International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Designing Highly Secured Speaker Identification with Audio Fingerprinting using MODWT and RBFNN

I. Rajani Kumari[1], Dr. Kalyan Babu[2]

Abstract: The process of matching Speech data with database entries is known as Speaker identification. This research processes a Novel technique to speaker recognition by merging neural networks with audio fingerprinting. In comparison to classical models the radial basis function neural network(RBFNN) combined with the hidden markov model(HMM) provides superior results. To facilitate the denoising process for both spoken and unvoiced speech, including ambient noise, the speech data is divided into low and high frequency segements.The maximal overlap discrete wavelet transform (MODWT) is used to denoise the split signals. When it comes to boundaries MODWT is more resilient. In the presence of white noise the Power normalized cepstral coefficients (PNCC) will provide more accurate results than Mel cepstrums.The results are accurately observed using this suggested strategy in relation to SNR.

Index Terms: *RBFNN-PNCC, HMM-LSTM,MODWT,Speaker identification, Audio fingerprinting.*

I INTRODUCTION (literature survey)

An audio signal is known for its ability to link to metadata example is song and artist name etc. these are having audio fingerprinting or content based metadata of audio that is stored in database. Various recordings identified as same audio content, and there is a necessity to compare the fingerprint with huge collection of data stored in the data base.more efficient implementation is using radial basis functions in combination with power-normalized cepstral coefficients leads to less noise an ideal finger printing system full fill several requirements it should be identify accurately regardless of level of distortion or compression or interference in the transmission of channel. Having a proper definition of an audio fingerprint we now focus on the different parameters of an audio fingerprint system. After Radial basis function neural network has a different structure than other neural structures. In other neural network structures there are many hidden layers which lead to nonlinearity in processing of input audio data. Here in the RBFNN contains one input layer and one hidden layer and one output layer (Ali bounassif,Et.Al,2022). This enables the linear transformation following by non linear transformation to achieve higher dimension in the hidden layer. This is a three layer network used to solve both Classification and Regression Problems.

Recognition accuracy which is one of the most challenging contemporary issues tends to decline significantly when the test environment differs from the training environment or when there are disturbances in acoustical environment such as

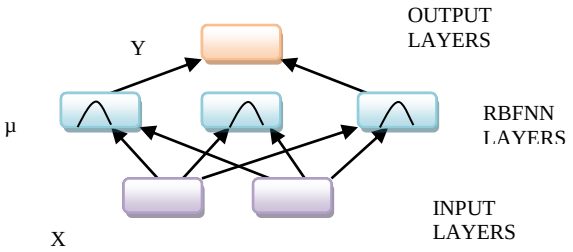

Fig1: Radial Basis Neural Network Layers

Noise distorted channels varied speakers, and reverberation. Over the years numerous algorithm have been introduced to tackle these challenges. While many conventional noise compensational algorithms have shown considerable improvements in accurately recognizing speech amidst quasi-stationary noise, theyoften fail to deliver significant enhancements in more demanding environments characterized by transient disturbances such as interference from a single speaker or background music. These poses a limitation for many existing systems designed to address these challenges. The field of automatic speech recognition and speaker identification relies on either Mel frequency Cepstral coefficients (MFCC) or perceptual linear prediction (PLP) coefficients as the basis for analysis(.B.Vimal,Et.Al,2021). Recently spectro temporal features have also been introduced and have shown promising results. Researchers have found that two dimensional filters can approximate the auditory cortex's spectro-temporal response fields, leading to different approaches for extracting features in speech recognition. This paper involves development of a new set of features called power normalized cepstral coefficients

[1]*Research Scholar of GITAM UNIVERSITY, Department of ECE, Vizag, AP, INDIA, & Assistant Professor of Geethanjali college of engineering and technology, Department of ECE,Hyderabad,INDIA.*
rajanikumari.ece@gcet.edu.in
rinapago@gitam.in
[2]*Professor, Departmentof ECE,GITAM ,Vizag,AP,INDIA.*
kkillana@gitam.edu

(PNCC) for speech recognition purposes (Xuechen Liu,Et.Al,2021). All the previous attempts at PNCC processing showed promise;they could not be easily applied to online applications without considering the entire sentence in advance. Furthermore earlier implementations of PNCC did not account for the impacts of temporary masking. However the current paper relates to the revised version of PNCC processing the addresses these limitations.This revised approach enables it to achieve superior accuracy(Ishwar chandra yadav ,Et.Al,2021) in recognizing various noise and reverberation conditions. Moreover the new Method utilizes features that can be computed in real time using online algorithms. The benefits offered by PNCC are particularly pronounced when the speech recognition system is trained on clean speech and tested in environments noise and /or reverberation. In cases where large data bases of speech with diverse environmental conditions are used for training and testing,PNCC processing also tends to outperform MFCC and PLP processing(Shalbbya Ali,Et.Al,2020),to a lesser extent for systems that are taught and tested using sizable speech databases with variety of environmental circumstances,though the results are less pronounced.Other feature extraction methods have also integrated elements of PNCC processing.we discuss the larger motivations and general structure of PNCC, we go into some detail on the main component in the processing we compare the recognition accuracy offered by PNCC processing that of alternative processing schemes under a range of scenarios and take into account the effects of different PNCCC components on these outcomes.

Broadcasting for audio fingerprint for acoustic playlist of the radio generation and web broadcast for other applications of various collection verification of programs is the most well known application people listen to broadcast monitoring system based on fingerprint consist of several monitoring sites and from where the servers located, i.e. central site, finger prints are extracted from all of the broadcast channels however, verification and metering of people broadcast monitoring is still a manual process(Jose juan garcia hernandez,Et.Al.,2019). The central site gather fingerprints from observational sites as well as a from a finger print server with a sizable finger print to create a play list of every channel that will be broad cast. Coupled audio is standard for consumer applications and is permanently coupled to supporting information; song identification is one example. This is done by a variety of business because to signal degradation caused by radio stations various processing during broadcast, the route between the loudspeaker and the microphone and speech coding during transmission via mobile networks. A universal linking system for audio material is highly helpful for many uses of audio, the additional examples with radios with identification or fingerprint buttons and speech recognitions in mobile devices. The efficiency of that particular fingerprint technique is known to the general public(Annapurna p Patil,Et.Al.,2021). Using a legal file sharing site one may use more sophisticated audio filter to remove copyrighted content or music of various types premium music and music that is prohibited. Audio fingerprinting can be seen as a dubious technological advancement if you take into account how consistently a music appears in search results when you use the accurate meta information of a fingerprint taken from a database. What is really being downloaded and what it claims are the same fingerprints. (Sungkyun chang,Et.Al.,2021). The Normalized hamming distance is

$$f(x,y) = 1/\mathrm{d}\sum_{n=1}^{d}(\mathrm{x(n)} - \mathrm{y(n)}), \mathrm{n} = 1,2,\ldots.\mathrm{d}$$

The recall and precision rates are compared and identified by

	Truth=0	Truth=1
Prediction=0	True Negatives	False Negatives
Prediction=1	False Positives	True Positives

$$\mathrm{FPR} = \frac{False\ Positive}{True\ Nagitive + False\ Positive}$$

$$\mathrm{TPR} = \frac{Recall\ True\ Positive}{True\ Positive + False\ Nagitive}$$

A Viterbi algorithm is central component of HMM-based speech recognition systems. The viterbi algorithm determines the optimal alignment between an input speech model and it self via dynamic programming.The main idea is to apply an update algorithm to improve the segement location after creating an initial population of segmentation vectors in the solution search place.A number of approaches to the representation of the particles and the two methodologies of the segmentation are investigated.The conventional segmentation approach is the first technique and its aims to optimize the probability function for each competing auditory model independently.Next in the technique a common tied segmentation and a global segmentation tied between many models, with the system attempting the maximum likelyhood.The findings demonstrate that these factors have a discernible impact on locating the global optimum while preserving system accuracy.(yang sung-hyun Et.Al.,2018) Tested on separate word recognition and phone categorization tasks. The concept demonstrate a note worthy performance in terms of both computaional complexity and accuracy.

The foudation of a number of effective methods for acoustic modelling in speech identification and recognition is the hidden markov model (HMM).The

MODWT(Maximal overlap discrete wavelet transform) speech signal denoising technique is evaluated and testedtoday a lot of research is being done on how to ensure that the voice signal can be understood in loud surroundings by isolating it from the background noise.However because of challenges in eliminating the background noise recovering the original speech from the noisy signal with little distortion is a problem.The signal in environmental noise situations can be interfered by a multitude of ways.This study uss several wavelet filters to experimentally examine the performance of various discrete wavelets transform methods(Selma ozaydin,Et.Al.,2018)The MATLAB environment is used to carry out the analytical carry out speech sounds with various ambient background noises as the input noise speech signal.Wavelet analysis was using during the testing to separate the noise signal from these message signal.The wavelet coefficients are generated by breakingdown the noisy voice signal input using various thresholding techniques.SNR measurements between the noisy and smooth output signals are measured in order to compare the reconstruction speech.

The primary factors contributing to this achievement are the models capacity for the speech phenomenon analysis and its accuracy in real world speaker identification applications (Rafizah mohd hanifa,Et.Al.,2021) The HMM's convergence and trust worthy parameter training process is another important feature vectors serves as the representation for speaker utterences.As such segmenting a speech sequence of feature vector serves as the representation for spoken utterences.As such segementing a speech sequence into stationary states is necessary for statistical evaluation of the speech sequence.a finite state machine is an HMM model.A single guassian or a multi model guassian mixure may be used to model each state.Since speech observations are continuous and finite state machines included.transitions between any state and its neighbours as well as to it self are permitted under this left-to right topology.Using adequate training data sets maximum likelihood based or discriminative based training techniques are typically used to estimate the parameters of HMM models during the training phase.The simple speech signal is

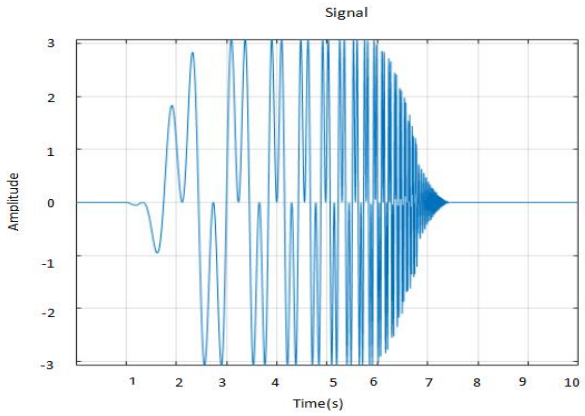

Fig1: Simple signal

The capacity of long short term memory(LSTM) to describe and predict non linear time variant system. Dynamics has lead to extensive research into LSTM in recent years. The study provides through analysis of network designs and LSTM cell derivatives that are currently available for time series prediction. It is suggested to characterized LSTM with interacting cell states and LSTM with optimal cell state representations.(Benjaminlindemann,Et.Al.,2021)The examined methods are assessed in light of the specified criteria that are necessary for a precise time series prediction. These include the ability to make multi model and multi step forward predictions. The behavior of short and long term memory and the corresponding error propagation. The best options to meet the requirements are sequence to sequence networks with partly conditioning, which perform better than bidirectional or associative networks. Neural networks have been used extensively to simulate and forecast the dynamics of intricate systems. Although the is a variety of network types the modeling accuracy is largely influenced by how well the network architecture fits the problem under consideration. which have been applied to dynamic system modeling in a variety of fields like including manufacturing autonomous systems, energy consumption, image processing, speech recognition and identification. Setting up prediction models based on time series data or data sequences to forecast nonlinear time variant system outputs is the unified goal across all examined challenges.

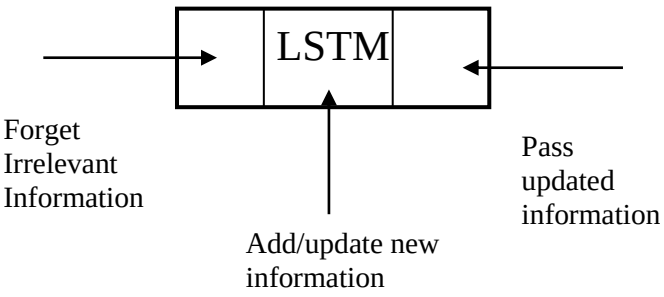

Fig2:LSTM Architecture

II PROPOSED SYSTEM BLOCK DIAGRAM

The input voice signal will be received by the filter component of the proposed system,which will seperates into low and high frequency next the voiced,unvoiced and silence data will be divided into groups and described baased on their frequency features.The adaptive range at low frequencies was then increased using sophisticated algorithms,which reduced the noise.The original noise free speech will eventually blend in with the noise free signal.the low signal will be strengthened and the high frequncy signals noise will be removed by applying the MODWT approach.

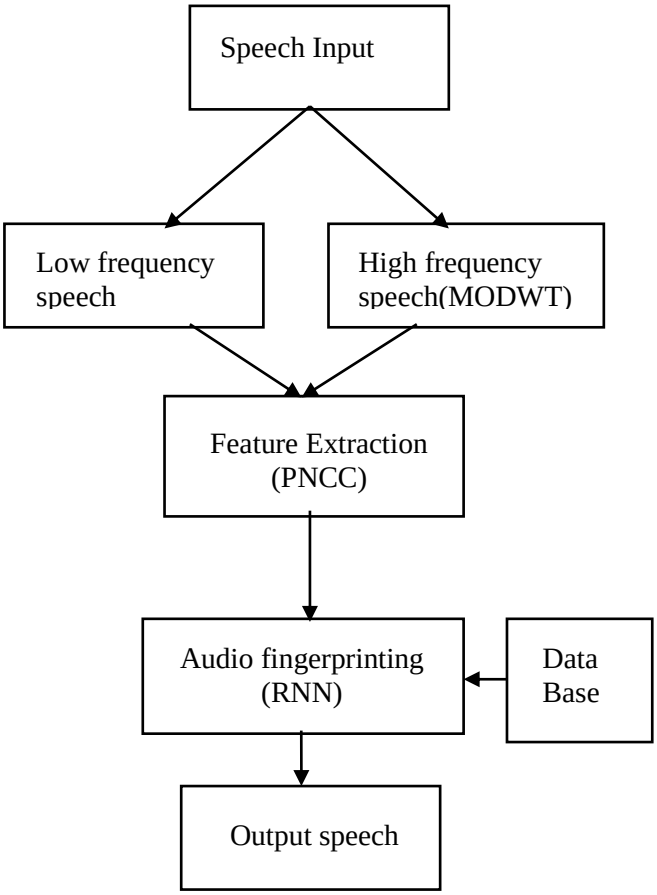

Fig3: Proposed Block diagram

The audio fingerprinting block that follows has the voice signal processing done to it.Using the power of the signal increases information accuracy.The cepstral coeficients and the comparision of mel frequency to PNCC as,

MFCC	PNCC
Mel filter bank(Triangular)	Gammatone filter bank
Logarthmic linearity	Power linearity
Less accurate	Better accurate in presence of white noise

Table1:Differences of MFCC and PNCC

In PNCC the procedures of mean power normalization and temporal frequency normalization Have been completed.Coefficient generation is the ultimate result of the PNCC process,which begins with the generation of the power function linearity and ends with the mean normalization via the discrete cosine transformation.The rate is higher and produces a more accurate signal when comparing the results with SNR to accuracy.

The Proposed Algorithm states that

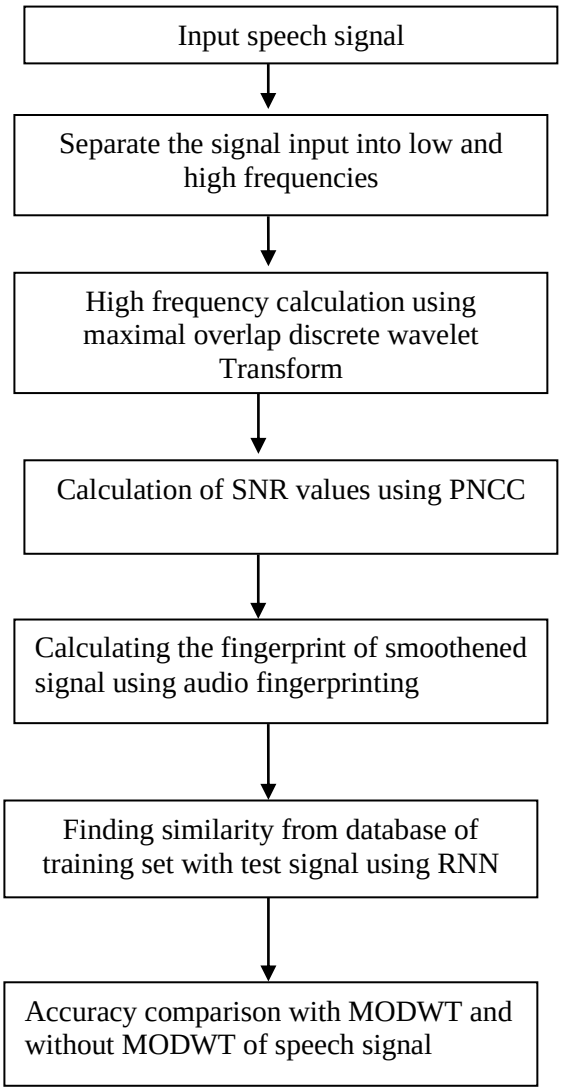

Fig4:Proposed Algorithm

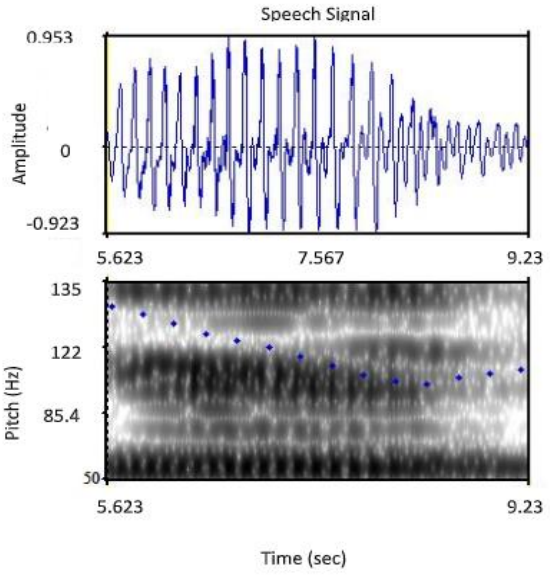

Fig5:Pitch calculation of signal

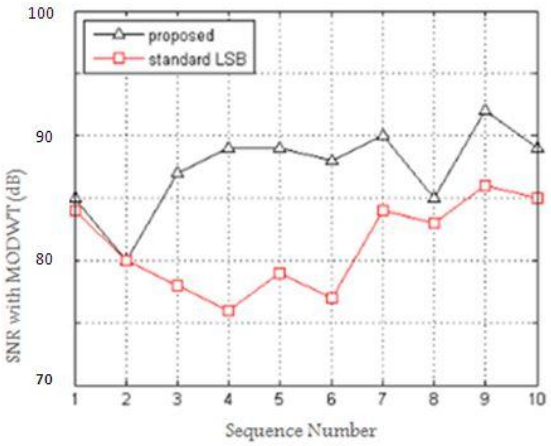

Fig6:SNR values of 10 sequences using MODWT in dB

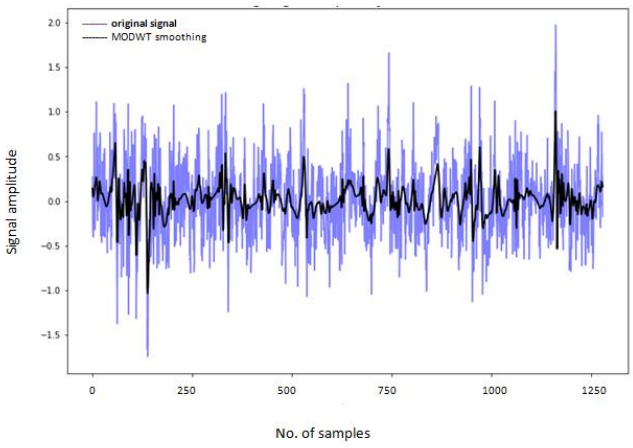

Fig7:MODWT smoothing for 1250 samples and comparision with original signal

III CONCLUSION

Speaker identification is the process of comparing speech data with database records. This study combines audio fingerprinting and neural networks to process a novel method for speaker detection. The hidden markov model (HMM) and radial basis function neural network (RBFNN) work better together than they do with conventional models. The speech data is separated into low and high frequency segments in order to make the denoising process easier for both spoken and unvoiced speech, including background sounds.To denoise the split signals, the maximal overlap discrete wavelet transform (MODWT) is applied. In terms of boundaries, MODWT is more adaptable. Mel cepstrums are less accurate than Power normalized cepstral coefficients (PNCC) in the presence of white noise.This recommended approach yields more accurate results.

REFERENCES

[1] Shiqing zhang,Xiaoming Zhao,Bicheng lei,"Spoken emotion recognition using radial basis function Neural Network","Advances in computer science,environment,ecoinformatics and education".pages 437-442.

[2] Ali bounassif,Noha alnazzawi,Ismail shahin,Said A.Salloum,Noor Hindawi,Mohammed Lataifeh,Ashraf Elnagar,"A novel RBFNN-CNN Model for speaker identification in stressful talking environments","Human-Computer interactions,May-2022.

[3] R.L.K.Venkateswarlu,R.Vasantha Kumari,G.Vani jayasri,"Speech recognition using radial basis function neural network",3rd international conference IEEE,2011.

[4] Kumud Arora,Dr.V.P.Vishwa Karma,Dr.Poonam Garg,"Radial basis function neural network trained withvariant spread learning","International journal of Engineering research and Technology",Volume3.Issue9.(sep-2014).

[5] D.Prabhakaran,S.Sriuppili,"Speech processing:MFCC based feature extraction techniques-an investigation","Journal of physics",2021.

[6] B.Vimal,Muthyam surya,darshan,V S sridhar,Asha ashok,"MFCC based audio classification using machine learning","12th international conference IEEE,2021.

[7] Shalbbya ali,Dr.Safdar Tanweer,Syed Sibtain Khalid,Dr.Naseem Rao,"Mel Frequency cepstral coefficient :A Review", "Proceedings of the 2nd International Conference",ICIDSSD,2020.

[8] Xuechen Liu,Md Sahidullah,Tom kinnunen,"Optimized power normalized cepstral coefficients towards robust deep speaker verification","Automatic Speech recognition IEEE,2021.

[9] Ishwar chandra yadav,Gavadhar pradhan,"Pitch and noise normalized acoustic feature for children's ASR","Digital signal processing",vol-10,feb-2021.

[10] Chanwookim,Member IEEE and Richard M.Stern,"Power Normalized Cepstral Coefficients for robust speech recognition","IEEE/ACM transactions on audio speech and language processing",Vol 24,No7,July-2016.

[11] Selma ozaydin,Iman khalil alank,"Speech enhancement using maximal overlap discrete wavelet transform","Gazi university journal of science",Dec-2018.

[12] Sasikumar gurumoorthy,Naresh babu muppalaneni,G.sandhya Kumari,"EEG signal denoising using Haar transform and maximal overlap discrete wavelet transform(MODWT) for the finding of

Epilepsy",”Etilogies instrumental diagnosis and treatment”Sep-2020.

[13] Davi v.q.rodrigues,Delong zuo,Changzhi,”A MODWT based algorithm for the identification and removal of jumps/short-term distortions in displacement measurements used for structural health monitoring”,Dec-2021.

[14] Annapurna p Patil,Lakshmi j itagi,ashika cs ,ambika g,mallika ravi,”Design and implementation of an audio fingerprinting system for the identification of audio recordings”,”9th region IEEE conference,2021.

[15] Sungkyun chang,Donmoon lee,”Neural audio fingerprint for high-specific audio retrieval based on contrastive learning”,”IEEE international conference on acoustics”,June-2021.

[16] Jose juan garcia hernandez,Juan jose gomez ricardez,”Hardware architecture for an audio fingerprinting system”,”Computers & electrical engineering,Vol 74,Mar-2019.

[17] Yuexing chen,Jiarun Li,”Recurrent neural networks algorithms and applications”,”2nd international conference IEEE,2021.

[18] NM Alharbi,”Evaluation of sentimental analysis via word embedding and RNN variants for amazon online reviews”,”Mathematical problems in engineering”,2021.

[19] Shirali kadyrov,Cemil turan,altynbek amirzhanov,cemal ozdemir,”Speaker recognition from spectrogram images”,”IEEE international conference”,2021.

[20] Rafizah mohd hanifa,Khalid isa,Shamsul mohamad,”A review on speaker recognition technology and challenges”,”Computer and electrical engineering”,Vol 90,Mar-2021.

[21] Mahdi barhoush,ahmed hallawa,anke schmenik,”Robust automatic speaker identification system using shuffled MFCC features”,”IEEE international conference,2021.

[22] Shabnam farsiani,habib izadkhah,Shahriar Lotfi,”An optimum end to end text independent speaker identification using conventional neural network”,”Computers and electrical engineering”,May 2022.

[23] Vincent roger,jerome farinas,julien pinquier,”Deep neural networks for automatic speech processing a survey from large corpora to limited data”,”EURASIP journal on sudio speech and music processing”,Aug-2022.

[24] Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, Michael Weyrich,” A survey on long short-term memory networks for time series prediction”,” 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME ‘20”,2021.