Gender Classification Based on Speech Analysis using DNN

Ajab Godhrawala[1] , Namita Bhor[2], Dheeraj Patil [3], Pallavi Sawant[4] ,Kokila Kasture[5],
[1,2,3] *Student,*[4,5] *Faculty, Department of Electronics and Tele-communicatio00206tszz gtgvfvb sdvgen Engineering, Smt.Kashibai Navale College of Engineeing,Vadgaon,Pune*
[1]*ajabgodharawala67@gmail.com* [2]*namitabhor1998@gmail.com* [3]*patildheeraj259@gmail.com*
[4]*pallavi.msawant_skncoe@sinhgad.edu* [5]*krkasture.skncoe@sinhgad.edu*

Abstract— *Language is one of the significant methods for correspondence; Speech is its fundamental medium by which two individuals can communicate. In this current period, technology based on speech are broadly utilized as it has limitless uses. Speech Recognition (SR), as the man-machine interface assumes a fundamental part in the field of AI and ML where precision is a significant challenge. In this proposed system, a model has been processed where to separate two voices by using DNN techniques. Conversation is made on various methods and approaches of the discourse acknowledgment measure utilizing the DNN. DNN can effectively extract low-dimensional features. Hence the accuracy of speech separation can be improved by the designed system. The main motive is to bring to light the progress made in the field of speech recognition.*

Keywords—*Deep Neural Network, Speech, Speech Separation, Speech Recognition*

I. INTRODUCTION

A reflexive multilayered mission organizer with an unaided discourse detachment system for blends of two inconspicuous speakers in a single channel setting dependent on deep neural network systems (DNNs). It depends on a key suspicion that two speakers could be isolated on the off chance that they are not like one another. A difference measure between two speakers is first designed to describe the partition capacity between contending speakers. At that point appear that speakers with the equivalent or various sexes can regularly be isolated. If two speaker bunches, with huge enough separations between them, for every sexual orientation gathering could be set up, bringing about four speaker groups. Next, a DNN-based sexual orientation blend identification calculation is designed to decide if the two speakers in the blend are females, guys or from various sexual orientations. This locator depends on a recently designed DNN design with four yields, two of them speaking to the female speaker groups and the other two describing the male gatherings. At long last the propose is to develop three autonomous discourse division DNN frameworks, one for every one of the female-female, male-male furthermore, female-male blend circumstances. Each DNN gives double yields, one speaking to the objective speaker gathering and the other portraying the meddling speaker group. Prepared and tried on the Speech Separation Challenge corpus, our trial results demonstrate that the designed DNN-based methodology accomplishes enormous execution increases over the best-in-class solo procedures without utilizing a particular information about the blended target and meddling speakers being isolated. Single-channel source detachment means to recoup at least one source sign of enthusiasm from a blend of sign.

II.LITERATURE SURVEY

There are various techniques for Speech Recognition and Separation based on utterance, vocabulary size and specker mode. Speech has evolved as a primary form of communication between humans. The advent of digital technology, gave us highly versatile digital processors with high speed, low cost and high power which enable researchers to transform the analog speech signals in to digital speech signals that can be scientifically studied. The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. Federico Cruciani el. gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. It helps in choosing the technique along with their relative merits & demerits[2] .Achieving higher recognition accuracy, low word error rate and addressing the issues of sources of variability are the major considerations for developing an efficient Automatic Speech Recognition system. In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase [3].The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, and data base and performance evaluation[1]. To overcome above issue, we use Deep learning concept in Automatic Speech Recongnition. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. There are different types of neural networks but they always consist of the same components: neurons, synapses, weights, biases, and functions. These components functioning similar to the human brains and can be trained like any other ML algorithm. DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network. For instance, it was proved that sparse multivariate polynomials are exponentially easier to approximate with DNNs than with shallow networks. Deep architectures include many variants of a few basic approaches. Each architecture has found success in specific domains. It is not always possible to compare the performance of multiple architectures, unless they have been evaluated on the same data set.The main objective of this paper is the comparing speech recognition accuracy of a target speech signal that was extracted from a mixture of two speakers and determine whether the two speakers in the mixture are females, males or from different genders.

III. MPLEMENTATION DETAILS OF MODULE

The basic system consists of 4 stages as shown in fig.1. such as Feature Extraction, Training, DNN, Output Speech Separations.

Gender Mixture Detection- To show the importance of the gender mixture detector and the effectiveness of the DNN-based approach, First introduce a method widely used in the speaker recognition community as a comparison in experiments. The alternative speaker representation and a form of Bayesian adaptation to derive the speaker models, two models representing male speakers and female speakers are trained and then used to determine the gender identities of mixed speech.

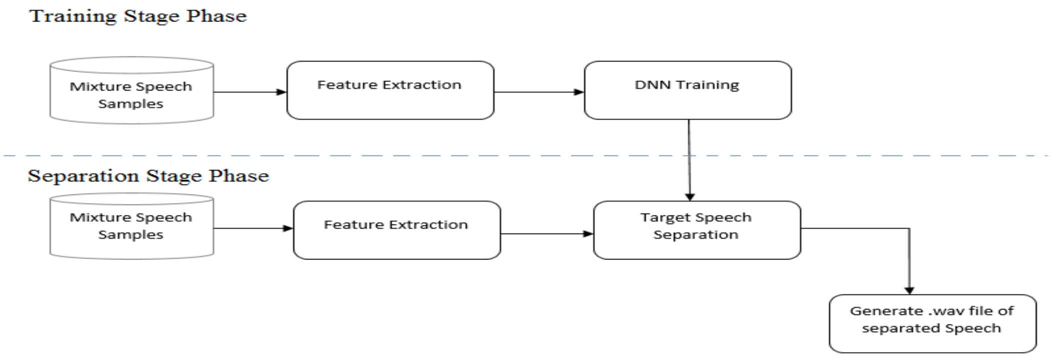
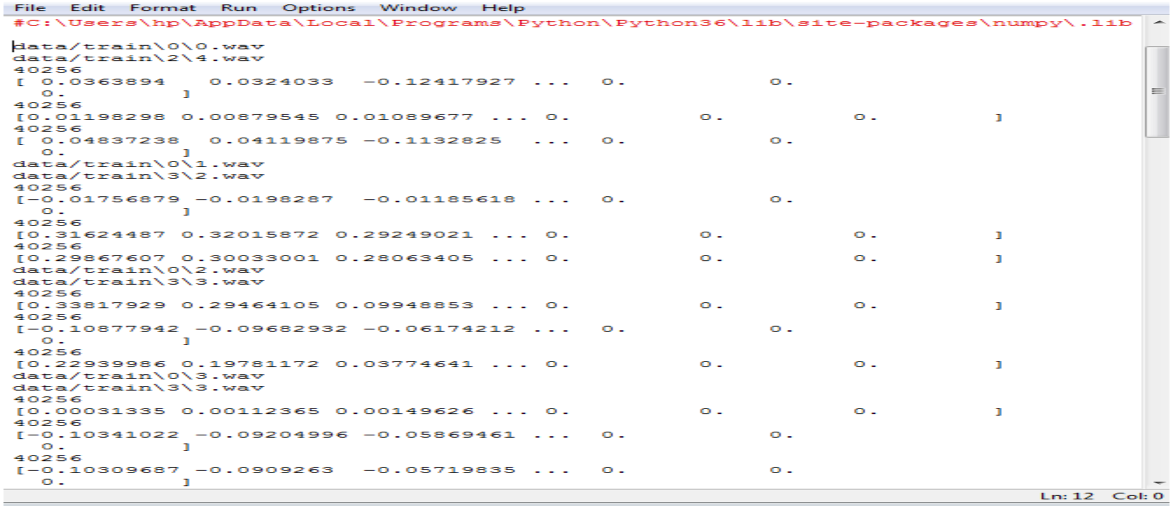


Fig.1 System Architecture

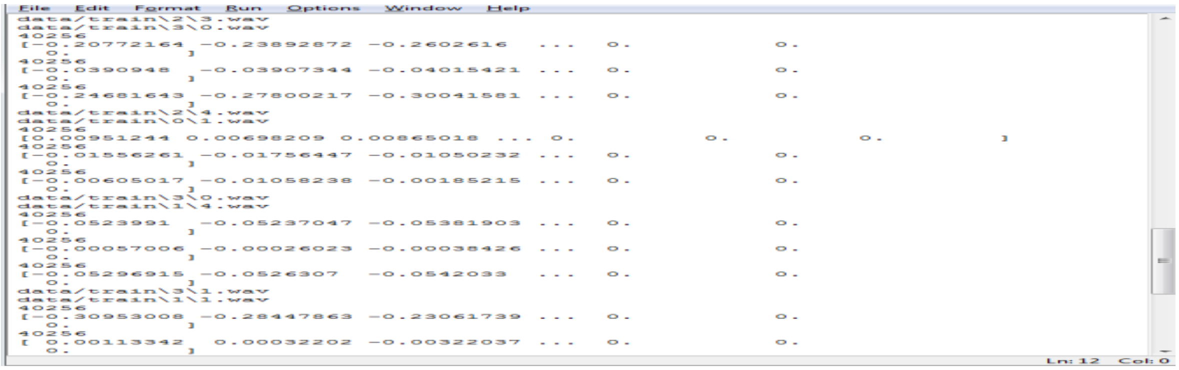
Speech Separation- Speech separation or segregation is the separation of a desired speech signal from a mix of environmental signals. These can include ambient room noise, other talkers and any other non-stationary noise. The majority of speech separation techniques try to reduce noise by replicating the signal processing performed naturally by human auditory sensory system. Speech segregation can be separated into two categories. The first is monaural approaches, which includes speech enhancement techniques and computational auditory scene analysis.



```
#C:\Users\hp\AppData\Local\Programs\Python\Python36\lib\site-packages\numpy\.lib

data/train\0\0.wav
data/train\2\4.wav
40256
[ 0.0363894   0.0324033  -0.12417927 ...  0.          0.
  0.        ]
40256
[0.01198298 0.00879545 0.01089677 ... 0.          0.          0.          ]
40256
[ 0.04837238  0.04119875 -0.1132825  ...  0.          0.
  0.        ]
data/train\0\1.wav
data/train\3\2.wav
40256
[-0.01756879 -0.0198287  -0.01185618 ...  0.          0.
  0.        ]
40256
[0.31624487 0.32015872 0.29249021 ... 0.          0.          0.          ]
40256
[0.29867607 0.30033001 0.28063405 ... 0.          0.          0.          ]
data/train\0\2.wav
data/train\3\3.wav
40256
[0.33817929 0.29464105 0.09948853 ... 0.          0.          0.          ]
40256
[-0.10877942 -0.09682932 -0.06174212 ...  0.          0.
  0.        ]
40256
[0.22939986 0.19781172 0.03774641 ... 0.          0.          0.          ]
data/train\0\3.wav
data/train\3\3.wav
40256
[0.00031335 0.00112365 0.00149626 ... 0.          0.          0.          ]
40256
[-0.10341022 -0.09204996 -0.05869461 ...  0.          0.
  0.        ]
40256
[-0.10309687 -0.0909263  -0.05719835 ...  0.          0.
  0.        ]
```

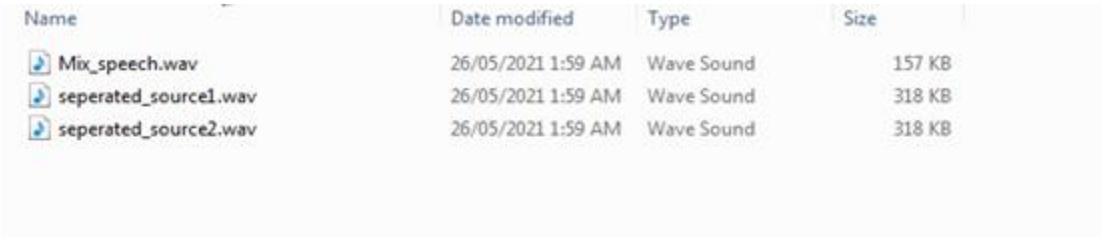
Fig.2 Training and Testing



```
data/train\2\3.wav
data/train\3\0.wav
40256
[-0.20772164 -0.23892872 -0.2602616  ...  0.          0.
  0.        ]
40256
[-0.0390948  -0.03907344 -0.04015421 ...  0.          0.
  0.        ]
40256
[-0.24681643 -0.27800217 -0.30041581 ...  0.          0.
  0.        ]
data/train\2\4.wav
data/train\0\1.wav
40256
[0.00951244 0.00698209 0.00865018 ... 0.          0.          0.          ]
40256
[-0.01556261 -0.01756447 -0.01050232 ...  0.          0.
  0.        ]
40256
[-0.00605017 -0.01058238 -0.00185215 ...  0.          0.
  0.        ]
data/train\3\0.wav
data/train\1\4.wav
40256
[-0.0523991  -0.05237047 -0.05381903 ...  0.          0.
  0.        ]
40256
[-0.00057006 -0.00026023 -0.00038426 ...  0.          0.
  0.        ]
40256
[-0.05296915 -0.0526307  -0.0542033  ...  0.          0.
  0.        ]
data/train\3\1.wav
data/train\1\1.wav
40256
[-0.30953008 -0.28447863 -0.23061739 ...  0.          0.
  0.        ]
40256
[ 0.00113342  0.00032202 -0.00322037 ...  0.          0.
  0.        ]
```

Fig. 3 Training and Testing

The figure 2 and 3 is the output of training and testing. Various voice were trained and calculated their train loss and vali_loss To know the loss of each sample we have used print() to print the status of train loss and vali_loss. The above fig indicates the loss made during validating and testing when sample like 0.wav , 1.wav, 2.wav,3.wav,4.wav were passed for training testing phase.



Name	Date modified	Type	Size
Mix_speech.wav	26/05/2021 1:59 AM	Wave Sound	157 KB
seperated_source1.wav	26/05/2021 1:59 AM	Wave Sound	318 KB
seperated_source2.wav	26/05/2021 1:59 AM	Wave Sound	318 KB

Fig.4 output

The above results i.e. Figure 4 states that the sample of mixed voice i.e. Mix_speech.wav was been passed to the system designed. The proposed DNN algorithm extracts the features and compare with trained dataset. The frames were separated, clustering was performed, if some frames were not needed to be supported those were kept as it is. Rearrange of frames with the sequence was performed. Reconstruct of waveform was done and later on the audio was separated into two different voice. i.e. Seperated_source1.wav and Seperated_source2.wav

IV CONCLUSION

A novel DNN-based gender mixture recognition also, discourse partition system for solo single channel discourse partition inspired by the investigation of the speaker dissimilarities. An extensive arrangement of trials also, examinations, including the significance of DNN-based finder also, the correlations among various blend mixes, are led. The designed DNN structure could reliably beat the cutting-edge approach in wording of various target measures. This investigation is an effective show of applying the profound learning innovation to unaided discourse detachment in a solitary channel setting which is as yet a difficult open issue. Later on, target refining the designed system by structuring better speaker gathering calculations and improving the exhibition of both locator and separators. Besides, intend to further build up our framework on bigger datasets and even some other dialects. The other neural system structures are likewise going to be investigated later on, for example, intermittent neural system for our framework. Another intriguing course is to consolidate the uniqueness measure with cost-capacities for DNN-based finder and separator.

REFERENCES

[1] Prarthana T V, Dr. B G Prasad, “Human Activity Recognition using Computer Vision based Approach – Various Techniques”, International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 06 | June 2020

[2] Federico Cruciani, Ian Cleland, Paul McCullagh, Konstantinos Votis, “Feature learning for Human Activity Recognition using Convolutional Neural Networks ”, Springer, 2020

[3] Antonio Bevilacqua li EL Moussati and Omar Moussaoui , “Human Activity Recognition with Convolutional Neural Networks”, Human Activity Recognition with Convolutional Neural Networks,

September 2018

[4] Naifan Zhuang, Jun Ye and Kien A. Hua," Group Activity Recognition with Differential Recurrent Convolutional Neural Networks",12th IEEE International Conference on Automatic Face and Gesture Recognition,2017

[5] Seyed Ali Rokni and Maryjane Nourollahi, "Personalized Human Activity Recognition Using Convolutional Neural Networks" ,Proceedings of AAAI Conference on AI, 2018

[6] Sina Mokhtrazadeh and Mina Ghadimi Atigh, "A Multi-Stream Convolutional Neural Network Framework for Group Activity Recognition", Computer Vision and Pattern Recognition , Cornell University , 2018

[7] Surya Dhanraj and Dinesh Dash, "Efficient Smartphone-Based Human Activity Recognition Using CNN", International Conference on Information Technology (ICIT), 2019

[8] Wen Qi and Hang Su , "A Fast and Robust Deep Convolutional Neural Networks for Complex Human Activity Recognition Using Smartphone",2019

[9] Ming Zeng and Bo Yu," Convolutional Neural Networks for human activity recognition using mobile sensors", IEEE

[10] Jiahui Huang and Ning Wang, "TSE-CNN: A Two-Stage End-to-End CNN for Human Activity Recognition", IEEE