A Comparative Study of Automated Speaker Recognition System

Navya.A
Research Scholar, School of Computer Science and Information Technology,
JAIN (Deemed to be University),
Bangalore, Karnataka, India

Dr.M.N.Nachappa
Professor and Head - IT implementation,
JAIN (Deemed to be University),
Bangalore, Karnataka, India

***Abstract*:**

Speaker Recognition uses characteristics extracted from the voices of users for computing the task of validating a user's identity. Two stages are involved in the whole process of speaker recognition. The first stage involves extraction of features from enrolled speakers. In the second stage, extracted features from the speech are used for recognition and to compare these to the speaker models. Since this technology can be put into use in several areas, lot of research is in progress as regards Speaker Recognition. However, automatic classification is already being executed with Machine Learning and Deep Learning classifiers such as RNN, CNN and Support Vector Machines amongst others. This paper, while taking up an extensive study of current research trends, applications and challenges, also looks into significant machine learning and deep learning models on Speaker Recognition.

Keywords: Speaker Recognition, Speech Recognition, Speaker Identification, Speaker Verification, Deep learning classifiers, Machine learning

I. INTRODUCTION

Speaker recognition refers to identifying a person from the features of his or her voice. Such recognition makes speech conversion easier in systems that have been customized to certain voices. As part of security screening, this procedure can also be used to check or verify a speaker's credentials. Speaker recognition dates back to around four decades. This technology makes uses of acoustic characteristics of speech which distinguish two individuals. These features are specific to both anatomy and learned behavioral patterns. The phrase **voice recognition** also means speaker recognition or speech recognition. This is also sometimes referred to as speaker verification or speaker authentication.

But, ***speaker recognition*** differs from *speaker-diarisation* (recognizing when the same speaker is speaking). Speaker verification indicates 1:1 analogue – that is to say, in this process, the voice of a speaker is compared to a particular template. But, speaker identification is a 1:N match since in this process a voice is compared against multiple parameters. There are two categories of speaker recognition. They are: (i) text-dependent, (ii) text-independent.

Speech is tool of communication for exchanging information or messages among human beings. We are bestowed with mechanisms of speech production and perception, but several times we fail to realize the complexity of the process involving speech production, pattern perception and the processing of auditory signals. Perhaps we get to understand the sophistication of this process only if a human being is replaced by a machine for the purpose of speech generation/speech recognition.

It is well-established that information is transmitted from one medium to the other by waves through a medium. In the same fashion, the intended message is carried by the speech signals along with characteristics of the language. The fundamental challenge faced in speaker recognition is to establish the user's claimed identity making use of characteristics in that voice of that particular user. While identifying with the system in the initial stages, the user utters a word or sentence which is captured with the help of a microphone. This captured voice is transformed into an analog signal representation.

The distinguishing features of this speech are extracted. During speech recognition, these features either match or do not match with the enrolled sample. A classifier will have to be made use of to determine the percentage of accuracy while comparing the speech utterances.

The following block diagram elucidates a general process of speech recognition.

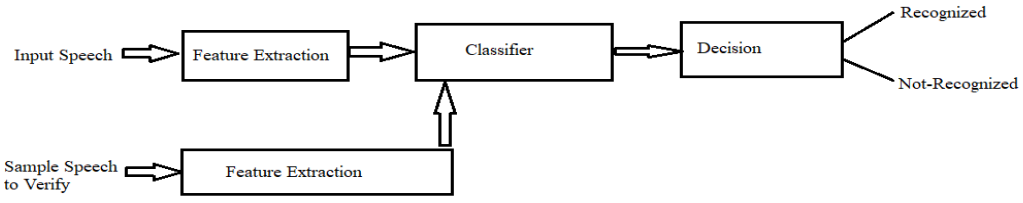

Fig. 1. A general Speaker Recognition System.

Input Speech: This is the first step of Speech Recognition Process. Here, the sample speech voice is recorded by using microphone and the same is the recorded voice is converted into signal. This signal is known as electrical signal and this will be converted to digital signal and is then passed to the preprocessing step.

Feature Extraction: In this process, the features of the voice signal are extracted using several feature extraction methods. Vocal tract acoustic characteristics are useful for this work; because of this features are extracted from the audio sample. This is performed during Training phase. However, in the testing phase, the speech sample is subjected to feature extraction method and the results are compared with those obtained during the training phase.

Classifier: In this step, different classification methods are applied to determine best recognition.

II. APPLICATIONS

We identify the speakers when we talk with each other face to face or even if speakers are in different places and not face each other. A blind person identifies the speaker solely on the basis of the vocal characteristics of the speaker on the other side. Even animals have this ability to identify their familiar people by using these characteristics. Speaker recognition uses speech sound for identifying someone based on the sample of that person's voice. A voice comprises a lot of acoustic information – referred to as voice prints. These voice prints are specific to each individual and it is these voice prints which help humans to differentiate individuals on the basis of voice. The keys, passwords are man-made traditional identification tools. But, voice prints are the biological characteristics. The benefit of this identification is that we may carry it with us wherever we go at any time, and it cannot be lost or leaked. The daily lives of people will be safer and more convenient with this system of recognition.

The following are some of the areas where speaker recognition system plays a vital role:

i. Credential authentication: Signature, Fingerprints, voice, facial feature are distinct features which helps differentiate an individual from the other. This mode of authentication is referred to as biometric person authentication. However, speaker recognition means establishing the identity of individuals on the basis of their voice. The voice comprises a lot of acoustic information, often called voiceprints. These voice prints are specific to each individual and it is these voice prints which help humans to differentiate individuals on the basis of voice. The keys, passwords are man-made traditional identification tools. But, voice prints are the biological characteristics. The advantage is that this speaker recognition application will provide human daily life safer. This will be convenient to use in future.

ii. Speaker recognition used for surveillance: Security agencies collection information from wide sources through various modes. One of these is the radio conversation. Speaker recognition plays a vigorous role in this intelligence gathering. The amount of huge data and certain filter mechanisms will be applied on these data on these data to fetch the required information [20].

iii. Forensic speaker recognition: Forensic methods play crucial role in establishing the commission or otherwise of a crime and thus are vital in criminal jurisprudence. By comparing a suspect's sample with a criminal's, a speech recognition system can determine who the suspect is. In case of crimes such as kidnap, terrorist activities or bribery, voice recordings are collected from telephone conversation or tapped recordings. Here the sample collected from suspect during investigation process and then this sample will be compared with crime scene happened time collected data [8].

iv. Security credential: Automatic speaker verification is a crucial safety precaution to protect the user's privacy and security such as passwords or keys, remote access to computers, cars, home appliances, voicemail, voice-dialling, phone-banking, telephone shopping, database access services, information services etc.[14][4].

v. Speech recognition: Sustained speech to text conversion and continuous automatic speech recognition are still a challenge owing to high fluctuation in speech signals. In this case speakers accents will be different, pronunciations are different for each person or region they live in, and speaker may speak in several different styles speed while talking that is also different [16].

vi. Multi speaker tracking: In this process the audio recording comprises voices of several speakers – such as a teleconference. Suppose, multiple speakers are involved in teleconference then particular person will be detected. In this detection process identify whether a known speaker involved in this teleconference or not [20].

vii. Personalized user interfaces: Voice-mail is the popular development. This will perform based on speaker recognition technique. There are certain issues are microphone could be different directions or may be different distances [20].

III. AVAILABLE DATASETS

Different datasets are being created for Speaker recognition. Some such datasets are listed below:

i. Speaker Recognition dataset: This dataset contains speeches of these prominent leaders; Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Tacher and Nelson Mandela. Acknowledge to American Rhetoric (online speech bank) making this speech available.

ii. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus: TIMIT has been developed with the joint efforts of several sites under sponsorship of the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-

ISTO). Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). While the speech was recorded at TI, it was transcribed at MIT, and maintained, verified, and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). TIMIT contains a total of 6,300 sentences, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the United States.

iii. Voxceleb1 audio wav files for India celebrity: These files contain subset of Voxceleb1 audio files for Indian Celebrities. This huge dataset has been prepared by Nagrani.

iv. LibriSpeech ASR corpus (clean): The LibriVox project's LibriSpeech corpus is a collection of roughly 1,000 hours of audiobooks. The majority of the audiobooks on this site are from Project Gutenberg. They offer the n-gram language models and the corresponding texts that have been taken from Project Gutenberg books, which have 977K unique words and 803M tokens.

v. Speaker Recognition Audio Dataset: This dataset consists of audio files for 50 speakers, each of which has a length of more than an hour. The data has been divided into 1-minute segments and converted to wav file around 16 KHz. For issues involving speaker recognition, this dataset can be used. This data set was extracted from Librivox and YouTube.

vi. The VoxCeleb1 Dataset: VoxCeleb1 compiles over a million quotes from 1,251 celebrities that have been taken from YouTube videos.

IV. RELATED WORK

Authors proposed several methods of Speaker Recognition in literature. All these methods comprise preprocessing, feature extraction, classification stages.

Sonal et.al [1] used Transfer Learning approach of Deep Learning. The spectral features are extracted for this work Spectral centroid, Spectral Rolloff, Speactral bandwidth, Zerocrossing rate, Mel-frequency Cepstral Coefficients (MFCC's), Chroma Feature. In this work they used Artificial Neural Network (ANN), Convolutional Neural Network (CNN). While performing CNN with two layer having filter 32, 64 and maxpooling performed. ReLu activation function used in this work and SVM used for classification.

Rania Chakroun et.al [2] provides an improved system for large population text-independent speaker recognition with short utterances. They used MFCC features, STZCR features. The i-vector modeling based on PLDA technique used for this work.

Qing Wang et.al [3] proposed adversarial training for multi-domain speaker recognition. The performance of speaker-recognition system reduces when there is no match amongst training and evaluation data. Domain adaptation is a common tool implemented to overcome the differences between domains so that the classifier trained on the source domain data generalizes well to the target domain data. Numerous such adaptation approaches have been used to solve the domain mismatch problem in speaker recognition. They used DAC13 dataset as the evaluation dataset. The proposed methods are Multi-domain adaptation by adversarial training and Adversarial training for multi-domain speaker recognition. In Multi-domain adaptation by adversarial training, model can be decomposed into three parts as a feature generation network, a speaker classifier, a domain discriminator. In Adversarial training for multi-domain speaker recognition, the data used to train the PLDA are usually supposed to share the same distribution with the evaluation data. This data is defined as target domain.

They used multiple source and target data to train the multi-domain adaptation neural network (MDANN). The object of the proposed adversarial training for multi-domain speaker recognition was to eliminate multi-domain mismatches among different subsets in speaker recognition. Compared to previous domain adaptation studies, they take inner dataset variance into consideration and extract multi-domain invariant and speaker-discrimination representations for the speaker recognition. The results suggested the effectiveness of the multi-domain-invariant and speaker-discriminative speech representations in speaker recognition.

Seong Min Kye et.al [4] presented supervised attention for speaker recognition. They used D-vector based feature. They used Convolutional Neural Networks and Recurrent Neural Networks are used. They used self-attentive pooling (SAP), attentive statistics pooling (ASP), learnable dictionary encoding (LSE) and Cross attentive pooling (CAP).

Maros Jakubec et.al [5] describes speaker recognition with ResNet and VGG networks. Several deep neural network (DNN) topologies have been introduced recently for the purpose of automatic speaker recognition (SR) tasks. Residual networks (ResNet) are one of the recently developed architectures which help in solving solve the vanishing gradient problem. They are extension of VGG nets (Visual Geometry Group), the deep concept of convolutional neural networks (CNN) with fixed-sized kernels. The proposed SR system is tested on the VoxCeleb1 database for two types of tasks (1) the text-independent speaker identification (2) verification (SV).

They made experiments with the below mentioned deep CNN architectures: VGG-M, ResNet-18 and ResNet-34. 257-dimensional spectrograms were used as in input to DNN in the said experiment. The spectrograms were generated by Short-Term Fourier Transform (STFT) with 25ms Hamming window duration and 10ms shift. Then DNN setup performed. Two metrics were used for experiments. The EER metric was used for SV task, and Top-1 and Top-5 precision were used in the case of SI. Environmental noise makes this task even more challenging. In this study, they compare two deep CNN architectures for SI and SV tasks. The performance of the ResNet solutions is compared with VGG neural networks. A neural network based on ResNet-34 achieved the best accuracy for both SI and SV tasks.

P. Rama Koteshwara Rao et.al [6], introduced the purging of silence for reliable speaker identification. Silence is eliminated during feature extraction, and pitch and pitch-strength metrics are also taken. The methods for feature extraction are multi-linear principle component analysis (MPCA), zero-crossing rate (ZCR), and endpoint detection algorithm (EDA). SVM, or support vector machines, are used to classify speakers.

Qinghua Zhong et.al [7], Text-independent speaker recognition based on adaptive course learning loss and deep residual networks was presented here. The log filter bank and MFCC are used for feature extraction. The Ghost VLAD and AM-Softmax are the foundation of a potent deep network for speaker recognition that has been proposed. A residual (ResNet) and a convolutional attention statistics pooling (CASP) layer made up the deep residual network.

A Fuzzy Approach to Statistical Models in Voice and Speaker Recognition was described by Dat Tran et.al [11]. This work proposes a unified fuzzy approach to statistical models for speech and speaker recognition. This fuzzy, EM method serves as the foundation for the fuzzy algorithms for Hidden Markov Models, Gaussian Mixture Models, and Vector Quantization.

An Innovative Method on Biometric and Voice Detection Using Recurrent Neural Networks was given by Shaik Mohammad Ayesha et al. [12]. This paper's primary goal is to demonstrate how Recurrent Neural Networks (RNNs) can be effectively used to address issues with speech recognition and electrocardiogram (ECG)-based biometrics verification. The main architecture in this paper is the deep bidirectional LSTM.

Qihang Xu et.al [13] proposed Speaker Recognition Based on Long Short Term Memory Networks. With the use of acoustic characteristics in the speaker's voice or audio for recognition, they are executing the task of recognizing the speaker's identity in this job. By simulating the speaker's vocal tract and the auditory system of a human, work in the fields of human daily life and the military can be improved. Acoustic feature extraction, which includes audio processing, LPC feature extraction, log-mel feature extraction, and the fusion of the two features, is the initial step in this speaker recognition procedure. The construction of neural networks and the choice of loss function and LSTM neural networks make up the next section. For this, speech recognition independent of text is appropriate. Deep neural networks may be trained to recognize speakers using acoustic data. Following that, the three preprocessing procedures are audio framing, windowing, and eliminating the unvoiced audio. The acoustic qualities of the human ear are perfectly suited to LPC. the feature level fusion algorithm based on LSTM was then carried out. the LPC speech characteristics and the log-mel spectral feature can be integrated with this. There has been a noticeable improvement in speech detection accuracy.Recognition is appropriate. Deep neural networks may be trained to recognize speakers using acoustic data.

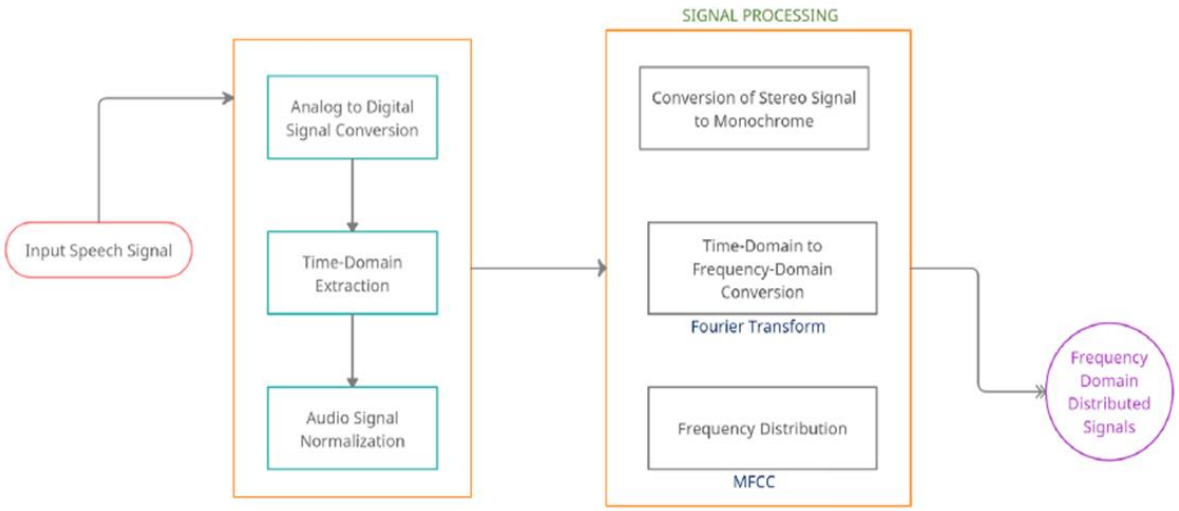

Fig. 2. Processing of Speech (Audio) Signals [21].

Framework for speaker recognition using machine learning models:

Preprocessing: Preprocessing of speech data is a crucial step in the development of speaker recognition.

The conversion from analog to digital consists of the below two processes:

i. Sampling is a technique used to turn a signal that changes over time, s(t), into a discrete succession of real numbers, x.(n). The term sampling period (Ts) refers to the space of time between two subsequent discrete samples. The inverse of the sampling time is the sampling frequency (fs = 1/Ts). The three most popular sampling rates are 8 kHz, 16 kHz, and 44.1 kHz. One sample is taken every second at a sampling rate of 1 Hz, hence higher sampling rates imply better signal quality.

ii. Quantization is the process of substituting an approximation for each real number produced by sampling in order to achieve a finite precision. (Defined within a range of bits). The representation of a single quantized sample is typically done using 16 bits per sample. As a result, the signal range for raw audio samples is typically between -215 and 215;

however, during analysis, these values are standardized to the range between (-1, 1) for easier model validation and training. Bits per sample are the only unit used to express sample resolution. Bits per sample are the only unit used to express sample resolution.

Preprocessing consists of removal of silence/unvoiced part of the speech. This silence and unvoiced does not give any information.

Feature Extraction: The speech is represented to analyze the features in the speech sample. During feature specific information can be obtained for speech signal. Later this gained information can be used to create a speaker model. The speaker model is used to find the similarity between the speech samples. The features can be extracted for speaker verification is Pitch analysis, Mel-frequency cepstral coefficient (MFCC), linear predictive coefficient (LPC), linear predictive cepstral coefficient (LPCC).

Classification models: This includes classification performed on the features which are extracted. The values obtained on features are applied on classification. The different classifiers are - Support Vector Machine(SVM), Vector Quantization (VQ), Gaussian Mixture Model (GMM), Dynamic Time Wrapping (DTW), Hidden Markov Model (HMM), Artificial Neural Network (ANN).

The general framework of speaker recognition using Deep learning methodologies:

Machine learning played vital role in research till few years ago. During machine learning audio signal could be analyzed with the help of traditional digital signal processing techniques to extract features. In recent years, Deep Learning becomes more and more useful because it has been observed that the traditional audio processing changed. Without manual generation of features now becomes standard data preparation. Sonal Ganvir et al. [1] built and trained two models, an Artificial Neural Network (ANN) and a Convolutional Neural Network (CNN), are using a deep learning approach, and thereafter compared the outcomes.

Maros Jackubec et.al [5] experiments on text-independent speaker recognition in the wild with two deep Convolutional DNN architectures: ResNet and VGG.

To increase the ability of log filter bank feature vectors to be recognized, Quinghua Zhong et al. [7] performed an analysis on a deep residual network model. A convolutional attention statistics pooling (CASP) layer and a residual network (ResNet) layer made up the deep residual network. The problem was reported using an adaptive curriculum learning loss (ACLL), and two distinct margin-based losses, AM-Softmax and AAM-Softmax, were introduced. The SincNet architecture was suggested by Dan Oneata et al. [9] as a method for the speaker recognition challenge. Recurrent Neural Networks (RNN) was employed by Shaik Mohammad Ayesha [12] to raise a production solution problem in speech recognition and electrocardiogram (ECG)-based biometric identification. A strong model for sequential data is RNNs. Several RNN architectures and parameter sets were assessed.

A speaker recognition technique based on Long Short-Term Memory Networks is proposed by Qihang Xu et al. [13]. (LSTM). The Softmax loss function is used to classify the d-vector output from the LSTM network. Recurrent neural networks were utilised by Sruthi Vandhana T et al [14] to analyse sequential data. In order to categorise the emotions of the speech, Balaji Dharamsoth et al. [15] experimented with a number of deep learning algorithms, including RNN and Bi-RNN, on voice data as well as its textual features. For automatic non-stop speech recognition, Prof. Kirti Rajadnya et al. [16] used deep learning methods along with Deep Neural Networks (DNN) and Deep Belief Networks (DBN).

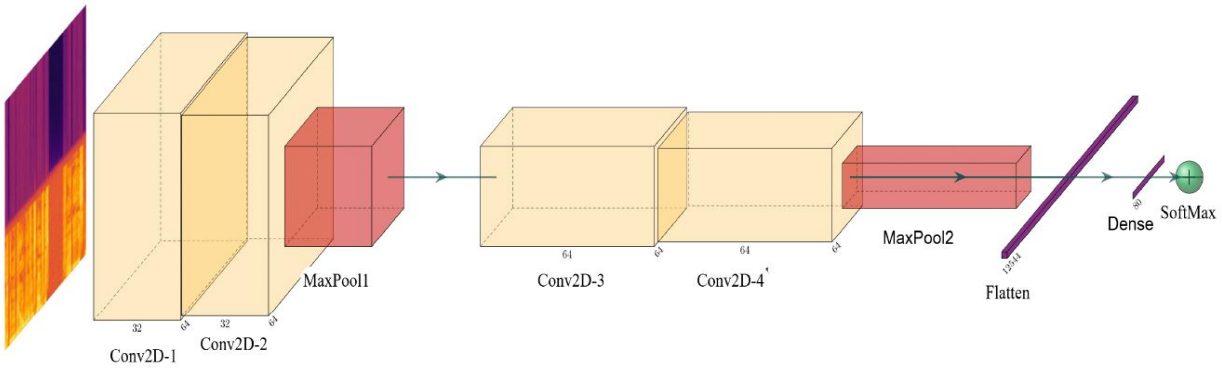

Fig. 2: CNN Model Architecture

Fig. 3. CNN Model Architecture [1].

V. CHALLENGES

a) The most important challenge is when we collect audio sample there can be inconsistency in their audio quality. This is due to channel mismatch. The data for training and testing could be collected through different channels - using different recording devices and different scenarios.

b) Difficult to handle audio of multiple speakers. This is known as overlapping. This is very difficult to identify who is speaking?

c) In certain situation where a person is speaking but the recording device i.e., microphone could be away from person.

d) Identify which feature extraction method to use to extract features from speech.

e) Identify the most appropriate model for speaker recognition is still a matter of confusion.

VI. CONCLUSION

This paper provides information about various speaker recognition technologies, applications and datasets. Machine Learning and Deep Learning provides various ways of leading researchers to discover solutions to problems.

REFERENCES

[1] Sonal Ganvir, Dr. Nidhi Lal "Automatic Speaker Recognition Using Transfer Learning Approach Of Deep Learning Models". Proceedings of the Sixth International Conference on Inventive Computation Technologies[ICICT 2021] IEEE Xplore Part Number:CFP21F70-ART; ISBN:978-1-7281-8501-9

[2] Rania Chakroun, Mondher Frikha "An Improved System for Large Population Text Independent Speaker Recognition with Short Utterences". 2021 International Wireless Communications and Mobile Computing (IWCMC)|978-1-7281-8616-0/21/$31.00©2021 IEEE|DOI:10.1109/IWCMC51323.2021.9498981

[3] Qing Wang, Wei Rao, Pengcheng Guo, Lei Xie "Adversarial Training for Multi-domain Speaker Recognition" 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) /978-1-7281-6994-1/20/$31.00©2021 IEEE/DOI:10.1109/ISCSLP49672.2021.9362053

[4] Seong Min Kye, Joon Son Chung, Hoirin Kim "Supervised Attention for Speaker Recognition". 2021 IEEE Spoken Language Technology Workshop (SLT)| 978-1-7281-7066-4/20/$31.00©2021 IEEE|DOI:10.1109| SLT48900.2021.9383579

[5] Maros Jakubec, Eva Lieskovska, Roman Jarina "Speaker Recognition with ResNet and VGG networks". 2021 31st International Conference Radio elecktronika(RADIOELEKTRONIKA)|978-1-6654-1474-6|20|$31.00©2021 IEEE|DOI:10.1109/RADIOELKTRONIKA2220.2021.9420202

[6] P.Rama Koteswara Rao, Sunitha Ravi, Thotakura Haritha. "Purging of Silence for Robust Speaker Identification in Colossal database". International Journal Of Electrical and Computer Engineering (IJECE) Vol.11, No.4, August 2021, pp.3084~3092 ISSN: 2988-8708, DOI:10.11591|ijece.V11i4.pp3084-3092

[7] Quinghua Zhong, Ruining Dai, Hanzhang, Yongsheng Zhu and Guofu Zhou "Text-Independent Speaker Recognition Based on Adaptive Course Learning Loss and Deep Residual Network". EURASIP Journal on Advances in Signal Processing, Zhong et.al.EURASIP Journal on Advances in Signal Processing (2021) 2021:45 https://doi.org/10.1186/$13634-021-00762-2

[8] Habib Saifuddin, Joko Sarwono, Miranti Indar Mandasari "Within- Class Covariance Normalization (WCCN) for Channel Normalization Based Of Indonesian Speaker Recognition System".2021 International Conference on Instrumentation Control and Automation (ICA) Bandung, Indonesia, August 25th-August 27th, 2021

[9] Dan Oneata, Lucian Georgescu, Horia cucu, Dragos Burileanu, Corneliu Burileanu "Revisiting SiincNet: An Evaluation Of Feature and Network Hyper parameters for Speaker Recognition" 978-9-0827-9705-3 EUSIPCO 2020

[10] Md.Monirul Islam,Fahim Hasn Khan, Abdul Ahsan Md.Mahmudul Haque "A Novel Approach for Text-Independent Speaker Identification Using Artificial Neural Network". International Journal Of Innovative Research in Computer and Communication Engineering, Vol.1, Issue 4, June 2013.

[11] Dat Tran, Michael Wagner and Tongtao Zheng "A Fuzzy Approach to Statistical Models in Speech and Speaker Recognition". 1999 IEEE International Fuzzy Systems-Conference Proceedings. August 22-25, 1999, Seoul, Korea.

[12] Shaik Mohammad Ayesha "An Innovate Approach on Biometric and Speech Recognition using Recurrent Neural Networks (RNN)" International Journal of Computer Science and Mobile Computiong. A Monthly Journal of Computer Science and Information Technology. ISSN 2320-088X, IMPACT FACTOR:7.056

[13] Qihang Xu, Mingjiang Wang, Changlai Xu, LuXu "Speaker Recognition Based on Long Short Term Memory Networks" 2021 IEEE 5th International Conference on Signal and Image Processing.

[14] Sruthi Vandhana.T, Srivibhushanaa.S, Sidharth.K, Sanoj.C.S "Automatic Speech Recognition Using Recurrent Neural Network" International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181, Vol.9 Issue 08, August 2020.

[15] Balaji Dhramsoth, Zia Uddin Mohammed, Dr.Surash Pabboju "Speech Emotion Recognition Using Deep Neural Networks" International Journal for Research in Applied Science and Engineering Technology (IJRASET). ISSN:2321-9653; IC Value: 45.98, Impact factor: 7.429 www.ijeraset.com

[16] Prof. Kirti Rajadnya, Purva Ingle, Vinit Tawsalkar, Shivani Teli "Speech Recognition Using Deep Neural Network Neural (DNN) and Deep Belief Network (DBN)" International Journal for Research in Applied Science & Engineering Technology (IJRASET). ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue V May 2020- Available at www.ijraset.com

[17] https://en.wikipedia.org/wiki/Speech_production

[18] https://in.images.search.yahoo.com/search/images;_ylt=Awrx1snR9C5i6EsAUgK7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3BpdnM-?p=human+vocal+tract+image&fr2=piv-web&type=E211IN826G0&fr=mcafee#id=15&iurl=https%3A%2F%2Fwww.translationdirectory.com%2Fimages_articles%2Fwikipedia%2Fhuman_vocal_tract.jpg&action=click

[19] S. K. Singh (Roll No. 03307409) Supervisor: Prof P. C. Pandey, FEATURES AND TECHNIQUES FOR SPEAKER RECOGNITION, M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03

[20] Nilu Singh, R.A.Khan, Raj Shree "Applications of Speaker Recognition" International Conference on Modelling, Optimisation and Computing(ICMOC 2012) Procedia Engineering 38(2012) 3122-3126

[21] https://towardsdatascience.com