International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Two Level Hierarchical Classification Model for Identification of Speaker using Deep Learning

Suyog Devi[1], Shubham Jadhav[2], Siddhi Deshmukh[3], Megha Wadvekar[4], Kshama Balbudhe[5]

[1,2,3,4]Student, Department of Information Technology, Pune Vidhyarthi Griha's College of Engineering and Technology and G. K. Pate (Wani) Institute of Management Pune, Pune-411009, Maharashtra, India.
[5]Professor, Department of Information Technology, Pune Vidhyarthi Griha's College of Engineering and Technology and G. K. Pate (Wani) Institute of Management Pune, Pune-411009, Maharashtra, India.

Abstract - *Speaker identification (SI) is a process of identification of human voice using different machine learning techniques. In the process of identifying speaker, extracting discriminative and salient features from speaker utterances is a critical task to accurately identify speakers. Currently available systems for speaker identification have utilized short-time features, Mel frequency cepstral coefficient (MFCC), shows the effectiveness in correctly identifying. However, the performances of these features decrease on complex speech datasets, and therefore, these features fail to accurately identify speaker characteristics. Due to this we are going to propose a Two-Level Hierarchal Classification Model for Speaker Identification using Deep Learning to improve the accuracy of text-independent speaker identification systems. Moreover, a two-level hierarchical classification model to identify speakers' gender and identity. The first level identifies the gender of the speaker (i.e., male or female), whereas the second level identifies the specific identity of the speaker. The extracted MFCC features were fed as input to a deep neural network (DNN) to construct the speaker identification model and for the identification of the gender same features are feed to Mel frequency cepstral coefficient-Gaussian mixture model (MFCC-GMM).*

Key Words: **Text-Independent, Speaker Identification, MFCC, Deep Learning, DNN, GMM**

1. INTRODUCTION

Speaker recognition is the process of automatically recognising who is speaking using speaker specific information from speaker's utterance. Speaker recognition is divided into two processes: speaker identification and speaker verification. Speaker identification involves the identification of a speaker utterance from a group of trained speaker utterances. In the process of identifying speaker, extracting discriminative and salient features from speaker utterances is a critical task to accurately identify speakers. Alternatively, speaker verification involves the process of determining whether a speaker of a test utterance belongs to a group of speakers. Speaker identification (SI) is the process of extracting the identity of a speaker by using a machine from a group of familiar speech signals. Speech signals Speech signals are powerful media of communication that always convey rich and useful information, such as emotion, gender, accent, manner of speaking, intonation style, pronunciation patterns, choice of vocabulary etc. Through such characteristics, machines can become familiar with the utterances of speakers, similar to humans. These unique characteristics enable researchers to distinguish among speakers when calls are conducted over phones although the speakers are not physically present. There is no individual with similar sound due to the differences in vocal tract shape, larynx sizes, and other sound production systems in the body (Ma, Zhanyu and Hong Yu [1]) (Selva Nidhyananthan, Senthur Selvi, [2]) ([Sreenivas Sremath Tirumala, Seyed Reza Shahamiri ,2017,[3]). In general, speakers can be identified using two different approaches: text independent and text dependent. In text dependent speaker identification system, the text being spoken during testing must be exactly the same as that spoken during the training of the system. By contrast, for the text independent speaker identification system, the speaker identification process does not depend on the text being spoken by the speaker (Jahangir, Rashid; TEh, Ying Wah,2020, [4]). Therefore, text independent approach is widely used for automatic speaker identification (ASI) system as it enables the use of different sample data for training and testing and independent evaluation of the system. Text independent speaker identification system imposes no boundary or limitation on the words or phrases that can be used for identifying the speaker. Since the speaker is provided with the freedom of using any utterance during testing irrespective of the utterance used during enrolment, this mode of speaker identification is comparatively complex and challenging. Speaker recognition has become an area of intense research due to its wide range of applications, including forensic voice verification to detect suspects by government law enforcement agencies, access control to different services, such as telephone network services, voice dialing, computer access control, mobile banking (G. S. Morrison, F. H. Sahito, 2016,[9]). Furthermore, speaker identification systems are extensively used to improve security, automatic speaker labeling of recorded meetings and personalized caller identification using intelligent answering machines. With the technology advancements in smart home sector, voice control and automation are key components that can make a real difference in people's lives.

The voice recognition technology market continues to involve rapidly as almost all smart home devices are providing speaker recognition capability today. In this paper, we propose real time two level hierarchal classification model to identify speakers' gender and identity. In first level, a deep neural network and different machine learning algorithms are used to construct an deep neural network to identify speakers based on speaker's utterance which has unique pattern and in second level MFCC-GMM model is used to identify the gender of the speaker (i.e., male or female) where the distance between different samples is calculated using maximum log likelihood ratio. The previous work on speaker recognition is done mostly on database-oriented environment. This work aims to compute results where utterance (speech data) is shorter and dialects in the real time. For the identification of speaker, we have created our own dataset whereas, for the training of gender classification model we have used publicly available AudioSet dataset.

2. LITERATURE SURVEY

Human voice is the most useful medium of communication due to its features of simplicity, uniqueness, and universality which is more beneficial than other biometric verification systems as it is easily accessible, easy to use and comparatively simpler for user to recognize speaker [1]. As speech recognition systems need to operate under a wide variety of conditions, therefore, such systems should be robust to extrinsic variations induced by a number of acoustic factors such as transmission channel, speaker differences and background noise (Selva Nidhyananthan, Senthur Selvi, [2]). In order to enhance classification performance, most of the speech applications perform digital filter, where the clean utterance estimation is learnt by passing noisy utterance through a linear filter ([Sreenivas Sremath Tirumala, Seyed Reza Shahamiri ,2017,[3]). In spectral subtraction of speech lots of valuable spectral features in the original speech can also destroy or can loss of some features. In order to overcome this issue, support vector machine (SVM) classifies speech features into various classes, aiming to minimize the difference among speech features of same class to enhance classification accuracy (Jahangir, Rashid; TEh, Ying Wah,2020, [4]). The major challenge in speaker identification is extraction of discriminative features that accurately characterize the speech signal and classification algorithms. In this regards, different features have been proposed namely, Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficient (LPCC), Power Normalized Cepstral Coefficient (PNCC), Spectral features, Time domain features and combination of these methods for enhanced recognition accuracy (G. Mujtaba, L. Shuib,2017, [5], S. S. Tirumala, S. R. Shahamiri, [6]). MFCC provide better identification accuracy with clean speech data. Also, to improve the identification accuracy delta (D) and delta–delta (D–D) coefficients of the features were used (X. Zhao and D. Wang,2013,[7]). However, Mel frequency cepstral coefficient use frequency bins to parametrize speech data and resolve the frequency linearly across audio spectrum using Fast Fourier Transform (FFT) or linear predictive coding. MFCC are widely used in audio analysis and speaker identification (Sadaoki Furui, 2010,[8]).

2.1 Different approaches of speaker identification

2.1.1 Histogram Transform-based Speaker Identification

A novel text-independent speaker identification (SI) method uses the Mel-frequency Cepstral coefficients (MFCCs) and the dynamic information among adjacent frames as feature sets to capture speaker's characteristics. In this dynamic information is utilized by cascading three neighbouring MFCCs frames together using super MFCCs features. The Probability Density Function (PDF) of these super MFCCs features is estimated in this Histogram Transform (HT) method, which generates more training data by random transforms to realize the histogram PDF estimation and recedes the commonly occurred discontinuity problem in multivariate histograms computing.

2.1.2 Noise Robust Speaker Identification Using RASTA– MFCC Feature with Quadrilateral Filter Bank Structure

Relative Spectra–Mel Frequency Cepstral Coefficients (RASTA–MFCC) were used for the feature extraction from the newly designed Quadrilateral filter bank structure and Gaussian Mixture Model–Universal Background Model (GMM– UBM) for improved text independent speaker identification under noisy environment. However, it uses neural network model which requires retraining of entire database when a new sample is added to it, but GMM–UBM model does not require retraining of entire database which leads to easier and faster processing. In this Quadrilateral filter bank structure with RASTA–MFCC feature and GMM–UBM modelling for speaker identification demonstrates supremacy over triangular and Gaussian filter banks.

3. METHODOLOGY

This section describes in detail the methodology used in identification of the speakers and gender classification. First, utterances of the several speakers were collected to create dataset for identification. Second, required features extracted from the collected speech to form a feature vector. Then, extracted features were fed as an input to deep neural network architecture to construct the speaker identification

model. After that for the gender classification model publicly available AudioSet dataset is used to train model. Finally, to check the performance of constructed model input was accepted in real time environment and existing speaker dataset. The details of these methods are discussed in following sections.

3.1 Feature Extraction

Features that are required were initially extracted using MFCC algorithm from input speech, followed by noise removal. The MFCC feature extraction technique basically subdivided into five phases such as frame blocking section in which the speech waveform is more or less divided into frames of approximately 30 milliseconds, windowing the signal which minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero, applying the DFT which converts each frame from the time domain to the frequency domain, taking the log of the magnitude, and then wrapping the frequencies where the signals are plotted against the Mel-spectrum, followed by applying the inverse DCT. Fig.1 shows the MFCC features of Male and Female and Table.1 shows the list of features which are required for the Identification and Classification.

Table -1: List of MFCC features

Label	MFCC Features
Meanfreq	mean frequency (in kHz)
Sd	standard deviation of frequency
Median	median frequency (in kHz)
Q25	first quantile (in kHz)
Q75	third quantile (in kHz)
IQR	interquantile range (in kHz)
Skew	Skewness
Kurt	kurtosis
sp.ent	spectral entropy
Sfm	spectral flatness
Mode	mode frequency
Centroid	frequency centroid
Meanfun	mean fundamental frequency
Minfun	minimum fundamental frequency
Maxfun	maximum fundamental frequency
Meandom	mean of dominant frequency
Mindom	minimum of dominant frequency
Maxdom	maximum of dominant frequency
Dfrange	range of dominant frequency
Modindx	modulation index

3.2 Deep Neural Network

An Artificial Neural Network (ANN) which consists multiple number of the hidden layers between the input and output layer is refers as Deep neural network. The hidden layers of DNN get the features from the input layer. The output layer computes the prediction of each class, and the results are applied to the input data through the series of functions of the hidden layers. Also, DNN classifier consists of neuron layers, which work using the activation function, called rectified linear unit (ReLU). Each neuron performs the normalization of the weighted sum by applying the transfer function to a simple weighted sum of the information it received from the input layer. To compute the output of the hidden layers and to return matrix of n elements, DNN uses the Neural transfer functions. However, the softmax neural transfer function is used differently in the output layer compared with that in the hidden layers to compute the predictions of each class. In this paper to identify the speaker, customized DNN was used as a classifier. The default DNN architecture consists of one input layer, one hidden layer, and one output layer. The customized DNN architecture which is used to classify speaker is consists of 1 input layer, 5 hidden layers, and 1 output layer. Input layer used 25 neurons, which are equal to the number of features extracted from speech of each speaker. Each hidden layer has 128 neurons, because the performance of neural networks depends on the number of neurons. A minimal number of neurons can contribute to underfitting, whereas a large number of neurons can lead to overfitting. Each hidden layer used the hyperbolic tansig transfer function to compute output from the input within the range of -1 and 1. However, the output layer used the softmax transfer function to compute the output values for multiclass classification. Later MFCC features were fed to the trained DNN to identify the speaker.

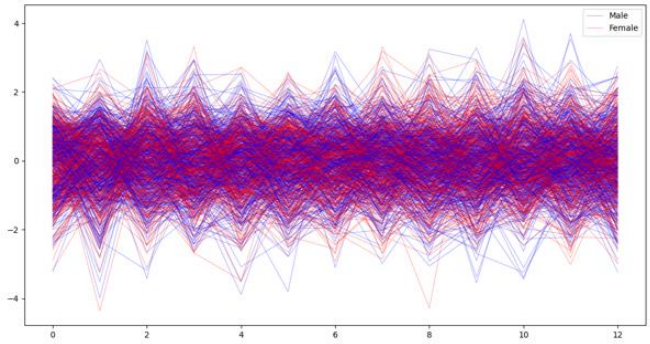

Fig-1. MFCC Features of Male and Female

3.3 Gaussian Mixture Model

A GMM is a parametric probabilistic density function model which assumes that all generated data points are from a mixture of a finite number of Gaussian distributions with unknown parameters. This model is used for the probability distribution of continuous measurements. Collecting each accent from the utterance of speech and finding the weight of the mean vector and mixture of each accent from speech is the most import aspect of any accent modelling. The iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation is used to estimate the maximum likelihood of the GMM model. Then maximum likelihood ratio of the training data and testing data is compared to recognize the speaker. The result will recognize as true, if the calculated likelihood ratio for training and testing data are close and less than a fixed threshold value.

3.4 Architecture

The Two-level hierarchical classification model approach was used to identify the speaker. Firstly, input speech is recorded using microphone, then it is stored into training dataset. After that Extraction of the feature is done using the MFCC Feature extraction method and the extracted features were fed to both DNN and GMM model respectively. In this, first level uses a deep neural network (DNN) and machine learning algorithms to identify speakers and second level, uses GMM model to identify the gender of the speaker (i.e., male or female).

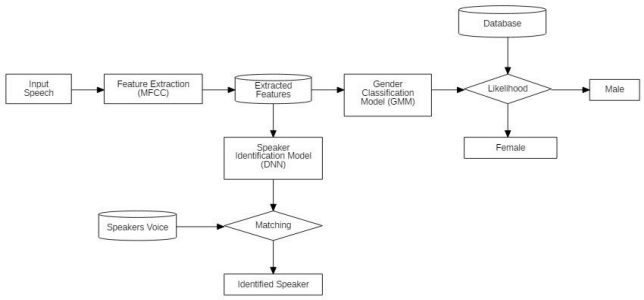

Fig-2. Architecture of System

4. RESULT

This section represents the results from the experiment performed in real time environment, in which the extracted MFCC features were fed to the DNN and GMM model. The overall accuracy of the text independent speaker identification model is 97% whereas, the accuracy of Gender classification model is 94%. The implementation of the project is done using the python3 idle and for the GUI we used the streamlit and results can be seen in the GUI. The GUI basically contains textbox and button named as Start Recording. In the textbox speaker has to enter the file name for training purpose and then press the button which will start the recording speech for 3 seconds and then it will display the "Training of model is completed" at the same time features from the speech were extracted and stored in excel sheet. After this, the GUI will contain button after the button is pressed it will record the voice of speaker and displays the output that who is speaker and Gender of the speaker. In the following Fig.3 (a) represents Training phase of the model whereas, Fig.3 (b) represents the Testing of the System.

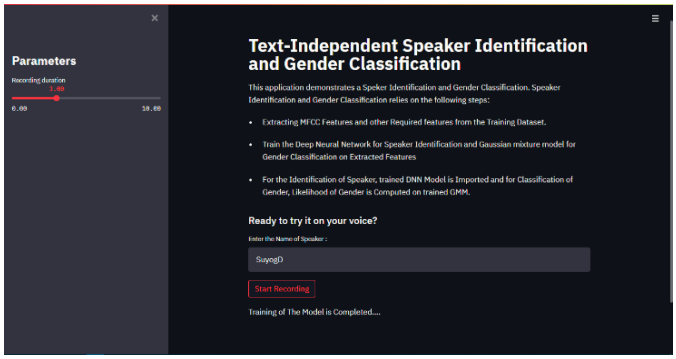

Fig-3(a). Training Phase

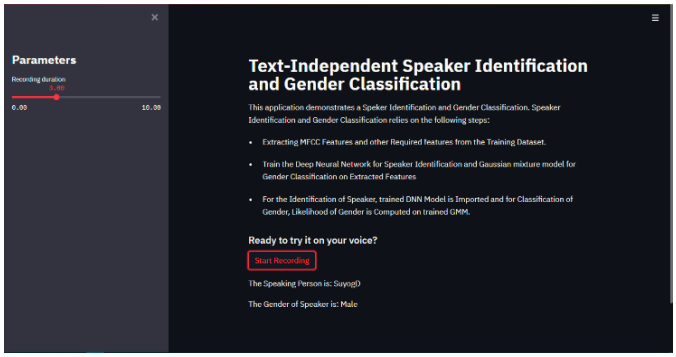

Fig-3(b). Testing Phase

5. CONCLUSION

In this paper we have proposed Two-Level Hierarchical Classification Model for Speaker Identification which got the accuracy approximately 96-97% for DNN and 93-94% for GMM in real time environment. The hierarchical classification approach is works in cascading style, where the first level identifies the speaker and the second-level

identifies the gender of speaker. This approach uses a MFCC, DNN and MFCC-GMM models which improves the text-independent speaker identification (SI) systems. Hence two-level hierarchical classification approach is used to obtained better results than the one-level classification model. Future work will consist of improving accuracy, further exploring speaker recognition as a sequence learning task using representation learning approaches like recurrent neural networks. Additionally, further investigations concerning the directions of the convolutional filters, pooling and stride are valuable to determine why 1D and 2D operations seem to perform comparable to each other, and how such filters could be interpreted in terms of auditory processing of the time-evolution of speech. Also, we will try to implement the system which identify the speaker when one or more speakers are interacting with system, identify speaker in noisy environment.

REFERENCES

[1] Ma, Zhanyu and Hong Yu. "Histogram Transform-based Speaker Identification." ArXiv abs/1808.00959 (2018): n. pag. Strunk, W., Jr., & White, E. B. (1979). The elements of style (3rd ed.). New York: MacMillan.

[2] Selva Nidhyananthan, S., Shantha Selva Kumari, R. & Senthur Selvi, T. Noise Robust Speaker Identification Using RASTA–MFCC Feature with Quadrilateral Filter Bank Structure. Wireless Pers Commun 91, 1321–1333 (2016) https://doi.org/10.1007/s11277-016-3530-3

[3] Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, Abhimanyu Singh Garhwal, Ruili Wang, Speaker identification features extraction methods: A systematic review, Expert Systems with Applications, Volume 90, 2017, Pages 250-271, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2017.08.015.

[4] Jahangir, Rashid; TEh, Ying Wah; Memon, Nisar Ahmed; Mujtaba, Ghulam; Zareei, Mahdi; Ishtiaq, Uzair; Akhtar, Muhammad Zaheer; Ali, Ihsan (2020), "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network," IEEE Access, 8(), 32187–32202. doi:10.1109/ACCESS.2020.2973541

[5] G. Mujtaba, L. Shuib, R. G. Raj, M. A. Al-Garadi, R. Rajandram, and K. Shaikh "Hierarchical text classification of autopsy reports to determine MoD and CoD through term-based and concepts-based features,'' inProc. Ind. Conf. Data Mining. Cham, Switzerland: Springer,2017, pp.209–222.

[6] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review,'' Expert Syst. Appl., vol. 90, pp. 250–271, Dec. 2017.

[7] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification,'' inProc. IEEEInt. Conf. Acoust, Speech Signal Process., May 2013, pp. 7204–7208.

[8] Sadaoki Furui, Chapter 7 - Speaker Recognition in Smart Environments, Editor(s): Hamid Aghajan, Ramón López-Cózar Delgado, Juan Carlos Augusto, Human-Centric Interfaces for Ambient Intelligence, Academic Press, 2010, Pages 163-184, ISBN 9780123747082, https://doi.org/10.1016/B978-0-12-374708-2.00007-3

[9] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, and C. G. Dorny, ``INTERPOL survey of the use of speaker identification by law enforcement agencies,'' Forensic Sci. Int., vol. 263, pp. 92_100, Jun. 2016.