iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

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 IV Apr 2020- Available at www.ijraset.com

Conversation Transcription and Speaker Identification in Group Environment

Aakarshan Rastogi[1], Ayush Gandhi[2], Tarang Vadodaria[3]
[1, 2, 3]*Department of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai, India*

Abstract: ***In face to face meetings or interactions, one can easily identify the person based on the features of the speaker's voice, due to our complex human brain. However, if we were to recollect each statement with precision and the speaker's name along with it, the difficulty increases manifold. In this paper, we compare two different approaches using KNN and SLGMM [9] algorithms for speaker identification that use Mel Frequency Cepstral Coefficients (MFCC's) [2] and pitch [4] as voice features as a part of a proposed system that generates a conversation transcript using the Azure speech to text API, giving this system has an edge over conventional methods in terms of ease and accuracy.***
Keywords: KNN, SLGMM, GMM, RNN, MFCC, ASR, API

I. INTRODUCTION

Identification and classification of a person based on the voice have always been a longstanding challenge in the domain of Machine Learning. Speaker Identification is a process of establishing the identity of a speaker using previously collected training data. The task of speaker identification is affected by variables like overlapping voices and ambient noise.

A unique form of speaker identification can also be used for biometric verification purposes which are known as Automatic Speech Recognition (ASR) [8]. The goal of an ASR system is to identify the speaker's identity by extracting, characterizing, and recognizing the information from the speech signal.

With the advent of automatic transcription of the conversation, various new business models have sprung up. Thus, the combination of automatic transcription with speaker identification has tremendous use cases in meetings or conference calls.

In this paper, we propose a system model that uses a supervised machine learning model to analyze the acoustic signal structure of the speaker in group dynamics to identify the speakers and utilizes the Azure Speech to Text API for transcribing speech chunks. A final transcript is prepared for the entire conversation by the system and emailed to the user. The paper discusses two models for speaker identification which are K Nearest Neighbour and (KNN) and Supervised Learning Gaussian Mixture Model (SLGMM) with the latter having better accuracy. The speech signal collected is broken into chunks based on a period of silence, this time frame referred to as pause-split duration in this paper is pivotal in determining the accuracy of the model, therefore a comparison of the model performance is drawn on this basis.

The rest of the paper is a description of how this system works, the next section is about sampling voice data correctly for supervised model training. Section III presents the KNN and SLGMM models for speaker identification, while section IV explains the results based on which a performance comparison is drawn. Finally, Section V gives a summary and conclusion. The system process can be visualized through the diagram below. SLGMM's and GMM [3] will be used interchangeably in this paper, reader's discretion is advised.

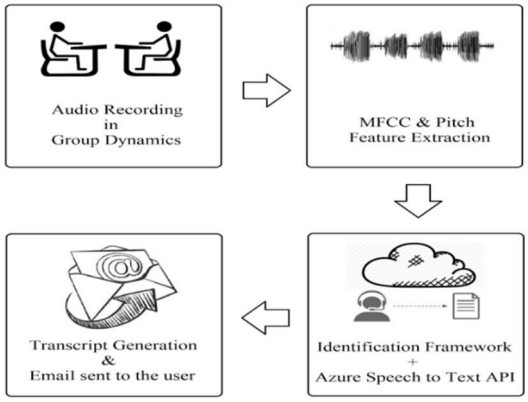


Fig. 1 Model Abstraction

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

II. SYSTEM MODEL

Explained below is the system design, in brief, the diagram acts like a golden source for data and tracks how it undergoes major steps until the transcript is created.

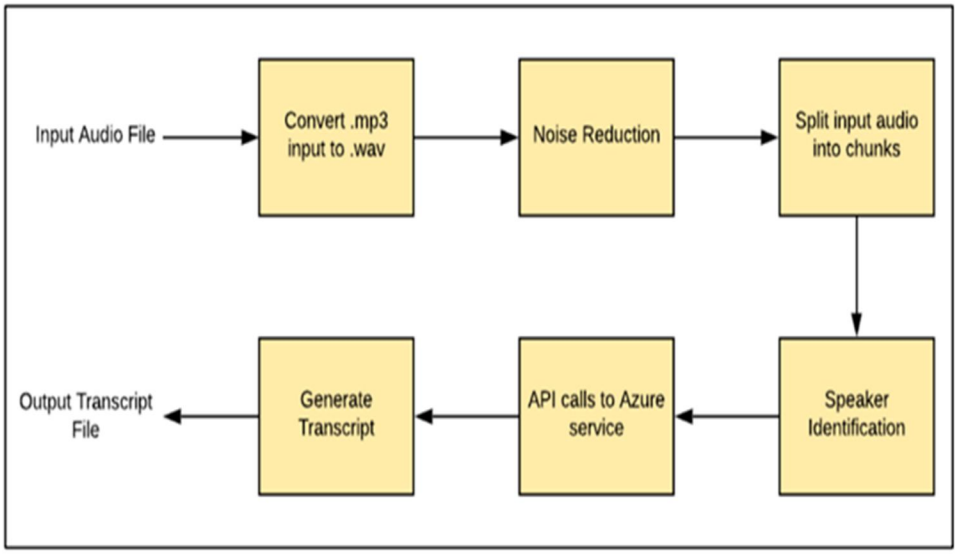


Fig. 2 System Process Model

The major steps are explained below, from voice sampling from users, denoising it to get better quality data and then splitting that data into chunks for collection of speech characteristics. Making the training data API calls using these chunks and getting the converted speech text from Microsoft Azure. Finally clubbing both the information and presenting the conversation transcript to the user through the web application as well as via an email.

A. Speaker Voice Collection

For the supervised training model [1], data has been collected by recording the voice files of fellow individuals.
The audio files for these individuals are first denoised [6] to obtain higher quality speech samples. This is then broken down on pauses to create multiple files with a duration averaging three seconds. A set of fifty voice samples for each speaker is made having 1-2 spoken words.

B. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. The shape of the vocal tract manifests itself in the envelope of the short-time power spectrum, and MFCC's represent this envelope. These features are collected from voice samples broken down into a window frame of 35ms each. A Hamming window is used as it gives good results. A vector matrix is created with 20 such features for each 35ms of the voice sample.

C. Pitch

Pitch represents the fundamental frequency of the sound. Pitch differs not a lot in the same gender but is used as an added feature to recognize the correct speaker. This feature is also extracted in windows of frame size 35ms.

D. Training Data

For each speaker, the MFCC data is combined with pitch data (horizontal stacking) and all these features vertically stacked together in matrix M having dimension n x 21, where n is the number of windows. The final training dataset is created by giving a label to each dataset point, which is the corresponding speaker name.

E. Azure Speech to Text API

All the generated voice chunks are sent for conversion using the Azure Speech to Text API [5]. The API returns a JSON response containing the transcription along with other parameters such as confidence, offset, duration, among others.
The speaker identification data and the transcription data are merged to create the final transcript.

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

III.SPEAKER IDENTIFICATION MODELS

A. K Nearest Neighbours (KNN) Model

The KNN algorithm is simple in nature and is often called a lazy learning algorithm. K denotes the training dataset items for classification, in the context of this paper K represents the number of speakers.

A KNN classifier employs distance between cluster centres as a metric to classify the data point. The KNN model used is an eager learner meaning based on the training data it creates its K clusters, and during testing, it just calculates the distance of the new points from cluster centres. The data point X having minimum D from a cluster centre C is classified as belonging to cluster C.

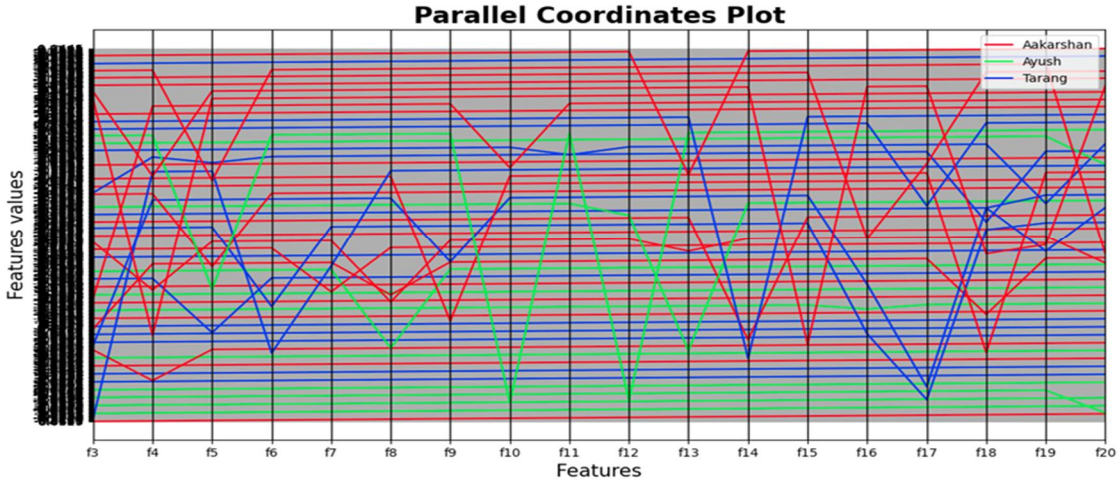


Fig. 3 Parallel plot showing variance in feature vectors for 50 samples

B. Supervised Learning Gaussian Mixture Model (SLGMM) Model

The main reason behind using the Supervised Learning Gaussian Mixture Model (SLGMM) simply because it improves the recognition accuracy of the traditional GMM in recognizing patterns. These are like KNN in the sense that they too use the clustering approach for classifying data points. The difference is unlike KNN, GMM uses a probabilistic model and does not hard assign to any cluster rather it keeps space for some uncertainty. So instead of having a fixed distance 'D' in a KNN model, GMM uses concepts of mean and covariance amongst the K probable clusters. To assign a specific data point to any one cluster the maximum likelihood is considered which depends again on mean and covariance. However, since determining such exact values is computationally exhaustive, in practice, the Expectation-Maximization (EM) function is used to find maximum likelihood.

The EM function is used in several optimization problems, what it does is to find optimum values for dependent variables by using an iteration approach. At the end of the iteration, the algorithm finds a local maxima for the GMM model. As the iterations progress, the Gaussians fit better to data points belonging to each cluster.

Spherical covariance matrix type is used for the GMM model since this shaped matrix type gives us the maximum performance.

SLGMM's are trained in an above-mentioned manner against training data for each user and are saved in the file system. For testing, we compare the incoming voice data points against all the models and calculate the logarithmic probabilities of them. These are calculated for every model, following which a total is calculated for each model. The model having the highest score is selected and so the corresponding speaker is identified.

The diagram below is a cluster graph made using 2 features on approximately 22000 data points for a 3-speaker system. Note: Data points towards the left are outliers.

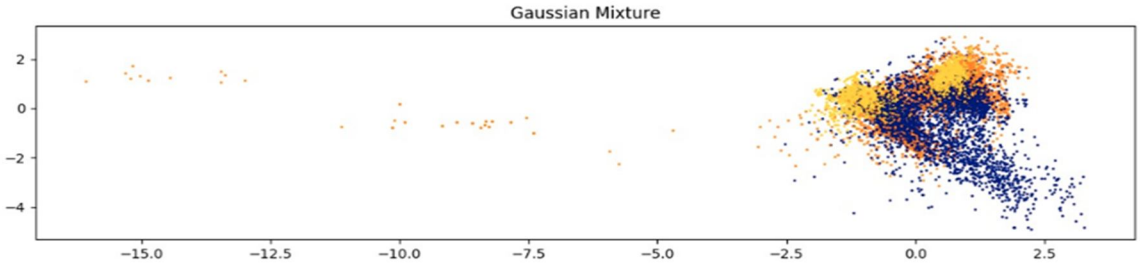


Fig. 4 Clustering graph for 3 speakers using GMM

Confusion Matrix for the SLGMM model - Tested for 28 voice samples

$$\begin{bmatrix} 10 & 0 & 0 \\ 2 & 8 & 2 \\ 1 & 0 & 10 \end{bmatrix}$$

The Accuracy for SLGMM is $84.\overline{84}\%$

TABLE I
KNN vs SLGMM speaker identification comparison drawn for a 3-speaker conversation

Observation Number	Number of speakers	Speaker Identification accuracy using KNN model on test sets	Speaker Identification accuracy using SLGMM model on test sets	Percentage Increase (KNN vs SLGMM)
1	3	43.18%	72.72%	63.41%
2	3	77.14%	88.88%	15.21%
3	3	66.66%	77.77%	16.66%

IV. RESULTS

The accuracy of the model as a whole is obtained by choosing an optimal pause split value, that balances results from speech to text transcription API and speaker identification using the discussed models, this is obtained iteratively and is found to depend on the speaker's dialect and cultural or regional variables (some regions may speak faster than the other).

The pause-split parameter can be changed depending on the pause duration (the duration of the audio below the silence threshold which is then defined as a pause) defined in the pause split code. We have considered the pause duration [7] to be 250ms and 350ms, and below are the results for speaker identification rates and rates of voice chunks returned accurately transcribed using Azure API calls.

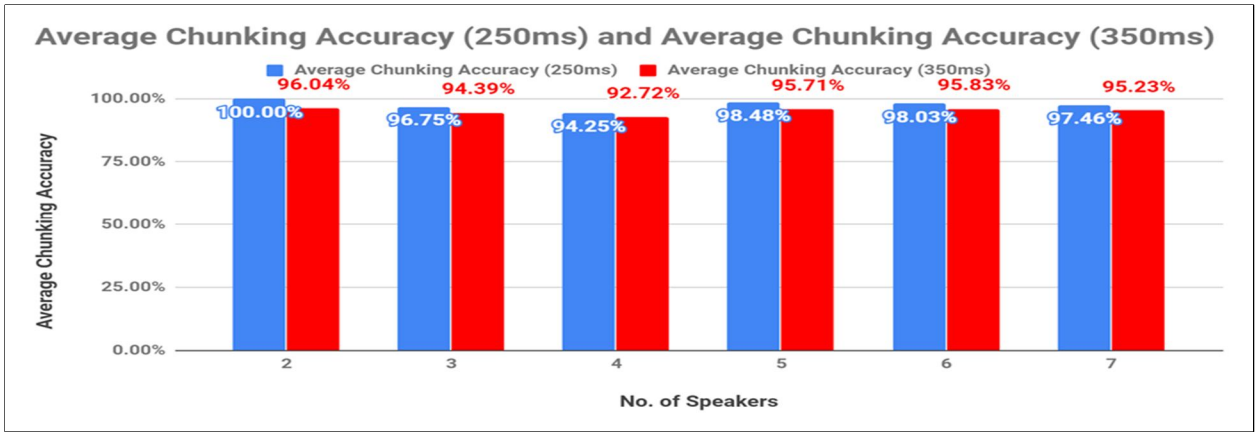


Fig. 5 Average Chunking Accuracy for 250ms and 350ms pause splits

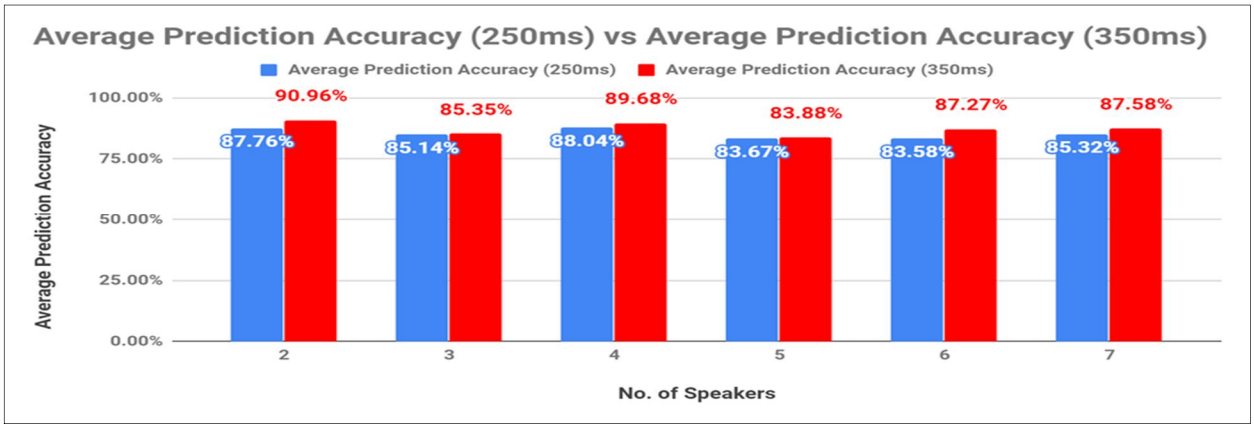


Fig. 6 Average Prediction Accuracy for 250ms and 350ms pause splits

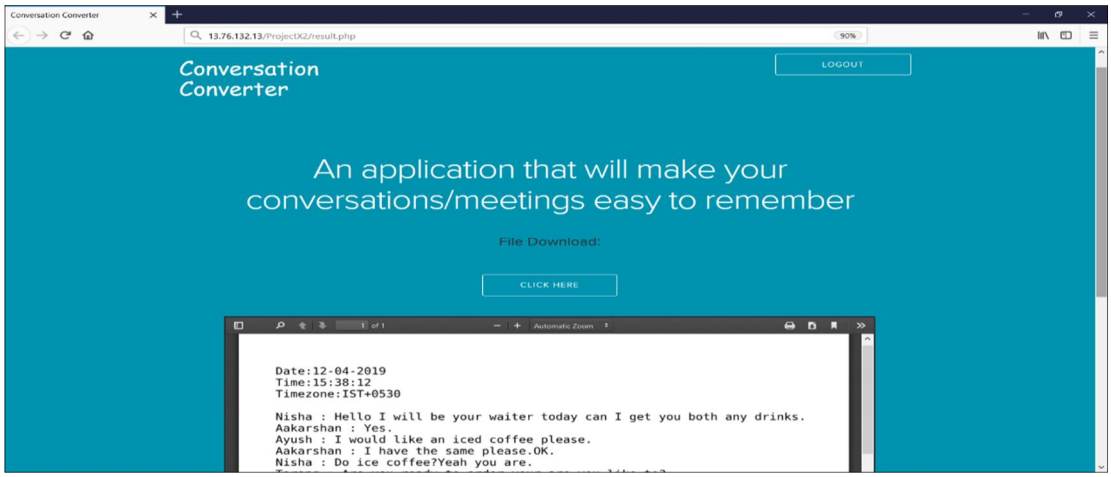


Fig. 7 Transcript generated by the model via the web interface

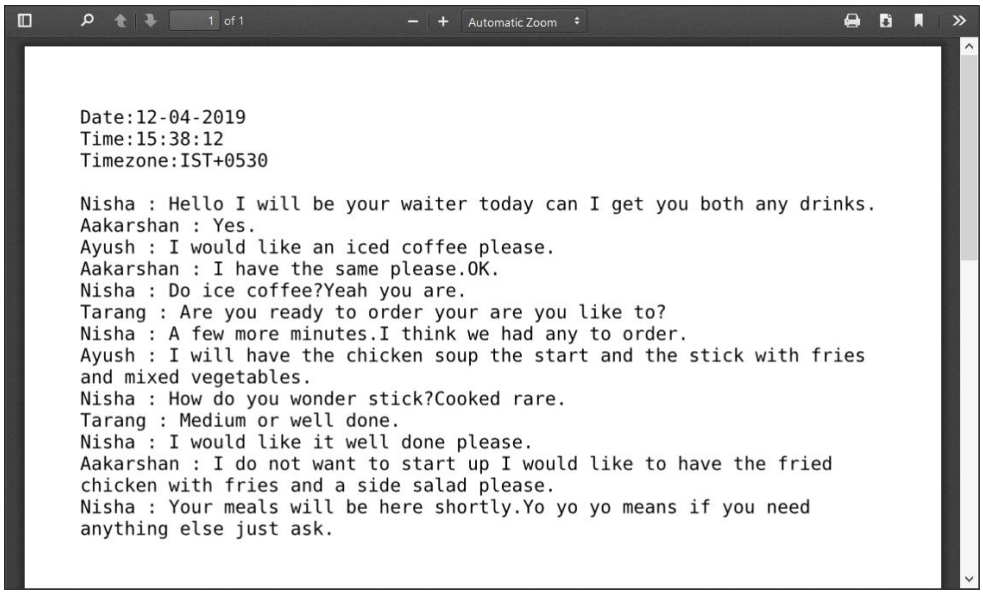


Fig. 8 A sample transcript for 3 speaker system

V. CONCLUSIONS & FUTURE WORK

The aim of this paper was to introduce a system that creates transcripts from a group conversation environment and is time sequenced. Gaussian Mixture Models proved to be more robust to varying testing environments and gave better accuracy in general. The JSON responses from the Azure API along with sequential speaker identification data were combined by the system to create a final transcript. The system sends this transcript as an attachment, in PDF format to the registered user via E-mail. The system has constraints on its need to know the speaker's identity and speech data apriori, similar functional models can be implemented on remote servers for use in limited or controlled scenarios.

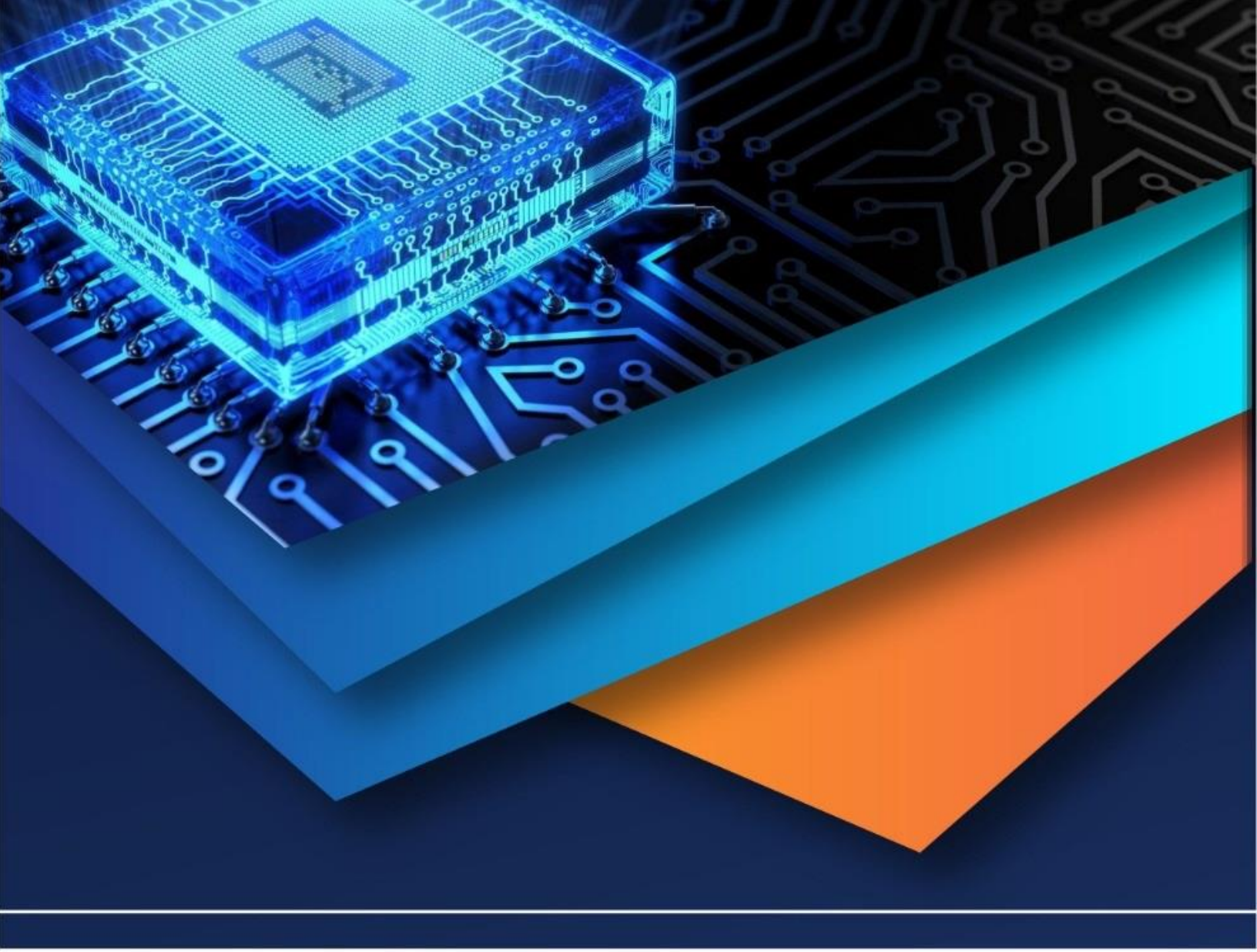
Considering the scalability and performance aspects of the current model, using ensemble models that use neural networks to greatly scale this to a diverse audience and have a wider use case, as they are promising in unsupervised learning domains. For generating transcripts, APIs that support region-specific transcribing capabilities will be incorporated which will greatly enhance the accuracy of the model. Incorporating the above will enable the current model to be dynamic and make it ready to be used on-spot basis.

VI. ACKNOWLEDGMENT

The authors express their sincerest regards to Dr. Kavita Kelkar, Prof. Smita R.Sankhe and Prof. Jyothi M. Rao for their valuable inputs, able guidance, encouragement, whole-hearted cooperation, and constructive criticism. The authors take this opportunity to thank all colleagues for building the audio corpus.

REFERENCES

[1] Mrunal Bhogte, Prof. Shanthi Therese, Prof. Madhuri Gedam, “Speaker Identification techniques in Overlapping Speech Analysis: A Review,” in ICEMTE,vol. 5[3], pp. 210-213, 2017.
[2] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, “SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS,” in ICECE, Dhaka, Bangladesh, 2004.
[3] D. A. Reynolds and P. Torres-Carrasquillo, “Approaches and Applications of Audio Diarization,” in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 953-956, 2005.
[4] Speaker Identification Using Pitch and MFCC on MathWorks MATLAB [Online]. Available: https://in.mathworks.com
[5] Speech to Text API - Microsoft Azure [Online]. Available: https://azure.microsoft.com/
[6] Jean-Marc Valin.(2017) “Recurrent neural network for audio noise reduction” [Online]. Available: https://people.xiph.org/~jm/demo/rnnoise/
[7] Gustafson-Capkova, S., & Megyesi, B, "A comparative study of pauses in dialogues and read speech. In Proc of EUROSPEECH," pp 931-934, Aalborg, Denmark, 2001.
[8] Rajalakshmi.P, Anju.L “Feature Extraction and Speaker Identification in Automatic Speaker Recognition System,” in IJIRCCE vol. 5[3], Apr. 2017.
[9] J Ma, W Gao “The supervised learning Gaussian mixture model,” Journal of Computer Science and Technology, 1998.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY