ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/321722803

Gaussian Mixture Model: A Better Modeling Technique for Speaker Recognition

Conference Paper · December 2017

CITATIONS
2

READS
262

5 authors, including:

Nilu Singh
K L University
114 PUBLICATIONS 281 CITATIONS

SEE PROFILE

Alka Agrawal
Babasaheb Bhimrao Ambedkar University
141 PUBLICATIONS 2,168 CITATIONS

SEE PROFILE

Prof. Raees Ahmad Khan
Babasaheb Bhimrao Ambedkar University
294 PUBLICATIONS 3,380 CITATIONS

SEE PROFILE

All content following this page was uploaded by Nilu Singh on 11 December 2017.

The user has requested enhancement of the downloaded file.

Gaussian Mixture Model: A Better Modeling Technique for Speaker Recognition

Nilu Singh[1], Alka Agrawal[2] and R. A. Khan[3]

[1,2,3] SIST-DIT, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, UP, India

Email: nilu.chouhan@hotmail.com, alka_csjmu@yahoo.co.in, khanraees@yahoo.com

Abstract—***Speech is a prime medium to communicate with each other. Development in technology has made human machine interaction possible. Human voice is used for automatic authentication of individual and this process is called speaker recognition. Since its inception, there have been a lot of developments in the area. Speaker recognition gradually becomes better with the new advancements including better modeling techniques. Gaussian Mixture Model (GMM) is a modeling technique which gives better results when combined with speaker recognition system. The paper discusses about various available modeling techniques and presents their comparative study. In addition, it also throws light on why GMM is better than the other modeling techniques.***

Index Terms— *Speaker Recognition, GMM, GMM Component.*

INTRODUCTION

Automatic Speaker Recognition (ASR) is a process of automatically recognizing a person by their voice. A speech signal is enrich with speaker-specific information [1, 2, 3]. Though a majority of speaker recognition system approaches are based on cepstral coefficient such as Mel Frequency Cepstral Coefficient (MFCC) [4, 5, 6] but in the last decade it is observed that researcher are keen to use prosodic features to develop speaker recognition system. The reason behind is that prosodic features produce better results as compared to MFCC's and more robust against noise [7, 8, 9, 10]. Commonly used prosodic features are pitch and energy contours [3] [11, 12]. Prosodic features are based on speakers speaking style and speaker's intonation. Prosodic features based systems require large amount of voice data to train speaker models. It is an established fact that prosodic features are related to phonemes and syllables (such as pitch and duration), hence these are less sensitive to channel distortion than cepstral features [8] [13]. According to the results obtained in [11, 14], the author claims that to obtain better results, prosodic features (duration, pitch and energy) are more suitable.

Gaussian Mixture Model modeling for text-independent speaker recognition process was introduced by Reynolds in 1992 and now it is a dominant approach for modeling methods [15, 16]. To extract the evidence of an individual's identity from his/her speech signal is known as speaker recognition. Speaker recognition tries to find out who is source of a particular utterance. There are many feature extraction techniques, algorithms and modeling techniques available such as HMM, NN, SVM and VQ are used for speaker recognition. These modeling techniques well perform under clean speech signal. The performance of speaker recognition degrades when speech signal is received in noisy condition. In addition, these modeling techniques also fail with corrupted signal, channel mismatch or with very small input data. To overcome from these kinds of problems GMM is used. GMM performs well and provide high classification accuracy. It is also very robust in noisy environment and with corrupted signal [2, 45].

The rest of the paper is organized as follows: the next section describes about selected development in speaker recognition area. Section III is about the available different modeling techniques. In section IV and V is presented the concept of GMM. Section VI describes about speaker recognition and GMM. Finally paper is concluded in section VII.

I. GROWTH OF SPEAKER RECOGNITION SYSTEM

With the rapid development in technology in 1960's, it became possible to develop autonomous speaker recognition system. The development made during this period covered a wide-range of disciplines in the field of speaker recognition system. In this row, Gunnar Fant in 1960 developed a physiological model of human voice production system. The model sets a basis for understanding speech analysis for speaker and speech recognition both. Fant's idea of physiological model of voice directed future researchers to characterize speech signal as a linear source filter model. Through Fant model, it became possible to make various advances in discovering human voice characteristics which is individually recognizable [3, 18- 19]. Figure 1 shows the timeline of crucial development in the field of automatic speaker recognition system.

In 1963, Bogert, Healy and Tukey have published a research article titled "The Quefrency Alanysis of the Time Series for Echos: Cepstrum, Pseudo-Auto-Covariance, Cross-Cepstrum, and Saphe Cracking" [20]. The article (oddly titled) has made a study about echo detection. In 1965, Cooley and Tukey published their research on digital implementation of the Fourier Transform and later it is known as Cooley-Tukey Fast Fourier Transform (FFT) [21]. Cooley and Tukey method has been considered as an efficient method for frequency analysis of digital signal. Inspired by [20], Michael Noll in 1969, has presented an idea for pitch detection of a human voice by using cepstrum [22]. Ronald Schafer who joined Oppenheim research has contributed to build on Noll' speech detection to be used for cepstral analysis to model speech signal. Later, the cepstral speech model has been used as an important tool for speaker recognition [23, 24].

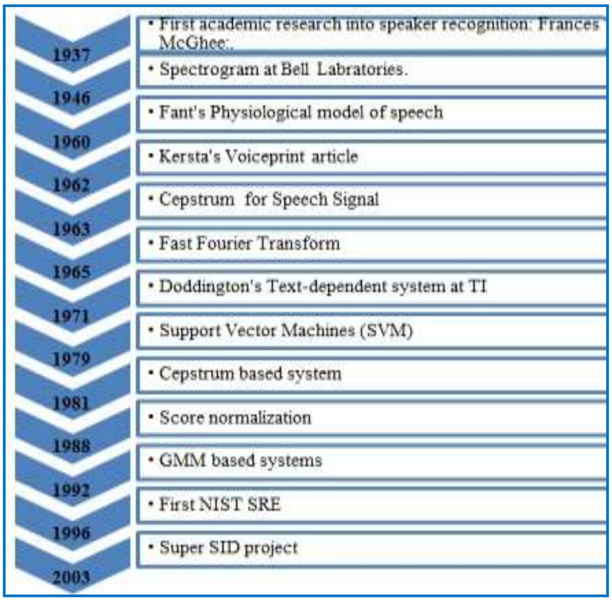

Fig: 1.Timeline of Major Speaker Recognition development system [25]

II. AVAILABLE APPROACHES FOR MODELING AND CLASSIFICATION

Speaker modeling is one of the important components of speaker recognition. For every registered speaker, speaker models are created during training and testing phases after the computation of speech feature vectors. Training phase involves creating voice database for registered speakers whereas in testing phase, claimed voice input is matched with the training database. During matching (classification), received speech (either known or unknown) is compared with the speaker model to evaluate a score value (match score). By using this score value, it is decided whether the speaker is accepted or rejected [2, 3].

Speaker models can be categorized as stochastic and template models also known as generative models and discriminative models respectively. In stochastic modeling, speaker models are created by using probability density function. During training phase, probability density function parameters are estimated from the given speech. And for matching a likelihood of the utterance is evaluated. Whereas in template modeling, training and testing models are directly compared with each other and the quantity of falsification between these two is the degree of similarity [2, 26-28]. Table1 shows the comparative study of different modeling techniques on the basis of different parameters. A short description of selected modeling techniques is given in the following subsections.

Table1: comparative study of different modeling techniques on the basis of different parameters

Speaker Modeling Methods	Comparison Based on Characteristics				
	Text-independent/ Text-dependent	Robustness	Purpose/Used For	Model Type	Approach used
GMM	Text-independent and text-dependent	• Not affected with time variability. • Robust against Noise [1, 2, 16].	• Useful for classification • Reduced computing (posterior probability) complexity. • Speaker Recognition. • Gives high recognition accuracy [1, 2, 16].	Stochastic models [2]	Generative classifiers [2]
HMM	Text-independent and text-dependent both	Robustness against utterance variations [1, 2, 16].	• Acoustic feature space. • Multiple-state ergodic HMM [1, 2, 16].	Stochastic models [2]	Generative classifiers [2]
SVM	Text-independent and text-dependent	Most robust classifiers in Speaker Verification. [1, 2, 16].	• Not useful for classification • Try to minimize the classification error on a set of training data. • Speaker recognition and pattern classification [1, 2, 16].	Classifier Method [1, 2, 16].	Discriminative approach [2]
VQ	Text-independent and text-dependent [1]	Affected by Time variability [1, 2, 16].	• Reduces storage requirements. • Speaker Verification [1, 2, 16].	Template models [2]	Clustering methods [2].
ANN	Text-independent and text-dependent.	Less robust classifiers [1, 2, 16]	• Used for classification methods. • Speaker recognition [1, 2, 16].	Classical pattern [1, 2, 16].	Discriminative approach [2].
DTW	Text-independent and text-dependent	Use for non-uniformity problem [1, 2, 16].	Resolve the matching problem [1, 2, 16].	Template Model [1].	Dynamic programming [1, 2, 16].

A. ***Support Vector Machine (SVM):*** It is a binary discriminative classifier. SVM uses boundary between two classes for creation of speaker models. One class contains training data vectors for target speaker's which are labeled +1 while the other class contains imposters training data vectors from a large data set and labeled as -1 [2, 29].

B. ***Hidden Markov Model (HMM):*** It is the most popular modeling methodology for text-independent and text-dependent speaker recognition. HMM, a doubly stochastic process, was developed in 1980s. The term hidden is use because it has an underlying stochastic process that is not observable. To observe the hidden process another stochastic process are used [30- 32].

C. ***Gaussian Mixture Model (GMM):*** It is comparatively more suitable method for speaker modeling. For creation of speaker model it uses speech features as a linear combination of finite mixture of multivariate Gaussian components. It uses Expectation-Maximization (EM) algorithm and maximum likelihood (ML) estimation for estimation of GMM parameters [33, 34].

D. ***Vector quantization (VQ):*** It is a classification method for speaker verification. It uses clustering methods e.g. K-means use to reduce training vector. Each cluster is represented by code vector and this code vector is the centroid of that cluster. Collection of centroid vectors is called codebook. During verification process the training data of a registered speaker is used to create a codebook which is the model for specific speaker. If the provided speech is of an unknown speaker then the matching is determined by evaluating distance between the testing data feature vector and the target speaker nearest vector codebook. The evaluated distance is called score value of a verified speaker [1, 2].

E. ***Artificial Neural Networks (ANNs)*****:** It is a discriminative methodology used for speaker classification. There are several types of neural networks, for speaker recognition generally Multi-Layers Perception (MLP) is used. MLP is a feed-forward network. In this network, multiple layers of nodes are achieved and collectively these are used for complex machine learning task. For each node, weighted sum are calculated for the inputs. Here, weights are adjustable parameters. After that transfer function is applied for calculating output of that node. Back propagation algorithm is use for determining the weight parameters [35].

III. WHY GMM IS BETTER?

The Gaussian mixture speaker model was introduced in 1990 by Rose and Reynolds [36]. Then Gaussian Mixture Modeling (GMM) technique for text-independent speaker recognition introduced by Reynolds in 1992 came into picture [37]. The specialty of GMM is that any distribution can be modeled by using Gaussian mixture modeling technique. The reason behind is that it is able to provide large number of mixture components of voice. This modeling method is useful in text-independent speaker identification as well as speaker verification [38]. The GMM provides high recognition accuracy and effective speaker representations which is also computationally inexpensive [36].

GMM has widely used for speaker modeling in text-independent speaker recognition system. It has the following characteristics which make it more useful for modeling [39, 40], such as:

- GMM has the ability to form smooth approximations to arbitrarily shaped density. It is based on a linear combination of Gaussian basis functions which are capable of representing a large class of arbitrary densities.
- In GMM, for each Gaussian component, an implicit realization of probabilistic modeling of speaker dependent acoustic classes to a broad acoustic class such as vowels, nasals and fricatives etc. is used.
- GMM is not susceptible to natural changes such as aging or cold.

GAUSSIAN COMPONENTS

The task of speaker recognition is done by using individual mixture components i.e. Gaussian components [42, 43]. For speaker recognition, features are obtained from the speech signal. The fundamental information of speaker discrimination can be characterized by Gaussians. For the expected GMM parameters i.e. covariance and Gaussian component, weight is associated to the location of formant, magnitude of speech signal and bandwidth of speech signal [42]. As discussed in [43], for good quality system performance at least 8 to 16 Gaussian components are mandatory where voice/speech is considered as noiseless. The GMM is created by using these components through diagonal covariance matrix. To build multi conditional robust systems, the minimum number of essential Gaussian components involving 64 and 128 are required. The main advantage of GMM involves its likelihood function being computationally inexpensive. GMM is collected from a finite mixture of Gaussian components. Since Gaussian components have potential to characterize discriminative information of speaker, it is widely used for speaker recognition [42] [44].

IV. GAUSSIAN MIXTURE MODEL AND SPEAKER RECOGNITION

GMM takes sequence of vectors provided by feature extraction technique and use it to create speaker model. These models are called Gaussian Mixture Models. The Gaussian mixture model is ‘mixture density’, categorized as a sum of M Gaussian component densities. Component density is a product of ‘mixture weight’ with a ‘Gaussian component’. Individual ‘Gaussian component’ represent acoustic classes and these classes reflect a speaker specific vocal tract information therefore is useful for modeling speaker identity [39, 40].

The GMM is used to represent speaker’s model in the speaker recognition systems. The distribution of feature vectors extracted from a speaker’s speech signals is modeled by Gaussian mixture density function [36]. Equation (1) for a D-dimensional feature vector denoted as x, the mixture density function P for speaker s is:

$$P\ (x|\lambda s) = \sum_{i=1}^{M} p_i^s b_i^s\ (x) \qquad \ldots\ldots(1)$$

In equation (2) the density is a weighted linear combination of M component Gaussian densities, $b_i^s(x)$ each parameterized by a mean vector, μ_i^s and covariance matrix, $\sum_i^s$

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i^s|^{1/2}} \times \exp\left\{-\frac{1}{2}(x-\mu_i^s)'(\Sigma_i^s)^{-1}(x-\mu_i^s)\right\} \quad \text{……………(2)}$$

The mixture weights are p_i^s and is represented as

$$\sum_{i=1}^{M} p_i^s = 1$$

And $\lambda_s = \{p_i^s, \mu_i^s, \Sigma_i^s\},\ i = 1, \ldots, M.$

For an input data X and a number of mixtures M (assume a priori), data can be fit using M Gaussian distributions. The figure 2 shows the component of a speech signal and figure 3 represents the process of computing the probability of a feature vector given a GMM model.

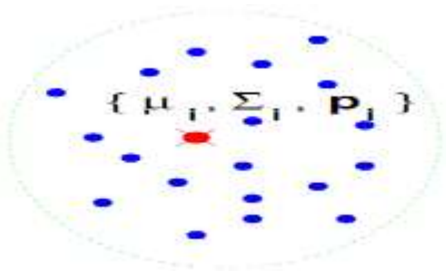

Fig. 2: One component of a GMM speaker model [40]

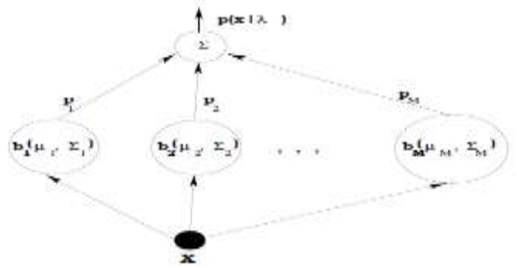

Fig. 3: the process of computing the probability of a feature vector given a GMM model [41]

V. CONCLUSION

In the past few years, speaker recognition has seen various advancements due to emerging technologies. Though robustness of ASR systems is improved a lot, there remains still some issues remain including noisy data, channel mismatch etc. To resolve the issues, GMM is used for a better recognition system even with noisy condition. The reason behind is that it creates more components for a given speech hence increases the possibility for better match. Hence, it provides compact representation for speaker recognition system. It is more effective for text-independent speaker recognition. It can be concluded that when GMM is combined with prosodic may produce better results.

REFERENCES

[1]. S Furui, “50 years of progress in speech and speaker recognition” ECTI Transactions on Computer and Information Technology (ECTI-CIT) 1 (2), 2005, pp. 64-74.

[2]. Bezawit Wubishet, “ Noise Robust Speaker Verification using SVM based GMM Supervector”, A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Electrical Engineering, Addis Ababa, Ethiopia, December-2012, pp.1-86

[3]. Nilu Singh, Alka Agrawal and R. A. Khan, “Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades”, Global Journal of Enterprise Information System. Volume-9, Issue-3, July-September, pp. 38-45.

[4]. D.A Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, no. 1-3, pp. 19–41, 2000.

[5]. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Factor analysis simplified,” in Proc. ICASSP 2005, Philadelphia, PA, Mar. 2005, pp. 637–640.

[6]. V. Wan and W.M. Campbell, “Support Vector Machines for Speaker Verification and Identification,” in IEEE International Workshop on Neural Networks for Signal Processing, Sydney, Australia, 2000, vol. 2, pp. 775–784.

[7]. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, “A Lognormal Tied Mixture Model of Pitch For Prosody-Based Speaker Recognition,” in Proc. EUROSPEECH, Rhodes, Greece, September 1997, pp. 1391–1394.

[8]. A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey, “Modeling Prosodic Dynamics For Speaker Recognition,” in Proc. ICASSP 2003, Hong Kong, 2003, pp. 788–791.

[9]. S. Kajarekar, L. Ferrer, A. Venkataraman, K. S¨onmez, E. Shriberg A. Stolcke, H. Bratt, and R.R. Gadde, “Speaker Recognition Using Prosodic and Lexical Features,” in Proc. IEEE ASRU, St-Thomas, December 2003, pp. 19–24.

[10]. S. Kajarekar, L. Ferrer, K. S¨onmez, J. Zheng, E. Shriberg, and A. Stolcke, “Modeling NERFs For Speaker Recognition,” in Proc. Odyssey 2004, Toledo, Spain, June 2004, pp. 51–56.

[11]. Najim Dehak, Pierre Dumouchel, and Patrick Kenny, “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech And Language Processing, April 20, 2007, pp.1-9.

[12]. Nilu Singh, Alka Agrawal, Raees Ahmad Khan, “A Critical Review on Automatic Speaker Recognition”, Science Journal of Circuits, Systems and Signal Processing. Vol. 4, No. 2, July 2015, pp. 14-17. doi: 10.11648/j.cssp.20150402.12

[13]. L. Ferrer, H. Bratt, S. Kajarekar, E. Shriberg, K. S¨onmez, K. Stocke, and A. Venkataraman, “Modeling Duration Patterns For Speaker Recognition,” in Eurospeech, Geneva, 2003, pp. 2017–2020.

[14]. R. Summerfield, T. Dunstone, C. Summerfield, “Speaker Verification in a Multi-Vendor Environment”, www.w3.org/2008/08/siv/Papers/Centrelink/w3c-sv_multivendor.pdf

[15]. D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Dig. Signal Process., vol. 10, no. 1-3, pp. 19 -41 2000.

[16]. Nilu Singh, Alka Agrawal, Raees Ahmad Khan, “Gaussian Mixture Model: A Modeling Technique for Speaker Recognition and its Component, International Journal of Computer Applications (IJCA), year 2014, pp. 28-31.

[17]. K Sreenivasa Rao, “Role of neural network models for developing speech systems”, Sadhana (Indian Academy of Sciences), Vol. 36, Part 5, October 2011, pp. 783–836.

[18]. Fant, G., “Acoustic Theory of Speech Production”, Mouton and Co., The Hague, Netherlands, 1970.

[19]. Clark D. Shaver and John M. Acken, “The Development of Text-Independent Speaker Recognition Technology”, Journal of Electrical Engineering, pp. 1-8.
[20]. Bogert, Healy, Tukey, “The Quefrency Alanysis of the Time Series for Echos: Cepstrum, Pseudo-Auto-Covariance, Cross-Cepstrum, and Saphe Cracking” in Time Series Analysis, ch.15, 1963, pp. 209-243.
[21]. Cooley, J.W., Tukey, J.W., “An algorithm for the machine computation of complex Fourier series” Math Computation, vol. 19, Apr. 1965, pp.297-301.
[22]. Noll, A.M., “Cepstrum Pitch Determination”, Journal of Acoustical Society of America, vol. 41, February 1969, pp. 293-309.
[23]. Oppenheim, A. V., Schafer, R. W., “Homomorphic Analysis of Speech”, IEEE, Trans. on Audio and Electroacousitcs, Vol. 16:2, June 1968, pp. 221-226.
[24]. Schafer, R. W., Rabiner, L.R., “Digital Representation of Speech”, Invited Paper in Proceedings of the IEEE, Vol. 63:4, April 1975, pp. 662-667.
[25]. T. Kohler, “The 2010 NIST Speaker Recognition Evaluation”, SLTC Newsletter, July 2010
[26]. T. Kinnunen, H. Li. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. Speech Communication. 2009, pp. 1-30.
[27]. Furui, S. Vector-Quantization-Based Speech Recognition and Speaker Recognition Techniques. IEEE Signals, Systems and Computers. 1991, Vol. 2, pp. 954-958.
[28]. Furui, S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics Speech and Signal Processing. 1981, pp. 254–272.
[29]. J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia and A. Malegaonkar. Relative Effectiveness of Score Normalisation Methods In Open-Set Speaker Identification. Odyssey04 -The Speaker and Language Recognition Workshop. 2004.
[30]. L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, 77 (2), pp. 257-286, 1989.
[31]. J. Ferguson, Ed., Hidden Markov models for speech, IDA, Princeton, NJ, 1980.
[32]. S Furui, “50 years of progress in speech and speaker recognition”ECTI Transactions on Computer and Information Technology (ECTI-CIT) 1 (2), 2005, pp. 64-74.
[33]. D. Reynolds, T. Quatieri, R. Dunn. Speaker verification using adapted Gaussian mixture models. Digital signal processing. 2000, Vol. 10, 1-3, pp. 19-41.
[34]. McLaren, Mitchell. Improving Automatic Speaker Verification Using SVM Techniques. Queensland University of Technology. 2009. PhD Thesis.
[35]. Haykin, S. S. Neural networks: A comprehensive foundation. 2nd. s.l. : Prentice Hall, 1999.
[36]. D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Processing, vol. 3 no. 1, 1995, pp. 72-83.
[37]. D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Dig. Signal Process., vol. 10, no. 1-3, pp. 19 -41 2000.
[38]. Juergen Luettin, “Visual Speech and Speaker Recognition”, Dissertation submitted to the University of Sheffeld for the degree of Doctor of Philosophy May 1997, pp. 1-156.
[39]. Qin Jin, “Robust Speaker Recognition”, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, PhD Thesis, 2007, pp. 1-177
[40]. Alexandre Majetniak, “Speaker Recognition Using Universal Background Model on YOHO Database”, Aalborg University, Master Thesis project, May 31, 2011, pp. 1-61.
[41]. Brett Richard Wildermoth, “Text-Independent Speaker Recognition Using Source Based Features”, PhD thesis, Griffith University Australia, 2001, pp. 1-101.
[42]. Sriram Ganapathy et.al., “Robust Feature Extraction Using Modulation Filtering of Autoregressive Models”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 8, August 2014, pp. 1285-1295.
[43]. Q. D’Almeida, Frederico , Francisco A. O. Nascimento, Pedro A. Berger, and Lúcio M. da Silva. "Automatic Speaker Recognition with Multi-resolution Gaussian Mixture Models (MR-GMMs)." The International Journal of FORENSIC COMPUTER SCIENCE, pp.9-21.
[44]. Reynolds, Douglas A, and Richard C Rose. "Robust Text Independent Speaker Identification Using gaussian Mixture Speaker Model." IEEE transactions on speech and audio processing Volume 3, pp. 72-83.
[45]. Nilu Singh, Khan R A, “Underlying of Text Independent Speaker Recognition”, in IEEE Conference (ID: 37465) (10th INDIACom 2016 International Conference on Computing for Sustainable Global Development), held on 16th -18th March, 2016 at BVICAM, New Delhi, pp. 11-15.

View publication stats