

An efficient Multilingual Speaker Recognition system using fusion technique

¹Mayur Rahul, ²Sonu Kumar Jha, ³Sarvachan Verma, ⁴Vikash Yadav, ⁵Devendra Kumar Dellwar

¹Department of Computer Application, UIET CSJM Kanpur Nagar, Uttar Pradesh, India

²Krishna Engineering College, Ghaziabad, Uttar Pradesh, India

³Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India

⁴Government Polytechnic Bighapur Unnao, Department of Technical Education, Uttar Pradesh, India

⁵SR Group of Institutions, Jhansi, Uttar Pradesh, India

Abstract

The robustness and performance of speech signal based framework depends on the quality of features. In the today's era of research, working of single feature might not be enough to cover both robustness and performance simultaneously. In order to resolve this problem, researchers use multiple sources by applying various fusion techniques. These fusion techniques are categorized into few categories: Model level, Feature level and Score level combination scheme. The documents available in previous research shows the features available from different sources are used to enhance the strengths and recognition rate of the system. Even though these fusion techniques enhance the strengths and recognition rate of the system, but they found some demerits in the system. This will helps us to investigate further. The aim of the work is to introduce a system for multilingual speaker system with the help of SVM using fusion technique. The objective is to explore the advantage of various fusion techniques and how these techniques are useful to build efficient system for multilingual speaker system. The results from our proposed system indicate goodness of our work.


Keywords

Multilingual speaker recognition, SVM, fusion techniques, model level

1. Introduction


The speech recognition can be classified into two categories: speech recognition and speaker recognition. These systems consist of extracting important information form speech signals and identifying the required results by machine. In the case of speaker recognition, the machine tries to retrieve information based on any specific criteria from given speech signals and in speech recognition, only textual information is extracted from speech signals. They are similar to the pattern recognition systems. The accuracy of the system is depending on the discriminating power of the features used in the process. The feature extraction generally depends on the type of tasks. In case of speaker recognition, the machine calculates linear prediction cepstral coefficients (LPCC) or mel-frequency cepstral coefficients (MFCC) characteristics which represents speaker based vocal information in precise form [1, 2, 3].

Proceedings Acronym: Smart Cities Challenges, Technologies and Trends, December 07, 2023, Rohini Delhi, India
✉ mayurrahul209@gmail.com (M. Rahul); sonukumarjha1990@gmail.com (S. K. Jha);
sarvachan.verma@gmail.com (S. Verma); vikas.yadav.cs@gmail.com (V. Yadav); devendra.kumar6@gmail.com (D. K. Dellwar)

 0000-0002-2394-865X (M. Rahul); 0009-0009-4378-7302 (S. K. Jha); 0009-0003-7588-9449 (S. Verma);
0000-0003-1348-1379 (V. Yadav); 0009-0007-8928-2321 (D. K. Dellwar)



© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The researchers also explore speaker based information as alternative proof using various fusion methods. These methods provide better performance as compared to the independent vocal based systems. Moreover, these systems are comparably robust against various conditions [4]. The MFCC characteristics retrieved from phoneme samples are used as main characteristics for speech recognition systems. These MFCC characteristics shows the spectral envelope design of various phonemes, which are used for speech recognition system. The speech recognition systems is a speaker independent procedure, therefore need huge amount of information to efficiently represent the phoneme based information. To remove those complications people use much information. Tripathi et al. proposed different kinds of source information and then incorporated with MFCC characteristics by using given fusion methods [5]. They have also showed that combination of source information and MFCC characteristics not only enhance accuracy rate but also improves the robustness of phoneme recognition process.

The information consists of source excitation is generally used as additional proof with tract information to get enhanced information in various speech recognition systems [5, 6, 7]. The purpose for using source based excitation information as additional proof has two reasons: people use excitation features like duration, intonation and pitch to identify speakers as well as the matter of the speech data [8, 9, 10, 11]. People have proven themselves powerful even in decadent conditions, representing the capability of the excitation source data [12]. The other reason is the approbative description of source and vocal information. This approbative description gives additional proof that is use to enhance the performance and robustness of the baseline framework. The researchers also observe that combination of source excitation information and vocal tract enhance the robustness and performance of the speaker and speech recognition framework [13,14,15].

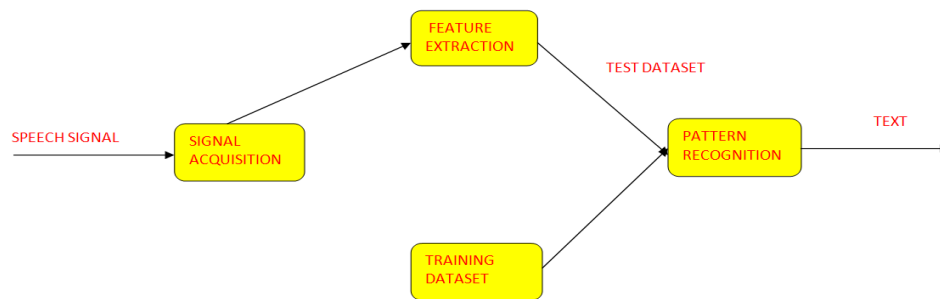


Figure 1: Block diagram of a generic speech based recognition system [16]

The performance of combined given system depends on importance of the features as well as on the suitable fusion methods. The optimized advantage can be found with the help of suitable fusion and effective features. Since excitation source and MFCC information are paramount for various speeches processing frameworks [17, 18, 19], getting optimized performance mostly based on the applied fusion technique. The fusion of features can be processed at the comparison, model or feature level. This could be better explained from the diagram of speech recognition shown in figure 1. The speech sample is processed to make them input for feature extraction step in pre-processing steps. The purpose of the feature extraction step is to calculate the required features by applying various signal processing

techniques. In feature level fusion, various features are calculated and fused for creating models. A similar technique is followed to calculate the test characteristics and used for matching. In case of Model based fusion, various models are created using individual feature sets. Further, the different models parameters are combined to create composite models. Finally, the comparison is created with test speech specimen and composite model. In score level fusion, different characteristics are obtained from given voice signal and used to create the corresponding models. During matching, the given features are matched with corresponding models, and calculate individual score. These score are combined to give final score.

In the speech recognition system, features represent the corresponding information about the job in a precise form. These given features are then used for creating blocks for various classes/pattern. For example, phonemes design in automatic speech recognition and speakers design for automatic speaker recognition. The existing work shows that instead of using single features, fusion of given multiple features gives optimized classes for speech based pattern identification tasks. Moreover, fusion of various features not only enhances the robustness but also performance of the systems. For example, recent researchers have shown the benefits of different features speech recognition systems, for automated speech recognition, and replay identification systems. In these researches, the fusion based techniques are limited to combination of features at every level. These techniques have their own advantages and disadvantages. A combined fusion technique could be created by utilizing the advantages of individual combination scheme which can be effective, efficient and useful for different speech processing systems. The target is to elaborate the advantages of different speaker recognition systems and apply them for the improvement of the effective recognition scheme. The main findings of the research work are as follows:

- (1) The paper introduces a literature review of various types of speaker identification systems with its historical background.
- (2) The paper summarizes the feature extrication, datasets, accuracy and demerits of existing work.
- (3) The paper introduces the SVM based multilingual speaker recognition using MPDSS, RMFCC, and MFCC features.
- (4) The paper introduces the combination of MPDSS, RMFCC, and MFCC in TIMIT, NIST 2003 SRE datasets.
- (5) The performance of our paper is best when compare with other existing work.

The remaining part of paper is sketched as follows. In section II, we explain the related works. Section III presents the research methodology used in multilingual speaker recognition. The experiments and results are demonstrated in Section IV. Finally, Section V concludes the paper with future works.

2. Related works:

The speech recognition system refers to extrication of important information from speech signal by using different signal processing techniques for some applications. People's speech

reflects effectively the textual content and speaker information recognition. The speech processing systems are generally categorized into two categories: speech recognition and speaker recognition. The extrication of textual data present in speech is called speech recognition system, and the speaker data is used to identify speaker is called as speech recognition system. We consider the systems related to above two fields as benchmark to represent the robustness of the proposed method. A detailed explanation about the speech and speaker recognition is given in present section.

The method of identifying people by machine using the data available in speech samples is called speech recognition systems (SRS). The SRS is broadly categorized in two categories: Automatic speaker verification system (SVS) and speaker identification system (SIS). In SIS, the objective of the machine is to detect the speaker from the given test samples, whereas, in SVS, the objective is to verify the particular identity with the help of given speech samples.

The entire SRS process consists of two parts: training and testing [20]. In training step, machine gathers the given speech sample from the speaker and register them by using SRS technique. The training step consists of feature extraction and creating models. The speaker based information is retrieved in feature extraction step from each and every sample by using various signal processing techniques and represent it in parametric form. These important features are then used in modeling stage to create model. In testing step, the machine calculates the speaker based features from test sample by using same feature extraction technique as used in training step, and used to compare with the existing model. Depend on the task, comparison processed in the comparison step. The comparison steps gives matching score that identify of the speaker for the speech samples.

The existing systems predominantly use cepstral computing technique for feature extrication and probabilistic technique like Gaussian Mixture Model (GMM) [21,22]. Based on the given speech samples the SRS are classified into two categories: Text independent and text dependent. In case of text dependent, the speakers kept for test are required to present same speech sample as given at enrollment process. There is no textual limitation on text independent model. They are used for real-time.

In the field of speaker recognition, additionally two research areas including limited data based speaker verification and replay attacks identification. In comparison with traditional speaker verification system, limited data based speaker verification requires less amount of data for testing and training processes. As smaller amount of data is used, the limited data based speaker identification is very challenging task in the area of speaker recognition. The replay attack is a kind of spoofing attack to automated speaker verification task, where the decisions can be changed by prerecorded speaker samples by recording and playback devices. It doesn't require any technical knowledge, only a smart phone is needed for spoofing. The existing reviews shows that replay attack is highly efficient and effective and easily accessible constitute a critical threat to automated speaker verification.

Speaker identification is a method to identifying the speakers by using speech samples. A set of well known speakers are enrolled by the machine and used as reference patterns for recognizing the unknown speaker. The speaker identification system is performed in two steps: testing and training steps. In training step, individual speaker based features are retrieved from the set of speakers and used to create respective reference models. In common excitation source based information and vocal tract are used for creating reference models. In

testing step, the same speaker based features are extricated from the test based speaker samples, and used for matching with the entire stored speaker design for recognition.

The speaker verification is the method of identifying the unknown applicant to a reference design by given speech samples. It is very clear that the applicant should be registered by the machine before placing the application. So, firstly applicant is asked to give speech samples for registration. Further, during verification, the voice samples are compared by matching with the corresponding samples. The decision is purely based on the threshold. The matching score is greater than threshold, it is accepted otherwise rejected.

The limited data based speaker verification refers to an identification task where the availability of testing and training data is very less say less than 10 sec. The forensic based investigation where data is less, performance is mostly affected. This is also affected due to inadequate coverage of the speech samples. So, effective and efficient technique is required for these conditions.

The automated speaker verification is generally applied without human directions. In that particular circumstance, it is possible that a fraud may fool the system by fake speech samples of any speaker. In the field of speaker recognition, fraud in automated speaker verification system by giving fake speech samples is called as spoofing. In case of speaker verification, spoofing can be processed with the help of four techniques: voice conversion, speech synthesis, replay attack, and impersonation. The impersonation is the method where a fraud try to generate the speech by voice mimicry [23,24]. The replay attack is the method of changing the decision of automated speaker verification with the help of pre-recorded speech samples through playback and recording devices [25,26]. The speech synthesis and voice conversion techniques requires deep speech processing and signal processing knowledge and also large amount of data to produce synthesized voice. On comparison, spoofing through replay and record do not need speech processing information. The replay attack could be easily obtained by using good quality playback and recording devices. An existing research reports on spoofing to automated speaker verification says that replay attack is highly effective and easily accessible.

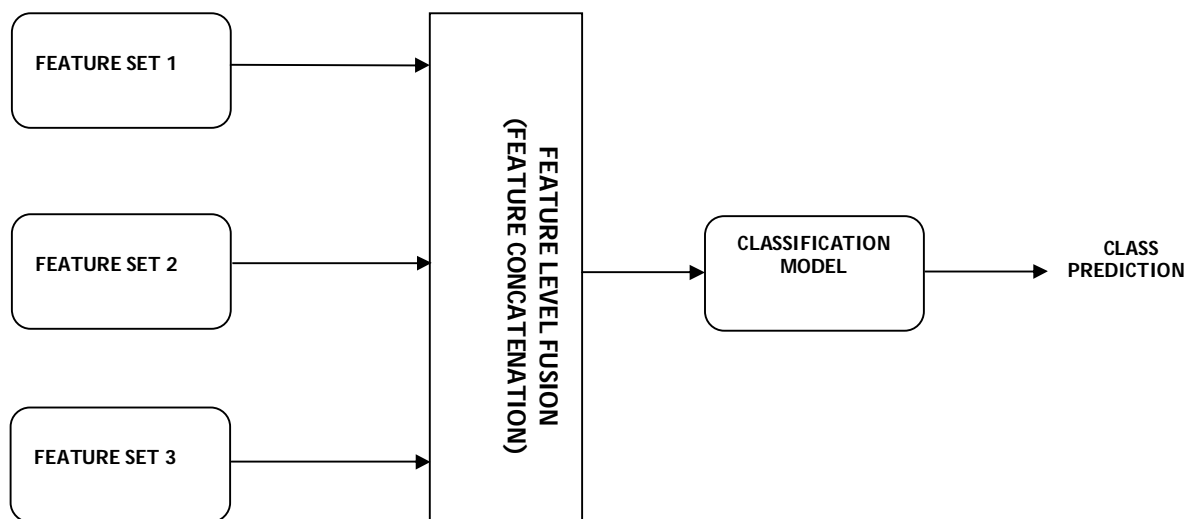


Figure 2: Block diagram of proposed speaker recognition system

3. Proposed Methodology: In speech processing systems, fusion of different features for the making of efficient and effective models is called as feature based fusion technique. The motive behind the feature based fusion technique is that every individual feature contains few important features that may be missed by different models. In feature based model technique, different features are merged and then used for creating models [27]. The general block diagram of speech oriented speech recognition system by applying feature based scheme is shown in figure 2. In training step, the input voice signal is proceed across the pre-processing step and then various features are calculated by using various signal processing techniques. These individual features are merged to build a combined feature, which is further used for creating reference models. The point to be noted that on concatenating the fusion of different features not required to be of same dimensions. At the time of simulation a same method is followed to create composite features for matching.

The first research was done by applying the feature based fusion scheme by Fururi et al. [28]. The author has proposed the concatenation of well-known features with the first and second order polynomials having form of DeltaDelta and Delta coefficients and then applied for speaker recognition systems. On comparison with cepstral features, the concatenation of various features reduces the error rate by 30% [29,30]. Further, fusion of different techniques reduces the speaker identification and verification process by 1.43% and 37.5% respectively [31,32,33]. The aforementioned research shows that fusion of different features helps in enhancing the robustness and performance of the different speaker recognition systems. So, due to this reason we have applied this technique in our proposed framework [34,35].

In this proposed technique, different feature sets are computed separately and corresponding models are created with the help of particular modeling method. The feature based model is defined by their corresponding modeling variables. The introduced combination technique produces composite model by padding variables of corresponding feature based models. The padding technique is used to reduce the dimensions of the features and also reduces the computational complexity of the model. Additionally, the difficulty arises due to mapping of different modeling parameters can be avoided by same modeling technique. During testing, feature vectors of different parameters are set in same manner and put before for evaluation. The proposed fusion scheme based method is different by the fact that they are based on combined opinions. The scores produced by introduced technique are exactly used for matching without using any weights. Further, the proposed technique is more appropriate for real time systems.

3. Experiments and results:

The three MFCC, RMFCC, and MPDSS features are broadly used as features to show excitation source data. We are able to give experimental recognition report in this section to pick the suitable source excitation feature, particularly in the situation of using it as additionally demonstrate the speaker recognition system. On the basis of performance and robustness, the specific feature that is used to give optimized performance is further chosen.

We carry out speaker recognition process by using GMM technique with TIMIT dataset, and the speaker verification process by using GMM-UBM with NIST-2003 SRE dataset. In

speaker recognition system, processing of signals takes place at 7500 samples per sec and unvoiced and voiced identification are done by thresholding based on energy. The features are calculated from 25 msec overlapping speech frames at the rate of 90 frames per sec with the help of most recent literatures. We have consider the suggestions of prasanna et al. to derive the residual signal LP to calculate the MPDSS from 25 LP residual power spectrum and 25 dimensional features are used as MPDSS features[30]. In the same way, the 13 mel-cepstral coefficients combine with 13 Delta and DeltaDelta is calculated from LP signals and speech to get RMFCC and MFCC features.

Table 1: GMM based speaker identification performance of MFCC, RMFCC, MPDSS features and their combined methods with TIMIT database.

Feature	Identification Accuracy (%) Proposed Method	Identification Accuracy (%) [29]
MPDSS	74.35	73.65
RMFCC	83.74	82.14
MFCC	96.19	95.39
MPDSS+ RMFCC	84.56	-
MPDSS+ MFCC	96.13	95.55
RMFCC+ MFCC	97.14	96.91

The speaker recognition performance of MFCC, MPDSS, and RMFCC features with TIMIT dataset are reported in table 1. The individual accuracy of these features is assessed using GMM modeling techniques. The MFCC feature produces the recognition rate of 96.14%, whereas an RMFCC and MPDSS feature gives the recognition rate of 83.74% and 74.35% respectively. It is observed that among these features, MFCC performs the best accuracy. The feature based fusion technique between all these features are given in the table 1, it is observed that fusion of RMFCC and MFCC produces the best recognition rate of 97.14%.

Table 2: Performance of MFCC, RMFCC, MPDSS features and their combined representation using GMM-UBM based speaker verification system using NIST-2003 SRE dataset.

Feature	EER (%) Proposed Method	EER (%) [29]
MPDSS	20.24	21.38
RMFCC	18.10	18.89
MFCC	6.94	7.54
MPDSS+ RMFCC	17.24	-
MPDSS+ MFCC	6.12	7.54
RMFCC+ MFCC	5.94	7.30

The speaker verification process (EER) result is calculated for all three MFCC, RMFCC, and MPDSS features are depicted in table 2. The same trend is observed for speaker verification system as observed in speaker recognition system. The error rate for the MFCC gives the beat error rate of 6.94% as compared to MPDSS and RMFCC having 20.24 and 18.10 respectively. The feature based fusion technique is used for all the three features and

found that combination of RMFCC and MFCC produces best error rate of 5.94% as compared to MPDSS+ RMFCC and MPDSS+ MFCC of 17.24% and 6.12% respectively.

4. Conclusion and future works:

The robustness and performance of speech signal based framework depends on the quality of features. In the today's era of research, working of single feature might not be enough to cover both robustness and performance simultaneously. In order to resolve this problem, researchers use multiple sources by applying various fusion techniques. These fusion techniques are categorized into three categories: Model level, Feature level and Score level combination scheme. We have used feature based fusion technique in our research. The SVM is used as a classification technique after combining different features. We have also proved that our speaker recognition and speaker verification framework works well with MFCC with TIMIT and NIST-2003 SRE dataset. Further, the fusion technique gives better results as compared to existing work. In future, more features will be added to enhance the recognition rate of speaker recognition and speaker verification system and try to incorporate some more deep learning methods.

References:

- [1] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.
- [2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [3] R. K. Das and S. Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.
- [4] D. Pati and S. M. Prasanna, "Speaker verification using excitation source information," *International journal of speech technology*, vol. 15, no. 2, pp. 241–257, 2012.
- [5] K. Tripathi and K. S. Rao, "Improvement of phone recognition accuracy using speech mode classification," *International Journal of Speech Technology*, vol. 21, no. 3, pp. 489–500, 2018.
- [6] B. Yegnanarayana, S. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on speech and audio processing*, vol. 13, no. 4, pp. 575–582, 2005.
- [7] D. Pati and S. M. Prasanna, "Speaker information from subband energies of linear prediction residual," in *2010 National Conference on Communications (NCC)*. IEEE, 2010, pp. 1–4.
- [8] P. Thévenaz and H. Hügli, "Usefulness of the lpc-residue in text-independent speaker verification," *Speech Communication*, vol. 17, no. 1-2, pp. 145–157, 1995.
- [9] D. Pati and S. R. M. Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," *International Journal of Speech Technology*, vol. 14, no. 1, pp. 49–64, 2011.
- [10] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2006.

- [11] T. C. Feustel, R. J. Logan, and G. A. Velius, "Human and machine performance on speaker identity verification," *The Journal of the Acoustical Society of America*, vol. 83, no. S1, pp. S55–S55, 1988.
- [12] T. C. Feustel, R. J. Logan, and G. A. Velius, "Human and machine performance on speaker identity verification," *The Journal of the Acoustical Society of America*, vol. 83, no. S1, pp. S55–S55, 1988.
- [13] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, no. 1, pp. 147–155, 2006.
- [14] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2012.
- [15] K. Manjunath and K. S. Rao, "Source and system features for phone recognition," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 257–270, 2015.
- [16] James, Praveen & Mun, Hou & Vaithilingam, Chockalingam & Chiat, Alan. (2020). Recurrent neural network-based speech recognition using MATLAB. *International Journal of Intelligent Enterprise*. 7. 56. 10.1504/IJIE.2020.104645.
- [17] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [18] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech, and Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.
- [19] A. Venturini, L. Zao, and R. Coelho, "On speech features fusion, _-integration Gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1951–1964, 2014.
- [20] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
- [21] [5] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.
- [22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [23] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry." in *Proc. Interspeech*, 2013, pp. 930–934.
- [24] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.
- [25] J. Lindberg and M. Blomberg, "Vulnerability in Speaker Verification-A Study of Technical Impostor Techniques," *Proceedings of European Conference on Speech Communication and Technology*, 1999.
- [26] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 10 workshop*, pp. 131–134, 2010.
- [27] D. Hosseinzadeh and S. Krishnan, "Combining vocal source and mfcc features for enhanced speaker recognition performance using gmms," in *Multimedia Signal Processing, 2007. MMSp 2007. IEEE 9th Workshop on. IEEE, 2007*, pp. 365–368.
- [28] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech, and Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[29] Krishna Dutta, "hybrid fusion scheme for different speech Processing tasks" DOE, NIT Nagaland, may 2020.

[30] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speakerspecific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[31] M Rahul, N Kohli, R Agarwal, S Mishra, "Facial expression recognition using geometric features and modified hidden Markov model", *International Journal of Grid and Utility Computing*10 (5), 488-496, 2019

[32] Jyoti, Amrita, Yadav, Vikash, Rahul, Mayur, "Blockchain Security Attacks, Difficulty, and Prevention", *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*,16. 10.2174/0123520965252489231002071659, 2023.

[33] Rahul, Mayur ,Tiwari, Namita , Prakash, Ayushi , Yadav, Vikash,"Garment Defect Detection System Based on Histogram Using Deep Learning". 10.1007/978-981-99-3716-5_22.,2023.

[34] Rahul, Mayur , Shukla, Rati , Tyagi, Devvrat, Tiwari, Namita, Yadav, Vikash," A New Hybrid Approach for Efficient Emotion Recognition using Deep Learning." *International Journal of Electrical and Electronics Research*. 10. 18-22. 10.37391/IJEER.100103,2022

[35] Rahul, Mayur, Pal, Parashuram, Yadav, Vikash , Dellwar, Devendra, Singh, Swarnima." Impact of Similarity Measures in K-means Clustering Method used in Movie Recommender Systems". *IOP Conference Series: Materials Science and Engineering*. 1022. 10.1088/1757-899X/1022/1/012101, 2021