IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 12, December 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

A Review on Auto Steno using Speech Recognition and Speaker Identification

Harsh Gahlot[1], Ayushi[2], Diksha Bharti[3], Dhanya N. Naik[4], Pallavi R[5]

B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[1]

B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[2]

B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[3]

B.E. Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, Bengaluru, India[4]

Associate Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology Bengaluru, India[5]

ABSTRACT: Stenographers use shorthand methods and a steno typing machine to transcribe the information as fast as people speak but this profession requires a high degree of skill and focus. This paper provides ways to create an Artificial Intelligence based Automatic Steno Transcription model which is less trained but more precise and faster as compared to a highly trained stenographer. This model can be effectively used for court hearings and medical proceedings at a very low cost as compared to the costs charged by skilled stenographers. The paper ends with a discussion of various Speech Recognition algorithms with good accuracy scores for word to word transcription of recorded proceedings and algorithms for identifying speakers over a recorded digital audio. Algorithms such as Mel Frequency Cepstral Coefficient(MFCC) for feature extraction, Gaussian Mixture Model(GMM) and classification algorithms such as Extra-trees algorithm, KNN algorithm, SVC algorithm, MLP classifier algorithm and Gaussian NB algorithm for personality identification are discussed.

KEYWORDS: Automatic Stenography, Speech Recognition, Speaker Identification

I. INTRODUCTION

With the advent of the modern era, everyone wants to get done with the work as fast as possible. Today in every field, automation is replacing the work that can be done manually by humans because automation always results in cost-effective and time saving solutions. So far stenographers are hired by law courts which are very fast typists and make detailed transcription reports of the proceedings in a live court trial. But the stenographers often charge a huge amount of money from their clients and they might make some typing mistakes too. Thus, keeping this thought in mind our proposed model comes into the picture. Auto Steno using Speech recognition and classification as the name says makes the transcription of voice data into text which is comparatively more accurate and less costlier. It uses speech classification technique which makes the transcript look like a real chat room thus one of the most intrusive techniques developed so far.

The methods and techniques involved for detecting and recognizing voice includes feature extraction as the most important step which includes many acoustic features such as meandom, peakf, median of the voice signal. These features can be trained using different coefficients and classifiers. The model's accuracy can be increased by including different types of voice training datasets and including voice data of different frequency ranges and variations. After capturing audio using python's pyaudio module through microphone and feature extraction from the voice data, a model can be trained using sklearn library over a larger dataset for a particular speaker. But rather than using the features directly derivatives stacked up can also be used to get better results. Thus python modules can be used to achieve better results in a short period of time.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

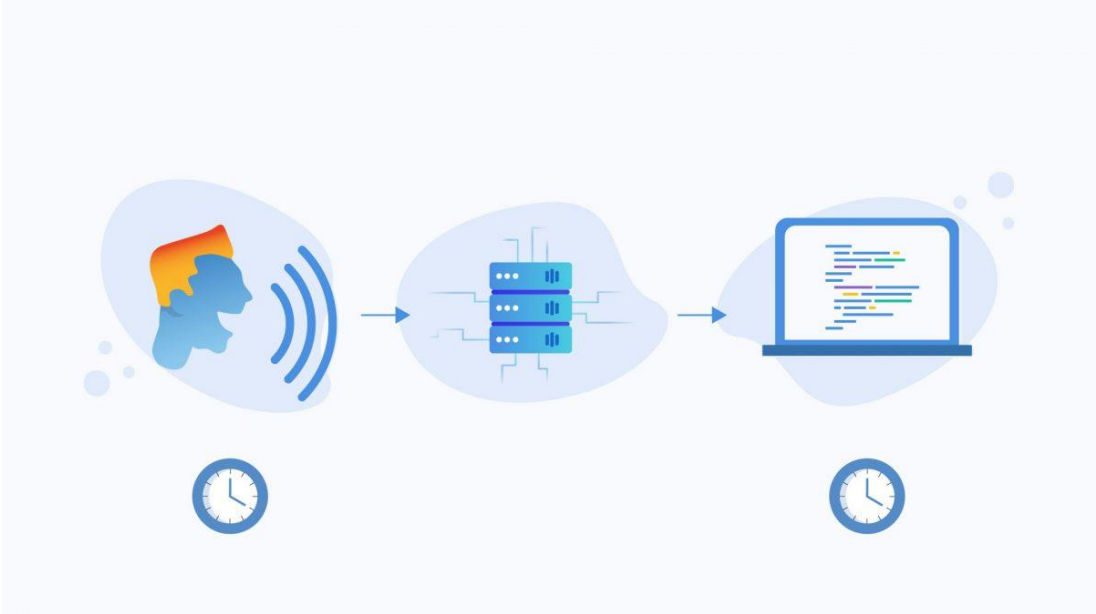

Figure 1. Audio Transcription and Script Generation

The rest of the paper is organized as follows. In section 2, we are describing various papers on how different voice parameters lead to the development of the voice recognition model and how different methodologies for the problem of speech recognition and classification can be used. In section 3, we are listing a summary of literature surveys over various research works . In section 4, there is a discussion about the methodology of the proposed model. And in section 6, there is discussion about future scope of the proposed model. Finally we conclude at the end of the paper.

II. LITERATURE REVIEW

Various papers describing how different voice parameters leads to the development and use of different methodologies for the problem of speech recognition and classification are illustrated and discussed below.:

Scaling uponline speech recognition using ConvNets by Vineel Pratap et al. **[1]** proposes a speech identification technique using Time-depth Separable (TDS) convolution and Connectionist Temporal Classification(CTC) to perform speech recognition in real-time using 1 million anonymous voice samples which aimed upon improving throughput, real-time factor and reducing the word error rate.They observed ~ 4% relative increase in WER[13] because of the changes they made to decrease the latency of the model.

***Minute meeting system using speech recognition by Ainon Sahirah et al*.** **[2]** proposes a web-based system that uses Laravel and speech recognition model to help the users record minute meetings.

Automatic Speech Recognition System Development in the wild by Anton Ragni et al. **[3].** In this paper the author scrapes off unrecognized out-of-domain data also referred as "wild" data to train the already existing acousting model on automatic speech recognition to see and compare if it correlates with the performance on the target domain using language model interpolation. The word error rate [21] came out in the range 30-40% for the above data.

Voice pathology detection and classification using convolutional neural network model by Mazin Abed Mohammed et al. **[4]** aimed at using a pre-trained CNN [17] model on the Saarbrucken voice database(SVD) , a voice pathology database to maximize the classification accuracy and generalizing its applications toa wide range of voice disorders. When combined with a SVM [29] classifier the model reached the most elevated correctness of 99.26%.

Voice classification with neural networks by Jurgen Arias **[5]** focuses on classifying the training data based on the gender of the speaker. He used two approaches- extracting features from audio files using librosa library and converting audio files to images to be fed to train a CNN [17] .He got 93% accuracy on test data using Neural Nets (NN) and 92% using Convolutional Neural Nets (CNN).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

Speech/Music classification using subband coding and AANN by R. Thiruvengatanadhan **[6]** extracts subband coding features from the voice data and then classifies using an auto associative neural network (AANN) model which leads to a better accuracy. This study shows that the AANN learning method achieves an accuracy rate of 92%.

Speech recognition in a high noise environment by Chunling Tang et al. **[7]** . This study aims at improving the existing ASR systems to be able to work efficiently in more noisy environments. They used speech enhancement techniques along with the discard feature model for data preprocessing . The algorithm used here is based on characteristic-abandon and speech recognition algorithms.

Voice Identification using classification algorithms by Mamyrbayev Orken et al. [8] does a comparative study of different classification algorithms -Extra-trees algorithm, KNN algorithm, SVC algorithm, MLP classifier algorithm and Gaussian NB algorithm for the personality identification problem and concluded that the SVM and multilayer perceptron was the most effective. They used the MFCC algorithm for speech preprocessing and discussed various preprocessing issues.The support vector machine and the multilayer perceptron showed the best results, 0.90 and 0.83, respectively

Combined speaker clustering and role recognition in a conversational speech by Nikolas Flemotomos et al. **[9]** focus on turn level and speaker level Speaker Role Recognition (SRR) problem. They combined an unsupervised speaker classification (SC) algorithm along with turn- level supervised role classifier algorithm to give an output to be fed to the final classifier. The final classification is made using Conditional Random Fields (CRF).

A neutral attention model for speech command recognition by Douglas Coimbra de Andradea et al. [10] uses a non-trainable keras layer for the preprocessing of data and then makes use of the Attention RNN model to achieve an accuracy of 94.5%.

III. SUMMARY ON LITERATURE SURVEY

Ref	Author	Title	Data Preprocessing	Algorithm used (Classification)	Results
[11]	S. Pavankumar Dubagunta et al.	Segment level training of ANNS based on acoustic confidence measures for hybrid HMM/ANN speech recognition (2015)	ASR experiments on Mediaparl and AMI data sets	Establishing a relationship between acceptor HMM framework and training of ANNs in HMM/ANN-based systems	segment level training of ANNs yields better ASR systems.
[12]	Tejashri Manohar Mhatre et al.	Birds Voice Classification using ResNet (2018)	Sampling voice samples to 22,050 HZ using the Linux command-line tool sox	Residual Neural Network	The accuracy achieved: 90%
[13]	Hengguan Huang et al.	Recurrent Poisson Process Unit for Speech Recognition(2018)	Acoustic hidden Markov models (HMM) based on Gaussian mixture model (GMM), LSTM, SRU and quasi-RNN	recurrent Poisson process (RPP) into a recurrent neural network (RNN)	RNN baseline systems achieve a WER of 2.8%, RPPU achieves WER of 2.5%, and 10.7% relative performance gain

[14]	Abhijeet Kumar	Spoken Speaker Identification based on Gaussian Mixture Models: Python Implementation (2017)	Not reqd.	Gaussian Mixture Model (GMM)	The accuracy achieved: 100%
[15]	Pala Mahesh Kumar	A New Human Voice Recognition System (2016)	Wavelet-based	RASTA (Relative Spectral Algorithm)	The accuracy achieved: 90%
[16]	Prerana Das et al.	Voice Recognition System: Speech-to-Text (2015)	Mel Frequency Cepstral Coefficients(MFCC	MFCC and VQ techniques	Comparison of feature extraction techniques and speech recognition approaches of different languages
[17]	S. Pavankumar Dubagunta et al.	Improving children speech recognition through feature learning from a raw speech signal (2018)	Kaldi toolkit and Keras with Tensorflow were used	cepstral feature-based ASR approach and CNN-based end-to-end acoustic modeling	CNN-based end-to-end acoustic modeling and augmenting children data with adult speech data yield better systems
[18]	Neha Jain et al.	Speech Recognition systems - A comprehensive study of concepts and mechanism (2019)	acoustic speech vectors representing utterances via a communication channel	Comparison between Dynamic type Warping and Hidden Markov model	_
[19]	Mohsin Uddin Anwar et al.	Boosting Neuro Evolutionary Techniques for Speech Recognition (2019)	downsampling operation to smoothen the signal and audio embedding via MFCC encoding	Genetic Algorithm (GA) over Deep Neural Network (DNN)	The accuracy achieved: 91.4%
[20]	Atul Anand Jha	Speaker-Identification System using Python(2019)	Mel Frequency Cepstral Coefficient(MFCC)	Gaussian Mixture Model (GMM)	The accuracy achieved: 100%: on VoxForge Dataset , 95.29 %: on the self prepared Dataset.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

[21]	Moritz et al.	Triggered Attention for End-to-End Speech Recognition (2019)	VGG based deep CNN component and deep BLSTM	Triggered attention (TA) system architecture for attention-based end-to-end ASR systems	CER % (dev) :6.6 CER %(test) :6.7 WER% (dev) :5.3 WER% (TEST): 5.4

IV. METHODOLOGY

The basic overview of the proposed system is that whenever a court trial is going on, the system does two jobs at a time: first it stores the voice data of speakers individually and secondly it makes word to word transcription of the collected voice data. And finally the system outputs a transcript containing the details of each speaker's speech tagged with the speakers information respectively similar to a chat room.
Two main jobs performed by this system are illustrated in figure 2 and explained below:

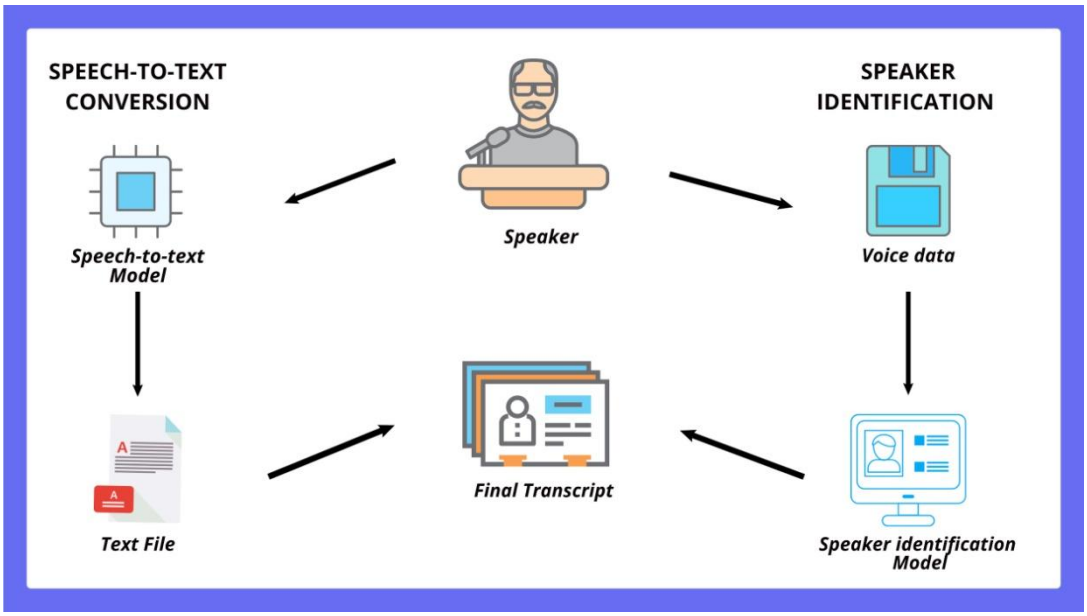

Figure 2. System Model

1. **Speech to text conversion**
   In this step, the voice data of the speakers is stored and then channelized to the speech-to-text model where using a higher accuracy algorithm results in a less error prone text script of the recorded data. Python's *PyAudio* library can be used for taking voice inputs through a microphone. And then the speech is transcripted word to word using *Speech Recognition algorithms*. It is a major step for getting a whole conversational transcript during a court proceeding.

2. **Speaker Identification**
   In this step, the system performs speaker identification[20] where the voice of the speaker is predicted by the system over a dataset of different speaker's voices using speaker identification algorithms. The prerequisite to this step is that before any proceeding, there is a requirement of training voice data. The speaker identification process is achieved in 3 phases and illustrated below in figure 3.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

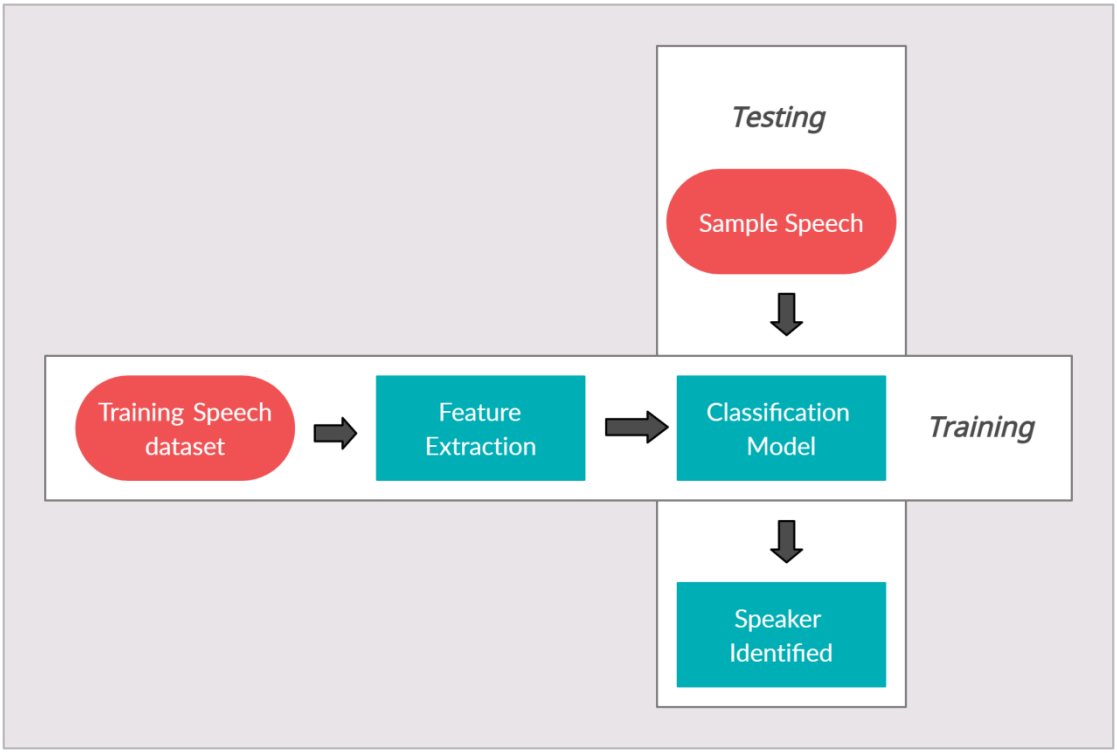

Figure 3. Speaker Identification Process

In the first phase, feature extraction is done. From the training speech data, speech features are extracted. Various feature extraction techniques like Mel Frequency Cepstral Coefficients(MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Perceptual Linear Prediction (PLP) and Discrete Wavelet Transform (DWT) as discussed above in survey can be used.

In the second phase, speaker models are trained for each speaker individually. The prerequisite for this phase are features which are extracted in the previous phase. These features act as input for the training of speaker models. The different algorithms that can be used for training speaker models are Extra-trees algorithm, KNN algorithm, SVC algorithm and Gaussian Mixture Models(GMM).

In the third and final phase, performance of the system is tested by providing a test dataset of voices as input to the model and in return it predicts the speaker. Technically when a new voice sample is to be identified then at first, features are extracted from the voice sample and then the trained speaker model is used to calculate the score of the features for all speaker models. The speaker model with the maximum score is predicted as identified speaker for the test data. And finally, the transcript from the speech-to-text conversion model is tagged with it's respective speaker where the speaker's information is predicted by the speaker-identification model.

V. FUTURE SCOPE

Currently, in our proposed system we have used basic classification algorithms and have reached 70% accuracy. And it can be further improved by training the model over larger datasets, using audio from different environments over larger datasets and also by using audio of different frequency ranges from different environments. More acoustic features can be used during feature extraction. Further we are planning to make the user interface of the system more user friendly with options like a live transcription window and a mail integrated system to generate reports for respective participants. This system can be put to further use in schools and for the people with deafness. Thus this can make drastic changes in the future of AI transcription.

VI. CONCLUSION

We tried to design a dedicated cost effective and comparatively faster software solution which is an alternative for stenographers at law courts to transcribe every word that is said during an examination for discovery, deposition, arbitration, live trial and more to prepare a transcript using various algorithms. Algorithms for speech-to-text

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 12, December 2020 ||

conversion and speaker identification were surveyed and adopted in the initial stages of the project implementation. We reviewed relevant research papers with different algorithms that can be adopted to identify the best methodology providing the highest accuracy and ensuring robust speech recognition. Also the solution is flexible enough to be used for numerous applications for getting real time transcriptions by doing minimal changes according to the requirement.

REFERENCES

[1] Pratap, Vineel & Xu, Qiantong & Kahn, Jacob & Avidov, Gilad & Likhomanenko, Tatiana & Hannun, Awni& Liptchinsky, Vitaliy & Synnaeve, Gabriel & Collobert, Ronan. (2020). Scaling Up Online Speech Recognition Using Convnets.

[2] Ainon Sahirah bt Mohd Jarkasi "Minute Meeting System Using Speech Recognition", 2020.

[3] Ragni, Anton & Gales, M.J.F.. (2018). Automatic Speech Recognition System Development In The "Wild". 2217-2221. 10.21437/Interspeech.2018-1085.

[4] Mazin Abed Mohammed , Karrar Hameed Abdulkareem , Salama A. Mostafa , Mohd Khanapi Abd Ghani, Mashael S. Maashi , Begonya Garcia-Zapirain , IbonOleagordia , HosamAlhakami And Fahad Taha Al-Dhief "Voice Pathology Detection And Classification Using Convolutional Neural Network Model", Applied Sciences, 2020.

[5] Jurgen Arias "Voice Classification With Neural Networks", 2020.

[6] R.Thiruvengatanadhan "Speech Recognition Using AANN", 2019, International Journal 0f Engineering Research And Technology (IJERT).

[7] Chunling Tang , Min Li " Speech Recognition In High Noise Environment", 2019, Ekoloji 14(55): 8-23.

[8] Orken, Mamyrbayev & Mekebayev, Nurbapa & Turdalyuly, Mussa & Oshanova, Nurzhamal & Medeni, Tolga & Yessentay, Aigerim. (2019). Voice Identification Using Classification Algorithms. 10.5772/Intechopen.88239.

[9] Flemotomos, Nikolaos & Papadopoulos, Pavlos & Gibson, James & Narayanan, Shrikanth. (2018). "Combined Speaker Clustering And Role Recognition In Conversational Speech". 1378-1382. 10.21437/Interspeech.2018-1654.

[10] Douglas Coimbra De Andradea, Sabato Leob, Martin Loesener Da Silva Vianac, Christoph "A Neural Attention Model For Speech Command Recognition", 2018, Engineering Applications Of Artificial Intelligence (EAAI), 1808.08929.

[11] S. P. Dubagunta And M. Magimai.-Doss, "Segment-Level Training Of ANNS Based On Acoustic Confidence Measures For Hybrid Hmm/Ann Speech Recognition," ICASSP 2019 - 2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6435-6439, DOI: 10.1109/Icassp.2019.8683513.

[12] Tejashri Manohar Mhatre, Srijita Bhattacharjee, "Birds Voice Classification Using RESNET", International Journal Of Engineering Development And Research (IJEDR), Issn:2321-9939, Vol.6, Issue 4, Pp.168-172, November 2018.

[13] Huang, Hengguang, Hao Wang And B. Mak. "Recurrent Poisson Process Unit For Speech Recognition." AAAI (2019).

[14] Abhijeet Kumar "Spoken Speaker Identification Based On Gaussian Mixture Models : Python Implementation", 2017, Applied Machine Learning Blogs 2017/11/14.

[15] Pala, Mahesh. (2016). A New Human Voice Recognition System. AISAT. 5. 23-30.

[16] Das, Prerana & Acharjee, Kakali& Das, Pranab & Prasad, Vijay. (2015). "Voice Recognition System: Speech-To-Text". Journal Of Applied And Fundamental Sciences. 1. 2395-5562.

[17] S. P. Dubagunta, S. Hande Kabil And M. Magimai.-Doss, "Improving Children Speech Recognition Through Feature Learning From Raw Speech Signal," ICASSP 2019 - 2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 5736-5740, DOI: 10.1109/Icassp.2019.8682826.

[18] Jain, Neha & Rastogi, Somya. (2019). Speech Recognition Systems – A Comprehensive Study Of Concepts And Mechanism. Acta Informatica Malaysia. 3. 01-03. 10.26480/Aim.01.2019.01.03.

[19] M. U. Anwar And M. L. Ali, "Boosting Neuro Evolutionary Techniques For Speech Recognition," 2019 International Conference On Electrical, Computer And Communication Engineering (Ecce), Cox's Bazar, Bangladesh, 2019, pp. 1-5, DOI: 10.1109/Ecace.2019.8679206.

[20] https://github.Com/Atul-Anand-Jha/Speaker-Identification-Python

[21] Moritz, N., Hori, T., Le Roux, J., "Triggered Attention For End-To-End Speech Recognition", IEEE International Conference On Acoustics, Speech, And Signal Processing (ICASSP), DOI: 10.1109/Icassp.2019.8683510, May 2019.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details