IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.10, October 2020 E-ISSN: 2347-2693

Speaker Recognition System Using Deep Learning with Convolutional Neural Network

Sandeep Kumar[1*], Samridhi Dev[2]

[1,2]School of Information and Technology C-DAC Noida

**Corresponding Author: sk136398@gmail.com,*

DOI: https://doi.org/10.26438/ijcse/v8i10.6064 | Available online at: www.ijcseonline.org

Received: 01/Oct/2020, Accepted: 08/Oct/2020, Published: 31/Oct/2020

Abstract- The task of identifying humans by their voice seems to be an easy task for human beings as people interact with a particular person, their mind is upskilled with that voice and the brain becomes proficient enough to easily recognize that particular voice next time. Using this human mind concept, the structure is designed and implemented. In the proposed system Convolutional Neural Network (CNN) has been used. 110 voice samples from 11 different participants/speakers have been collected. These voice signals were converted into the form of an image of the signal spectrogram. 90% of data were used for training and the remaining 10% was used for testing. Implementation was done in RStudio with R programming language. The system achieved 82% accuracy. The proposed system is facile and lucrative.

Keywords- Convolutional neural network, speaker recognition, Keras, voice signal spectrogram, tuneR.

I. INTRODUCTION

Recognition of speakers is the practice of identifying who is speaking spontaneously based on specific knowledge found in speech waves. Diagnosis of the speaker's voice can be utilized in authentication management in the fields of sensitive information and Providing managed access to facilities such as voice dialing, database access services, information services, remote machine access, and a range of other fields of importance for defence.

Speech is a dynamic signal primarily the result of several changes occurring at various levels: semantic, linguistic, articulatory, and acoustic. The variations in the acoustic properties of the speech signal are expressed in the variations in these transformations. All these differences are factored into the equation for speaker identification and are used to differentiate among speakers. The forthcoming information elucidates the procedure for developing a superficial, skill full, and self-executing speaker recognition system. The system designed has the potential of being utilized in several security applications which includes, obtaining access to the laboratory they work in.

The area of speech science is broken down into two parts: one is speech recognition, and the second is the recognition of speakers. what a person is speaking has to be determined in speech recognition. So, the task of speech recognition is to distinguish specific words from the speech of the speaker. And

In speaker recognition, the speaker has to be identified that means who is speaking. So, speaker recognition's task is to identify the speaker.

Speaker identification is the maneuver of identifying who speaks and confirming speakers automatically based on information derived from their speech signals. Speaker recognition is divided into two areas [1]:
A. Speaker Identification
B. Speaker verification

The number of decision alternatives is the basic distinction between recognition and verification. In recognition, the number is proportional to the population size, while there are only two options, approval or denial, irrespective of the population size, in verification.

Furthermore, as population size expands, recognition efficiency drops, while authentication efficiency reaches a constant regardless of the population size unless the distribution of physical characteristics of speakers is highly skewed.

A. speaker Identification:

An utterance from an anonymous speaker is studied and correlated with speech models of recognized speakers in speaker recognition. As the one whose model better suits the input utterance, the unknown speaker is named.
In the identification, the number of decision alternatives is equal to the size of the population.

B. Speaker Verification:

An identity is asserted by an anonymous speaker in speaker verification, whose utterance is compared to a model for the registered speaker (customer) whose identity is being asserted. The argument is acknowledged when the match is good enough, that is, above a threshold.
A high threshold makes it strenuous for the system to consider multiple speakers, but it carries the risk of wrongly refusing members [2].

© 2020, IJCSE All Rights Reserved

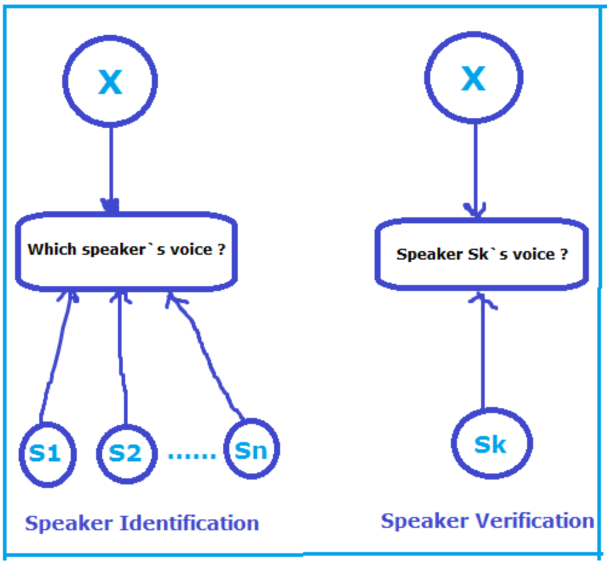

Figure 1. Classification of Speaker Identification and Verification process.

II. RELATED WORK

A. Patil, et. al [3] have developed a database for speaker recognition for automatic recognition of speakers via telephone. They reported that the results of the automatic speaker recognition system rely on the correctness of the database used and the database should be collected with proper recording conditions. They have discussed the methodology and experimental setup used for the development of the speech database for Indian languages. The languages considered by them were Hindi, Marathi, Oriya, and Urdu.

B. Waghmare, et. al [4] compared the techniques available to create the emotional speech databases and reported that the superiority of the database depends on the intensity of emotions. As many as seven databases were studied, in which some databases used the professional actors to speak and record the emotions, while the other used people without acting background. As per their study, the method involved in designing the Serbian database is commendable with an accuracy of 95%. Finally, they have created a new database for the Marathi language by recording the speech emotions from the movies. The emotions considered were Anger, Happy, Sad, Neutral, and Afraid.

Ibon Saratxaga et.al [5] have developed an emotional speech database of Basque. They have developed a database that can serve two purposes. The first purpose is to serve a synthesis and the other is the study of prosodic features of emotions. The designed database is very large consisting of the same sentence recorded in six good quality different emotions for synthesis. The speech emotions were recorded from one male artist and one female dubbing artist. They have explored the different stages involved in the development of the database.

Kim. H, et al [6] have developed a speech database of dysarthric language. The total number of speakers in the database was 19. All the speakers considered in the study had a movement disorder called cerebral palsy problem. The number of words per speaker was 765. The speech was recorded using an array of 8 microphones. This database provided a means for speech recognition of speakers having a neural inability. This research database would benefit clinical treatments. The database is kept open for the researchers and can be used freely upon a request.

Pooja V. Janse and Ratnadeep R. Deshmukh [7] have developed a Marathi database with speech information, which was gathered from 100 speakers. Every speaker speaks 124 words with 3 articulations and hence 372 expressions of words were recorded from each speaker summing up to 37200 articulations of words. The database was recorded using high-quality microphones like Sennheiser PC 350 by using PRAAT speech software. Even though the speech data were collected with noise, it is still a robust speech system was developed.

P.J. Castellano, et. al [8] have developed multiple binary classifier model and Gaussian mixture model (GMM) solutions and evaluate it for both automatic speaker verification and automatic speaker identification difficulties containing text-independent telephone-speech from the King speech datasets.

III. METHODOLOGY

The proposed method requires a maneuver. The method suggested is facile and encompasses lucid
operations described below.

Steps:
A. Data Collection
B. Noise Reduction
C. Spectral Image of Signal
D. CNN Model

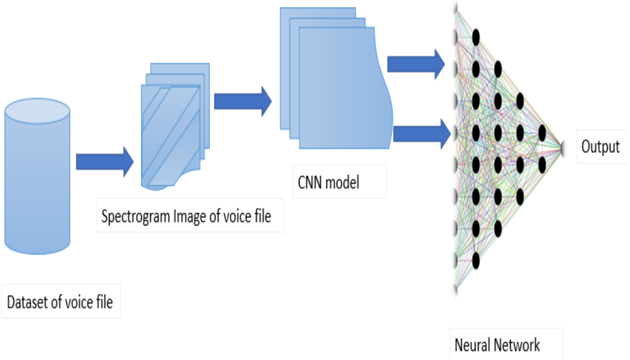

Figure 2. The proposed methodology of Speaker Recognition System

A. Data collection
A Voice sample from different participants was collected. Total 200 voice sample was collected from 20 participants.

10 voice samples were taken from each participant. 90 percent of the dataset was utilized for training and the rest 10 percent of the dataset was used in testing.

B. Noise Reduction

Initially, the voice signal has to undergo pre-processing steps. It is a procedure of eliminating background noise that interferes with signal processing and identification. Random sounds weaken voice signals [9]. The voice signals are difficult to process with the background noises. eliminating overall background noise will enhance the consistency and intelligibility of the signals, and therefore it is convenient to process.

Yeldener and Rieser reported that the voice coding system with typically low bit rate does not have its mechanism to eliminate voice signal background noises. That is due to the complexity and uncertainty of the speech signal in the context of speech coding systems. It is therefore vital to have a mechanism of eliminating background noises. In MATLAB (SIMULINK) software, the elimination of background noises or any unwanted signal can be achieved by passing it through the Digital Filter Architecture block [10].

C. Spectral Image of Signal

After noise reduction from voice signal data samples. It is necessary to collect all voice samples images. By using the R package, we collect all spectral image of each voice sample and store in a particular folder with the same size and dimensions. So that the convolutional neural network performs well on those images.

D. Convolutional neural network (CNN) model

The model deals with input images and produces patterns and decreases the dimensionality of image features so that neural networks perform well with fewer image features and by using coevolutionary features of deep neural networks, it also reduces the complexity of the neural network.

The input images can move through a series of convolution layers with filter (Kernals), pooling, fully connected layers (FC), and apply Softmax to classify an object with probabilistic values between 0 and 1. Theoretically, Convolutional Neural Network (CNN) is utilized to train and evaluate. The figure below is a full CNN stream for processing an input image and classifying the objects based on values.

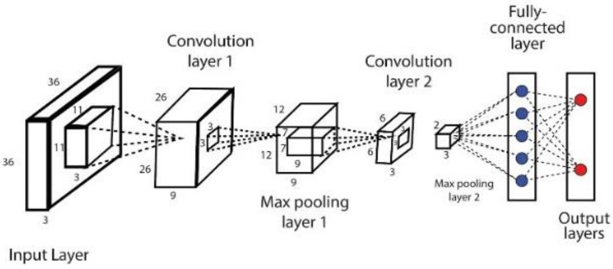

Figure 3. The architecture of Convolutional Neural Network

CNN steps:

- Provide input images to the convolution layer of the model.
- Select parameters, apply stride filter, and padding if required. Perform the image convolution and add the activation function ReLU (Rectified Linear Unit) to the matrix.
- Perform pooling to reduce the dimensional size
- More convolutional layers can be added until a satisfying result is obtained.
- Flatten the output and feed into a fully connected layer.
- Use an activation function to classify images.

IV. IMPLEMENTATION

It is not simple for the computer system to identify the speaker from its voice sample. By this model's strategy, initially, voice samples need to be collected in .mp3/.wav file format using WO Mic version 4.6.2 followed by a retrieval spectrograph of the Speech signal from the audio file. ***turnR*** and ***seewave*** library of R programming language succor to extract the frequency spectrum and spectrogram image of voice signal. A convolutional neural network works on stored images and gets the features from the image.

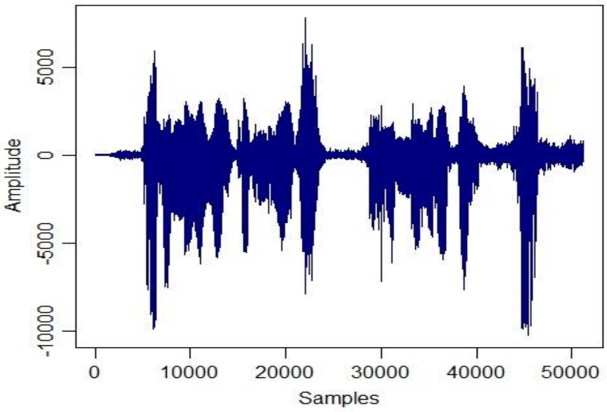

Figure 4. Frequency signal of voice sample

The above figure shows a particular voice in form of a frequency signal of an audio file in samples and amplitude. Which is extracted from the voice file in Rstudio with help of *seewave* and *tuneR* libraries.

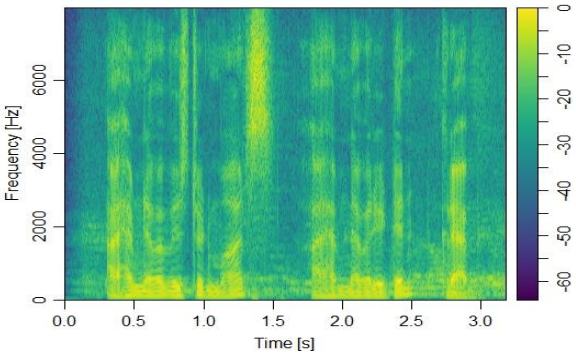

Figure 5. Spectrogram Image of voice sample

The above figure shows a spectrogram of a particular input voice file in the frequency and time domain. Which is extracted from the voice file in Rstudio with help of *signal* and *oce* libraries.

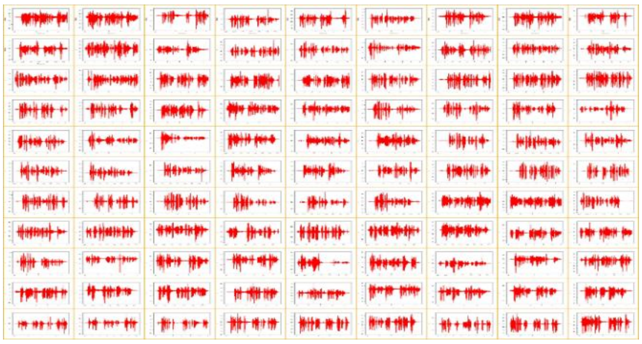

Figure 6. Combined training image set

The above figure shows all the training images each with the dimension of 99x100x100x3 (99 is the no. of train images and 3 is RGB). Using the *combine()* function.

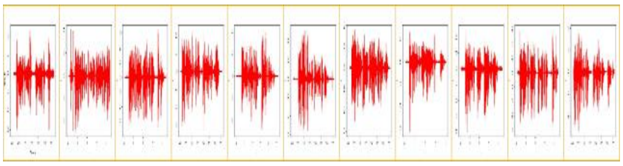

Figure 7. Combined testing image set

The above figure shows all the testing images each with the dimension of 99x100x100 (99 is the no. of train images). Using the *combine()* function.

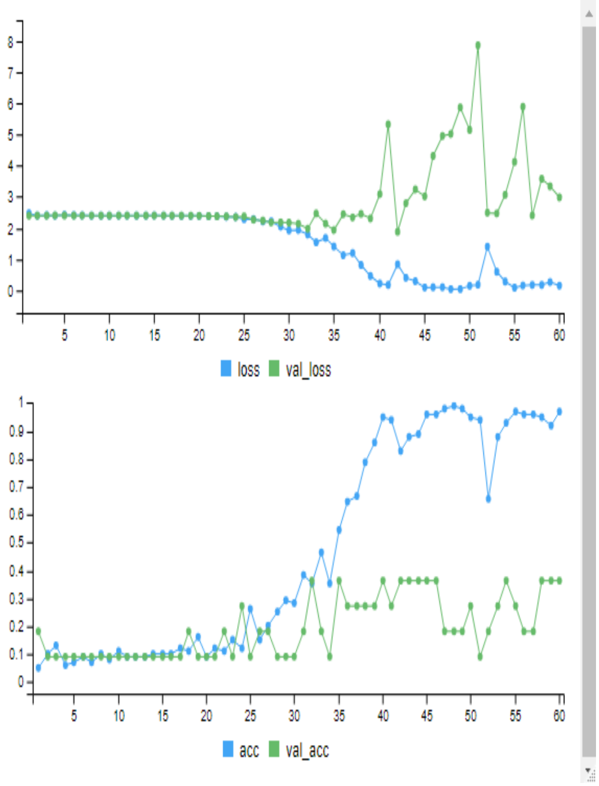

Figure 8. Process of model fitting

The above figure shows the accuracy and loss of the CNN trained model. Using the fit function of Keras library upto 60 epochs.

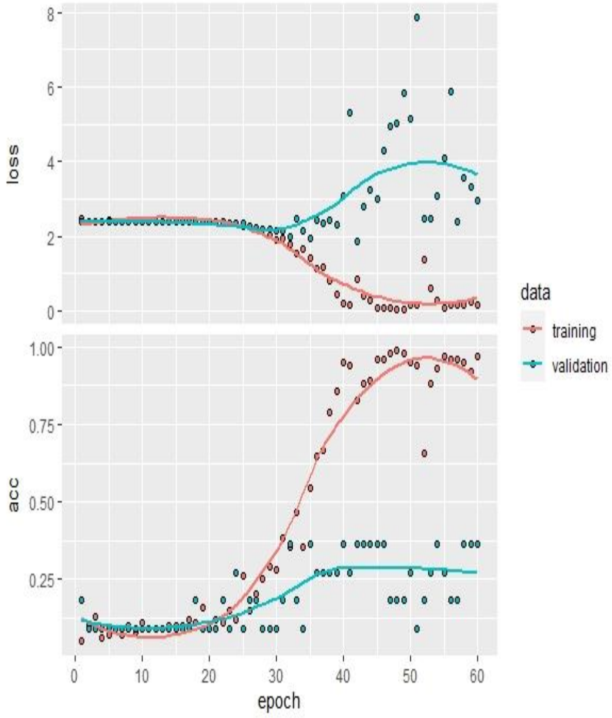

Figure 9. A plot of evaluation in the form of accuracy and loss.

The above figure shows the accuracy and loss of CNN trained model with 0.2 validation_split. Using the evaluation function of Keras library.

V. RESULT

The model evaluates the accuracy and the loss of given training and testing sets. With the help of the Softmax activation function, it gives the probabilistic value between 0 and 1.

the proposed system was tested under meticulous conditions. The system managed to Give 82% accurate results. A tailored dataset has been created which encompasses 110 samples of voice. Out of 99 samples, 81 samples were recognized correctly with 0.1798 loss. CNN proved to be an efficient algorithm for speaker recognition. The overall time taken by the system to complete its task is 84 seconds. The proposed system can be used to develop equipment for IoT devices and for those who are physically handicapped (hearing impaired). Table 1 shows metrics for analyzing the efficiency of the system.

Table 1. Result Metrics

S.N.	Actions	Value
1.	Accuracy	0.82
2.	Precision	0.86
3.	Recall	0.84
4.	Loss	0.18

VI. CONCLUSION AND FUTURE SCOPE

Based on the Convolutional Neural network, the proposed model has given 82% accurate result with a loss of 18%. To reduce the loss and to increase the accuracy, the sample dataset needs to be increased.

A more powerful computer process like GPU instead of a CPU is needed for training and testing of a large amount of dataset so that Convolutional Neural Network will perform well and give the desired result. ReLU and Softmax activation function play an important role in the deep learning model. The proposed system is facile and performs real-time operations. The system tends to have a manageable time complexity. The proposed system can be used for further research to provide a solution for smart IoT devices as well as a physically challenged person. For speaker recognition, CNN proved to be an efficacious choice.

REFERENCES

[1]. Rajsekhar G., "Real-Time Speaker Recognition using MFCC and VQ", Ph.D. Thesis, Department of Electronics & Communication Engineering, National Institute of Technology Rourkela, **pp. 9-71**, **2008**.

[2]. S. Furui, "An Overview of Speaker Recognition Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, **pp. 1-9, April 1994.**

[3]. Hemant A. and T. K. Basu, "Advances in Speaker Recognition: A Feature-Based Approach," Int. Conf. Artificial Intelligence and Pattern Recognition, AIPR'07, Orlando, Florida, USA, July 9-12, **pp. 528-537, 2007.**

[4]. Waghmare, et. al., "Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques" **2014**.

[5]. P. L. De Leon, et. al., "Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. **20**, no. **8, pp**. **2280-2290**, **Oct. 2012**.

[6]. Kim, et. al., "Dysarthric speech database for universal access research". Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. **1741-1744**. **2008**.

[7]. Shrishirmal, et. al., "Development of Marathi Language Speech Database from Marathwada Region" **2015**.

[8]. P. J. Castellano, et. al., "Telephone-based speaker recognition using multiple binary classifiers and Gaussian mixture models," IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, 1997, **pp. 1075-1078 vol.2. 1997**.

[9]. G. Doddington, "Speaker Recognition – Identifying People by their Voice", Proceedings of IEEE, **vol.73, 1651-1664, Nov. 1985.**

[10]. Yeldener, S. & Rieser, J.H., "A background noise reduction technique based on sinusoidal speech coding systems. Acoustics, Speech, and Signal Processing", International Conference on. 3. 1391 - 1394 vol.3. 10.1109/ICASSP.2000.861840.

[11]. Ch. Srinivasa Kumar, P. M. Rao., "Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm", International Journal of Computer Sciences and Engineering, **Vol. 3, No. 8, pp.2942-2954, 2011**.

[12]. Parmar Dharmistha R, "a survey on speaker recognition with various feature extraction techniques, "International Journal of Computer Sciences and Engineering, **Vol. 7, Issue. 8, pp.884-887**, **2019**.

AUTHORS PROFILE

Sandeep Kumar received Bachelor's Degree in Computer Science in the year 2017 from PGDAV College (University of Delhi), and a Master's Degree in Computer Science and Engineering in the year 2020 from C-DAC NOIDA (GGSIP University New Delhi). He is interested to do a Ph.D. in the field of Machine Learning and Deep Learning from Central University near Delhi-NCR.

Samridhi Dev received Bachelor's Degree in Computer Science & Engineering in the year 2018 from GRD Institute of Management of Technology Dehradun, affiliated to Uttrakhand Technical University, and a Master's Degree in Computer Science and Engineering in the year 2020 from C-DAC NOIDA (GGSIP University New Delhi). She is interested to pursue a Ph.D. in the field of Image Processing from Central University near Delhi-NCR.