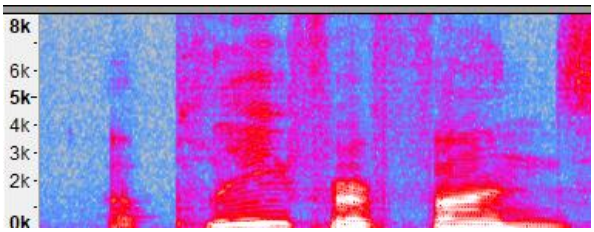

Fig.8. The spectrogram of the original signal

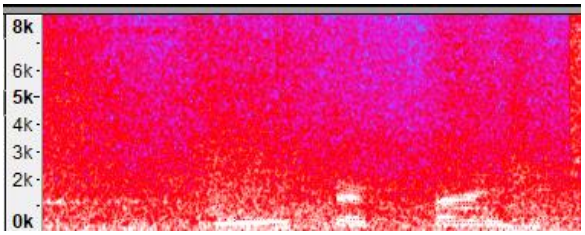

Fig.9. The spectrogram of the mixed signal

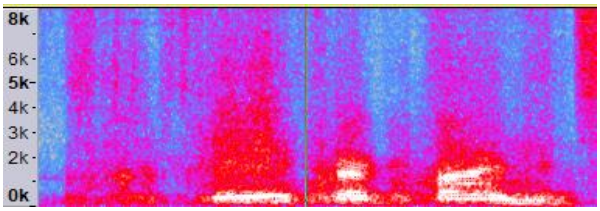

Fig.10. The spectrogram of the separated signal

V. CONCLUSION

In the area of speech recognition, a monaural speech separation is much competitive task.th source separation is merge with DNN which appear as a recent trend. The magnificent enhancement is attaining low SNR and mobile noise conditions. The most approved training targets are IRM and IBM. Where IRM increased the speech intelligibility along with improved quality of speech. The feature MFCC, RASTA plays an important role in the extraction of important feature of audio signal. The DNN module used which has multiple hidden layer which is deep network. Speech is separated using DNN. Where STOI shows the proper evaluation about separated signal from model. The proposed network will show better result for both supervised and semi-supervised network. The further, improvement can be including the multichannel speech separation. Where the voice from noise is separated. Whereas each individual voice also can be separate out using DNN.

RFERENCES

[1] Hu, Ke, and DeLiang Wang. "An unsupervised approach to cochannel speech separation." *IEEE Transactions on audio, speech, and language processing* 21.1 (2012): 122-131.

[2] Huang, Po-Sen, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. "Deep learning for monaural speech separation." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1562-1566. IEEE, 2014.

[3] Chen, Jitong, Yuxuan Wang, and DeLiang Wang. "A feature study for classification-based speech separation at low signal-to-noise ratios." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, no. 12 (2014): 1993-2002.

[4] Xia, S., Li, H. and Zhang, X., 2017, December." Using optimal ratio mask as training target for supervised speech separation." In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*(pp. 163-166). IEEE.

[5] Wang, Yuxuan, Arun Narayanan, and DeLiang Wang. "On training targets for supervised speech separation." *IEEE/ACM transactions on audio, speech, and language processing* 22, no. 12 (2014): 1849-1858.

[6] Dave N. "Feature extraction methods LPC, PLP and MFCC in speech recognition". International journal for advance research in engineering and technology. 2013 Jul;1(6):1-4.

[7] Prithvi, P., and T. Kishore Kumar. "Comparative Analysis of MFCC, LFCC, RASTA-PLP." *International Journal of Scientific Engineering and Research* 4, no. 5 (2016): 1-4.

[8] Williamson, Donald S., Yuxuan Wang, and DeLiang Wang. "Complex ratio masking for monaural speech separation." *IEEE/ACM Transactions*

on Audio, Speech and Language Processing (TASLP) 24, no. 3 (2016): 483-492.

[9] Liu, Yuzhou, and DeLiang Wang. "Speaker-dependent multipitch tracking using deep neural networks." *The Journal of the Acoustical Society of America* 141, no. 2 (2017): 710-721.

[10] Wang, Yannan, Jun Du, Li-Rong Dai, and Chin-Hui Lee. "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25, no. 7 (2017): 1535-1546.

[11] Noda, Kuniaki, Naoya Hashimoto, Kazuhiro Nakadai, and Tetsuya Ogata. "Sound source separation for robot audition using deep learning." In Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference on, pp. 389-394. IEEE, 2015.

[12] Luo, Yi, Zhuo Chen, and Nima Mesgarani. "Speaker-independent speech separation with deep attractor network." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, no. 4 (2018): 787-796.

[13] Hu, Ke, and DeLiang Wang. "An unsupervised approach to cochannel speech separation." IEEE Transactions on audio, speech, and language processing 21, no. 1 (2013): 122-131.