FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 848-852 | e-ISSN: 2347-470X

Speaker Identification Analysis Based on Long-Term Acoustic Characteristics with Minimal Performance

Mahesh K. Singh[1*], S. Manusha[2], K.V. Balaramakrishna[3] and Sridevi Gamini[4]

[1,2,3,4]Department of ECE, Aditya Engineering College, Surampalem, India, [1]mahesh.singh@accendere.co.in, [2]sunkavallimanusha9977@gmail.com, [3]balaramakrishna_ece@acoe.edu.in, [4]sridevi_gamini@yahoo.com

***Correspondence:** Mahesh K. Singh; mahesh.singh@accendere.co.in

ABSTRACT- The identity of the speakers depends on the phonological properties acquired from the speech. The Mel-Frequency Cepstral Coefficients (MFCC) are better researched for derived the acoustic characteristic. This speaker model is based on a small representation and the characteristics of the acoustic features. These are derived from the speaker model and the cartographic representation by the MFCCs. The MFCC is used for independent monitoring of speaker text. There is a problem with the recognition of speakers by small representation, so proposed the Gaussian Mixture Model (GMM), mean super vector core for training. Unknown vector modules are cleared using rarity and experiments based on the TMIT database. The I-vector algorithm is proposed for the effective improvement of ASR (Automatic Speaker Recognition). The Atom Aligned Sparse Representation (AASR) is used to describe the speaker-based model. The Short Representation Classification (SRC) is used to describe the speaker recognition report. A robust short coding is based on the Maximum Likelihood Estimation (MIE) to clarify the problem in small representation. Strong speaker verification based on a small representation of GMM super vectors. Strong speaker verification based on a small representation of GMM super vectors.

Keywords: GMM super vector, Robust sparse coding, MFCC, Speaker recognition, Sparse representation.

ARTICLE INFORMATION

Author(s): Mahesh K. Singh, S. Manusha, K.V. Balaramakrishna and Sridevi Gamini
Received:19/07/2022; **Accepted**:03/10/2022; **Published**: 18/10/2022;
e-ISSN: 2347-470X;
Paper Id: IJEER-RDEC1;
Citation: 10.37391/IJEER.100415
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100415.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Speaker recognition is located on the speaker's speech recognition is extensively used in business interactions, debate, and law enforcement [1]. Speaker recognition involves identifying the speech of the person from graphic images. To improve the method of voice recognition, it is better to select a method of extraction of features that fully combines performance and precision [2]. The argument changes gradually over time. For powerful acoustic properties, much of the actual speaker recognition system uses short-term acoustic properties such as cepstral coefficients (MFCCs), predictive Gamma-tone Frequency Cepstral Coefficients (GFCCs), etc. The signal is processed into frames that overlap one another. The frame length is 20-40msec, with a 30-70 percent overlap [3]. To capture the knowledge of how these acoustic vectors, change over time, the conventional approach is to calculate derivatives of cepstral coefficients in the first and second order [1, 4, 5]. For experimental purpose original recorded speech sample is shown in *figure 1* and for their frequency domain analysis, their Fourier transform is shown in *figure 2*.

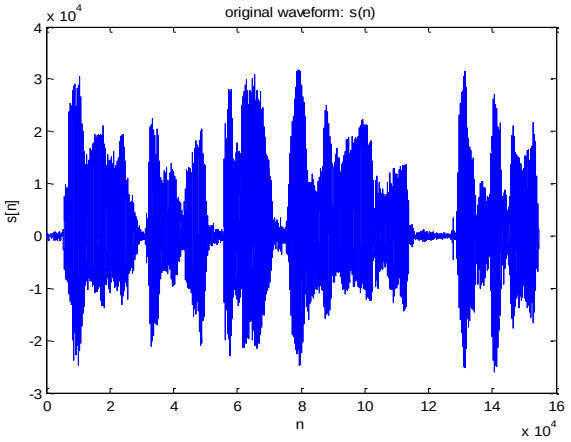

Figure 1: Recorded speech sample for speaker recognition

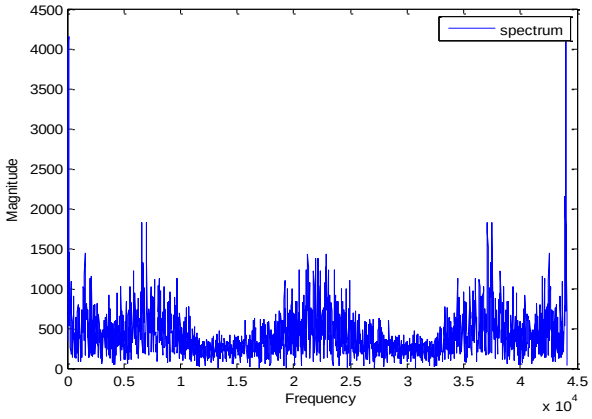

Figure 2: Fourier transform of the windowing signal

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 848-852 | e-ISSN: 2347-470X

The sparse representation of acoustic features such as I-vector features, GMM-UBM features tensor features, MFCCs, and mean super vectors for speaker recognition with sparse representation models has been implemented recently [4, 18, 19]. In the field of sparse representation, the focus is growing due to the application of speech de-noising, source separation, speech encryption, and speech classification based on the sparse model analysis shown in *figure 3*.

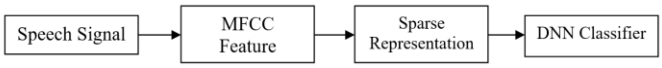

Figure 3: Speaker identification model

This paper proposed a new speech recognition system is given for the training of a DNN classifier [6]. The MFCC frame features are extracted from the speech frames and the super MFCC frames are built to capture the static and dynamic information in the speech signal [7]. The merger super MFCC frames into MFCC maps to learn how to evaluate the speaker model [8, 20, 21]. It is contained the speaker model and uses them as the LTA features to train the DNN classifier. The outline of the technique is shown in the *figure* also. In the training phase, all voice signals are split into a series of overlapping time frames [9]. The extracted MFCCs are then converted to a super MFCCs map, the MFCCs map can be used as the sparse model analysis for training data [10].

DNN is a popular former method that achieves better recognition performance in speaker recognition [13, 14, 22]. A large number of hidden layers which must be linear or nonlinear are included in this DNN These hidden layers represent the data in the encoded form [11]. The main concept of DNN is activating the current output layer to the input of the next hidden layer [12, 23]. Identification capability is enhanced by using enormous hidden layers [13]. In this section, the DNN training using an adaptive HHO is elaborated. The major goal of DNN is speaker recognition, for such, it uses the features that are extracted from the speech signal [14].

2. RELATED WORK

Speaker recognition is based on the acoustic characteristics of speaker recognition systems. In 2019 short-term acoustic characteristics were derived from single speech pictures. The most famous are MFCCs. This speaker model is based on the sparse depiction and LAT features are derived from the speaker model's MFCC map [1]. The novel sparse representation classification algorithm identifies the speaker problem. Hence, speaker recognition can now be regarded as a typical question of pattern classification [2]. For independent speaker identification, the MFCC is constructive feature extraction. The novel sparse representation with the subspace algorithm, define the speaker recognition problem [3]. DNN-based end-to-end speaker verification technique. This technique shows the efficient outcome for the short utterances of both text-dependent and independent tasks. Recently machine learning-based recognition techniques attains huge demand. In this, an end-to-end speaker verification system is introduced [15]. The training process for this system is performed end-to-end but in regularized form, so that, it won't deviate much long from the primary system. Due to this, the over-fitting can be minimized as it retards the effectiveness of these end-to-end methods. It provides better outcomes for both short and long-duration than the i-vector + PLDA baseline. It consists of two major components-extraction functions and speaker recognition challenge (SRC). In the extraction of a feature, a PPCA-super vector is constructed using PPCA and the number of own values is then obtained using the Bartlett test [4]. The robust verification of speakers is based on the minimal representation of GMM super vectors. when compared to I-vector PLDA the NIST cost is increased to 14%. Super vectors are constructed by connecting the mean vectors of GMM to describe the speaker's MODEL [5]. Speaker verification is the work to know the speaker from the speech of a person. Recently, speaker verification is done very effectively but there is an issue with background speakers that need more examples for each speaker, so to avoid this issue GMM sparse coding was launched [6]. It is explained as stressed speech recognition using sparse representation. The dissimilarities between natural speech and stressed speech are identified by linear prediction coefficients (LPCs) [7].

The class label of the speaker is not maintained properly in traditional dictionary learning. In recent works applying Learned Language Recognition Dictionary (LR). So, it is analyzing SRC in this work on a discriminative learned LR dictionary [8].

The arrival direction of the speaker is evaluated using an acoustic vector sensor (AVS) meaning inter-sensor data ratio (ISDR) [16]. When the announcement of deep visual networks into traditional voice recognition pipelines is widely demonstrated to help system performance [17]. Text-Independent verification of speakers is based on a traditional visual network construction. In this, for speaker verification, a text-independent speaker embedding system was suggested [10]. The novel approach presents the problem of visual speech recognition. It is based on the lip action of the person. Using special temporal features visual lip information is extracted and the process is done by removing the noise and adding to the comparison of images in every video and it is used to build kernel sparse representation classifier (KSRC) [11].

3. THE IDENTIFICATION OF SPEAKERS WITH ANALYSIS OF SPARSE REPRESENTATION

Figure 1 it's described the overall recognition of a conversationalist. In the training phase, the signals are split into time frames with window combinations. The MFCCs are collected from the frame and converted to super MFCCs to create the MFCCs diagram. It is used to train the data for the development of a sparse model. It can be used as a classifier coefficient of a DNN classifier for speaker recognition as an LTA feature within the speaker.

3.1 MFCC

This feature is normally applied for both speaker identification and speech recognition. The main objective of this MFCC

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 848-852 | e-ISSN: 2347-470X

method is to obtain significant information from speech waveform by eliminating redundant information. After each frame has been turned into a single "N" dimensional feature vector, this is done frame by frame. The total number of samples in a frame is less than N. The back-end system receives data in a way that minimizes the amount of data processing required. In feature extraction, the audio input is converted into a vector sequence. The procedures listed below are used to find the MFCCs. The first step is pre-emphasizing the voice signal. The hamming window is applied to each of the speech signal's many frames, each of which has a period of 20 milliseconds and an offset of 10 milliseconds. MFCCs map is used to represent speech signals. These MFCCs are obtained through cepstral analysis and corrupt by the frequency scale of Mel. The MFCC feature extraction technique is presented in *figure 4*.

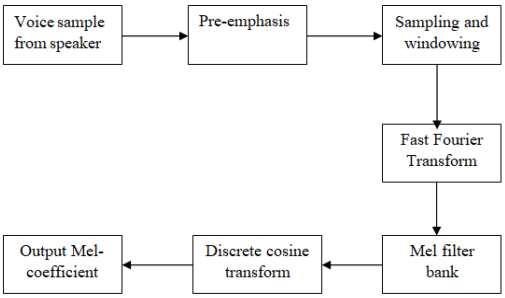

Figure 4: MFCC feature extraction technique

that went into creating a tiny folded rectangular strip antenna. The antenna is in miniature form. In all cases ground plan is fixed, only variation is done in the upper part.

Sparse representation model:
According to the sparse representation speaker model, LTA features of the speaker are obtained from the MFCCs map for simplicity it is denoted as

$$X \in R^{M*1}, M = 12l \quad (1)$$

DNN classifier:
In speaker recognition, DNN is used as a classifier. It Uses four layers each of which has 'R' sigmoid units. The units in the input layer shall be equal to the input vector dimensions 'M'. The output lot DNN is the *softmax* layer of 'l'.

3.2 Features Extraction and GMM Super Vector

In the feature extraction method, MFCC works well for signal representation. Voice activity detection is used to decrease silence noise. The delta features can obtain dynamic information from the frame.

GMM super vector:
GMM is an effective method for designing text-independent speaker recognition. It is represented as:

$$P(x) = \sum_{i=1}^{M} P_i f\left(\frac{x}{m_{i,}}\right) \quad (2)$$

Campbell invented the GMM super vectors for speaker verification. The above *figure* shows the process of GMM super vectors. To get the GMM super vector GMM-UBM is taken as a database. This method has speaker verification done by sparse coding shown in *figure 5*.

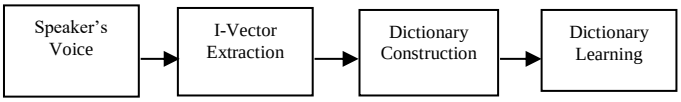

Figure 5: Dictionary creation for speakers' recognition

Dictionary Construction:
In sparse representation using I-vectors. Non-target speakers are used to constructing dictionaries. They are two methods of constructing a dictionary and K-SVD is exemplary.

Exemplar Dictionary:
$D = \left[N_{qtar,} D_{bg}\right] = [q_{tar}, 1, \ldots\ldots, q_{tar,} N_{tar,} q_{bg,} 1, \ldots\ldots, q_{bg,} N_{bg}]$ (3)

K-SVD Directory:
The K-SVD is obtained by a clustering algorithm for speaker representation. It is represented as

$$\min D, X \{\|Y - DX\|_F^2\}, s.t. \forall_{i,} \|x_i\| 0 \leq T_0 \quad (4)$$

4. RESULTS

The proposed LTA is used for speaker recognition. The LTA's robustness is determined when white noise is present. There are four databases for speaker recognition. Which is the TIMIT database, Vox forge database, THCHS30 database, and Libri speech database. The speaking content sampling rate was 16 kHz, and the sample size is 16 bits. In prior research, the efficiency in the future LTA is exposed by comparing the presentation of the same recognition with a changed description. The acknowledgment of the concert of the projected speaker identification systems is established by comparing the projected classification with the speaker acknowledgment method. In this immediate experimentation, the forcefulness of the description is established in the occurrence of noise. Here employed four changed databases to examine the concept of a speaker identification scheme. Libri Speech database, THCHS30 database, Vox Forge database, and the TIMIT database are used for experimental purposes. 10 combined speakers had been selected for the TIMIT database. In those three female speakers and seven male speakers, by means of ten English dialogue sentences of every speaker. Here derived twelve speakers, from them 4 female and 8 male speakers, commencing the online Vox Forge website. It is considering eight English speech samples per speaker. All details about the speakers are given in *table 1*.

Table 1: Different databases for speaker recognition

Database	Total speakers	Male speakers	Female speakers
TIMIT	10	7	3
Vox forge	12	8	4
THCHS30	10	3	7
Libri speech	10	5	5

These experimentations were intended to train and test the identification performance in the projected speaker recognition

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 848-852 | e-ISSN: 2347-470X

system. The phase of training consists, of every speaker's voice first pre-processed, pre-emphasized, and segmented. It is considering the 799 frames with a 20 ms duration of each frame by using Hamming windowing size of 12 ms. For frequency, domain conversion takes the Fast Fourier transform of each frame. It was transformed commencing a power range to the Mel-scale by twenty-four triangular Mel-filters. These are computed by applying 'log' solidity first. Based on traditional and DNN-based MFCC the performance rate is addressed in *table 2*. From the table, it is shown that the DNN-based speaker recognition gives a better recognition rate.

Table 2: Performance rate of traditional and DNN-based MFCC

No. of Speakers	Performance Rate (%)	
	Traditional MFCC	DNN based MFCC
10	91	93
12	86	90
15	76	79

The proposed long-term acoustic characteristics for speaker recognition obtain average classification accuracy (ACA) as follows.

$$ACA=\frac{number\ of\ correct\ samples}{number\ of\ testing\ samples} * 100\% \quad (5)$$

The accuracy rate in terms of the centroid is given in *table 3* is shown that the if the number of speakers is increasing the accuracy rate decreases. But advantages are that if the number of the centroid is increasing then the accuracy rate is increasing. Experimental results in terms of accuracy are taken from the original speech sample shown in *figure 6*.

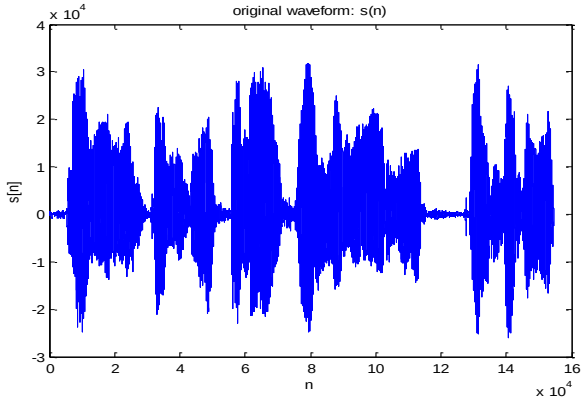

Figure 6: Original speech sample for speaker recognition

Table 3: Speaker recognition accuracy rate

No. of Centroid (k)	Accuracy Rate		
	No. of speakers 10	No. of speakers 12	No. of speakers 15
2	84%	82%	80%
4	90%	87%	85%
8	93%	91%	77%
16	94%	89%	74%

Here an algorithm is projected for the relevance of accuracy rate using different centroid techniques. DNN-dependent MFCC feature extraction techniques for different accents equally distributed for testing and training databases are extracted effectively. Two types of accent identification systems used which are based on traditional MFCC and DNN classifier based on MFCC feature extraction is also analyzed in this article.

5. CONCLUSION

The speaker model is built based on the LTA features obtained from the MFCC 's speech map with the speaker model. The LTA functions contain both dynamic and static information. The DNN classifier is used as a baseline method. Today identification of speakers is seen as a typical question of classification of trends. N this paper well proposed an algorithm to classify speakers on sparse representation. The proposed algorithm is checked in the TIMIT database and the study is carried out on a state-of-the-art speaker we planned to develop a novel speaker-dependent emotion benefit technique ASSR to incorporate the emotional I-vectors algorithm. Robust sparse coding is used to identify the robust speaker. Auditory features are much dependent on the speaker's pitch and formants. Acoustic feature extraction using MFCC is an efficient move toward accent recognition for the reason that it is combined with the sensing feature of speech.

REFERENCES

[1] Lin, T., & Zhang, Y. (2019). Speaker recognition is based on long-term acoustic features with an analysis of sparse representation. IEEE Access, 7, 87439-87447.

[2] Naseem, I., Togneri, R., & Bennamoun, M. (2010, August). Sparse representation for speaker identification. In 2010 20th International Conference on Pattern Recognition (pp. 4460-4463). IEEE.

[3] Xu, L., & Yang, Z. (2013, August). Speaker identification based on sparse subspace model. In 2013 19th Asia-Pacific Conference on Communications (APCC) (pp. 37-41). IEEE.

[4] Chin, Y. H., Wang, J. C., Huang, C. L., Wang, K. Y., & Wu, C. H. (2017). Speaker identification using discriminative features and sparse representation. IEEE Transactions on Information Forensics and Security, 12(8), 1979-1987.

[5] Singh, M., Nandan, D., & Kumar, S. (2019). Statistical Analysis of Lower and Raised Pitch Voice Signal and Its Efficiency Calculation. Traitement du Signal, 36(5), 455-461.

[6] Priya, B., & Dandapat, S. (2016, November). Sparse representation of LPC for analysis of stressed speech in lower-dimensional subspace. In 2016 IEEE Region 10 Conference (TENCON) (pp. 661-666). IEEE.

[7] Singh, M. K., Singh, A. K., & Singh, N. (2019). Multimedia analysis for disguised voice and classification efficiency. Multimedia Tools and Applications, 78(20), 29395-29411.

[8] Singh, O. P., & Sinha, R. (2017, November). Sparse representation classification over discriminatively learned dictionary for language recognition. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 2632-2636). IEEE.

[9] Singh, M. K., Singh, A. K., & Singh, N. (2018). Acoustic comparison of electronics disguised voice using different semitones. Int J Eng Technol (UAE), 7(2), 98.

[10] Zou, Y., Guo, Y., Zheng, W., Ritz, C. H., & Xi, J. (2014, July). An effective DOA estimation by exploring the spatial sparse representation of the inter-sensor data ratio model. In 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP) (pp. 42-46). IEEE.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 848-852 | e-ISSN: 2347-470X

[11] Singh, M. K., Singh, A. K., & Singh, N. (2018). Disguised voice with fast and slow speech and its acoustic analysis. Int J Pure Appl Math, 118(14), 241-246.

[12] Zhang, C., Koishida, K., & Hansen, J. H. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9), 1633-1644.

[13] Singh, M. K., Singh, A. K., & Singh, N. (2019). Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement. Multimedia Tools and Applications, 1-16.

[14] Frisky, A. Z. K., Wang, C. Y., Santoso, A., & Wang, J. C. (2015, September). Lip-based visual speech recognition system. In 2015 International Carnahan Conference on Security Technology (ICCST) (pp. 315-319). IEEE.

[15] Siddiqa, S. K., Apurva, K., Nandan, D., & Kumar, S. (2021). Documentation on smart home monitoring using the internet of things. In ICCCE 2020 (pp. 1115-1124). Springer, Singapore.

[16] Singh, M. K., Singh, N., & Singh, A. K. (2019, March). Speaker's Voice Characteristics and Similarity Measurement using Euclidean Distances. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 317-322). IEEE.

[17] Punyavathi, G., Neeladri, M., & Singh, M. K. (2021). Vehicle tracking and detection techniques using IoT. Materials Today: Proceedings.

[18] Veerendra, G., Swaroop, R., Dattu, D. S., Jyothi, C. A., & Singh, M. K. (2021). Detecting plant Diseases, quantifying and classifying digital image processing techniques. Materials Today: Proceedings.

[19] Priya, B. J., Kunda, P., & Kumar, S. (2021). Design and Implementation of Smart Real-Time Billing, GSM, and GPS-Based Theft Monitoring and Accident Notification Systems. In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (pp. 647-661). Springer, Singapore.

[20] Kiran, K. S., Preethi, V., & Kumar, S. (2022). A brief review of organic solar cells and materials involved in its fabrication. Materials Today: Proceedings.

[21] Haris, B. C., & Sinha, R. (2015). Robust speaker verification with joint sparse coding over learned dictionaries. IEEE Transactions on Information Forensics and Security, 10(10), 2143-2157.

[22] Sreeram, G., Haris, B. C., & Sinha, R. (2015, November). Improved speaker verification using block sparse coding over joint speaker-channel learned dictionary. In TENCON 2015-2015 IEEE Region 10 Conference (pp. 1-5). IEEE.

[23] Sudeep, S. V. N. V. S., Venkata Kiran, S., Nandan, D., & Kumar, S. (2021). An Overview of Biometrics and Face Spoofing Detection. ICCCE 2020, 871-881.