International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 2, April 2020, pp. 1859~1867
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2.pp1859-1867

High level speaker specific features modeling in automatic speaker recognition system

Satyanand Singh[1], Pragya Singh[2]
[1]School of Electrical and Electronics Engineering, Fiji National University, Fiji Island
[2]School of Public Health and Primary Care, Fiji National Universisty, Fiji Island

Article Info

Article history:

Received Apr 19, 2019
Revised Oct 29, 2019
Accepted Nov 6, 2019

Keywords:

Automatic speaker recognition (ASR)
Extreme learning machine (ELM)
Gaussian mixer model (GMM)
Hidden markov model (HMM)
Linear discriminant analysis (LDA)
Support vector machines (SVM)
Universal background model (UBM)

ABSTRACT

Spoken words convey several levels of information. At the primary level, the speech conveys words or spoken messages, but at the secondary level, the speech also reveals information about the speakers. This work is based on the high-level speaker-specific features on statistical speaker modeling techniques that express the characteristic sound of the human voice. Using Hidden Markov model (HMM), Gaussian mixture model (GMM), and Linear Discriminant Analysis (LDA) models build Automatic Speaker Recognition (ASR) system that are computational inexpensive can recognize speakers regardless of what is said. The performance of the ASR system is evaluated for clear speech to a wide range of speech quality using a standard TIMIT speech corpus. The ASR efficiency of HMM, GMM, and LDA based modeling technique are 98.8%, 99.1%, and 98.6% and Equal Error Rate (EER) is 4.5%, 4.4% and 4.55% respectively. The EER improvement of GMM modeling technique based ASR systemcompared with HMM and LDA is 4.25% and 8.51% respectively.

Corresponding Author:

Satyanand Singh,
School of Electrical and Electronics Engineering,
Fiji National University, Fiji Island.
Email: satyanand.singh@fnu.ac.fj

1. INTRODUCTION

Most of ASR application modeling techniques make various mathematical assumptions about speaker-specific features. If voice data does not satisfy these attributes, incompleteness will occur at ASR modeling stage. Therefore, the mathematical model fits the features and is forced to derive recognition scores based on these models and test speech data. Converting audio segments into the functional parameter, after that modeling process started in ASR. In ASR modeling is a process flow to categories all speakers based on their characteristics. The model should also provide its meaning for comparison with unfamiliar speaker utterances. ASR modeling is called as robust when its speaker specific feature characterization process is not significantly affected by unwanted maladies, although these features are ideal if such features can be designed in such a way that interspeaker discrimination is maximum, then no intraspeaker variation exists and simple modeling methods can be sufficient. In short form, the non-ideal properties of the speaker specific feature extraction phase require different compensation techniques during the ASR modeling phase so that the effect of the disturbance variation present in the speech signal can be reduced during the testing of the speaker

recognition process. Most of the ASR modeling techniques do different mathematical hypotheses about the speaker-specific features. If assumed properties are not met from the speech data, then we are basically presenting flaws even during the ASR modeling phase.

The normalization of speaker-specific features can reduce these problems to some extent, but not completely. As a result, mathematical models are compelled to adopt the characteristics and speaker recognition scores are obtained based on these models and test speech data. Thus, in this process, the properties of detecting artifacts are introduced and a family of score standardization techniques has been proposed which is proposed to complete this final stage mismatch [1]. In essence, the decline in acoustic signal affects the speaker-specific features, patterns, and scores. Therefore, it is important to improve the robustness of ASR systems in all three domains. It has been mentioned recently that speaker modeling techniques have improved and score normalization techniques are not much effective [2].

Probabilistic modeling techniques such as GMM and HMM are widely used for the speaker, language, emotion, and speech recognition. In the probabilistic model, each speaker/language/emotion is modeled as a probability source with an unknown but fixed probability density function. The training phase is a parameter that estimates the probability density function from a sufficient number of training samples. For ASR recognition, the possibility of test utterances on the model is calculated. GMM is a linear combination of multivariate Gaussian distributions that simulate $P(X/C)$. GMM can be converted to a post classifier using Bayesian rules [3]. There are other advantages, such as being able to train the model for a large amount of speech data and adapt it to the new data format. When using a model for ASR application such as GMM, the speaker-independent Universal Background Model (UBM) first uses voice data for training. UBM represents the distribution of feature vectors independent of speakers. When a new speaker is registered in the ASR system, the parameters of the background model are adapted to the feature distribution of the new speaker. The adaptive model is then used as an ASR speaker's model.

Statistical Language Modeling (LM) is the science of building a model to estimate the prior probability of word strings. Successful use of language model to model the rhythm of speaker and language. The fundamental frequency F_o and energy profiles are labeled as discrete classes and then modeled using two bigrams or trigrams [4]. Hidden Events LM contains special words that appear in the model's N-gram. Instead, they correspond to the state of the HMM and can be used to simulate language events such as boundaries of unmarked sentences. Alternatively, these events may be associated with unnatural possibilities for adjusting LM (eg, rhythm) for other sources of knowledge. A special type of hidden event LM can simulate a nonsmooth speech by letting hidden events modify the word history [5].

Decision trees are also successfully used in prosodic modeling for ASR application [6]. The decision tree model "progress" by system-generated question to the speaker at once. The features of the questions in each question and then the thresholds in the questions (eg normalized pitch greater than threshold value) preferably distinguish the class of nodes in the tree. In the test phase, the decision tree estimates the posterior probability of each class C of each sample X, resulting in $P(X/C)$ [7]. One of the main drawbacks of decision trees is the greedy build process: at each step, the combination selects a single best variable and the best breakpoint, but considering multi-step prefetching of variable combinations than a good result. Another disadvantage is the fact that continuous variables are implicitly discretized by the partitioning process and information is lost along the way. The advantage of decision trees for other machine learning methods is that they are not black-box models, but can easily be represented as rules. In many applications, these models are more important than disadvantages, so these models are widely used in ASR application.

Discriminant models such as Artificial Neural Networks (ANN) [8] and Support Vector Machines (SVM) are also used for prosodic modeling [9]. Deep Neural Network (DNN) [10], Extreme Learning Machine (ELM), and DNN-ELM have proved useful for prosodic-based speaker recognition [11]. The SVM model is an algorithmic implementation of the idea from the statistical learning theory [12] and focuses on the problem of constructing a consistent estimator from the speech data. Model performance and training set estimation method for unknown data set when only model characteristics are given Performance? Regarding the algorithm, the support vector machine establishes an optimal separation boundary between data sets by solving the constrained quadratic optimization problem [13]. By using different kernel functions, different degrees of nonlinearity and flexibility can be included in the model. Support vector machines are gained from advanced statistical ideas and can calculate the range of generalization error for them, so we have gained considerable research interest over the past few years. The performance of other machine learning algorithms equal to or better than those of other machine learning algorithms are reported in the medical literature. A disadvantage of the support vector machine is that the classification result is purely dichotomous and there is no possibility of giving class membership [14].

2. MODELING BASED ON PROSODY IN AUTOMATIC SPEAKER RECOGNITION SYSTEM

Prosody uses the appropriate method to obtain the global statistics of the speaker's fundamental frequency F_o value and the ASR system recognizing the task. The dynamics of the F_o contour reflecting the person's talking style has been shown to be able to help the speaker recognition the task. The F_o motion of the speaker is modeled by fitting a piecewise linear model to the F_o orbit to obtain a stylized F_o profile. Using median F 0, the slope and duration represent each linear F_o segment. These features are modeled by log-normal distribution, normal distribution, and shift exponential distribution, respectively. In order to investigate the possibility of speaker recognition using rhythm and idiom, NIST introduced extended data task telephone talk based on exchange corpus. Unlike traditional speaker recognition tasks, the extended data task provides multiple complete session planes (4/8/16 sides) for speaker training and testing the ASR system.

In [15] the focus is on investigating various prosodic features. Fundamental frequency based on segment period and pause period. Periodic characteristics, or word characteristics, telephone periods and period sequences have been used to model the period. In [16], duration, pitch, and energy characteristics are calculated for each estimated syllable region. Syllable boundary obtained from the ASR system. These features are quantized and used to form N-grams called N-gram based syllable non-uniform extraction region features.

In [17], continuous prosodic features were modeled using Joint Factor Analysis (JFA) for speaker recognition. The prosodic feature used is the pitch and energy profile over units of similar syllables, represented using bases of Legendre polynomials. Standard GMM is used for modeling. In addition, the effect of the speaker and session change is modeled in the same way as conventional JFA. Legendre polynomial coefficients of pitch and energy, together with the length of the segment, constitute a 13-dimensional prosody feature set for GMM and factor analysis modeling [17].

2.1. Eigenvoice consideration in hidden markov models

In the standard eigenvoice approach, voice data is collected from the number of speakers with the diverse scenario. When each HMM state is modeled as a mixture of Gaussian distributions, a set of speaker-dependent HMMs are formed from each speaker. The speaker's voice is represented by the super vector composed of the concatenation of the mean vectors of all Gaussian HMM distributions. Therefore, the i-th speaker supervector is composed of R components, one Gaussian per distribution, and is expressed as $x_i = \left[x_{i1}^{'}, x_{i2}^{'}, \ldots . x_{iR}^{'}\right]^{'} \in \mathbb{R}^{d_2}$. The similarity between any two speaker supervectors x_i and x_j is measured by their dot product as follows.

$$x_i^{'} x_j = \sum_{r=1}^{R} x_{ir}^{'} x_{jr} \quad (1)$$

Principal component analysis (PCA) is then performed on the training speaker supervector and the resulting eigenvector is referred to as eigenvoice. In order to adapt to the new speaker, his/her supervector process deals with a linear combination of the top M eigenvoices $s = s^{(ev)} = \sum_{m=1}^{M} [\{w_1, w_2, \ldots . w_M\}]^{'} V_m$. Usually, only a less than ten eigenvoices are taken into consideration so that few second of adaptation speech will be required. The mathematically computed eighteen eigenvoices are as: 0.180696, 0.168936, 0.082378, 0.065117, 0.058677, 0.027971, 0.020124, 0.017375, 0.016086, 0.008081, 0.007063, 0.004332, 0.003474, 0.003072, 0.002031, 0.001976, 0.00112, and 0.001062. The adaptation data $o_t, t = 1, \ldots \ldots , T$ to estimate unique eigenvoice weights by maximizing the likelihood of o_t. In mathematically one can find w by maximizing the Q function as follows:

$$Q(w) = \sum_{r=1}^{R} \gamma_1(r) log(\pi_r) + \sum_{p,r=1}^{R} \sum_{t=1}^{T-1} \xi_t(p,r) log\left(a_{pr}\right) + \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) log\left(b_r(o_t, w)\right) \quad (2)$$

State r initial probability and posterior probability of observation is represented by π_r and $\gamma_t(r)$ respectively at time t. State p posterior probability of observation sequence is represented by $\xi_t(p, r)$ at time t and at state r at time $t+1$. b_r is the r^{th} Gaussian probability density function. Further $Q_b(w) = \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) \log\left(b_r(o_t, w)\right)$ is related to the new speaker supervector s as follows:

$$Q_b(w) = -0.5 \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) [d_1 \log(2\pi) + \log|C_r| + \|o_t - s_r(w)\|^2 C_r] \quad (3)$$

Covariance matrix of the Gaussian in eqn. (3) at state r is represented as C_r. Here the estimation of eigenvoices is generalized by performing kernel PCA in its place of linear PCA. Subsequently, let $k(.,.)$ be a kernel with a corresponding mapping φ. This maps the pattern x of the specific speaker supervector space χ to the $\varphi(x)$ in the speaker specific feature space $\mathcal{F}$. Given a set of N patterns speaker supervectors $(x_1, x_2, \ldots \ldots x_{N-1}, x_N)$ denote the mean of the φ -mapped feature vectors by $\bar{\varphi} = \frac{1}{N} \sum_{i=1}^{N} \varphi(x_i)$ and the centered map with $\tilde{\varphi} = \varphi(x) - \bar{\varphi}$. Next step Eigen decomposition is performed on $\tilde{K}$ where

$K = [k(x_i, x_j)]_{i,j}$. v_m is the m^{th} orthogonal eighnvector of NXN dimension covariance matrix in the feature space is represented as $v_m = \sum_{i=1}^{N} \frac{\alpha m_i}{\sqrt{\lambda_m}} \bar{\varphi}(x_i)$ by considering $K = U \Lambda U'$ where $U = [\alpha_1, \ldots \ldots \alpha_{N-1}, \alpha_N]$ with $\alpha_i = [\alpha_{i1}, \ldots . . \alpha_{i(N-1)}, \alpha_{iN}]'$ and $\Lambda = diagonal(\lambda_1, \ldots \ldots \lambda_{N-1}, \lambda_N)$. A computer generated $8X8$ orthogonal eighnvector v_m is represented in Table 1. Two-dimension representation of utterances from TIMIT database evaluation using KPCA+linear solution and non-linear SVM shown in Figure 1.

Table 1. A computer generated 8X8 orthogonal eighnvector v_m

	C1	C2	C3	C4	C5	C6	C7	C8
R1	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R2	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R3	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R4	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R5	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R6	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R7	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000
R8	-1.0000	-0.8571	-0.7143	-0.5714	-0.4286	-0.2857	-0.1429	0.0000

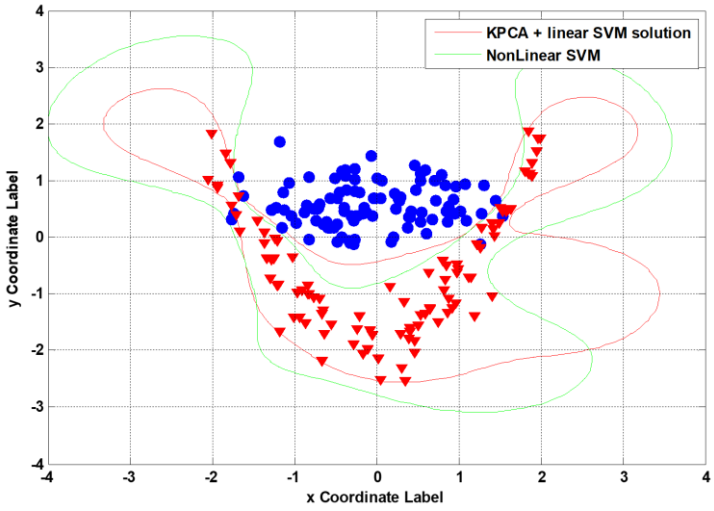

Figure 1. Two-dimensio representation of utterances from TIMIT database evaluation using KPCA+linear solution and non-linear SVM

2.2. Gaussian mixture model (GMM) based high label feature modeling

GMM has become the leading generation statistical model in the state of the art ASR system. GMM is an attractive statistical model because it can represent various probability density functions when estimating a sufficient number of parameters. The GMM, in general, contains a set of N multivariate Gaussian density functions represented by the index k. The resulting probability density function for a particular speaker model i is a convex combination of all density functions. GMM is built using standard multivariate Gaussian density,but introduces component index k as a latent variable with discrete probability $p(k/i)$. The weights are represented as $w_k^i = p(k/i)$. Complies with the GMM density function and the conditions that characterize the past contributions of the corresponding component as $\sum_{k=1}^{N} w_k^i = 1$. Each Gaussian density represents a conditional density function $p((x_t|k,i))$. According to Bayes' theorem, the joint probability density function $p((x_t|k,i))$is given by the multiplication of the two. The sum over all densities results in the multi-modal probability density of GMMs as follows:

$$p(x_t|\ominus_i) = \sum_{k=1}^{N} p(k|\ominus_i) \cdot p(x_t|k,\ominus_i) = \sum_{k=1}^{N} w_k^i \cdot \mathcal{N}\{(x_t|\mu_k^i, \Sigma_k^i)\} \quad (4)$$

Where μ_k is the mean vector and Σ_k is the covarience matrix. Each component density is completely determined by μ_k and Σ_k. The parameter set $\ominus_i = \{w_1^i, w_2^i, \ldots . , w_N^i, \mu_1^i, \mu_2^i, \ldots . \mu_N^i, \Sigma_1^i, \Sigma_2^i, \ldots . \Sigma_N^i\}$ where eighting factor including specific speaker model i of mean vector and covariance matrix.

Figure 2 illustrates the likelihood function of the GMM, including seven Gaussian distributions with covariance matrices of two dimensional mean and feature vectors are chosen x_1 and x_2 denote the elements of the feature vector. Computer generated log-likelihood completed training speaker 1 model is represented as -6.067379, -4.288333, -4.253459, -4.241043, -4.230592, -4.218451, -4.203952, -4.188224, -4.173566, -4.161955, -4.153866, -4.148612, -4.145268, -4.143124, -4.141712, -4.140738. A computer generated 8X8 training feature vectors of a speaker by Gaussian Mixture Models is represented in Table 2 and Table 3 represent testing feature vectors of same speaker with different text. Figure 2 shows a likelihood function for a GMM with seven Gaussian densities.

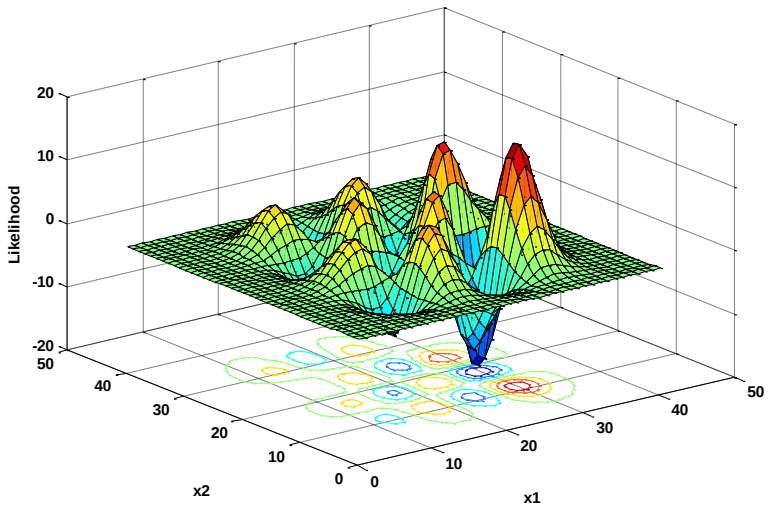

Figure 2. A likelihood function for a GMM with seven Gaussian densities

Table 2. A computer generated 8X8 training feature vectorsof a speaker by Gaussian mixture models

	C1	C2	C3	C4	C5	C6	C7	C8
R1	4.0646	2.7960	3.3696	2.5665	1.4115	1.4582	1.3393	0.7637
R2	4.8317	3.5756	3.3678	2.8608	0.9304	0.8075	0.9295	1.1848
R3	3.7562	3.4273	3.8380	2.7522	1.3471	0.9934	1.4731	1.6576
R4	5.0021	3.3969	3.4032	2.2354	0.4914	0.8931	2.0563	1.4244
R5	4.1528	3.3462	3.8148	3.4006	1.8268	1.0450	1.5436	1.1512
R6	3.8352	3.1605	4.3616	2.8652	1.7510	1.0464	1.6336	1.3007
R7	4.1610	3.3430	4.4114	1.7857	1.1003	1.5388	1.3885	1.6549
R8	3.5921	3.7265	4.1634	2.5118	1.8623	1.5231	1.5569	1.4148

Table 3. 8X8 testing feature vectors of a speaker by Gaussian mixture models

	C1	C2	C3	C4	C5	C6	C7	C8
R1	3.2927	2.0086	4.7630	3.1760	1.4675	0.9331	1.7318	1.3194
R2	3.6418	2.6172	5.1925	2.5124	0.5417	1.2929	1.9916	0.9756
R3	2.9897	1.6382	5.2565	4.0006	1.3647	1.8824	1.9576	1.0245
R4	3.4203	2.3760	4.4596	2.5434	1.0803	1.4107	1.8440	1.3208
R5	3.4864	2.9604	3.9410	3.2120	1.5138	1.5098	2.2160	1.2051
R6	4.0004	2.2980	4.2781	3.0504	1.8364	1.0121	1.2600	1.1491
R7	3.0806	2.0417	4.0331	3.6395	1.9743	1.8195	1.3774	1.0800
R8	2.9109	2.3116	4.6019	3.5167	2.3270	1.1858	2.6674	1.3994

2.3. Linear discriminant analysis (LDA) based high label feature modeling

LDA is a commonly employed technique in statistical pattern recognition that aims at finding linear combinations of feature coefficients to facilitate discrimination of multiple classes. It finds orthogonal orientation in place of most effective functions in class discrimination. By introducing the original features in these guidelines, the accuracy of classification improves. Let us indicate the set of all development utterances by D, utterance features indicated by $w_{s,i}$, these features obtained from the ith utterance of the speaker s, the total number of utterances belonging to s is indicated by n_s and the total number of speakers in D is indicated by S. Class covariance matrices between S_b and within S_w are given by

$$S_b = \frac{1}{S}\sum_{s=1}^{S}(\overline{w}_s - \overline{w})(\overline{w}_s - \overline{w})^T \quad (5)$$

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_{s,i} - \overline{w}_s)(w_{s,i} - \overline{w}_s)^T \quad (6)$$

Where the speaker dependant mean vector is given by $\overline{w}_s = 1/n_s \sum_{i=1}^{n_s} w_{s,i}$ and speaker independent mean vector is given by $\overline{w} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s} w_{s,i}$ respectively. The LDA optimization is therefore to maximize between class variance, whereas reducing within the class variance. The exact estimation can be obtain from this optimization by solving generalized eigenvalue problem:

$$S_{bV} = \wedge S_w v \quad (7)$$

The diagonal matrix containing of eignvector is indicated by ∧. If the matrix S_w in eqn. (6) is invertible then the solution can be easily found by $S_w^{-1}S_b$. A_{LDA} matrix of dimension R × k is as follows:

$$A_{LDA} = [v_1 \dots \dots \dots v_k] \quad (8)$$

k eigenvectors $v_1 \dots \dots \dots v_k$ obtained by solving eqn. (7). Thus, the LDA change of the utterance feature w is obtained in this way:

$$\Phi_{LDA}(w) = A_{LDA}^T w \quad (9)$$

A computer generated 8X8 $\Phi_{LDA}(w)$ matrix of dimension RXk by LDA Models is represented in Table 4.

Table 4. A computer generated 8X8 $\Phi_{LDA}(w)$ matrix of dimension RXk

	C1	C2	C3	C4	C5	C6	C7	C8
R1	-0.5302	-0.6328	-0.6402	-0.5861	-0.5306	-0.5137	-0.5403	-0.5678
R2	-0.6601	-0.7932	-0.8189	-0.7774	-0.7347	-0.7332	-0.7773	-0.8138
R3	-0.6949	-0.8420	0.8846	-0.8622	-0.8389	-0.8565	-0.9219	-0.9783
R4	-0.6594	-0.8031	-0.8484	-0.8308	-0.8124	-0.8399	-0.9289	-1.0271
R5	-0.6314	-0.7653	-0.7968	-0.7584	-0.7169	-0.7325	-0.8374	-0.9885
R6	-0.6698	-0.8029	-0.8170	-0.7446	-0.6615	-0.6450	-0.7462	-0.9332
R7	-0.7548	-0.8985	-0.9072	-0.8157	-0.7044	-0.6588	-0.7423	-0.9333
R8	-0.7876	-0.9328	-0.9467	-0.8688	-0.7722	-0.7314	-0.8065	-0.9806

LDA assumes normal distribution data for all classes, statistically independent features and the same covariance matrix. However, this only applies to LDA as a classifier. If these assumptions are violated, the dimensionally reduced LDA can work reasonably. Even for classification tasks, LDA seems powerful enough to be used for data distribution in ASR applications. The speaker feature modeling histograms with normal fit eigenvector obtained from the LDA is illustrated in Figure 3.

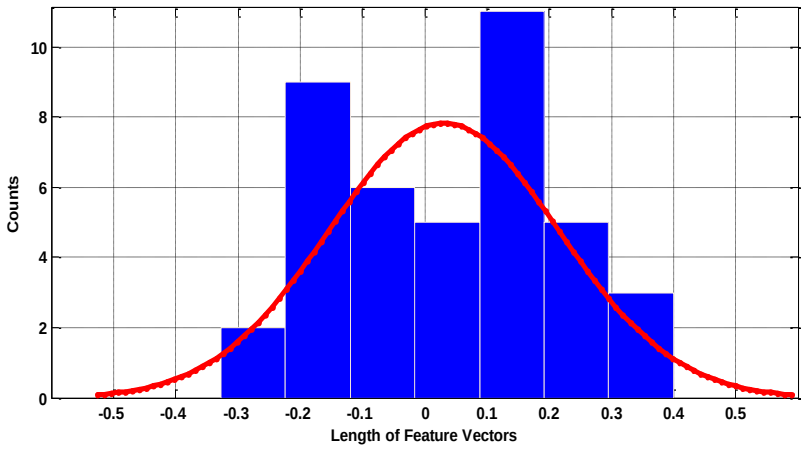

Figure 3. The speaker feature modeling histograms with normal fit eigenvector with LDA

3. ACOUSTIC DATA FEATURE EXTRACTION

The speaker specific features refer to parameters extracted from phrase segments/periods within a 20-25 ms frame. The most common short-term acoustic features are Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding (LPC) based features [18,19,20]. In order to obtain these coefficients from the speech recording, the speech samples are first divided into short overlapping segments. The signals obtained at these segments / frames are then multiplied by a window function (e.g. Hamming and Hanning) to obtain a Fourier power spectrum. In the next step, the logarithm of the spectrum is calculated and a mel-space filter bank analysis of non-linear intervals is performed. Logarithmic operations expand the range of coefficients and break up the multiplicative components into additional components [21]. In filter bank analysis, spectral energy (also called filter bank energy coefficient) is generated for each channel to represent different frequency bands.

Filterbanks, like the human auditory system, are designed to be more sensitive to frequency changes at the bottom of the spectrum. Finally, the MFCC is obtained by performing a discrete cosine transform (DCT) on the filter bank energy parameters and retaining many preamble coefficients [22, 23]. DCT has two important properties. (i) to compress the energy of the signal into multiple coefficients, and (ii) to be highly correlated with the coefficients. For these reasons, using DCT to remove specific dimensions improves the efficiency of the model and reduces some harmful components [24]. Furthermore, the uncorrelated properties of the DCT help to assume that the models of feature coefficients are not relevant. In summary, the following sequence of operations-power spectrum, logarithm, DCT-produces a signal with a well-known cepstral representation [25].

4. EXPERIMENTAL SETUP

The experiment uses the TIMIT set of database. The proposed algorithm implemented in MATLAB and results were compared with those of the Eigenvoice consideration in HMM, GMM and LDA. A total 1000 utterances of the TIMIT database of 6 sec, 4 sec and 2 sec voice were put to train and test the ASR system. For the above cases, ASR recognition efficiency has been calculated "Efficiency" = Number of utterance correctlyidentified/Total Number of utterance under test. Table 5 shows that the efficiency of the ASR system for HMM, GMM and LDA respectively. It can be observed from this table that use of GMM has highest efficiency compared to other modeling techniques. Figure 4 show the equal error rate (EER) of HMM, GMM, and LDA based modeling technique. The ASR efficiency of HMM, GMM, and LDA based modeling technique are 98.8%, 99.1%, and 98.6% and EER are 4.5%, 4.4% and 4.55% respectively. The EER improvement of GMM modeling technique based ASR system compared with HMM and LDA is 4.25% and 8.51% respectively.

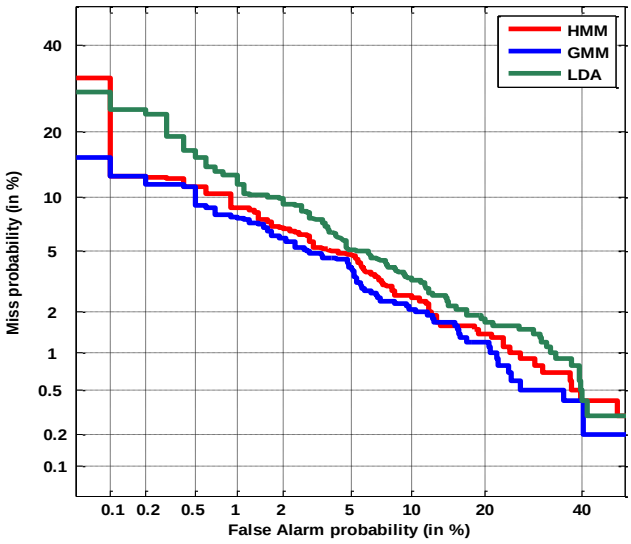

Figure 4. Equal Error Rate of ASR system of HMM, GMM and LDA based modeling technique for 2 sec of voice data

Table 5. Efficiency of the ASR system for HMM, GMM and LDA respectively

	HMM		GMM		LDA	
	Efficiency in %	EER in %	Efficiency in %	EER in %	Efficiency in %	EER in %
6 sec	99.6	4.9	99.9	4.7	99.1	5.1
4 sec	98.8	4.9	99.5	4.7	98.2	5.1
2 sec	98.8	4.9	99.1	4.7	98.6	5.1

5. CONCLUSION

This paper presented the research, development and evaluation of ASR system based on HMM, GMM and LDA modeling techniques. GMM models provide a simple but effective representation that offers inexpensive and high recognition accuracy for a wide range of speaker recognition tasks. An experimental evaluation of the performance of the speaker recognition system has been done on publicly available TIMIT database. For the 1000, voice samples of the TIMIT database spaker recognition accuracy 99.1%, 98.8% and 98.6 for GMM, HMM and LDA was obtained for 2 sec of voice length. The EER improvement of GMM modeling technique based ASR system compared with HMM and LDA is 4.25% and 8.51% respectively.

As experimental results showed that, speaker recognition performance is at practically usable levels for specific applications such as access control authentication. The main limiting factor in less controlled situations is the lack of robustness to transmission impairments such as noise and mic variability. Much more to address these limitations, such as exploring areas such as understanding and modeling the impact of impairments on spectral characteristics, applying more sophisticated channel compensation techniques, and exploring features that are less sensitive to channel degradation efforts are underway.

REFERENCES

[1] S. Singh, "Forensic and Automatic Speaker Recognition System" *International Journal of Applied Engineering Research*, Vol. 8, No. 5, 2018, pp. 2804-2811, 2018.

[2] S. Singh and Ajeet Singh "Accuracy Comparison using Different Modeling Techniques under Limited Speech Data of Speaker Recognition Systems," *Global Journal of Science Frontier Research: F Mathematics and Decision Sciences, vol* 16(2), pp.1-17, 2016.

[3] S. Singh. "Bayesian distance metric learning and its application in automatic speaker recognition systems" *International Journal of Electrical and Computer Engineering*, Vol, 9, No. 4, 2019.

[4] S. Singh. "The Role of Speech Technology in Biometrics, Forensics and Man-Machine Interface" *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 1, pp.281-288, 2019.

[5] S. Singh. "High Level Speaker Specific Features as an Efficiency Enhancing Parameters in Speaker Recognition System," *International Journal of Electrical and Computer Engineering*, Vol, 9, No. 4, 2019.

[6] S. Singh, Abhay Kumar, David Raju Kolluri, "Efficient Modelling Technique based Speaker Recognition under Limited Speech Data," *International Journal of Image, Graphics and Signal Processing(IJIGSP)*, Vol.8, No.11, pp.41-48, 2016.

[7] Shriberg, E., & Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," In Speech Prosody, Nara, Japan 2004.

[8] Mary, L., & Yegnanarayana, B, "Prosodic features for speaker verification," In *Proceedings of Interspeech*, Pittsburgh, Pennsylvania, pp. 917- 920, 2006.

[9] Ferrer, L., Shriberg, E., Kajarekar, S., & Sonmez, K, *"Parameterization of prosodic feature distributions for SVM modeling in speaker recognition*," In Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol. 4, pp. 233-236, 2007.

[10] Han, K., Dong, Y., & Tashev, I, "*Speech emotion recognition using deep neural network and extreme learning machine*," In Proceedings of Interspeech, pp. 223-227, 2014.

[11] Wang, Z. Q., & Tashev, I, "*Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks*," In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[12] Vapnik V. "*An Overview of Statistical Learning Theory,"* IEEE Transaction on Neural Networks, Vol. 10, No. 5, pp. 988-999,1999.

[13] S.Singh, "Support Vector Machine Based Approaches For Real Time Automatic Speaker Recognition System," *International Journal of Applied Engineering Research*, Vol. 13, No. 10, pp. 8561-8567, 2018.

[14] Scholkopf B, Smola A, "Learning with kernels: support vector machines, regularization, optimization, and beyond," Cambridge, MA: MIT Press; 2002

[15] Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D., et al., "*Using prosodic and conversational features for high-performance speaker recognition*," Report from JHU WS'02, In Proceedings of ICASSP, Hong Kong, China, Vol. 4, pp. 792-795, 2003.

[16] S.Singh, Mansour H. Assaf, Sunil R.Das, Emil M. Petriu, and Voicu Groza, "Short Duration Voice Data Speaker Recognition System Using Novel Fuzzy Vector Quantization Algorithms," 2016 IEEE International Instrumentation and Measurement Technology Conference, May 23-26, Taipei, Taiwan, 2016.

[17] Najim, D., Dumouchel, P., & Kenny, P, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15 7, 2095-2103, 2007.
[18] S.Singh, "Speaker Recognition by Gaussian Filter Based Feature Extraction and Proposed Fuzzy Vector Quantization Modeling Technique," *International Journal of Applied Engineering Research*, Vol. 13, No. 16, pp. 12798-12804, 2018.
[19] S.Singh, "High Level Speaker Specific Features Modeling in Automatic Speaker Recognition System," *International Journal of Electrical and Computer Engineering*, Vol. 10, No. 2, 2018, pp. 2804-2811, 2020.
[20] S.Singh, "Speaker Recognition System for Limited Speech Data Using High-Level Speaker Specific Features and Support Vector Machines" *International Journal of Applied Engineering Research*, Vol. 12, No. 9, 2018, pp. 8026-8033 2017.
[21] S.Singh, MH Assaf and Abhay Kumar, "A Novel Algorithm of Sparse Representations for Speech Compression/Enhancement and Its Application in Speaker Recognition System," *International Journal of Computational and Applied Mathematics*, Vol. 11, No. 1, pp. 89-104, 2016.
[22] S.Singh, "Evaluation of Sparsification algorithm and Its Application in Speaker Recognition System" *International Journal of Applied Engineering Research*, Vol. 13, No. 17, pp. 13015-13021, 2018.
[23] S.Singh and Mansour H. Assaf "A Perfect Balance of Sparsity and Acoustic hole in Speech Signal and Its Application in Speaker Recognition System" *Middle-East Journal of Scientific Research*, Vol. 24, No.11, pp. 3527-3541, 2016.
[24] S.Singh and Dr. E.G. Rajan, "MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors," *International Journal of Engineering Research & Technology, International Journal of Computer Applications*, Vol. 21, No. 6, pp. 1-6, 2011.
[25] S.Singh and Dr. E.G. Rajan, "Application of Different Filters In Mel Frequency Cepstral Coefficients Feature Extraction And Fuzzy Vector Quantization Approach In Speaker Recognition," *International Journal of Engineering Research & Technology*, Vol. 2 Issue 6, pp-3171-3182, 2013.