Speaker Recognition System From Audio Speech Signals Using Deep Learning

T.V.Vamsi Krishna[1], T.Harika[2], U.Jyothi[3], Z.Manoj Kumar[4], Y.Kishita[5]

Associate Professor, Department of CSE, Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur(DT), AP[1]

IV CSE, Department of CSE, Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur(DT), AP[2,3,4,5]

***Abstract*—Speaker recognition is a subfield of research within the broader area of digital signal processing, which deals with the analysis, manipulation, and interpretation of signals such as speech, images, and video. speaker recognition, also referred to as voice biometrics, is the identification of a person based on their voice. The speech signals convey many levels of information to the listener. Speaker recognition research involves developing algorithms and techniques to identify individuals based on their voice characteristics, and it involves signal processing, pattern recognition, and machine learning techniques. So, we introduced a deep learning model which aims to extract, describe, and identify the details about the speaker based on the audio speech signals. We upheld a Convolutional neural network (CNN) model with a dataset of five speakers. The implemented approach uses 1D CNN to extract discriminative features for speaker identification. In the speaker classification stage, a fully connected layer is used to classify the extracted features into different speaker classes. The proposed system is tested and shows a 97% accuracy rate, exceeding the standard operating procedures. The results demonstrate the effectiveness of using 1D CNNs for speaker recognition from audio speech signals. Our project research offers useful insights into the use of voice as a biometric identifier and the creation of successful voice-based biometric systems.**

Keywords—Speaker Recognition; Convolutional neural network; Audio speech signals

I. INTRODUCTION

Many levels of information are communicated through human speech. It mostly conveys a message through words. But on other levels, it communicates details about the speaker's language, dialect, emotion, gender, and identity. While speech recognition systems seek to recognize the words used in speech, speaker recognition systems seek to identify the speaker who is speaking a given signal of speech. Speaker identification is the process of identifying an individual based on their unique vocal characteristics or voiceprint. It is an important task in various applications such as security, surveillance, forensics, and speech recognition. The goal of speaker identification is to automatically determine the identity of a speaker from a given speech signal or recording.

There are two additional basic activities that fall under the general category of speaker recognition. The task of confirming if a person is who they say they are is known as speaker verification, also referred to as speaker authenticity. Finding the speaker among a list of recognized speakers is the task of speaker identification. The algorithm must do a 1:N classification because the unidentified speaker does not attempt to establish their identity. These tasks can be further divided into text-dependent and text-independent categories. The text being spoken to is already known to the recognition system in a text-dependent system. The text associated with the recognition is irrelevant in a text-independent system.

Text-dependent speaker recognition is a type of speaker recognition where the speaker is required to speak a specific phrase or set of phrases for identification. Text-dependent speaker recognition is often used in security applications, such as access control systems, where the speaker is required to speak a password or a passphrase to gain access to a secure area or system. The system compares the speaker's voice with a pre-recorded voiceprint of the speaker for verification. The process of text-independent speaker recognition involves extracting acoustic features from the speaker's voice, such as pitch, tone, and spectral characteristics. These features are then used to create a voiceprint, which is a mathematical representation of the speaker's voice. The voiceprint is then compared with a pre-recorded voiceprint of the speaker for identification.

Text-independent speaker recognition is often used in applications such as forensic analysis of recorded conversations or surveillance systems, where the speaker may not be speaking a predetermined phrase. One of the advantages of text-dependent speaker recognition is its higher accuracy compared to text-independent speaker recognition, as the system has more information to compare the speaker's voice with. Additionally, text-dependent speaker recognition is more resistant to spoofing attacks, where someone tries to impersonate the speaker by using a recording of their voice.

There are different approaches to speaker identification, including statistical modeling techniques such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Neural Networks. In recent years, deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used for speaker identification with great success.

Gaussian Mixture Models (GMMs) are a popular statistical modeling technique for speaker identification. GMMs model the probability distribution of the speech features for each speaker using a mixture of Gaussian distributions. The GMM parameters are estimated using an iterative algorithm, such as the Expectation-Maximization (EM) algorithm. During testing, the likelihood of the speech features for each speaker is computed using their respective GMMs, and the speaker with the highest likelihood is identified.

Hidden Markov Models (HMMs) are another statistical modeling technique used for speaker identification. HMMs model the temporal variations in the speech signal using a sequence of hidden states. Each hidden state is associated with a probability distribution of the speech features. During training, the HMM parameters are estimated using the Baum-Welch algorithm, and during testing, the likelihood of the speech signal given each HMM is computed using the Viterbi algorithm. The speaker with the highest likelihood is identified.

Vector Quantization (VQ) is a technique for clustering speech features into a codebook, which is a set of representative vectors. During training, the codebook is generated using a clustering algorithm, such as the k-means algorithm. During testing, the speech features are quantized to the closest vector in the codebook, and the speaker with the closest codebook is identified. These techniques model the probability distribution of the speech features or temporal variations in the speech signal to identify the speaker.

Convolutional Neural Networks (CNNs) are a type of deep neural network that is commonly used for image recognition tasks. However, they can also be used for speech recognition tasks, including speaker identification. CNNs can learn hierarchical representations of the speech signal by applying convolutional filters over the input signal, followed by pooling operations to reduce the dimensionality of the features. The resulting features are then fed into a fully connected neural network for classification. CNNs have been shown to be effective in identifying speakers from raw speech signals.

Recurrent Neural Networks (RNNs) are another type of deep neural network that can be used for speaker identification. RNNs can capture the temporal dynamics of the speech signal by processing the input sequence of speech features in a sequential manner. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are variants of RNNs that have been shown to be effective in modeling the temporal variations of the speech signal. These models can be trained on the raw speech signal or on preprocessed speech features, such as Mel-Frequency Cepstral Coefficients (MFCCs). These techniques can learn hierarchical representations of the speech signal or capture the temporal dynamics of the speech signal to identify the speaker.

II.LITERATURE SURVEY

The paper[1] proposes a novel approach for text-dependent speaker verification (SV) using attention-based models. The authors argue that conventional SV systems rely heavily on fixed-length feature vectors that may not capture important temporal information, resulting in suboptimal performance. They propose to use attention mechanisms to selectively focus on informative parts of the input sequence, allowing for more effective feature extraction and classification.The proposed approach consists of two main components: a convolutional neural network (CNN) for extracting high-level features from the input sequence, and an attention mechanism for selectively weighting the features based on their relevance to the speaker verification task. The attention mechanism is implemented using a self-attention network, which computes a set of attention weights based on the similarity between different parts of the input sequence.

One of the most enhanced papers [2] proposes a new neural network architecture called SincNet for speaker recognition from raw waveform signals. The authors argue that traditional methods for speaker recognition rely on handcrafted features extracted from the signal, which can be time-consuming and suboptimal. SincNet aims to learn suitable filters from the raw waveform directly, reducing the feature extraction step and improving performance. The SincNet architecture consists of a series of trainable sinc filters followed by a nonlinear activation function and pooling layers. The sinc filters are designed to be band-pass filters, which extract different frequency bands from the input signal. The authors train the SincNet architecture using a standard cross-entropy loss function on a large-scale dataset of speaker verification tasks. The experimental results demonstrate that SincNet outperforms traditional feature-based methods and other neural network architectures for speaker recognition from raw waveform signals. SincNet achieves state-of-the-art performance on several benchmark datasets, including the TIMIT dataset and the VoxCeleb dataset.

The Recent paper[3] Automatic Speaker Recognition (ASR) is a field of study at the intersection of signal processing, machine learning, and pattern recognition. ASR systems aim to automatically identify who is speaking by analyzing voice recordings. Optimized machine learning algorithms are often used in ASR systems to improve accuracy and efficiency. The first step in an ASR system is to extract features from the voice recording. These features can include things like the frequency of the voice, the energy of the signal, and the duration of certain sounds. These features are then used to create a feature vector for each recording. In order to train a machine learning algorithm, a large dataset of voice recordings is needed. This dataset should include recordings of all the speakers that the system needs to recognize.

The related paper[4] explains about Speaker identification using a convolutional neural network (CNN) is a popular approach for identifying the speaker of a given audio sample. This technique involves training a CNN on a dataset of audio samples to learn to recognize the unique features of different speakers' voices. The trained CNN can then be used to classify new audio samples based on their speaker identity. The use of CNNs for speaker identification has been shown to be effective for both clean and noisy speech samples. In the case of clean speech, the CNN can be trained using the raw audio waveform or a spectrogram representation of the audio. The use of spectrograms can help to extract the relevant frequency features from the audio, which can improve the accuracy of speaker identification.In the case of noisy speech, various techniques can be used to preprocess the audio before feeding it into the CNN. These techniques include denoising algorithms, such as spectral subtraction or Wiener filtering, which can remove noise from the audio signal.It is important to carefully consider the preprocessing techniques and training strategies used to ensure the best possible performance.

III.METHODOLOGY

A. Data preprocessing:

Speaker identification is the process of identifying the unique characteristics of a speaker's voice, such as their accent, tone, and pitch, to determine their identity. In order to perform speaker identification accurately, the data must be preprocessed to ensure that it is of high quality and consistency.

Audio file cleaning: The audio file should be cleaned to remove any background noise or artifacts that may interfere with the analysis. This can be done using filters and noise reduction algorithms and using noise files to eliminate it.

Audio segmentation: The audio file should be divided into smaller segments that correspond to individual speakers. This can be done by removing background noise in each sample.

B.Feature Extraction:

The features can be extracted from the audio signal using a technique called Mel-frequency cepstral coefficients (MFCCs). The MFCC method begins with pre-emphasis in which the audio signal is pre-emphasized to amplify the high-frequency components of the signal. This can be done by applying a first-order high-pass filter to the signal.Then it follows frame segmentation which divides audio into frames. A window function (such as a Hamming or Hanning window) is applied to each frame to reduce spectral leakage and improve frequency resolution.

The FFT is applied to each frame to convert the signal from the time domain to the frequency domain. A bank of Mel-scale filters is applied to the frequency spectrum to extract the relevant frequency components. The Mel-scale filters are designed to mimic the non-linear frequency response of the human ear. Discrete cosine transform (DCT) is applied to the compressed Mel-spectrum to obtain the MFCCs.

The DCT decorrelates the Mel-spectrum, producing a set of coefficients that are uncorrelated and have lower dimensionality. Once the MFCCs have been extracted, they can be fed into a 1D convolutional layer as input features. The convolutional layer can learn the relevant patterns and relationships between the input features to identify the speaker. The output of the convolutional layer can then be passed through one or more fully connected layers to make the final speaker identification prediction.

C. System Architecture:

Firstly, the input of our proposed system will take data directories that have various files. These samples will go through a procedure to set the training settings. The data directories contain audio recordings with noise and a small number of noise files; these go through various processes to produce the data set. To identify the speaker among these speakers, feature extraction will be performed on this dataset. The speaker is thus identified by our model with a lot more precision, producing effective outcomes.

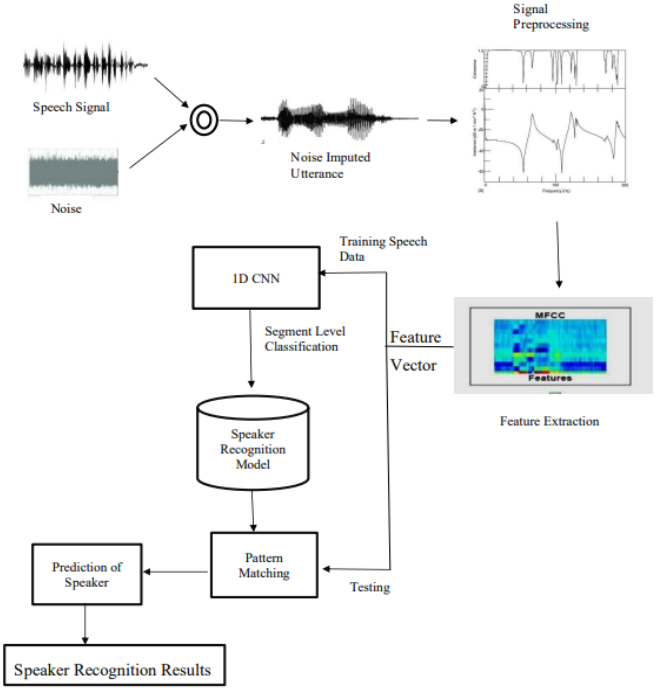

Fig1: System Architecture

D. Our Model:

We proposed a 1D convolution model to recognize the speaker within the audion speech signals.It consists of the following layers associated with it.

Input layer: The input to the model is a sequence of audio samples that are passed through a pre-processing step, such as the extraction of MFCCs.

Convolutional layer: The convolutional layer applies a set of learnable filters to the input signal to detect local patterns and features. Each filter slides over the input signal with a fixed stride, and the output is computed as the dot product of the filter weights with the input signal in each position. The resulting output is a feature map that encodes the presence of each filter in the input signal.

Pooling layer: The pooling layer downsamples the output of the convolutional layer by applying a pooling function, such as max pooling or average pooling, to each feature map. This reduces the spatial dimensions of the feature maps while preserving the most relevant features.

Fully connected layers: The output of the pooling layer is flattened and passed through one or more fully connected layers, which learn to map the high-level features extracted by the convolutional layers to the speaker identification task. The final layer produces a probability distribution over the possible speakers.

Output layer: The final layer is a softmax layer that produces a probability distribution over the possible speakers. The speaker with the highest probability is chosen as the predicted speaker identity.

Layer (type)	Output Shape	Param #	Connected to
input (InputLayer)	[(None, 8000, 1)]	0	[]
conv1d_15 (Conv1D)	(None, 8000, 128)	512	['input[0][0]']
activation_10 (Activation)	(None, 8000, 128)	0	['conv1d_15[0][0]']
conv1d_16 (Conv1D)	(None, 8000, 128)	49280	['activation_10[0][0]']
activation_11 (Activation)	(None, 8000, 128)	0	['conv1d_16[0][0]']
conv1d_17 (Conv1D)	(None, 8000, 128)	49280	['activation_11[0][0]']
conv1d_14 (Conv1D)	(None, 8000, 128)	256	['input[0][0]']
add_4 (Add)	(None, 8000, 128)	0	['conv1d_17[0][0]', 'conv1d_14[0][0]']
activation_12 (Activation)	(None, 8000, 128)	0	['add_4[0][0]']
max_pooling1d_4 (MaxPooling1D)	(None, 4000, 128)	0	['activation_12[0][0]']
average_pooling1d (AveragePooling1D)	(None, 1333, 128)	0	['max_pooling1d_4[0][0]']
flatten (Flatten)	(None, 170624)	0	['average_pooling1d[0][0]']
dense (Dense)	(None, 256)	43680000	['flatten[0][0]']
dense_1 (Dense)	(None, 128)	32896	['dense[0][0]']
output (Dense)	(None, 5)	645	['dense_1[0][0]']

Total params: 43,812,869
Trainable params: 43,812,869
Non-trainable params: 0

Fig 2: Structure of Proposed CNN Model

IV.EXPERIMENTS

A. *Dataset Preparation:*

The Our research focuses on reading data from folders that contain a variety of files. The data directories include a few noise files and audio recordings with noise. We have a collection of five distinct speakers, and we have 1500 samples of .wav format files for each speaker. Now we use a set of parameters to train so that each data sample has noise removed. In order to remove background noise from the set of samples, we first separate the noise samples into various chunks. Then, a dataset is created.

B. Training and Testing the model :

Training: The model is trained on the training set using the categorical cross-entropy loss function and stochastic gradient descent and Adam optimizer. During each epoch, the model is updated based on the gradients of the loss function with respect to the model parameters.

Validation: The model is evaluated on the validation set to monitor its performance and prevent overfitting. The validation accuracy and loss are used to adjust the hyperparameters of our model. The hypermerters are as valid_split=0.1; sample_rate=16000; scale=0.5; batch_size=128; Epochs=15.

Testing: The final model is evaluated on the testing set to estimate its performance on unseen data. hence in our model, we recognized each speaker with an accuracy of 97%. Hence, we can say that our model predicted more speakers correctly.

During training, it is important to monitor the training and validation loss and accuracy to ensure that the model is learning and not overfitting the training data. Dropout layers can be added to the model to reduce overfitting. Early stopping can also be used to stop training when the validation loss stops improving.

Perform predictions: The saved model weights and architecture can be loaded into memory. The test audio data is preprocessed in the same way as the training data, i.e., by extracting relevant features such as MFCCs. The preprocessed test data is fed into the trained model to make predictions about the speaker's identity. The output of the model is typically a probability distribution over the possible speakers, which can be used to make a final prediction. The predicted speaker identities are compared to the ground truth labels for the test data to evaluate the performance of the model. Common metrics for evaluation include accuracy, precision, recall, and F1 score.

C. Baseline Model:

The baseline model has been trained using a cross-entropy loss function and stochastic gradient descent optimizer. The model can also include dropout layers to prevent overfitting.It architecture of the baseline contains, the input as a sequence of 39 MFCCs for a given audio sample. The first convolutional layer has 64 filters with a kernel size of 3, followed by batch normalization and activation. The output is then downsampled using a max pooling layer with a pool size of 2, and then a dropout layer is added to prevent overfitting. The flattened output is passed through two fully connected layers with 64 and 20 units, respectively, and a dropout layer is added after the first fully connected layer. Finally, the output layer uses a softmax activation function to output a probability distribution over the possible speakers. This baseline model can be further improved with more convolutional layers, regularization techniques, and hyperparameter tuning, depending on the specific requirements and resources of the task.

V.RESULTS:

The proposed speaker recognition system using CNN technology has the following results:

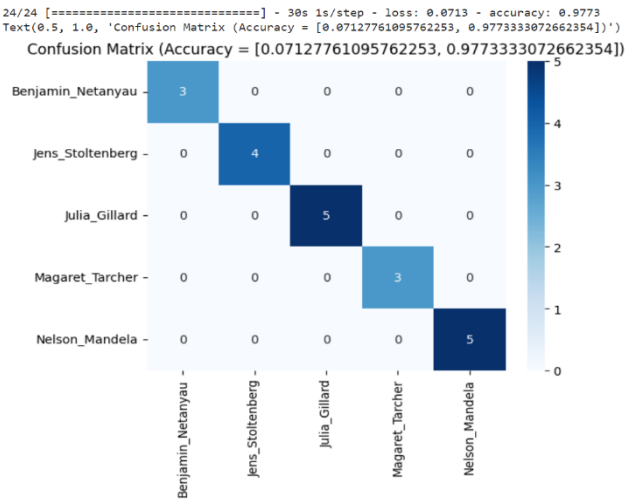

Fig 3: Experimental results of our model

The classification outcome from the proposed CNN-based model is shown in terms of confusion matrix in Fig. 3. It is an outcome of the testing set's performance. There are 20 test samples for the CNN 1D model. The overall success rate for categorisation is 97%.Each Epoch of proposed model shows accuracy of more than 90%.

VI.CONCLUSION

In conclusion, a speaker identification model with a 1D convolution layer is an effective tool for automatically locating speakers in audio recordings. The model is trained on preprocessed audio data, which is often translated into pertinent features like MFCCs, and the model architecture is created to efficiently learn and extract patterns in the data using convolutional layers. A proper optimizer, loss function, and hyperparameters has been used, and the model's performance on a validation set have carefully monitored to avoid overfitting. After the model is trained, its performance and generalizability can be assessed on a different test set. Overall, a 1D convolution layer-based speaker identification model has a wide range of possible applications in industries like forensics, security, and law enforcement. This kind of model can automate speaker identification, enhancing accuracy and consistency while saving time and resources. Our model outperforms with best accuracy to identify speaker.

VII.FUTURE SCOPE

Future scope of the work is to increase the efficiency of speaker recognition systems in terms of accuracy, work has to be performed using various datasets formed under uncontrolled conditions.The future scope of speaker recognition is vast and promising, and it has the potential to transform many industries. Speaker recognition technology could be used to enhance security by identifying authorized personnel and preventing unauthorized access to secure areas. It can also be used as Speaker recognition technology that could be used to monitor patients' vocal patterns and detect changes that may indicate health problems, such as neurological disorders or respiratory illnesses.

VIII.REFERENCES

[1] F A Rezaur Rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, Li Wan: ATTENTION-BASED MODELS FOR TEXT-DEPENDENT SPEAKER VERIFICATION. arxiv.org/pdf/1710.10470v3.pdf

[2] Mirco Ravanelli, Yoshua Bengio: SPEAKER RECOGNITION FROM RAW WAVEFORM WITH SINCNET. arxiv.org/pdf/1808.00158v3.pdf

[3] Tumisho Billson Mokgonyane; Tshephisho Joseph Sefara; Thipe Isaiah Modipa; Madimetja Jonas Manamela: Automatic Speaker Recognition System based on Optimised Machine Learning Algorithms.
research space. sirco.za/dspace/bitstream/handle/10204/11592.

[4] Ali Muayad Jalil; Fadhil Sahib Hasan; Hesham Adnan Alabbasi: Speaker identification using convolutional neural network for clean and noisy speech samples. ieeexplore.ieee.org/abstract/document/9075461/authors#authors.

[5] L. Rabiner L, B. H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice-Hall International, 1993.

[6] S. V. Ault, R. J. Perez, C. A. Kimble, and J. Wang, "On speech recognition algorithms," International Journal of Machine Learning and Computing, vol. 8, no. 6, pp. 518-523, 2018.

[7] T. Kinnunen, and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, pp. 12-40, 2010.

[8] J-C. Liu, F-Y. Leu, G-L. Lin, and H. Susanto, "An MFCC-based text-independent speaker identification system for access control," Concurrency Computat: Pract Exper., vol. 30, no. 2, Article ID e4255, 2018.

[9] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," Expert Systems with Applications, vol. 90, pp. 250-271, 2017.

[10] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities" IET Biometrics, vol. 7, no. 2, pp. 91-101, 2017.

[11] S. Bose, A. Pal, A. Mukherjee, and D. Das, "Robust speaker identification using fusion of features and classifiers," International Journal of Machine Learning and Computing, vol. 7, no. 5, pp. 133-138, 2017.

[12] Z. Ma, H. Yu, Z. H. Tan, and J. Guo, "Text-independent speaker identification using the histogram transform model," IEEE Access, vol. 4, pp. 9733–9739, 2016.

[13] K. W. Godin, S. O. Sadjadi, and J. H. L. Hansen, "Impact of noise reduction and spectrum estimation on noise robust speaker
identification," INTERSPEECH, 2013, pp. 3656–3660.

[14] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in Proc. 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing, 2016, pp. 1–6.

[15]A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in Proc. of Interspeech, 2011, pp. 2341–2344.

[16] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "Avoiding Speaker Overfitting in End-toEnd DNNs using Raw Waveform for Text-Independent Speaker Verification," in Proc. of Interspeech, 2018.

[17] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baumwelch statistics for speaker recognition," in Proc. Of Speaker Odyssey, 2014.

[18] Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," IEEE Signal Processing Letters, vol. 22, no. 10,
pp. 1671–1675, 2015.

[19] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in Proc. of ICASSP, 2015, pp. 4814–4818.