iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6140

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

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 VI June 2020- Available at www.ijraset.com

Investigation of Deep Neural Network for Speaker Recognition

A. V. N. S. Bhavana[1], Dr. A. S. N. Murthy[2]
[1]*U.G Student,* [2]*Professor, Electronics and Communication Department, B.V Raju Institute of Technology, Narsapur, Telangana.*

Abstract: ***In this paper, deep neural networks are investigated for Speaker recognition. Deep neural networks (DNN) are recently proposed for this task. However, many architectural choices and training aspects that are made while building such systems haven't been studied carefully. We perform several experiments on a dataset consisting of 10 speakers,100 speakers and 300 speakers with a complete training data of about 120 hours in evaluating the effect of such choices. Evaluations of models were performed on 10,100,300 speakers of testing data with 2.5 hours for every speaker utterance. In our results, we compare the accuracy of GMM, GMM+UBM, ivectors and also time taken for the various modelling techniques. Also, DNN outperforms the fundamental models indicating the effectiveness of the DNN mechanism.***
Keywords**: deep neural network, Joint Factor Analysis (JFA), Speaker identification Feature Extraction.***

I. INTRODUCTION

In our daily lives, there occur many forms of communication, for instance: speech, pictorial language, textual language, and body language, etc. However, amongst those forms speech is always regarded as the most powerful form because of its rich dimensions character. Except for the speech text (words), the rich dimensions also refer to the gender, attitude, emotion, health situation, and identity of a speaker. Such information is very important for effective communication. From the signal processing view, speech can be distinguished in terms of the signal transmitting message information. The waveform could be the representation of speech, also this kind of signal has been most useful in practical utilization. Extracting from the speech signal, we could get three main kinds of information: Speech Text, Language, and Speaker Identity, shown in Fig.1.1.

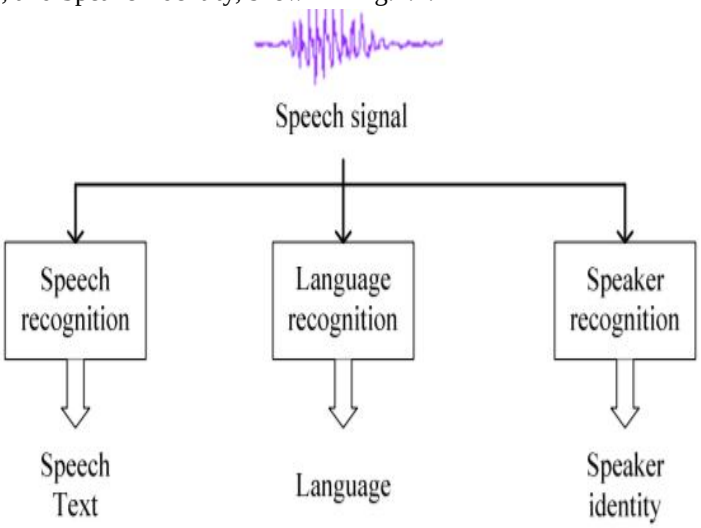


Figure 1Automatically extract information transmitted in speech signal

Speaker recognition is a biometric system that performs the computing task of validating a user's claimed identity using the characteristic features extracted from their speech samples. Speaker identification is one of the two integral parts of a speaker recognition system with speaker verification being the other one.

The field of speaker recognition has gained immense popularity in various applications ranging from embedding recognition in a product that allows a unique level of hands-free and intuitive user interaction, automated dictation and command interfaces, etc. The several phases of our project will begin to an in-depth understanding of several speaker recognition principles being employed while becoming associated with the speaker recognition community.

II. LITERATURE SURVEY

Speech information contains information pertaining to particular user his emotions and messages [1]. To identify and verify the speaker this speaker specific information will be useful. Many speaker recognition techniques were proposed and out of them the most famous ones include enhancement indecision-making approaches [4] ,kernel-based discriminative learning [3] and generative modelling like GMM [2]

Out of all these techniques Deep Learning approaches gave better results in processing speech information and in computer vision [7], [8]. To estimate the features or characteristics of a speaker deep learning methodologies are being used and the aim is to get the better results by employing a mathematical model in both unsupervised and supervised learnings as in [5] and [6]. The existing algorithms could not map speaker model in the output space as they lacked non-linearity and the depth. There are certain architectures like independent Component Analysis (ICA) [5], Support Vector Machines (SVM) [3],Nearest Neighbor Classifiers [12],Principal Component Analysis (PCA),Multilayer Perceptrons (MLP) [11] they are called the shallow architectures. These approaches were used in speaker recognition tasks but they haven't given promising results on data sets that are huge and varying.

In Deep neural networks we have multiple layers involved in a hierarchical manner where each layer is depends on previous layers and each layers are trained for task-specific informatiion.

There are three types of DNN-based speaker verification approaches that are:

1) End-to-end system. These three approaches are different in many ways.
2) d-vectors embeddings learned from DNN and used for scoring
3) Bottleneck features from DNN+MFCC or PLP and i-vector back end

These three approaches are different in many ways.

The first phase is the training phase where the features are trained either in unsupervised fashion or supervised fashion and its is combined with other data. In order to grab the features d-vectors are used and they are trained separately they act as front end part. where as in the end-to-end system back end and front end are trained together. To differentiate the features the D-vectors and the bottleneck features are used where as the end to end systems can identify whether the utterances are from different speaker or the same speaker. According to [9], For Smaller data sets d-vectors gives promising results where as end to end systems do not do well where as for larger sets if data end to end systems do well than the other two.

Therefore end to end systems requires large sets of training examples to get the favorable performance if not it will lead to over fitting to the training data and will not give the promising results[33].

A. Block Diagram

This system will first collect the audio samples of 10,100 ,300 speakers using 'audio read' command, after that it will separate training and testing data. Once the data is collected and separated. Now the enrolled data's MFCC feature extraction is performed. The output of the MFCC is acoustics vector. The speaker features obtained from MFCC is then done modelling using Deep Neural Networks. These features are then compared with the test utterance features. Scoring is performed.

We perform this process using different modelling techniques like GMM, GMM+ivectors and also using DNN. The performances of the different modelling techniques are compared.

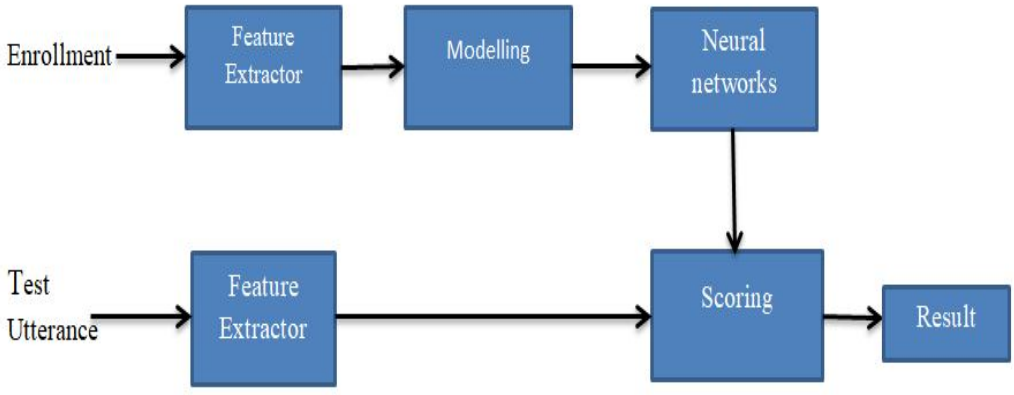


Figure 2 Block diagram

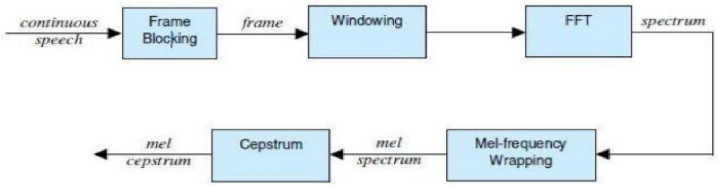
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

III.METHODOLOGY

Speaker recognition mainly consists of feature extraction and have modeling.

Feature extraction: the aim of this can be to convert the speech waveform into a group of features or rather feature vectors which are used for further analysis. this can be pointed because the signal processing front-end.

In this paper, we've got used MFCC –Mel frequency Cepstrum coefficients for the feature extraction. MFCC is predicated on the known variations of the human ear's critical bandwidths with frequency. Filter spaced linearly at low frequencies and logarithmically at high frequencies are accustomed capture the important features of speech. This can be represented within the Mel-frequency range, which may be a logarithmic spacing above 1000Hz and linear frequency spacing below 1000Hz.



A. Speaker Modeling

Using Cepstral coefficients and MFCC as illustrated within the previous section, a spoken syllable are often represented as a group of feature vectors. an individual uttering the identical word but at a special time, instant are going to be having similar still differently arranged feature vector sequence. the aim of voice modeling lies in building a model that may capture these variations in an exceedingly set of features extracted from a given speaker.

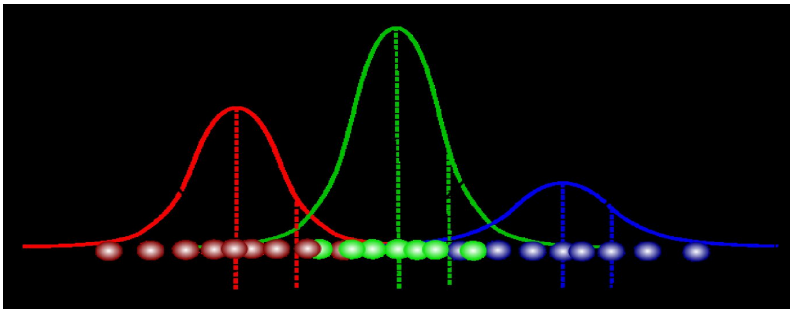
There are usually two sorts of models that are extensively utilized in speaker recognition systems:

1) Stochastic models
2) Template models

The stochastic model exploits the advantage of applied math by treating the speaking process as a parametric random process. It assumes that the parameters of the underlying model are often estimated precisely, in an exceedingly well-defined manner. In parametric methods usually, the idea is created about the generation of feature vectors but the non-parametric methods are free from any assumption about data generation. The template model (non-parametric method) attempts to get a model for the speaking process for a selected user in an exceedingly non-parametric manner. It does so by using sequences of feature vectors extracted from multiple utterances of the identical word by the identical person. Template models accustomed dominate early add speaker recognition because it works without making any assumption about how the feature vectors are being formed. Hence the template model is intuitively more reasonable. However, recent add stochastic models has revealed them to be more flexible, thus with the generation of higher models for the speaker recognition process. The state-of-the-art feature matching techniques utilized in speaker recognition incorporate Vector Quantization (VQ), Gaussian Mixture Modeling (GMM), and Dynamic Time Warping (DTW).

B. Gaussian Mixture Model

1) A Gaussian Mixture may be a function that's comprised of several Gaussians, each identified by k ∈ , where K is that the number of clusters of our dataset. Each Gaussian k within the mixture is comprised of the subsequent parameters:
2) A mean μ that defines its center.
3) A covariance Σ that defines its width. this may be such as the size of an ellipsoid in an exceedingly multivariate scenario.
4) A mixing probability π that defines how big or small the Gaussian function are going to be.



C. ivectors

Relevance MAP adaptation may be a linear interpolation of all mixture components of UBM to extend the likelihood of speech from a selected speaker. Supervectors carries with it the speaker-dependent GMM means components. Problem: Significance MAP adaptation adapts to not only channel and other nuisance factors but also speaker-specific characters of speech. Hence, super vectors produced during this way are non-ideal.

A supervisor for a speaker should be decomposable into speaker-dependent,speaker-independent, residual components, and channel-dependent.• Each components are often described by a low-dimensional set of things, which operate along the principal dimensions (i.e. eigen dimensions) of the similar component.

A given speaker GMM super vector s are often disintegrated as follows:

$$s = m + Vy + Ux + Dz$$

"Ideal" speaker supervector — Speaker-independent component — Speaker-dependent component — Channel-dependent component — Speaker-dependent residual component

where:

1) Vector m is a speaker-independent supervector (from UBM)
2) Matrix V is the eigen voice matrix
3) Vector y is the speaker factors. Assumed to have N(0,1) prior distribution
4) Matrix U is the eigenchannel matrix – Vector x is the channel factors. Assumed to have N(0,1) prior distribution
5) Matrix D is the residual matrix, and is diagonal.
6) Vector z is the speaker-specific residual factors. Assumed to have N(0,1) prior distribution

D. The i-vector approach

An i-vector system uses a set of low-dimensional total variability factors (w) to represent each conversation side.

Each factor controls an eigen-dimension of the total variability matrix (T), and are known as the i-vectors.

$$s = m + Tw$$

Conversation side supervector — Total-variability matrix — i-vector

To train T, run exact training procedure used to train V, but treat all conversation sides of all training speakers as belonging to different speakers

1) Given T, obtain i-vectors (w) for each conversation side
2) For channel compensation of i-vectors, perform LDA then WCCN (techniques empirically determined to perform well) on i-vectors. Denote channel-compensated i-vectors as ω.
3) Perform cosine distance scoring (CDS) on channel-compensated i-vectors ω for a pair of conversation sides:

$$score(\omega_1, \omega_2) = \frac{\omega_1^* * \omega_2}{\| \omega_1 \| * \| \omega_2 \|} = \cos(\theta_{\omega_1, \omega_2})$$

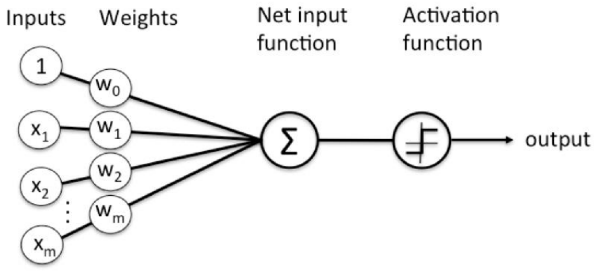
E. Deep Neural Networks

Neural networks are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabelled data according to similarities among the example inputs, and they classify data when they have a labelled dataset to train on.

Deep learning is the name we use for "stacked neural networks"; that is, networks composed of several layers.

The layers are made of *nodes*. A node is just a place where computation happens, loosely patterned on a neuron in the human brain, which fires when it encounters sufficient stimuli. A node combines input from the data with a set of coefficients, or weights, that either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn; e.g. which input is most helpful is classifying data without error? These input-weight products are summed and then the sum is passed through a node's so-called activation function, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome, say, an act of classification. If the signals passes through, the neuron has been "activated."



Deep-learning networks are distinguished from the more commonplace single-hidden-layer neural networks by their **depth**; that is, the number of node layers through which data must pass in a multistep process of pattern recognition.

Earlier versions of neural networks such as the first perceptron's were shallow, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So *deep* is not just a buzzword to make algorithms seem like they read Sartre and listen to bands you haven't heard of yet. It is a strictly defined term that means more than one hidden layer.

In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer

IV. RESULTS

In this paper, we are comparing the performance of the different modelling techniques.

S.no	No of Speakers	Total variance of Dimensions	Accuracy (%)
1	300	100	50
		400	65
2	100	100	47.5
		200	52
		300	63.5
		400	6
3	10	100	15
		400	25

Table 1 GMM Performance

In Table 1, we are first keeping the Number of speakers as 300 keeping the Total variance of Dimensions as 100 and noting the accuracy. After that keeping Number of speakers as constant and changing the total variance of Dimensions to 400 and noting the accuracy. The same procedure is followed for 100 speakers and 10speakers. If we observe the table the accuracy increases as the number of speakers increases and also accuracy increases as the tvd increases for particular No. of speakers.

s.no	No of speakers	Tv Dimensions	LD Dimensions	Accuracy(%)
1	50	300	250	55
2	50	300	300	56
3	100	100	200	49.5
4	100	300	200	56
5	100	300	250	61
6	100	300	300	61

Table 2 GMM performance check varying no. of speakers, tvd, lda

In Table 2, we are measuring the performance of the speaker by varying Tv dimensions, LD dimensions. From the above table, we conclude that with constant no. of speakers, constant Tvd, and varying LDA the accuracy is increasing as the LDA increases.

A. Time taken in GMM+iVector

S.no	Tv Dimension	Time/iteration
1	250/300	3-6min
2	200	2-6min
3	100	1min

Table 3 Time taken in GMM+iVector

B. DNN

1) Single layer Perceptron

S.no	HiddenLayer1	Accuracy (%)
1	32	70
2	64	80
3	128	80
4	256	100

Table 4 Performance table for Single layer Perceptron

In table 3, we are measuring the performance with a single hidden layer, and by increasing the number of neurons. With an increasing number of neurons, the accuracy increases.

2) Multi-Layer Perceptron

S.no	HiddenLayer1	HiddenLayer2	HiddenLayer3	Accuracy (%)
1	64	32	-	80
2	64	32	32	100
3	200	100	-	84
4	600	400	-	90
5	600	200	-	60
6	128	64	32	90

S.no	HiddenLayer1	HiddenLayer2	HiddenLayer3	HiddenLayer4	Accuracy (%)
1	256	128	64	32	90

Table 5 Performance table for Multilayer Perceptron

In Table 5, we are measuring the performance with multi-layer perceptrons like 2,3,4. As the number of layers increasing the accuracy increases. For a particular layer, as the number of neurons increases the accuracy also increases respectively.

C. Time Taken for DNN

S.no	No. of Hidden layers	Time/Step
1	1layer	200-600us
2	2layer	200-600us
3	3layer	200-600us
4	4layer	300-500us

Table 6 Time taken for DNN

From the above tables, we can conclude that the performance of speaker recognition with DNN is more accurate than the GMM, GMM+UBM, ivectors.

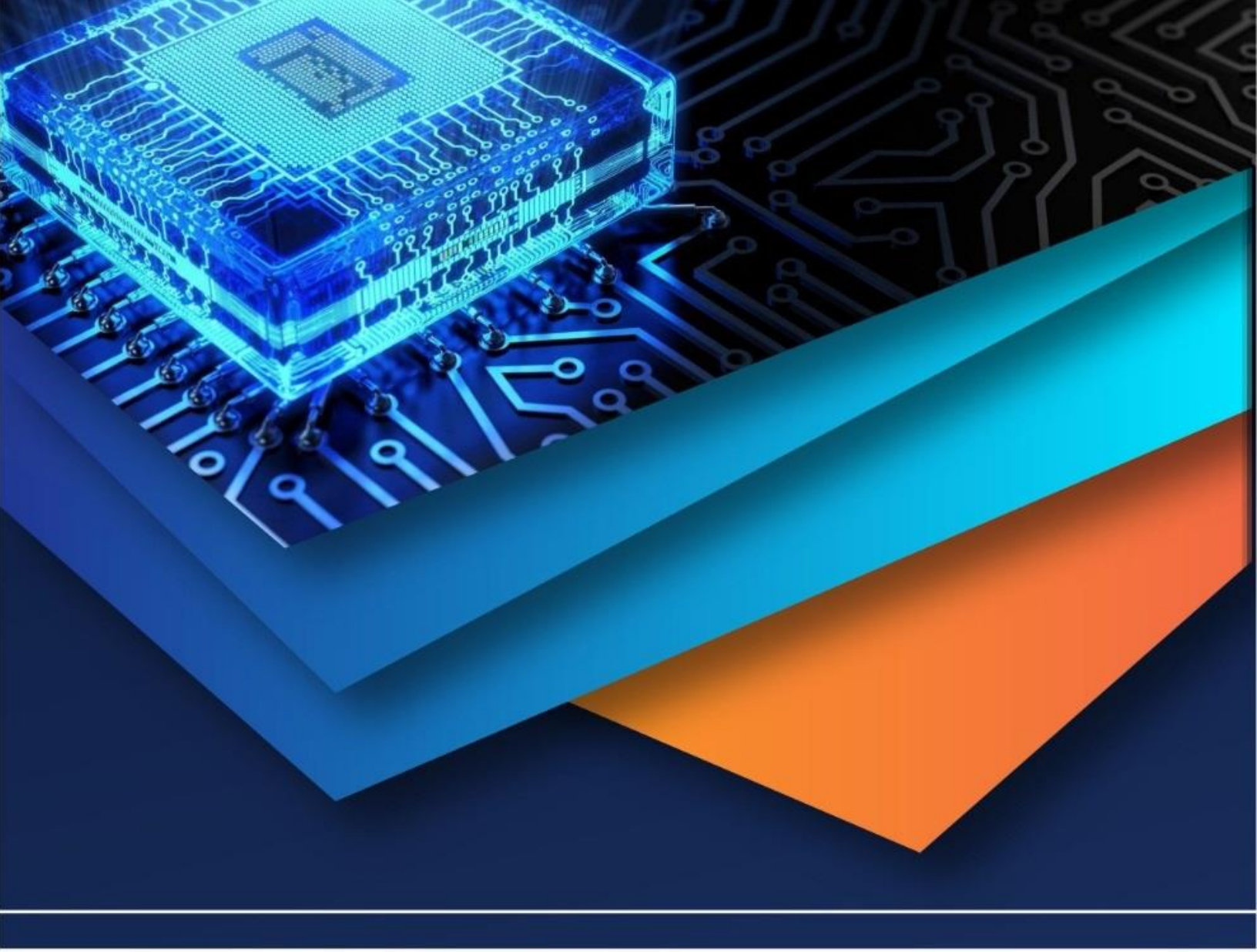
V. CONCLUSIONS

We have compared GMM and GMM+i-vector with DNN and it is evident that DNN is more accurate than the other two especially when there are more no of speakers. Accuracy of DNN is also improved when the no of layers in a perceptron is increased. DNN is more suitable than gmm and gmm plus i-vector when there are more no of speakers.

In addition to the low-level spectrum features used by current systems, there are many other sources of speaker information in the speech signal that can be used. These include idiolect (word usage), prosodic measures and other long-term signal measures. This work will be aided by the increasing use of reliable speech recognition systems for speaker recognition R&D. High-level features not only offer the potential to improve accuracy, they may also help improve robustness since they should be less susceptible to channel effects.

REFERENCES

[1] Salman and K. Chen, “Exploring speaker-specific characteristics with deep learning,” in Proc. IJCNN, pp. 103-110, 2011.
[2] A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,”Digital Signal Processing, vol. 10, pp. 19-41, 2000.
[3] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” Comput. Speech Lang., vol. 20, pp. 210-299, 2006.
[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Trans. Audio, Speech and Lang. Processing, vol. 15, pp. 1435-1447, 2007.
[5] Jang, T. Lee, and Y. Oh, “Learning statistically efficient feature for speaker recognition,”in Proc. ICASSP, pp. 437-440, 2001
[6] N. Malayath, N. Hermansky, S. Kajarekar, and B. Yegnanarayana, “Data-driven temporal filters and alternatives to GMM in speaker verification,” Digital Signal Process, vol. 10, pp. 55-74, 2000.
[7] Y. Bengio, “Learning deep architectures for AI,” Foundations and Trends in Machine Learning, vol. 2, pp. 1-127, 2009.
[8] Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, “Exploring strategies for training deep neural networks,” J. Machine Learning Research, vol. 17, pp. 1-40, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY