International Journal of Scientific Research in Computer Science and Engineering
Research Paper
Vol.5, Issue.5, pp.75-82, October (2017)
E-ISSN: 2320-7639

Comparison Studies of Speaker Modeling Techniques in Speaker Verification System

K. Sarmah

Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Goalpara, India

[*] *Corresponding Author: kshirodsarmah@gmail.com, Tel.: +91-94018-00625*

Available online at: www.isroset.org

Received 11th Sep 2017, Revised 20th Sep 2017Accepted 19th Oct 2017, Online 30th Oct 2017

Abstract— In this paper a brief comparison studies on the performance of different speaker modeling techniques in robust and reliable speaker verification (SV) system has been discussed. In text-independent speaker verification, lots of states of art speaker modeling techniques have been developed in different scenarios to upgrade its performance. The performance of SV system is not only depended on the fusion of different feature vectors but also it is highly depended upon the fusion of various speaker modeling techniques. In this work, an automatic SV system has been developed using the Mel-Frequency Cepstral Coefficients (MFCC) combined with the Prosodic feature vectors. The baseline of the SV system has been trained with speaker modeling techniques separately and fusions namely Vector Quantization (VQ), Gaussian Mixture Model (GMM), GMM-Universal Background Model (GMM-UBM), Support Vector Machine (SVM) and Joint Factor Analysis (JFA) to analyze its performances. The results reported here, have been evaluated using the multilingual speech database, namely Arunachali Language Speech Database (ALS-DB). From the experimental point of view we observe that the best performance of SV system shows by JFA with GMM-UBM modeling technique with its EER value of 4.76% and MinDCF value of 0.0872. Comparing with other modeling techniques VQ shows its poor performance with its EER value of 11.08% and MinDCF value of 0.2010. SVM shows of approximately 2.8% improvement of verification rate with comparison to that of GMM-UBM. Here, finally, we conclude that the fusions of both generative and discriminative models highly improve the performance of SV system.

Keywords— Speaker Verification,MFCC,Prosodic,GMM-UBM,SVM,JFA.

I. INTRODUCTION

Speaker verification also known as speaker detection is the task of determining whether the unknown voice is from the particular speaker or not [1]. It is also a biometric task and binary decision where the challenge is to find whether or not utterances come from a target speaker. The main problem arises from the biometric domain is to identify individuals for security purposes.

SV systems aim to extract, characterize and recognize the information enclosed in the speech signal conveying the identity of a speaker. The area of speaker recognition can be categorized into two fundamental tasks: speaker identification and speaker verification or detection [1,2]. In its simplest form speaker identification is the task of assigning an unknown voice to one of the speakers known by the system: it is assumed that the voice must come from a fixed set of speakers. Thus, the system must solve a *n*-class classification problem and the task is often referred to as closed set identification. A closely related problem is that of speaker detection, this is the task of determining whether an unknown voice is from a particular enrolled speaker [1]. The performance of SV system has been enhanced by combining the spectral features with prosodic features. Prosodic cues include stress, rhythm and intonation that are expressed using three acoustic parameters namely pitch, energy and duration that capable of conveying many more information of speakers and it is relatively less affected by channel variations and noise [3].

The objective of speaker modeling technique is to generate speaker models using speaker–specific feature vectors, which models will have enhanced speaker-specific information at a reduced data rate. This is achieved by exploiting the working principle of the classifiers or modeling techniques. A speaker is characterized by a speaker model such as VQ, GMM or SVM. An unknown speech sample is first represented by a collection of feature vectors may be Linear Predictive Cepstral Coefficients (LPCC), MFCC or Prosodic or a supervector which is a concatenation of multiple vectors, then evaluated against the target speaker models [5].

Early methods for speaker recognition included non-parametric techniques like Vector Quantization (VQ) and

Dynamic Time Warping (DTW) [4]. Recently, classification methods for speaker recognition have given importance on statistical as well as parametric approaches. The structures and choice of a classifier depends on the application and the scope of the features used. Classical speaker models can be divided into template model and stochastic models also known as nonparametric and parametric models respectively.

Template models are used to train and to test feature vector directly and compared with each other with assuming that either one is imperfect replica of the other. For example VQ and DTW. But in the case of stochastic model, each speaker is modelled as a probabilistic source with an unknown fixed probability function [4]. Training is done to estimate the parameters of the probability density function from the training sample of data. And matching is done by evaluating the likelihood of the test pattern with respect to the trained model. For example HMM and GMM are the most popular stochastic models for text-dependent and text-independent recognition respectively. In this paper, section I contains the introduction of speaker verification system and state-of-art speaker modeling techniques, section II describes the different speaker modeling techniques namely VQ and GMM-UBM, SVM and JFA. Sections III describes about the experimental setup, Performance evaluation and result analysis has been elaborated in the section IV and V respectively. Finally, Conclusion and future research scope has been explained briefly in the section VI.

II. Speaker Modeling Techniques

Based on the training paradigm classifiers or speaker models can also divide into two categories namely generative and discriminative models. The discriminative models like artificial neural networks (ANN) and SVM models the boundary between speakers whereas in case of generative model like VQ and GMM, estimating the feature distribution within each speaker [5]. In previous studies, we observed that GMM-UBM showed better performance while applying different normalization techniques in model level, score level and finally feature level [6,7,8]. In this paper we concentrate only on various speaker modeling techniques and its fusion of both generative and discriminative categories to get better performance in SV system.In this section, we are going to discuss different state of art speaker modeling techniques.

A. Vector Quantization(VQ)

First, VQ model is also known as centroid model which is one of the simplest text-independent speakers modeling technique [5]. It was introduced in 1980s and utilized in speaker recognition. VQ is also known as generative classifier like GMM because it estimates the feature distribution within each speaker. VQ provides good accuracy when it is combined with background model adaptation.

The average quantization distortion can be defined as follows:

Let the test utterance feature vectors denoted by X=$\{x_1, x_2, x_3, \ldots\ldots x_T\}$ and reference vector by R=$\{r_1, r_2, r_3, \ldots\ldots r_k\}$ Then the average quantization distortion is

$$D_Q(X,R) = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le k\le K} d(x_t\,, r_k\,) \qquad (1)$$

Where $d(.,.)$ is the Euclidean distance defined as $\|x_t - r_k\|$ (2)

Note to remember that

$$D_Q(X,R) \neq D_Q(R,X)$$

B. Gaussian Mixture Model (GMM)

Over the last decade, the Gaussian Mixture model GMM has become established as the standard classifier for text-independent speaker recognition [9]. Gaussian Mixture model (GMM) often to be used to the speaker verification because this model has good ability of recognition [10]. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped distributions [11]. GMMs have unique advantages compared to other modeling approaches because their training is relatively fast and the models can be scaled and updated to add new speakers with relative ease [12].

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a sum of Gaussian components densities. GMMs are commonly used as a parametric model of the probability distribution of a continuous measurement of features in a biometric system [11].

A GMM is a weighted sum of M component densities is given by the form

$$P(x|\lambda) = \sum_{i=1}^{M} w_i\, b_i(\mathrm{x}) \qquad (3)$$

Where x is a dimensional random vector, b_i(x), i=1,2……M, is the component densities and w_i i=1,2,….,M, is the mixture weights.

The Gaussian Function can be defined of the form

$$b_i(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}\, |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)' \sum_i^{-1}(x-\mu_i)\right\} \qquad (4)$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weight satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weight from all component densities.

These parameters can collectively represented by the notation:

$$\lambda \ = \{w_i, \mu_i, \textstyle\sum_i\} \qquad \text{for i= 1,2 ……, M.} \qquad (5)$$

In speaker verification system, each speaker can be represented by such a GMM and is referred to by the above model λ.

For a sequence of T test vectors X= {$x_{1,}$ $x_{2,}$ $x_{3,........}$ x_T } the required standard way to calculate the GMM likelihood in the log domain as follows:

$$\mathrm{L(X|\,\lambda) = log(X|\,\lambda)} \ = \textstyle\sum_{i=1}^{T} \mathrm{log(x_i|\lambda_i)} \qquad (6)$$

Once a model is trained then (6) can be used to compute the log-likelihood of model λ for an input test set of feature vector, X can be defined as

$$\mathrm{log\,p(X|\,\lambda)} = \textstyle\sum_{i=1}^{T} \mathrm{log\,p(x_i|\,\lambda)} \qquad (7)$$

It is also important to note that because the component Gaussian is acting together to model the overall feature densities, full covariance matrices are not necessary even if the features are not statistically independent. The linear combination of diagonal covariance basis Gaussians is capable of modeling the correlations between feature vector elements. The effect of using a set of M full covariance matrix Gaussians can be equally obtained by using a larger set of diagonal covariance Gaussians.

(i) Maximum Likilihood Parameter Estimation

For a given training vectors and a GMM configuration, we have to estimate the parameters of the GMM, λ, for the best matches the distribution of the training feature vectors. The most popular and well-known method is maximum likelihood (ML) estimation.

The main purpose of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors X= $x_{1,}$ $x_{2,}$ $x_{3,........}$ x_T , the GMM likelihood can be defined as

$$\mathrm{p(X|\,\lambda)} \ = \textstyle\prod_{t=1}^{T} p(\mathrm{x_t}|\lambda). \qquad (8)$$

The speaker-specific GMM parameters are estimated by the Expectation-Maximization (EM) algorithm using training data spoken by the corresponding speaker. The basic idea of the EM algorithm is, beginning with an initial language model λ, to estimate a new model λ′ such that $P(X|\lambda')$≥ $P(X|\,\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached [11].

On each EM iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value,

Mixture Weights:

$$\mathrm{w_i} \ = \frac{1}{T}\textstyle\sum_{t=1}^{T} \mathrm{pr(i|x_t,\lambda)} \qquad (9)$$

Means:

$$\mu_i = \frac{\sum_{t=1}^{T} \mathrm{pr(i|x_t,\lambda)x_t}}{\sum_{t=1}^{T} \mathrm{pr(i|x_t,\lambda)}} \qquad (10)$$

Variance (diagonal covariance):

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} \mathrm{pr(i|x_t,\lambda)}x_i^2}{\sum_{t=1}^{T} \mathrm{pr(i|x_t,\lambda)}} - \mu_i^2 \qquad (11)$$

The a posteriori probability for component i is given by

$$\mathrm{Pr(i|x_t,\lambda)} = \frac{w_i b_i(\mathrm{x})}{\sum_{k=1}^{M} w_k b_k(\mathrm{x})} \qquad (12)$$

(ii) Maximum A Posteriori (MAP) Parameter Estimation

GMM parameters can also be estimated using Maximum A Posteriori (MAP) estimation. MAP estimation is used to derive speaker model by adapting from a Universal Background Model (UBM). Like the EM algorithm, the MAP estimation is a two step process. The first step is similar to the "Expectation" step of the EM algorithm that sufficient statistics of training data are computed for each mixture in the prior model. In the second step, the new sufficient statistics from training data are used to update the prior sufficient statistics for mixture i to create the adapted parameters for mixture i.

The specifics of the adapting are defined as for given a prior model and training vectors from the desired class X= {$x_{1,}$ $x_{2,}$ $x_{3,........}$ x_T }. Here we first compute the probabilistic alignment of the training vectors into the prior mixture components. We compute $\mathrm{Pr}\big(\mathrm{i}\big|\mathrm{x_t},\lambda_{\mathrm{prior}}\big)$ as in Equation (12). Then we compute the sufficient statistics for the weight, mean and variance parameters as follows.

$$\mathrm{n_i} = \textstyle\sum_{t=1}^{T} \mathrm{Pr}\big(\mathrm{i}\big|\mathrm{x_t},\lambda_{\mathrm{prior}}\big) \quad \text{weight} \qquad (13)$$

$$\mathrm{E_i(x)} = \frac{1}{\mathrm{n_i}}\textstyle\sum_{t=1}^{T} \mathrm{Pr}\big(\mathrm{i}\big|\mathrm{x_t},\lambda_{\mathrm{prior}}\big)\mathrm{x_t} \quad \text{mean} \qquad (14)$$

$$\mathrm{E_i}(x^2) = \frac{1}{\mathrm{n_i}}\textstyle\sum_{t=1}^{T} \mathrm{Pr}\big(\mathrm{i}\big|\mathrm{x_t},\lambda_{\mathrm{prior}}\big)\mathrm{x_t^2} \ \text{variance} \quad (15)$$

Next, the new sufficient statistics from training data are used to update the prior sufficient statistics for mixture i to create the adapted parameters for mixture i. with the following equations:

Adapted mixture weight,w_i' = [$a_i^w \mathrm{n_i}$ /T +(1- a_i^w)$\mathrm{w_i}$]š (16)

Adapted mixture mean $\mu_i' = a_i^m E_i(\mathrm{x}) + (1 - a_i^m)\,\mu_i$ (17)
Adapted mixture variance

$$\sigma_i'^2 = a_i^v E_i(x^2) + (1 - a_i^v)(\sigma_i^2 + \mu_i^2) - \mu_i'^2 \quad (18)$$

The adaptation coefficients controlling the balance between old and new estimates are $\{a_i^w, a_i^m, a_i^v\}$ for the weight, means and variances, respectively. The scale factor š, is computed over all adapted mixture weights to ensure they sum to unity.
For each mixture and each parameters, a data-dependent adaptation coefficient $a_i^{\tilde{n}}$, ñ € {w,m,v} ,is used in the above equation defined as

$$a_i^{\tilde{n}} = \frac{\mathrm{n_i}}{\mathrm{n_i} + \mathrm{r}^{\tilde{n}}}, \quad (19)$$

Where $\mathrm{r}^{\tilde{n}}$ is a fixed “relevance” factor for parameter ñ
.
It is common in speaker recognition application to use one adaptation coefficient for all parameters $(a_i^w = a_i^m = a_i^v = \mathrm{n_i}/(\mathrm{n_i} + \mathrm{r}))$ and further to only adapt certain GMM parameters such as the mean vectors.

(iii) Universal Background Models (UBM)

A UBM or World Model is a model in a speaker verification system to represent general, person-independent, channel independent feature characteristics to be compared against a model of speaker-specific feature characteristics when making an accept or reject decision. Here, the UBM is a speaker-independent GMM trained with speech samples from a large set of speakers to represent general speech characteristics. The UBM also use when training the speaker-specific model by acting as a prior model in MAP parameter estimation. In state-of-the-art speaker verification system the UBM is used for modeling the alternative hypothesis in the likelihood ratio test. Assuming that a GMM distribution best represent the distribution of feature vectors for hypothesis H_0 so that λ_p denoting the weight, means and covariance matrix parameters of a GMM. The alternative hypothesis H_1 is likewise represented by a model $\lambda_{p'}$. The likelihood ratio statistic is then defined as

$$\mathrm{LR(X)} = \frac{\mathrm{p(X|\,\lambda p)}}{\mathrm{p(X|\,\lambda p')}} \quad (20)$$

For given a set of N background speaker models { $\lambda_1, \lambda_2, \lambda_3, \ldots\ldots\ \lambda_N$ } then the alternative hypothesis is represented by

$$\mathrm{p(X|\,\lambda p') = F(p(X|\lambda_1)\ p(X|\,\lambda_2) \ldots \ldots p(X|\lambda_N))} \quad (21)$$

Where F() is some function, such as average or maximum, of the likelihood values from the background speaker set.
Typically, GMMs are used for distribution models and a speaker specific model is derived by using MAP estimation with the UBM acting as the prior model. In GMM-UBM system we use a single, speaker-independent background model to represent $\mathrm{p(X|\,\lambda p')}$.

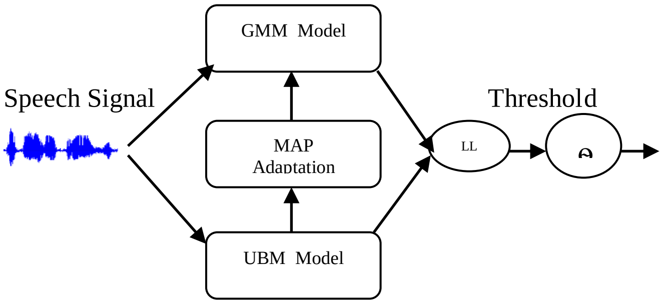

Figure 1. Training module of GMM-UBM for SV

The theory explains for determining the statistic from a single feature vector observation sample from the target or non-target speaker classes. This test statistic deals with two speaker classes identified as the target speaker and non-target (UBM) speaker set specified by models, $\lambda_{\mathrm{target}}$ and λ_{ubm} . For a given T independent and identically distributed observations, X= {$x_1, x_2, x_3, \ldots\ldots\ x_T$ }. The joint likelihood ratio may be determined. A more robust of measure for speaker verification is the expected frame-based log-likelihood ratio measure can be defined as follows.

$$\mathrm{E\,[LLR(x)] = E\,[\log p(x|\lambda_{target}) - \log p(x|\lambda_{ubm})\,]}$$
$$= \frac{1}{T}\sum_{t=1}^{T}(\log p(x_t\,|\lambda_{\mathrm{target}}) - \log p(x_t\,|\lambda_{\mathrm{ubm}})) \quad (22)$$

The UBM is a large GMM (1024 mixtures) trained to represent the speaker-independent distribution of features. The simplest approach is to train the UBM is merely pool all the speech data from the equal number of male and female speakers utilizing through the EM algorithm.
MAP adaptation integrates coupled target and background speaker model components is an effective way of performing speaker recognition. A significant advantage of a fully coupled system is that the coupling enables discrimination between regions of space that the GMM has learned from training speech. The mixture component will remain un-adapted, if there is no adaptation observation in the region nearby a mixture component. But due to applying adaptation, mixture components near training observation will be adjusted towards the speech data. As a result adapted regions will be more discriminative [13].

C. Support Vector Machine (SVM)

A support vector machine (SVM) is a versatile discriminative classifier that has gained considerable popularity in recent years that adopted in speaker recognition

[5]. It is a two-class discrimination technique which involves finding a hyperplane for effective separation of the two classes considered. An SVM is a discriminative model which determined the boundary between a speaker and a set of imposters and which has been applied with spectral, prosodic and high-level features vectors. SVM can also be successfully combined with GMM to get better performance. The typical methods employed in SVM speaker recognition is based upon comparing speech utterances using sequence kernels. In this case we train a target model with the target speaker utterances as well as a set of background speaker's utterances which have the characteristics of impostor population. Each speech sample from a target or background speaker becomes a point in the SVM space.

An SVM is a two-class classifier constructed from sums of a kernel function K**(. , .)**

$$F(x) = \sum_{i=1}^{N} \lambda_i \, t_i \, K(x, x_i) + d \tag{23}$$

Where the t_i are ideal outputs, $\sum_{i=1}^{N} \lambda_i \, t_i = 0$ and $\lambda_i > 0$.

The vectors x_i are support vectors and obtained from the training set by an optimization process. The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector is in class 0 or class 1, respectively. For classification, a class decision is based upon whether the value, F(x), is above or below a threshold.

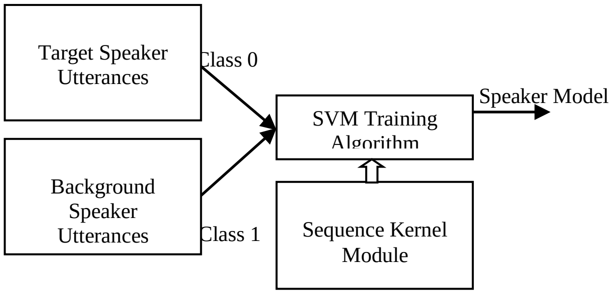

Figure 2. Training module of SVM for SV

The kernel **K(. , .)** is constrained to have certain properties (the Mercer condition), so that **K(. , .)** can be expressed as

$$K(x, y) = b(x)^t \, b(y) \tag{24}$$

where b(x) is a mapping from the input space (where x lives) to a possibly infinite dimensional space. The kernel is required to be positive semi-definite. The Mercer condition ensures that the margin concept is valid, and the optimization of the SVM is bounded.

The focus, then, of the SVM training process is to model the boundary, as opposed to a traditional GMM-UBM which would model the probability distributions of the two classes.

D. Joint Factor Analysis (JFA)

In state of the art modeling technique of speaker recognition, speaker variability is assumed to be of primary importance but it has long been recognized that session variability as well as channel variability is a serious problem. The speaker factors and the channel factors as well as session factors play different roles for a given speaker, the values of the speaker factors are assumed to be the same for all recordings of the speaker but the channel factors and session factors are assumed to vary from one recording to another.

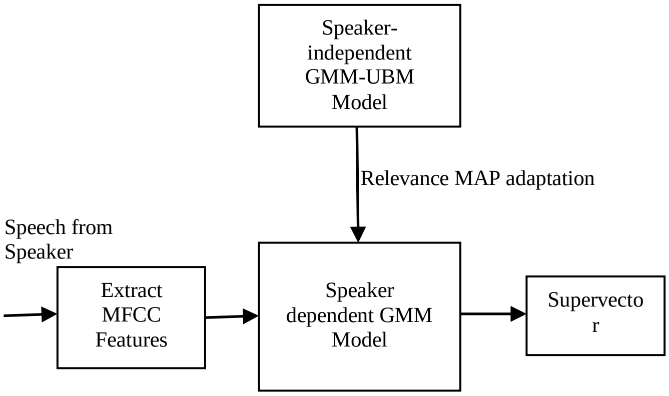

Figure 3. Supervector for speakers in JFA.

The joint factor analysis model is quite similar to feature mapping. The basic assumption is that each speaker- and channel-dependent supervector is a sum of a speaker-dependent supervector and a channel-dependent supervector [14]. A supervector for a speaker should be decomposable into speaker independent, speaker dependent, channel dependent and residual components. Supervectors consist of the speaker-dependent GMM mean components. Relevance MAP adaptation is a linear interpolation of all mixture components of UBM to increase likelihood of speech from particular speaker.

JFA model to produce a model of speaker and session variability originally formulated by Kenny in 2005 which can be integrated with standard models of speaker variability, namely classical MAP and eigenvoiceMAP [4,15,16]. The factor analysis model combines the priors underlying classical MAP, eigenvoiceMAP and eigenchannel MAP.

Let M(S) be the speaker supervector for a speaker S and let m denote the speaker- and channel-independent supervector. (The simplest way to estimate m is to take the supervector from a Universal Background Model (UBM).) In classical

MAP it is assumed that, for a randomly chosen speaker S, M(S) is normally distributed with mean m and a diagonal covariance matrix d^2. It is convenient to describe this prior in terms of hidden variables as follows:

$$M(S) = m + dZ(S) \qquad (26)$$

Where $Z(S)$ is a hidden vector distributed according to the standard normal density, N(Z|0, I).It is easily seen that, under this assumption, the expectation of M(S) is m and its covariance is d^2.
The major difference is that the factor analysis model treats the channel space as a continuum whereas in channel effects are quantized so that there is a discrete set of channel supervectors. For this approach, the second question above presents no particular difficulty since it can be tackled by applying the appropriate type of channel compensation in enrolment as well as in testing [20].

III. Experimental Setup

In the experiment of SV system a simple baseline system has been developed. In this case, the SV system has been developed using different speaker modeling techniques namely VQ, traditional GMM, GMM-UBM, SVM, JFA with GMM-UBM. The coefficients were extracted from a speech sample at 16 KHz with 16 bits/sample resolution and frame rate 100 Hz with frame size 30 msec. A pre-emphasis factor of 0.97 has been applied. The filterbank used in deriving the cepstral coefficients consists of 23 triangular filters and was constrained into a frequency band of 300-3200 Hz. A 39-dimensional feature vector has been used, made up of 13 MFCC and their first order and 2nd order derivatives. On the other hand, six dimensional prosodic features make the total of 45 dimensional hybrid feature vectors. Cepstral Mean Subtraction (CMS) and has been applied on all features to reduce the effect of channel mismatch. The speech data for all the experiments that have been carried using ALS-DB database [17,18,19].

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with speaker's data with Expectation Maximization (EM) algorithm and finding the average of all these models. The target speaker models were created by adapting only mean parameters of the UBM using maximum a posteriori (MAP) algorithm approach with the speaker specific data.

IV. Performance Evaluation

The performance measures of the baseline system in the experiment are same as the metric used in the 2010 NIST Speaker Recognition Evaluation plan. The primary SV performance metrics are FAR (false alarm rate) and MDR (miss detection rate). A popular one is the equal error rate (EER) which corresponds to the operating point where FAR = MDR Graphically, it corresponds to the intersection of the detection error trade-off (DET) curve with the first bisector curve.

The cost function is defined as a weighted sum of miss detection and false alarm probabilities. According to the NIST Detection Cost Function (DCF) can be defined as

$$C_{DET} = (C_{Miss} * P_{Miss|Target} * P_{Target}) + (C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * P_{NonTarget}) \qquad (27)$$

Where $P_{Miss|Target}$ and $P_{FalseAlarm|NonTarget}$ are the miss (false rejection) probability and the false alarm (false acceptance) probability respectively.

Parameter Values are

Cost of miss	$C_{Miss} = 10$
Cost of a false alarm	$C_{FalseAlarm} = 1$
Probability of a target	$P_{Target} = 0.01$
Probability of a non-target	$P_{NonTarget} = 1 - P_{Target} = 0.99$.

V. Results Analysis

DET curves showing the performance of spectral with prosodic based SV system with respect to different speaker modeling techniques.

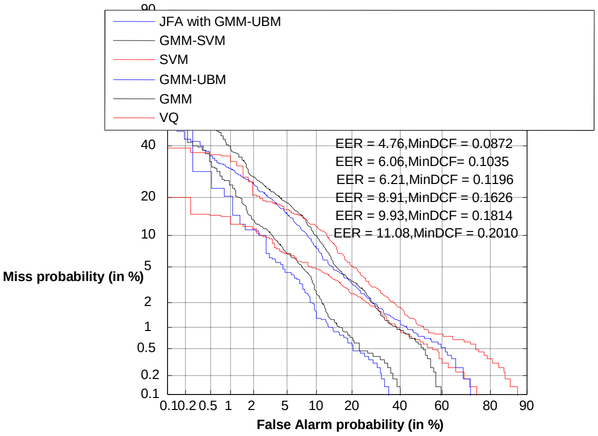

Figure 4. DET curves for the SV system using MFCC and Prosodic Features with different speaker modeling techniques.

TABLE1. EER and MinDCF values for the SV system using Mfcc and Prosodic Features with Different Speaker Modeling Techniques.

Speaker Modeling Techniques	EER%	Recognition Rate%	MinDCF Values
VQ	11.08	88.92	0.2010
GMM	9.93	90.07	0.1814
GMM-UBM	8.91	91.09	0.1626
SVM	6.21	93.79	0.1196
GMM-SVM	6.06	93.94	0.1035
JFA with GMM-UBM	4.76	95.34	0.0872

From the experimental point of view we observe that the best performance of SV system shows by JFA with GMM-UBM modeling technique with its EER value of **4.76**% and MinDCF value of **0.0872.** Comparing with other modeling techniques VQ shows it's the poorest performance with its EER value of **11.08**% and MinDCF value of **0**.**2010**. On the other hand fusions of GMM with SVM enhances the performance of **0.7%** that of single SVM technique. Similarly GMM-UBM shows better performance than traditional GMM and SVM shows of approximately of **2.8%** improvement with comparison to that of GMM-UBM. Above all, SVM shows its performance **3.72%** more improvement than that of its correspondence modeling techniques GMM.

VI. Conclusions AnD Future Research Scope

In this paper we have discussed a brief overview of the speaker modeling techniques in SV system using spectral and prosodic features. From the above studies, we conclude that the fusion modelling technique of GMM-UBM with JFA shows the best performance with **95.34%** and VQ shows the very poor performance with **88.92%** verification rate. Here, finally, we conclude that the fusions of both generative and discriminative models highly improve the performance of SV system. Speaker modeling continues to be a strong component of the SV problem. State-of art methods which deal with channel variability, phonetic mismatch, session variability will no doubt lead to significant improvements. In future, SV will continue to exploit more advance in speech processing area which continues to give innovative ideas into feature that characterize speakers in speaker dialect, speaker idiolect as well as vocal and articulator characteristics.

References

[1] F. Bimbot, et. al., "*A tutorial on text-independent speaker verification*," EURASIP Journ. on Applied Signal Processing, pp. 430-451, 2004.

[2] D.A.Reynolds, "*An overview of automatic speaker recognition technology*",. In: ICASSP, IEEE international conference on acoustics, speech and signal processing, vol 4, pp 4072–4075, 2002.

[3] L. Mary and B.Yegnanarayana, "*Extraction and representation of prosodic features for language and speaker recognition*",Speech communication,pp.782-796, 2008.

[4] D.A.Reynolds, T.F.Quateri and R.B. Dunn, "*Speaker verification using adapted Gaussian mixture models",* In Digital Signal Processing, Vol.10, pp.19-41, 2000.

[5] T.Kinnunen and H. Li, " *An overview of Text-independent Speaker Recognition: from Features to Supervectors*",Speech Communication,pp. 12-40**,** 2010

[6] K. Sarmah and U. Bhattacharjee, "*Speaker Modeling Distance Normalization Technique in Multilingual Speaker Verification*", International Journal of Electrical and Electronics Engineering Research, Vol.3, Issue-2, pp.319-326, 2013.

[7] K. Sarmah and U. Bhattacharjee, "*Improvement of Speaker Verification System with Feature Level and Score Level Normalization Techniques*", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol.2, Issue 2, pp. 3119-3126, 2014.

[8] K. Sarmah and U. Bhattacharjee, "*Text-independent multi-sensor speaker verification system", International Journal of Computer Science and Engineering,* Vol. 4, Issue 5,pp.7-16, 2015.

[9] D.A. Reynolds, "*Speaker identification and verification using Gaussian mixture speaker models*," Speech Communication, vol.17, pp. 91-108,1995.

[10] N.Malayath, , H. Hermansky, S.Kajarekar, and B.Yegananarayan, "*Data –driven temporal filters and alternatives to GMM in speaker verification*", In Digital Signal Processing, pp.55-74, 2000.

[11] D.A.Reynolds, "*Gaussian Mixture Models*". In Encyclopedia of Biometric Recognition, Springer, Journal Article, 2008.

[12] A. Fazel, and S.Chakrabartty, "*An overview of Statistical Pattern Recognition Techniques for Speaker Verification*",. In IEEE CIRCUITS AND SYSTEMS MAGAZINE. 2011.

[13] J. Pelecanos, R.Vogt and S.Sridharan, "*A study on standard and iterative MAP adaptation for speaker recognition*",. In Proceeding on the 9th Australian International Conference on Speech Science & Technology Melbourne, December 2 to 5, 2002.

[14] W. Campbell, J. Campbell, D.A. Reynolds, E. Singer, and P.Torres-Carrasquillo, "*Support vector machines for speaker and language recognition",*. Computer Speech and Language 20, pp.210–229, 2006.

[15] P.Kenny, "*Joint factor analysis of speaker and session variability: Theory and algorithms,* Tech. Report CRIM-06/08-13, 2005.

[16] P. Kenny and P.Dumouchel, "*Experiments in speaker verification using factor analysis likelihood ratios*," in Proc. Odyssey04, pp. 219-226, 2004.

[17] U. Bhattacharjee and K. Sarmah, "*A Multilingual Speech Database for Speaker Recognition*", Proc. IEEE, ISPCC, 2012.

[18] U. Bhattacharjee and K. Sarmah, "*GMM-UBM Based Speaker Verification in Multilingual Environments*",International Journal of Computer Science Issues.Vol. 9,Issue 6,No.2, pp.373-380,2012.

[19] U. Bhattacharjee and K. Sarmah, "*Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment*",International Journal of Soft Computing and Engineering. Vol.2, Issue-6, pp. 443-446, 2013.

[20] D.A.Reynolds, et..al, "*The SuperSID project: exploiting high-level information for high-accuracy speaker recognition*". In Proc. Int.

Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003 (Hong Kong, China, pp. 784–787, April 2003.

Authors Profile

Dr. Kshirod Sarmah pursed Master of Science in Computer Science from Gauhati University, Assam, India in 2003 and Ph.D. in Computer Science and Engineering from Rajiv Gandhi University (A Central University),Itanagar, India in the year 2015. He is currently working as HOD and Assistant Professor in the Department of Computer Science at Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Science Model College), Goalpara, Assam, India. He has published more than 10 research papers in reputed UCG approved international journals and IEEE international conferences. His main research work focuses on Speech Processing, Speaker Verification System. He has 10 years of teaching experience and 5 years of Research Experience.