e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

# A TRANSFER LEARNING APPROACH FOR AUTOMATING THE SPEAKER RECOGNITION FOR THE KANNADA LANGUAGE

**Shridhar Allagi*1, Anusha Hebbar*2**

*1Department Of Computer Science & Engineering K. L. E. Institute Of Technology, Hubballi, Karnataka 580027, India.

*2PG Student, Department Of Computer Science & Engineering, K. L. E. Institute Of Technology, Hubballi, Karnataka 580027, India.

## ABSTRACT

Automatic speaker recognition is the most popular and challenging problem in the area of AI & ML. It is used in the field of recognizing the human voice and it is mainly used for user authentication for safety majors and finding a particular human from a lot of speakers. It is difficult to work to put in a machine the dissimilarity persons' voices mainly with a variety of audio samples like accents, genders, language, etc. This paper includes the deep learning approach for setting up CNN and the neural networks, which were put on multiple takeout characteristic from audio samples, and is trained with various spectrograms. Transfer learning approaches are included to obtain a proper output utilizing a specific dataset i.e Kannada Kali. The proposed model gave an accuracy of up to 70%.

**Keywords:** CNN, Transfer Learning, Speaker Recognition, Deep Learning.

## I. INTRODUCTION

In this modern era, speaker recognition is widely utilized in biometrics and audio-based applications i.e implementing artificial intelligence applications such as Siri and Alexa. Speaker recognition is identifying the speaker using the speech indicators which are divided into speaker identity and also correspond to speaker verification. Speaker identity and verification have nevertheless developed literature, because of their significance in speech technology. It is a widespread topic with numerous packages, which include security, forensics, biometric authentication, speech reputation, and speaker. Due to the excessive quantity of research within the field, several methods have come up, so state-of-the-art inside the subject is quite mature, but also versatile.

Nowadays, as the popularity of deep learning (DL) methods is compatibly increasing due to speed and less priced hardware and reasonable software program solutions, it is available in infiltrating for every subject matter, where device getting to know-how is protected. So, it's far only natural, that specialists and scientists begin to apply deep learning models for speaker identification (SI). The deep learning models that are carried out in speaker identification and verification tasks from the earliest to the modern-day solutions.

The Speaker is associated with different variations and features with respect to voice density, depth, and pitch values. Some speakers may spell the word much faster whereas some might spell it at a lower pace. Other dominating features with respect to the speaker are accents, language used, the rate at which the user is speaking, etc. Hence the data associated with the individual speaker is highly dense and scattered in its nature and hence poses several challenging aspects in identifying the speaker. The deep learning models can be entrusted here to gain the knowledge at each level and used in the later stage for accurate classification of the speaker.

Another optimized way of addressing these challenges is to make use of transfer learning models. Transfer learning term explains about analyzing while solving the issue, cache the acquired know-how and the usage of it to clear the one greater related issues. In the deep mastering method, switch gaining knowledge of allows to utilize pre skilled networks with massive facts put that to work like typing on a pocket-size data.

In this research, we propose of experimenting with the Kannada Kali dataset with a convolutional neural network and an augmented transfer learning approach with a convolutional neural network. However, the dataset is limited, and the transfer learning approach is not fully explored on larger datasets, we limited the work to a specific feature set and a limited number of speakers.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

## II. RELATED WORKS

In[1], have proposed for categorizing component size sentiment. and the thing of locating the text in a given report and classifying it in step with words and the sentiment polarity with the objective. Switch capsule community model is that is used to switch the know-how acquired at the document level to the component degree to categorize, this is carried out with the help of the switch mastering framework.

In [2], has explained the interplay with the humans i.e. the commonplace manner of humans and they may become aware of the voice in their recognized one and identical factor achieved thru the tune if we pick out the voice of the artist is thought then popularity will smooth if the voice isn't cased to the listener it's a far hard process to perceive the voice with a tune this consists of some songs to create the education records and through this neural network is trained.

In [3], have proposed the deep mastering technique inside the field of speaker identification and verification and deep getting to know is the maximum popular solution for each speaker verification and identification. They have suggested the potential of the Artificial neural Network (ANN) and more than one linear regression (MLR) to find the best of spreadable gouda cheese at some point of 20`c,eight`c and 30`c .MLR fashions have been the most enormous ANN used 5 things decided on through principle factor analysis which enables to input data for ANN calculation.

In[5], have explored very interesting studies in the scope of music data retrieval and audio signal processing, they're using a novel approach for song style reputation with an ensemble of convolution lengthy brief time period reminiscence based totally on neural networks.

In [6], has suggested improvements in voice control, smart domestic segment, and so on. It consists of CNN and calculates the effectiveness of the speaker's popularity and then incorporates the use of the switch learning technique with a problem of confined training records.

In[7], have proposed speaker clustering, in this job of separating the speaker in recording with the aim to find "who spoke when" in the audio recording this is done with help of feature extraction from the recording to MFCC features and neural networks carried out by clustering and audio processing in the search of the same accuracy as 'state of the art' methods.

In[8], have suggested class is the essential discipline of research inside the region of information mining along with neural networks is a maximum used generation for the category and it includes the ANN algorithms and its kinds also their use in class.

In[9] have proposed deep neural networks (DNN) in the area of NLP, Recurrent neural network (RNN) and Convolutional neural network ( CNN) are two types of architecture are meant for handling the NLP tasks many NLP jobs switches due to the field of CNNs and RNNs .this is the systemic comparison of RNN and CNN on different range of representative NLP jobs and target to give guidance for DNN selection.

In[10], has explored architecture for fixing pc vision troubles and deep gaining knowledge of era has supplied a compelling opportunity that is automatically getting to know troubles with selected traits. In precise it consists of the deep network in laptop vision with CNN and "Alexnet" is base CNN.

## III. PROPOSED SYSTEM

This proposed model consists of the custom-made information set in which initially the audio statistics have aggregated from Kaggle, a global studies facts, open-access of audio datasets. The recordset consists of audio samples of various speakers in form of language, gender and accents, and so forth. These audio statistics grow to be further labeled and converted to spectrogram pics of the audio samples.

Fig.1 indicates the spectrogram pictures of unique speakers. It represents spectrographic variations of the audio system. A spectrogram photo is a time-frequency graph and it shows the complicated indicators along with audio in a simple to analyze and interpret, here X-axis represents time and Y-axis represents a frequency variety. After that, we put spectrogram photographs into a network along with hidden convolutional layers which act like a feature extractor. Then CNN model, the output is established by an SVM classifier for final elegance. For the use of a small information set, this method offers the high universal performance of SVM mixed with function gaining knowledge of CNN.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

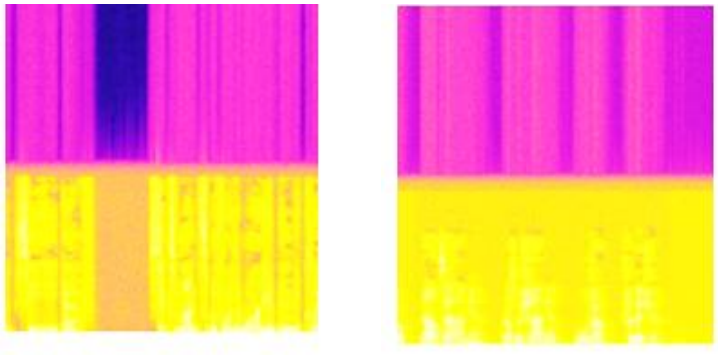

**Figure 1:** Spectrogram images of different speakers.

**The CNN Model**

Convolutional neural networks are used for speaker identity, natural language processing, and computer vision for characteristic extraction. We devised a CNN that is a modified version of CIFAR-10 architecture and represented it on spectrograms from 19 unique speakers. Spectrograms are graphs at the manner to assist us to visualize the energy of the audio signal over time along with several frequencies having a particular waveform. The vertical axis represents the frequency i.e pitch with the horizontal axis as time. The amplitude or loudness of the signal is visualized as 0.33 measurement in form of a color warmth-map wherein violet hue area act as low amplitude at the same time as extra brighter colors like yellow display immoderate amplitude place. So if we train our community on those images for a specific speaker, it's going to discover the underlying relationship between diverse amplitude and frequency through the years for the label.

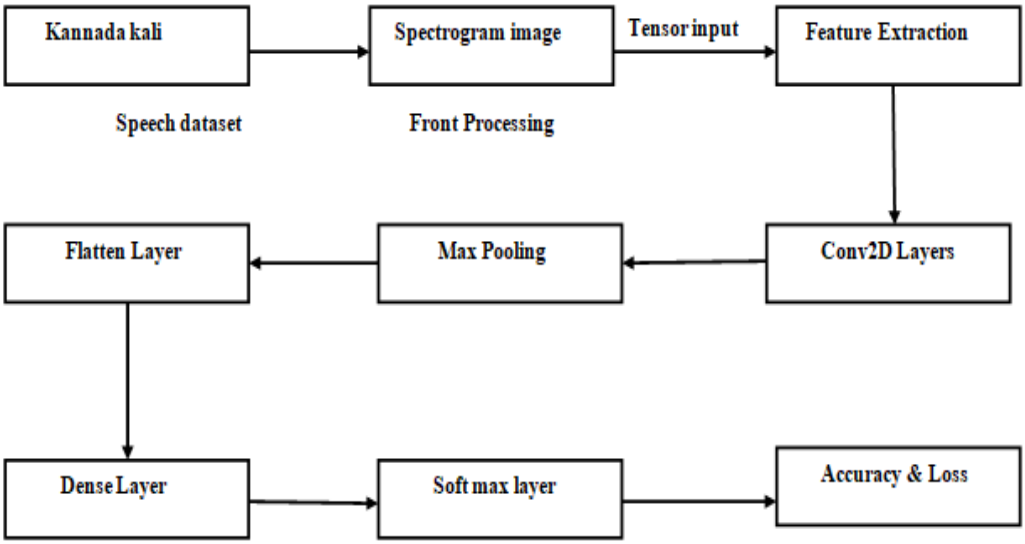


**Figure 2:** Architecture of Proposed Methodology

A 4 layer deep CNN version is used with two layers having clear out 32 and the other having sixty-four filters respectively (3x3 kernel). Among the layers of the one max pooling is completed with each iteration. After each layer, ReLu activation characteristics changed into a dense layer.

For transfer learning, the fourth convolutional layer's output is flattened out to create a representational vector. Figure.2 shows the structure of the CNN model with extent estimates for each layer. The model is skilled with Adam optimizer and 25% dropout is completed for the convoluted layers with regularisation.

## IV. EXPERIMENTS & RESULTS

As mentioned in the above methodology, our audio dataset is dependent on the total series of audio samples aggregated from Kaggle having 19 audio systems. Those 19 audio systems have generalizable variations together with gender, language, accent, and many others. The audio is split into 5-sec segments, so we get 43 /42 samples according to each speaker. Because of region and computing constraints, educated to our CNN model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

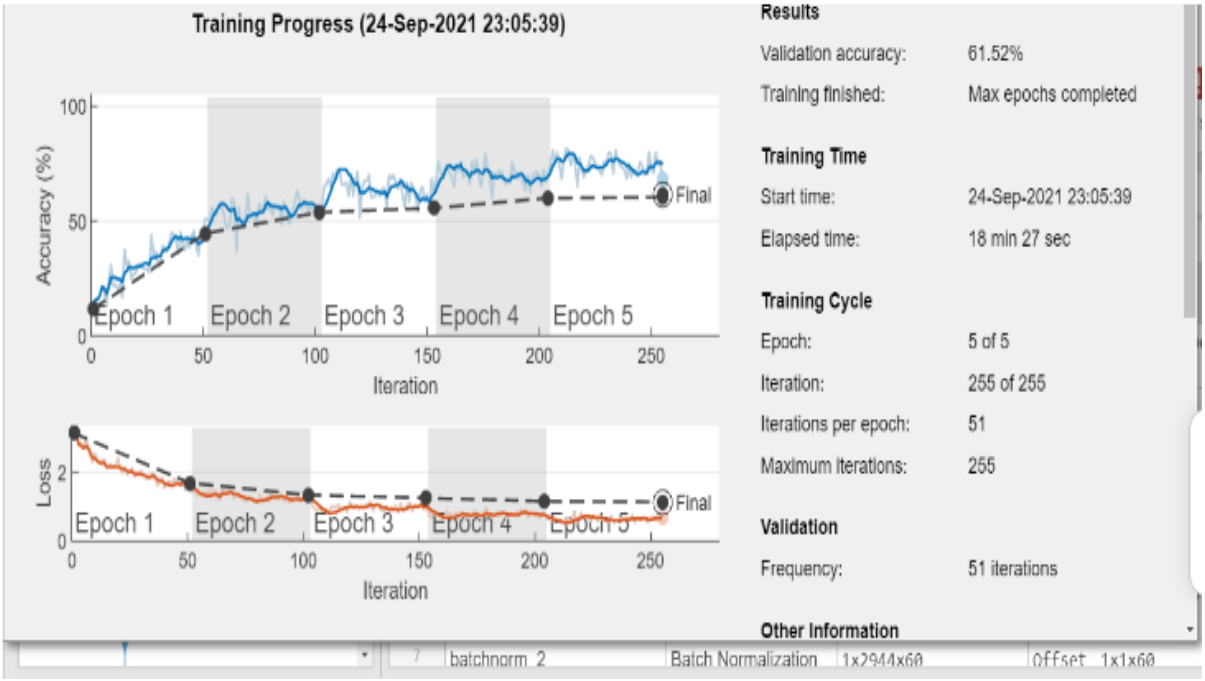


**Figure 3:** CNN model Result with accuracy and loss

Figure 3 shows the training process with model accuracy and model loss also it includes the validation accuracy of the CNN model, training time, training cycle consisting of the epoch, iteration, iteration per epoch, maximum iteration, at last, it having the validation frequency, and other information like hardware resources, learning rate schedule with learning rate.

## V. TRANSFERLEARNING IN CNN

The audio samples have been converted and saved as spectrogram pictures in .png format with a view to serve as training and validation units for our CNN having a shape (64,3) and RGB layout. As take over in the preceding phase, The records have 1048 training sets and 262 validation sets of spectrogram pics (of 80 instructions). The records have been trained for 5 epochs on the training and validation samples through batch processing of batch length 128. After 1 epoch of training, we finished with 20% accuracy. We imply it on 19 audio devices and after 5 epochs we have been given 59.96% accuracy.

Now we train with a neural network, that could become aware of the maximum of the audio device. It takes approximately 45min to 1h to generate the audio pattern. The final layer i.e. dense layer is flattened out to feed into an SVM that acts as a supervised classifier which has been used for numerous domains like text and photograph categories. With the assistance of CNN as characteristic extractors, we used SVM for the very last specific category of the audio system. A radial basis feature acted as a kernel for the SVM.

We trained audio segments of 20 seconds from five special speakers for testing, which were then examined and ended with 80% accuracy. We fed this into the SVM to get a type output. normal, SVM helped us lessen the time it generally takes for the version to analyze someone's voice while preserving accuracy nonetheless very high. suggest the accuracy and loss for 5 epochs for each training and validation device.

Figure 4 explains the result of the CNN model Accuracy of 61.52 % with the final iteration. It takes 20 min 8 sec, uses 5 epochs with 255 iterations, and also it has a validation frequency of 51 iterations. More training and a good audio sample give this model get approximately maximum output of 65 % to 70 %.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com


**Figure 4:** Result of CNN model

## VI. CONCLUSION

This challenge suggests the CNN model with the Deep learning technique within the field of speaker popularity, CNN features are trained on spectrographic images and brought by way of spectral features. The CNN deep learning features are attempted using the transfer learning method and the final output is put into an SVM classifier. Then CNN version is trained and tested with their samples and their performance and accuracy are computed. The proposed model gave an accuracy of up to 70% and the model performance can be increased with training for a larger dataset also use of ensemble models can optimize the models.

## VII. REFERENCES

[1] Biswas, S., Solanki, S.S. Speaker recognition: an enhanced approach to identify singer voice using neural network. Int J Speech Technol (2020).

[2] Bala, R. and Dr. Dharmender Kumar. “Classification Using ANN : A Review.” (2017).

[3] Ghosal, Deepanway & Kolekar, Maheshkumar. (2018). Music Genre Recognition Using Deep Neural Networks and Transfer Learning. 2087- 2091. 10.21437/Interspeech.2018-2045.

[4] Jumelle, Maxime, and Taqiyeddine Sakmeche. “Speaker clustering with neural networks and audio processing.” arXiv preprint arXiv:1803. 08276 (2018).

[5] M. Wang, T. Sirlapu, A. Kwasniewska, M. Szankin, M. Bartscherer and R. Nicolas, “Speaker Recognition Using Convolutional Neural Network with Minimal Training Data for Smart Home Solutions,” 2018 11th International Conference on Human System Interaction (HSI), Gdansk, 2018, pp. 139-145, doi: 10.1109/HSI.2018.8431363.

[6] Sztaho´, Da´vid, Gyo¨rgy Szasza´k, and Andra´s Beke. ”Deep learning methods in speaker recognition: a review.” arXiv preprint arXiv:1911.06615 (2019).

[7] Stangierski, J., Weiss, D. & Kaczmarek, A. Multiple regression mod- els and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable pro- cessed Gouda cheese. Eur Food Res Technol 245, 2539–2547 (2019).

[8] Sungheetha,Akey,and RajeshSharma “TransCapsule Model for Sentiment Classification.” Journal of Artificial Intelligence 2, no. 03 (2020): 163-169.

[9] Srinivas Suraj, Sarvadevabhatla Ravi Kiran, Mopuri Konda Reddy, Prabhu Nikita, Kruthiventi Srinivas S. S., Babu R. Venkatesh, ”A Taxonomy of Deep Convolutional Neural Nets for Computer Vision”, Frontiers in Robotics and AI Journal, Volume 2, 2016, Pages369144.

[10] Yin, Wenpeng, et al. ”Comparative study of cnn and rnn for natural language processing.” arXiv preprint arXiv:1702.01923 (2017).