International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0058 ISSN: 2581-4621

A MFCC-CNN BASED VOICE AUTHENTICATION SECURITY

V Anupama[1],Ch Amrutha[2], G Amrutha Varshini[3], G Sai Gowtham Nandan[4], GVLN Satya Sai Vivek[5]

[1,2,3,4,5]*Dept. of CSE, Lendi Institute of Engg & Tech*

Abstract

In this paper, a novel architecture is proposed using a convolutional neural network (CNN) and mel frequency cepstral coefficient (MFCC) to identify the speaker in a noisy environment. This architecture is used in a text-independent setting. The most important task in any text-independent speaker identification is the capability of the system to learn features that are useful for classification. We are using a hybrid feature extraction technique using CNN as a feature extractor combined with MFCC as a single set. For classification, we used a deep neural network which shows very promising results in classifying speakers. We made our dataset containing 60 speakers, each speaker has 4 voice samples. Our best hybrid model achieved an accuracy of 87.5%. To verify the effectiveness of this hybrid architecture, we use parameters such as accuracy and precision.
Keywords—Convolutional Neural Network, Mel Frequency Cepstral Coefficients, Feature Extraction, Text Independent, Speaker Identification, Deep Neural Network

I.INTRODUCTION

Speech processing is the investigation of speech signals and the processing techniques for signals. The signs are normally handled in a computerized portrayal, so speech processing can be viewed as an extraordinary instance of advanced sign processing, applied to speech signals. Speech recognition is the distinguishing proof of an individual from qualities of an individual's speech.Voice is a basic or you can say the most important part of a human's everyday routine. It utilizes one of a kind organic qualities to check a person's personality speech validation is otherwise called voice confirmation or voice recognition that investigates an individual's voice to check their character.

The question that we are solving in this paper is, who is speaking? Speaker identification is the process of consequently distinguishing the individual talking in the voice sample. Speaker identification is a topic that gained a lot of attention in the research community. It is the most challenging task because every speaker is different in terms of accents, speaking style, frequency of words and vocal tract. The presence of noise, background chatter and music also makes the task even more difficult. The conditions such as faulty recording devices also affect classification accuracy. In speaker identification closed-set and text independent setting, the voice must be from an enrolled speaker and does not depend upon the said words of the speaker. The main approaches include i-vector,hidden Markov model, Gaussian Mixture Model (GMM) with Universal Background Model (UBM), vector quantization, neural network and support vector machine. These approaches use various types of the dataset for speaker identification tasks. Some datasets are recorded in a laboratory environment where there is no noise and the others have little noise or chatter. Recent advancements are done in using a convolutional neural network for speaker identification tasks because they can easily handle noisy datasets and there is no need for feature engineering, as feature extraction and classification both can be done by CNN. Different classification and feature extraction techniques have been used in the speaker identification task, one of them is using CNN for both feature extraction and classification. Some researchers used

CNN only as a feature extractor and used other classifiers for classification. Researchers also used MFCC features for speaker identification but they are unreliable in noisy environments.
Speaker Identification has many applicable services such as authentication of speakers in telephone banking, control for confidential information, information services, access to remote computers and database access services.

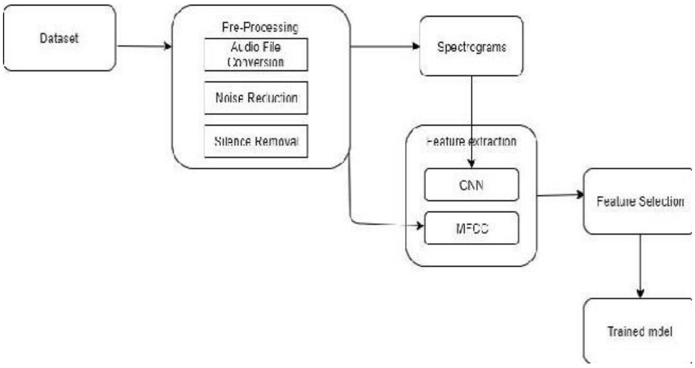


Fig:1.Component for speaker identification

The speaker identification task still has to cover a lot of milestones to make it state of the art. If this system becomes state of the art, it can replace the existing verification system. This system can also be used to add an extra layer of security in the existing system. Components for speaker identification. contains the block diagram of the whole speaker identification system. The proposed system contains dataset, pre-processing of voice samples reducing noise and remove silence, then making spectrograms of voices and use CNN as a feature extractor combine feature from CNN and features from MFCC combine and apply a feature selection technique on them and then pass these selected features to a DNN for classification. The fundamental objective of this paper is to structure and execute a speaker acknowledgment framework utilizing a neural network. The following sections cover the main components of this approach, including dataset and preprocessing, proposed approaches, results and in the end conclusion.

II.DATASET AND PRE-PROCESSING

The following subsections include information about the dataset and its characteristics and pre-processing of the voice samples

A. Dataset

The dataset that we used is made in house. The reason to use this dataset is that we want to collect voice samples that are more close to the voices that occur in a real-life environment. Other great datasets are also available presented by some great researchers, the reason for not using them is either they are recorded in laboratory environments where there is no noise at all or they are composed of hours and hours of voice samples from each speaker. This dataset is collected in the wild dataset, containing background noises and chatter. The dataset that we have to make is collected from our classmates and audio from multiple videos from YouTube. This dataset contains a total of 2 speakers, all speaking in Urdu. We collect 50 voice samples from each speaker, each voice sample is 3 seconds long. Speakers' names are put as the file name of the voice sample. The format that we follow is speaker name_number of audio samples like gowtham_01. All the voice samples are collected according to this format. Dataset is pre-divided into two portions, 3 voice samples from each speaker are used for training and the remaining one is used for testing. These voice samples are saved in a different folder named training and testing.

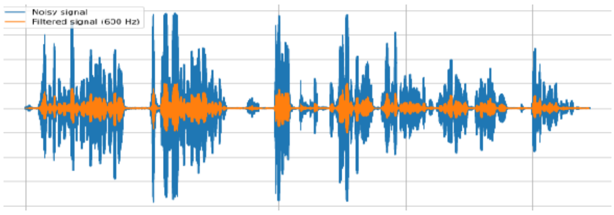


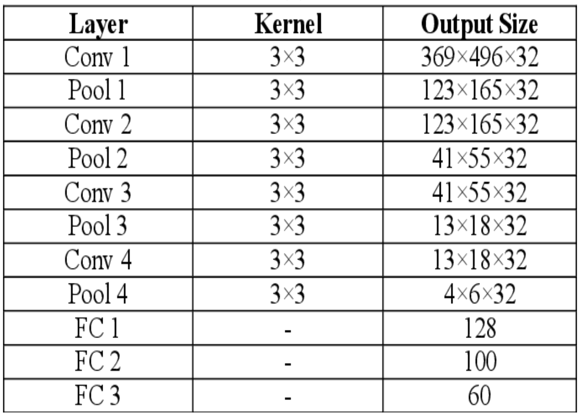
Fig:2. Actual signal and filtered signal B.Pre Processing

All the raw audio data is pre-processed before sending it to the neural network. To save time in preprocessing, we use the pysox library for pre-processing the voice samples. First, we check the format of the voice sample file. All the voice sample files are converted in the .wav format. Noise reduction is applied using the pysox noise profiling technique. This technique reduces static noises such as hiss, hum that can be caused by the surrounding environment or by the electrical wiring the voice travels through. (fig.2)Actual signal and filtered signal.(fig.2) It shows the noise that is removed from the voice sample, the original voice sample is in the blue and filtered noise in the red. The silence that is greater than 5 seconds is also removed from the voice samples. For CNN, we have to make a spectrogram from voice samples first, for that we average each voice sample channel to mono and a sampling rate of 16 kHz is used.

III.PROPOSED APPROACHES

We are proposing a comparison of two approaches first, before combining them into a hybrid approach. The first approach is using CNN for both feature extraction and classification. Second, use MFCC as a feature extractor and DNN as a classifier.

TABLE 1. CNN Architecture



Layer	Kernel	Output Size
Conv 1	3×3	369×496×32
Pool 1	3×3	123×165×32
Conv 2	3×3	123×165×32
Pool 2	3×3	41×55×32
Conv 3	3×3	41×55×32
Pool 3	3×3	13×18×32
Conv 4	3×3	13×18×32
Pool 4	3×3	4×6×32
FC 1	-	128
FC 2	-	100
FC 3	-	60

CNN Based Approach Convolutional neural networks extended the well known idea from image processing of filters by applying the filter on the pixel, followed by the pooling layer(table.1) to extract the values from that filter and reduce the dimensionality. In CNN based approach, we make a neural network composed of multiple hidden layers, layers containing convolution layer, pooling layer, batch normalization, and dropout. It is a hybrid approach we test the model by selecting the multiple combinations of filter 3*3 and 5*5 filters for the pooling and convolution layer respectively. An additional batch normalization technique is also used. Input to this network is a spectrogram of size 369×496 made from a voice sample.

Convolutional layer

The function of the convolutional layer is to apply a filter on the image. Image is presented in pixels, so a filter of size say 2×2 is started from left and applies the convolutional operation on the selected filter and slides to the right until the whole image is traversed. In convolutional operation, a mathematical operation is performed on two values, to measure the third value to show how the shape of one value affects others. The filter size for a convolutional layer we are using is 3×3. This layer is used to extract high-level features such as edges and colors from spectrograms.

Pooling layer

The pooling layer is applied to the convolved feature and the filter size of the pooling layer is 3×3. The function of the pooling layer is to decrease the computational complexity and to do dimensionality reduction. We are using max pooling, which means when we apply a filter of 3×3 on the convolved feature, it picks the top value from the feature.

Fully Connected Layer

The function of a fully connected layer is to learn non-linear combinations from the output of the convolutional layer when the image is converted into a suitable form, then we flatten the image. Such as if an image is represented 12×12×3, then we convert that image to a 256×1 column vector. Over many epochs, the neural network then learns to classify images using the softmax technique. The function of softmax is to convert values that are given to an image by the neural network into probabilities. Rectified linear unit (RELU) is used as an activation function.

DNN Based Approach

In a deep neural network (DNN) based approach, MFCC features are extracted(fig. 3) from each voice sample. 39 MFCC features are extracted from each voice sample and then these features were given as an input to the DNN. DNN has three layers, an input layer that takes the MFCC feature as an input. The hidden layer is logically found between the input and output layer where all the computation is performed and the main part as a classifier. An output layer makes the result for the input that is given in the input layer and computed in hidden layers. Rectified linear unit (RELU) is used as an activation function.

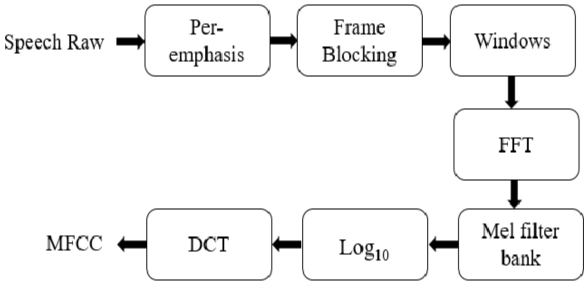


Fig.3.MFCC feature extraction

In MFCC feature extraction(fig.3), in which the audio sample is passed through a filter which increases the frequency. Then the voice signal is divided into a small duration of 22-32 milliseconds which are known as frames. Then a hamming window is used that is multiplied with every frame in ode rot to join the signal. Then a Fast Fourier Transform (FFT) is performed on every frame to get the magnitude frequency. Then magnitude frequency is multiplied with triangular band filters to help in size reduction of the selected feature. DNN Architecture Layer Output Shape Input Layer

70 Hidden Layers - Output Layer 60 C. Hybrid Approach In this approach, we combine both the properties of CNN based approach and DNN based approach, in such a way that we trained our

data on both neural network models and both models did not give us a promising accuracy in terms of identifying an unknown speaker. So what we are going to do is to combine features of both models into a single feature file and then train a deep neural network on it. From a convolutional neural network, we get the features from the convolutional layer before the last fully connected layer and that gives us a feature vector against every class. And From the DNN based approach, we get features from MFCC. These features are then combined in a single file and we apply a feature selection technique. The feature selection technique is proved to be helpful to reduce model overfitting and reduce model complexity. It also reduces the training time by selecting the best feature from a list of features. In the feature selection technique, the chi-square test is performed to obtain the best features from the extracted features. DNN model is again used to classify the best-selected features, this technique shows more promising results than the other two techniques separately.

IV.CONCLUSION

This work demonstrated a hybrid architecture that has a better performance than the other two proposed architectures. The performance is tested using a new in house dataset containing voices from 60 speakers, it is recorded in the wild dataset containing noise and background chatter. Comparing the performance for this task is difficult to compare because there are a lot of dataset conditions and scenarios. 87.5% is still good accuracy regarding the noise and the amount of background chatter in the dataset. This accuracy is achieved by combining the features from CNN and MFCC. Both feature extraction techniques have different characteristics so it covers a lot of properties of a voice sample that make it distinguish from other voice samples. There is still a lot to improve in this approach. First, the architecture can be made less computationally intensive and made deeper to extract more features from voice samples. Second, this approach can be combined with other feature extraction approaches to enhance the performance of the model.

REFERENCES

[1] P. Dhakal, P. Damacharla, A. Y. Javaid, V. Devabhaktuni, “A Near Real-Time Speaker Recognition Architecture for Voice-Based User Interface,” in Machine Learning, vol. 1(1), 2019, pp. 504-520.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, May 2011.

[3] S. Kanrar, “Speaker Identification by GMM based i Vector, ” in eprint arXiv:1704.03939, April 2017.

[4] T. K. Das, K. Nahar, “A Voice Identification System using Hidden Markov Model,” in Indian Journal of Science and Technology, vol. 9(4), January 2016.

[5] Rong Zheng, Shuwu Zhang and Bo Xu, "Text-independent speaker identification using GMM-UBM and frame level likelihood normalization," 2004 International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 2004, pp. 289-292. C. Ting, Sh-Hussain Salleh, Tian-Swee Tan and A.

K. Ariff, "Text independent Speaker Identification using Gaussian mixture model," 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, 2007, pp. 194-198.

[6] F. Soong, A. Rosenberg, L. Rabiner and B. Juang, "A vector quantization approach to speaker recognition," ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, USA, 1985, pp. 387-390.

[7] S. S. Tirumala, S. R. Shahamiri, "A review on Deep Learning approaches in Speaker Identification," the 8th International Conference on Signal Processing Systems, pp. 142-147, November 2016.

[8] M. McLaren, Y. Lei and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 4814-4818.

[9] Y. Lukic, C. Vogt, O. Dürr and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Vietri sul Mare, 2016, pp. 1-6.

[10] X. Zhao, Y. Wang and D. Wang, "Deep neural networks for co channel speaker identification," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 4824-4828.

[11] V. R. Apsingekar and P. L. De Leon, "Support vector machine based speaker identification systems using GMM parameters," 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2009, pp. 1766-1769.