Grenze International Journal of Engineering and Technology, Jan Issue

An Overview of Speaker Recognition: Conceptual Framework and CNN based Identification Technique

Avinash Dhole[1] and Vijaylaxmi Kadroli[2]
[1-2]Terna Engineering College, Mumbai University, Navi Mumbai, Maharashtra, India
Email: avinashpdhole@gmail.com

***Abstract*—One of the major characteristics of an individual is to recognize a human voice. He is capable to do so, via digital devices or over the telephone. Acquiring this human trait, multiple technologies based on voice recognition have been developed to fulfil the purpose of biometrics and authentication. One such speech analysing technology is automatic speaker recognition (ASR) method that has been introduced to extract characteristics from the speaker's voice and identify him as a genuine source of input. The primary aim of speaker recognition is to identify and verify a person using audio signals. Hence, this has become a dominating field of research in the domain of biometrics. However, several deep learning approaches based on convolutional neural networks have been adopted to enhance the overall system of biometrics. Therefore, in this paper we review feature components on speaker recognition using CNN and feature extraction methods such as MFCC. The major advantage of using this method over traditional identification is its representation ability to extract feature inputs including audio signals and networking structures. Further, we briefly describe all the main pieces of ASR related methodologies followed by evaluation metrics to enhance the overall recognition of the system. Finally, a few relatable challenges and future expansion of speaker recognition are mentioned at the closure of this review.**

***Index Terms*— Automatic speaker recognition, convolutional neural network, feature extraction, MFCC, performance measures.**

I. INTRODUCTION

The voice of a speaker generally depicts human traits with regards to his pronunciation and speaking manner, such as accent, frequency of words and rhythm. Therefore there is a high possibility of identifying the voice behind a speaker. This technology of identifying voices is termed as voice recognition and can further be categorized as speech recognition and speaker recognition. The approach for analysing only the contents of a speech by converting sound waves into useful data is known as speech recognition. In this method, algorithms are used to generate textual forms of output which are further interpreted by the machine. Whereas, speaker recognition is considered to be more of a fundamental task and involves categories of speech processing. Since every speaker inhibits a pattern of speaking his characteristics such as intonation style and choice of vocabulary is monitored to identify and is used to differentiate a person from his voice [1]. With its popularity in developing valuable biometric recognition technology, the concept has drawn massive attention from research scholars in broader fields of information security for many years. Hence, this field is widely used in real world applications such as building smart devices for voice based authentication systems. This concept has also created a domain in

Grenze ID: 01.GIJET.9.1.335

forensics and is used for investigation purposes and identity tagging.
With much broader concepts and fundamentals, the study of speaker recognition can be seen as utilization of statistical methods to identify human voices based on their unique characteristics. These human characteristics are fed to the machine in the form of audio signals and are later encoded to a sequence of samples over a span of time. With the actual implementation of the concept, this method is witnessed to possess two modes, namely: speaker verification and speaker identification. The former study, tries to highlight whether the claimed speaker is a true speaker or not, while the latter aims to focus in identifying the speaker.
The process of identifying a speaker is determined through their speech signals which are selected using feature extraction techniques and are compared against templates. These templates are responsible to perform a 1: N match processing of extracted features to later form speech signals. This is considered to be a pivotal task as human speech signals are a powerful medium of communication; inhibiting emotional traits and accents. These traits enable research scholars to conduct voice print recognition. All the collected utterances from the speaker are fed as input for training purpose [2]. Once all the extracted features are matched with their respective speaker, the identification system updates this information in the model library; and the speaker with highest probability of utterance is defined as the target speaker. However, this recognition of human voice heavily depends on the length of the audio signal, the environment in which the audio is recorded and the physical condition of the speaker. In addition to this, it is worthy to note that the overall performance of the recognition system decays in a noisy environment and hence becomes unsatisfactory [3]. Thereby, it is necessary to adapt an automated system that could serve its purpose to identify the speaker's voice. With a constant rise of accessible software and affordable hardware, the paradigm of automated systems has massively shifted to deep learning techniques. Figure below depicts an overview of deep learning in speaker recognition.

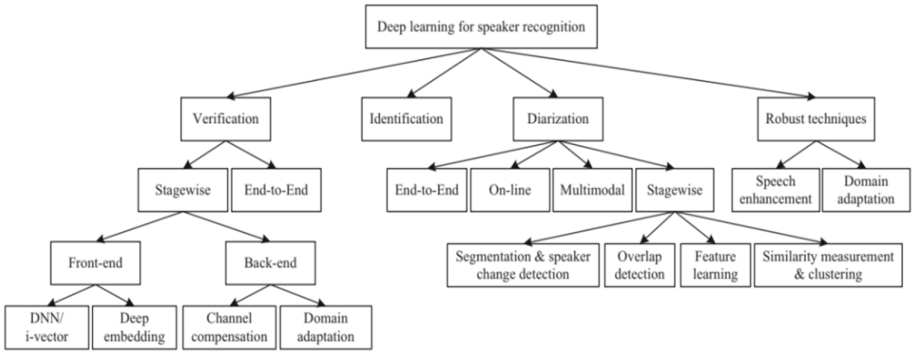

Figure- -1 An overview of deep learning concept for speaker recognition [2, 3]

Both the concepts of speaker verification and identification require its own use cases and thereby share similar recognition techniques. However, motivated by powerful concepts of deep learning techniques, we propose a review to examine CNN based speaker recognition techniques and methods in terms of research topic including, verification, identification, performance metrics and feature extraction method such as MFCC.
Therefore, the primary aim of the review is to provide resourceful information for speaker recognition community. The main contributions of this work are as follows:

- To provide a summarized literature review on speaker identification, verification and recognition
- To clearly differentiate between speaker recognition and speech recognition
- To mention the fundamental theory of deep learning based CNN algorithm
- To describe MFCC based feature extraction method
- To evaluate the proposed study based on hyper-parameters
- To briefly discuss challenges and future scope of the proposed study

The implementation of a speaker recognition model generally goes through multiple challenges. In data-driven strategies, intra-speaker changeability is the most common challenge. Apart from this, the model undergoes certain challenges as mentioned below:

- Limited data and constrained lexicon: the implementation of enrolment lexicon used for industrial purpose is generally comprised of repetitious duplications and further involves a trial period of unique replication. This replication is a sub-portion of the entire speech consisting an audio signal of 4-5 seconds. However, there are certain obligations associated with this challenge with majorly comprising of constraints that are illustrated on the customers end during enrolment and testing sessions

- Forward compatibility: the primary aim of this principal is to revise the enrolee database and create an application based on its underlying structures. However, the application produced on the security layer is based on the lexicon of the speaker's name. Hence, this becomes a major limitation as this feature becomes an obligatory part of the speaker model and any alterations made in this model might later harm the efficiency of the entire system
- Speaker-based variability: this limitation of the model is based on how the speaker speaks and how his voice shall affect the overall accuracy of the ASR model. However, this can also be regarded as an inherent variability of the speaker and further involves determinants such as verbal style and emotions involved while speaking

II. Literature Review

With advancements in deep learning technology, multiple research scholars have extended their study in voice recognition. Some authors have implemented convolutional neural networks as the base model whereas others have utilized feature extraction methods in identifying the same. However, the generated results depict the probability that the same concept can be applied to different audio domains and serve as a biometric authentication model. Therefore this section of the paper highlights research study on identification, verification and extraction methods used for voice recognition.

A. Speaker Recognition

In a survey conducted by Tirumala et al in [4], the author presented a systematic review on feature extraction approaches for voice recognition. The major highlight of the review was based on feature extraction methods adopted in the past. However, the major drawback mentioned by the authors was regarding the presence of noise in the audio signals that were received as inputs. This redundant and irrelevant noise in the data created distortion while training the model. Hence, the author focused on research questions that could improve the speaker recognition domain and optimized foundations based on the process involving feature extraction. The review also briefed on various challenges experienced by the domain. In a similar work [5], the authors discussed various speaker identification techniques that were based on deep learning. They tried to characterize and bifurcate its implementation using detection techniques and identification algorithms. The authors also contributed their work in describing the taxonomy and the architecture of deep learning. However, the primary aim of the authors was to bridge the gap between speaker recognition techniques and deep learning architectures.
In a research study by author Qawaqneh et al in [6], he proposed voice recognition using modalities such as text-independent and text-dependant. The working model of his system made use of only voice characteristics for text-independent case whereas the system combined voice as well as spoken words in text-dependent case. His work was further expanded by Rozi et al in [7]. It was keenly observed that the recognition system further involved three principle steps of feature extraction, modelling and scoring. Modelling is considered to be one of the most important criteria in voice recognition. This concept was further used to distinguish between two classical speaker models namely; template models and stochastic models.
The primary aim of speaker recognition is to convert the given input of audio signals into computer understandable format. To achieve this purpose, recognition systems implement their function in two phases of training and testing. In the training process, speech signal is taken as input and feature extraction is done to represent voice characteristics of the user. Whereas, in the testing phase the same audio signal is matched with the referral model using matching techniques and templates. Based on the recognition criteria, automatic speaker recognition (ASR) can be classified using multiple approaches as displayed in the figure below:

B. Speaker Verification

The fundamental of speaker verification is to identify the voice of a legit user and grant him authorization to access the information he requires. Hence, this aims at verifying whether the audio signal pronounced by the speaker is genuine or not based on his pre-recorded utterances. This verification process of speaker's voice can be categorized as end-to-end and stages ones. A stage wise algorithm usually contains a front end and a back end to extract speaker features and calculate similarity of speaker features respectively. The front end of the system is responsible to convert audio signals into high dimensional feature vectors and the back end accounts for calculation of similarity score between enrolments of test features to that of threshold values. In contrast to the working of a stage wise algorithm, the end-to-end process takes a pair of audio signals as input and directly converts them to similarity score.

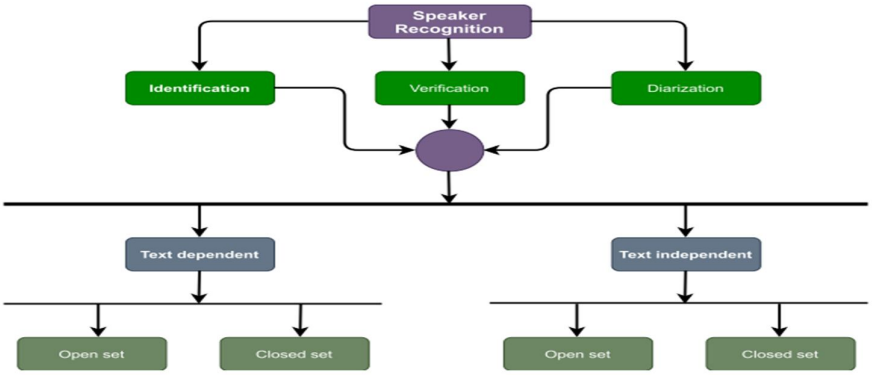

Figure -2 Approaches for speaker recognition [6, 7]

C. Speaker Identification

This step of speaker recognition aims to detect the speaker's identity from the enrolled database. It tends to do so by identifying a specific voice from previously recorded utterances of the speaker. For this purpose, it utilizes feature extraction technique and finds the match of the exact speaker from a set of recognised voice of speakers. This approach is commonly known as the 1: N match method wherein a particular exclamation is compared against N templates. In a research work presented by authors in [8] the authors proposed speaker verification model using text independent mode and were executed based on similarity measurement calculated using Cosine distance. However, in noisy conditions, the results were not good enough as much as they were in clean conditions. That was expected because it is difficult to look for pertinent information in corrupted feature vectors. The purpose of verifying the speaker's voice was later implemented on open source channels by Bouziane et al in [9] and a filtered speech was obtained using i-vector/PLDA and GMM-UBM.

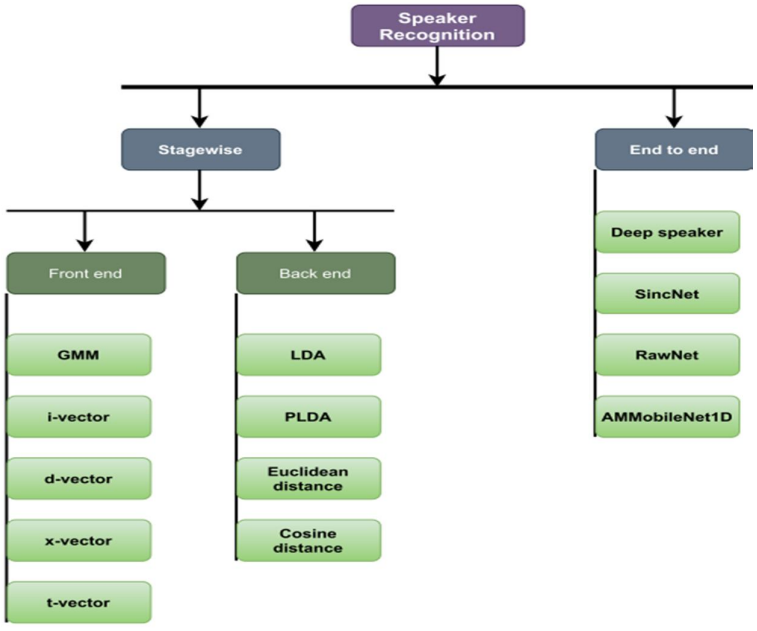

Figure -3Flow of Speaker Verification [8]

D. Feature Extraction techniques

The process of feature extraction is done to extract features and interpret an audio signal by applying mathematical tools to pre-set the signals number of components. Since majority of the speech signals in a dataset

are heavily acoustic, the data becomes inappropriate for processing and hence needs to be filtered for performing speaker recognition tasks. This section of the paper briefs on three commonly used feature extraction techniques.

Linear Predictive Coding (LPC)
The primary assumption of this technique is that a speech signal is produced by a buzzer at the end of a tube. Therefore LPC, analyses for any such formants and further estimates the intensity of the remaining buzz. The process used for removing such formants is commonly termed as inverse filtering and the remaining signal is called as the residue [10]. However, a major drawback of this technique is degrading performance of the system in presence of noise.

Linear Predictive Cepstral Coefficients (LPCC)
All the drawbacks of a conventional LPC technique are majorly discarded using LPCC algorithm [11]. Hence, implementation of this technique gives a smoother spectral envelope and stable representation when compared to LPC.

Mel-Frequency-Cepstral Coefficients (MFCC)
MFCC is the most commonly and widely used technique for application of speech signal from feature extraction. However, this tends to form the first step in ASR and is a means to identify all the components of an audio signal [12]. This algorithm is further responsible to implement only the required features for voice authentication and discards all other features that are redundant and tries to carry background noise along with it.

TABLE I. COMPARISON

Authors	Dataset	Methodology	Accuracy
Purnima Pandit.et.al [15]	Worked on the dataset on Gujarati Language	Implementation done using MFCC Feature Matching Using the concepts of Dynamic Time Warping	94.43%
Satyanand Singh.et.al [16]	Worked on a dataset of 70 speakers to recognize utterances	Implemented using MFCC, Inverted MFCC using Triangular Filters and Gaussian Filters	97.36%
Naveen Kumar [17]	Worked on the dataset of stuttered speech	Implemented using Mel Frequency Wrapping Discrete Cosine Transform (DCT) technique.	96.43%
Mousmita Sarma et al [18]	Worked on the dataset of Assamese Language	Implementation was performed using RNN and Multilayer perceptron.	93.24%
V. Anantha Natarajan [19]	Worked on the dataset of TV broadcasting which consisted of male and female voices	Implementation done using Hamming window	92.36%
Kishori R. Ghule et al [20]	Worked on the dataset of 100 speakers that involved the voice of different plant names	Implementation was done using Discrete Wavelength Transformation (DWT)	96.65%
Harpreet Kaur and et al [21]	Worked on the dataset of 21 Punjabi speakers	Implementation was done using MFCC Algorithm	97.56%

III. METHODOLOGIES

A. CNN

The fundamentals of a deep neural network are categorized as CNN, RNN and LSTM structures. However, the networking structure adopted for speaker recognition is quite flexible in nature and can exhibit properties of each structure as required. Each component associated with the deep learning network possesses multiple candidates within itself. For example, the conventional hidden layer of a neural network might comprise of a convolutional layer, a dilated convolutional layer, an LSTM layer, and FC layer or a combination of any of these mentioned layers. A typical neural network also consists of activation functions such as Sigmoid and Rectified Linear Unit

(ReLU). Apart from the mentioned layers and the topology of the existing network all the modes which are present in between the layers are considered to be variables. These variables and the hidden units at every layer are further responsible to affect the overall performance of the system. Hence, the generated output such as confusion matrix,
Accuracy of the system and precision factors are heavily dependent on the numerical values and activation functions of the system.
On the other hand, the fundamentals of a Convolutional neural network (CNN) are considered to be one of the most commonly used deep learning techniques used by research scholars in multiple domains of security and authentication. The basic implementation of this method follows the operation of convolutional layers. The fundamentals of a CNN are based on three networks, namely; sparse interaction, parameter sharing and equivalent representation. However, the process of feature extraction from inputs occurs using the convolutional stage, non-linearity stage and the pooling stage. All these layers are connected to each other via hidden layers in the network and are consecutively extracted using deep features. The initial layers of the network are responsible for extracting features from the input. These extracted features are later trained in the learning process and send to the next layers of the network for classification stage. The last layer of a CNN is present in the classification stage and is termed as the Fully Connected layer (FC). It is in this layer that the Softmax function is applied and the final output is generated. In the implementation model of a CNN, all the existing layers as mentioned above are arranged in order to form a network so that a learning process of feature extraction can be developed. Other stages of the network also involve dropout and normalization [13]. Next, all the resultant extracted features obtained from the learning stage are arranged in the form of vector matrix and FC layers are connected with the Softmax layer. Finally, in the last layer which is present before the output, the Softmax function is obtained and applied to transform the output values of the network. Once the entire network is trained using these function, evaluation parameters are applied to modify its performance.

B. Deep Features Extraction

The implementation phase of a deep learning model is executed in two steps. The first one involving a generative phase and the second one involving a discriminative phase. The overall training of the working model takes place in the generative phase in an unsupervised manner, whereas the supervised training of each feature vectors are assigned in the discriminative phase [14]. However, in the entire process of feature vector conversions, MFCC is commonly used for feature mining in speaker recognition tasks. A standard MFCC based extraction process is depicted in figure below:

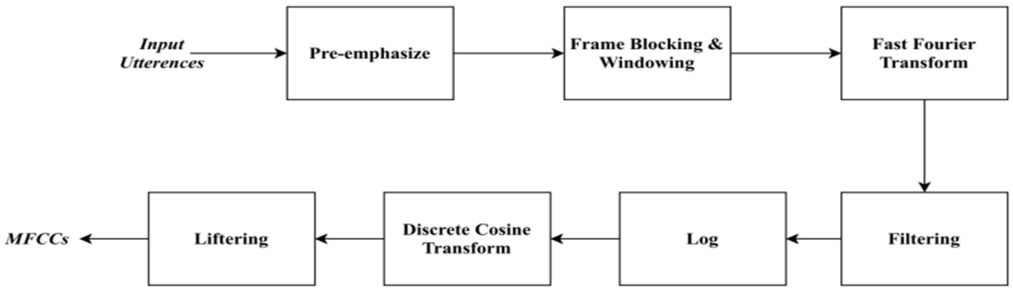

Figure -4 Block Diagram of steps depicting MFCC feature extraction [14]

In the initial phase, a pre-emphasis input is applied to the signal so that all the high frequencies are expanded. The audio signal which comes as an input x is implemented on this filter in the following equation:

$$y(t) = x(t) - \propto x(t-1) \qquad \textit{Equation 1}$$

Where, α is the filter coefficient.
x is audio signal input
Once the pre-emphasis phase is completed, the audio signal is converted into low time frames and further split into short time frames, using Hamming functions. In the next stage, a fast Fourier transform is applied and mapping from linear frequency to MFCC is performed. In the final stage, the MFCC is calculated using the formula given below:

$$MFCCi = \sqrt{2/N} \sum_{j=1}^{N} mj \cos\left(\frac{\pi i}{N}(j\text{-}0.5)\right) \qquad \textit{Equation 2}$$

Where, N is the number of bandpass filters;
mj is the log bandpass filter output amplitudes.

IV. Evaluation Parameters

A. ROC (Receiver Operating Characteristics) Curves

In order to measure all the recognition problems occurring at numerous thresholds, evaluation parameters such as area under the curve (AUC) and receiver operating characteristics (ROC) is utilized to demonstrate the measure of seperability. This generally represents the probability curve and the architecture used to represent the class. The threshold value is used to classify items as positive that in turn increases true positive followed by false positive values. However, there are two ways to generate ROC curves:

- True Positive Rate
- False Positive Rate

B. Equal Error Rate

The algorithm used to accurately foretell all the values of thresholds for false receiving and false rejection rates is known as the equal error rate (ERR). This rate is considered to be as a standard value which is similar to that of the error rates obtained from the threshold value. Further, this value is a representation of false acceptance and false rejections and tends to be equal in terms of percentage. Fundamentally, it is a mathematical method of detecting errors and error margins in inaccurate results.

C. Detection Trade-Off

A visual representation used to classify binary error margins is called as a detection error trade off. It is responsible to plot false rejection values to that of false acceptance values. However, a curve is used map error curves on a regular deviate scale and is referred to as the DET curve. The obtained curves are plotted on x and y axis and are later measured on regular natural scale. Hence, this result in more linear trade-offs than the generated ROC curves. Finally, this method is witnessed to observe sequential values and later emphasize on the significance so created in the critical operating zone.

V. Conclusion

The domain of speaker recognition is one of the most widely investigated numerous system to identify legit speaker's from their voice. However, a very limited survey has been conducted until now in this research and yet needs more techniques to authenticate the process. Therefore, this survey paper aims to focus on this renowned research domain in multiple aspects such as fundamentals, extraction techniques, performance evaluation and challenges involved. Finally, the paper ends with challenges faced by the ASR system. The paper suggestively upholds the existing defaults and variations of ASR systems that would benefit new researchers quickly adapt the research domain concepts.

References

[1] N. Singh, A. Agrawal, and R. Khan, ''The development of speaker recognition technology,'' Int. J. Adv. Res. Eng. Technol., vol. 9, no. 3, pp. 8–16, 2018

[2] N. Singh, ''A study on speech and speaker recognition technology and its challenges,'' in Proc. Nat. Conf. Inf. Secur. Challenges. Lucknow, India: DIT, BBAU, 2014, pp. 34–37

[3] C. D. Shaver and J. M. Acken, ''A brief review of speaker recognition technology,'' Elect. Comput. Eng. Fac. Publications Presentations, Portland State Univ., Portland, OR, USA, Tech. Rep. 350, 2016

[4] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, ''Speaker identification features extraction methods: A systematic review,'' Expert Syst. Appl., vol. 90, pp. 250–271, Dec. 2017

[5] S. S. Tirumala and S. R. Shahamiri, ''A review on deep learning approaches in speaker identification,'' in Proc. 8th Int. Conf. Signal Process. Syst. (ICSPS), 2016, pp. 142–147

[6] Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed mfccs for speaker's age and gender classifcation. Knowledge-Based Systems, 115, 5–14

[7] Rozi, A., Wang, D., Zhang, Z., & Zheng, T. F. (2015). An open/free database and benchmark for uyghur speaker recognition. In: Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/ CASLRE), 2015 International Conference, IEEE, (pp. 81–85)
[8] Hourri, S., & Kharroubi, J. (2019). A novel scoring method based on distance calculation for similarity measurement in text- independent speaker verifcation. Procedia Computer Science, 148, 256–265
[9] Bouziane, A., Kadi, H., Hourri, S., & Kharroubi, J. (2016). An open and free speech corpus for speaker recognition: The fscsr speech corpus. In Intelligent Systems: Theories and Applications (SITA), 2016 11th International Conference on, IEEE, (pp. 1–5)
[10] R. Kumar, R. Ranjan, S. K. Singh, R. Kala, A. Shukla, and R. Tiwari, ''Multilingual speaker recognition using neural network,'' in Proc. Frontiers Res. Speech Music (FRSM), 2009, pp. 1–8
[11] A. M. Ahmad, S. Ismail, and D. F. Samaon, ''Recurrent neural network with backpropagation through time for speech recognition,'' in Proc. IEEE Int. Symp. Commun. Inf. Technol. (ISCIT), vol. 1, Oct. 2004, pp. 98–102
[12] S. Narang and M. D. Gupta, ''Speech feature extraction techniques: A review,'' Int. J. Comput. Sci. Mobile Comput., vol. 4, no. 3, pp. 107–114, 2015
[13] Chen, Y.-H.; Ignacio, L.-M.; Sainath, T.N.; Mirkó, V.; Raziel, A.; Carolina, P. Locally-connected and convolutional neural networks for small footprint speaker; recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015
[14] Singh, S., & Rajan, E. (2011). Vector quantization approach for speaker recognition using mfcc and inverted mfcc. International Journal of Computer Applications, 17(1), 1–7
[15] Purnima Pandit, Shardav Bhatt "Automatic Speech Recognition of Gujarati digits using Dynamic Time Warping", International Journal of Engineering and Innovative Technology (IJEIT), Vol.3, Issue 12, ISSN: 2277-3754, June 2014, pp.69-73
[16] Satyanand Singh and Dr. E.G. Rajan, "Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC", International Journal of Computer Applications, Vol.17, No.1, ISSN: 0975 – 8887, March 2011, pp.1-7
[17] V.Naveen Kumar, Y Padma Sai and C Om Prakash, "Design and Implementation of Silent Pause Stuttered Speech Recognition System", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.4, Issue 3, ISSN (Online): 2278 – 8875, March 2015 , pp.1253-1260
[18] Mousmita Sarma, Krishna Dutta and Kandarpa Kumar Sarma, "Assamese Numeral Corpus for Speech Recognition using Cooperative ANN Architect", International Journal of Electrical and Electronics Engineering,Vol.3,Issue8,Nov 2009,pp.456-465
[19] V. Anantha Natarajan S. Jothilakshmi, "Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies", International Journal of Computer Applications, Vol. 56, No.15, ISSN: 0975 – 8887, October 2012, pp.24-27
[20] Kishori R. Ghule and Ratnadeep R. Deshmukh, "Automatic Speech Recognition of Marathi isolated words using Neural Network", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6(5), ISSN: 0975-9646 2015, pp.4296- 4298
[21] Harpreet Kaur and Rekha Bhatia, "Speech Recognition System for Punjabi Language", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 8, ISSN: 2277 128X, August 2015, pp. 566-573.