A Language Dependent Speaker Verification System Using Different Normalization Techniques

Dr. Kshirod Sarmah[1], Dr. Swapnanil Gogoi[2], Prof. Utpal Bhattacharjee[3]

[1] *Department of Computer Science, PDUAM Amjonga, Goalpara, India*
[2] *Department of Computer Science, Gauhati University IDOL, Guwahati, India*
[3] *Department of Computer Science and Engineering, Rajiv Gandhi University, Arunachal Pradesh, India*

E-mail: kshirodsarmah@gmail.com, swapnanil@gauhati.ac.in, utpalbhattacharjee@rediffmail.com

Abstract— At present most of the state-of-the-art speaker verification systems extract speaker dependent features from the short-term spectral information which ignores long-term information that can convey supra-segmental information like prosodic and speaking style. Many powerful methods for speaker verification have been introduced now a day's such as high-level features, novel classifiers and channel compensation techniques. Although different level normalization techniques like feature level, score level and modeling level had been applied to improve the performance but still it is not sufficient as expected a real time SV system. In this paper, we propose a hybrid approach language dependent speaker verification system. We first constructed acoustic-prosodic features based language model for each language using language specific speech data using GMM modeling techniques by applying Expectation Maximization (EM) algorithm. The result reported here, have been evaluated using speech data from the multilingual speech database, namely Arunachali Language Speech Database (ALS-DB).GMM-UBM modeling technique has been applied for constructing target speaker models in the SV system. The performance of the speaker verification (SV) system using proposed approach improves up to 4.18% and 4.80% of EER levels for language matching and mismatching conditions respectively. In this case approximately 1.00% and 3.70% improvements has been observed by applying the LID module with T-Norm and D-Norm in the SV system than that of the traditional without deploying LID module for language matching and mismatching conditions respectively.

Keywords— Multilingual Speaker Verification, Language Identification, MFCC, Prosodic, GMM-UBM,T-Norm, D-Norm

I. INTRODUCTION

In recent time, different innovative techniques have been applied to prove effective for text-independent speaker recognition and have reduced error rates dramatically. These techniques have seen some of the main issues in speaker recognition: speaker modeling, feature level solution, channel compensation and multiple linguistic cues to speaker identity. In this paper, our main aim is to discuss a successful fusion techniques and tradeoffs in complexity in providing low error rates. In the framework of speaker verification system, the most popular speaker modeling technique is based on Gaussian Mixture Models with short term cepstral features [1]. The most obvious advantages of GMMs are their simplicity and robustness to short-length recordings. These characteristics reflect the model's assumption that every 10–20 msec frame of speech can be considered independently. This works well when the test recording is only a few seconds long. However, as the amount of test and training data increases it becomes attractive to make use of speaker-specific characteristics which involve larger time scales, such as prosodic patterns and high level features.

Characteristics of a speaker can be represented using short term and long term features [2]. Short-time features are capable of reflecting the physiological difference among the speakers. The long term features mostly represent the habitual attributes of a speaker such as *prosody* and *idiolect* [3]. *Prosody* is a term used for representing characteristics such as intonation, timing and stress in a collective manner which is less sensitive to channel effects than cepstral features [9]. Prosodic systems are especially effective when large amounts of data are available to train speaker models [7] [8]. The complementary nature of the prosodic and spectral features helps to improve the overall performance of speaker verification, while combining the evidences [3].

Prosodic information in speech signal having mainly three categories: *Linguistic*, *paralinguistic* and *non-linguistic*. Linguistic information includes lexical stress, sentence modality, focus structure and segmentation; the paralinguistic information comprises speaker attitude, intention, dialect and sociolect. The non-linguistic deals with emotions and health [10]. It is not an easy task to extract properly feature vectors related with linguistic, paralinguistic, and non-linguistic for robust speaker verification performance in multilingual environments.

One important area of improvement in speaker verification has been in direct modeling of the spectral content of speech. Two most significant innovations have been the introduction of discriminative techniques like support vector machine (SVMs) that fused well with standard

Gaussian mixture models (GMMs) [4] and advanced channel compensation methods such as latent factor analysis (LFA) as well as nuisance attribute projection (NAP) have been significantly reduced error rates by supervised modeling of channel and session variation [5]. A disadvantage of the approaches that evolved from of multiple cepstral and high-level systems is that many systems were run independently without regard for system complexity [6]. On the other hand, focusing on many high-level systems creates a system which may be vulnerable to multilingual speaker verification system. In this paper, a two stages solution for the speaker verification in multilingual system has been designed with an objective to improve the performance of the baseline system in multilingual environment.

The first stage of the system is a language identification (LID) system that identifies the language spoken by the speaker. The second stage contains a bundle of language dependent speaker models for each speaker. Once the LID system identifies the language spoken, it will redirect the verification task to that particular language dependent model of the claimed speaker which belongs to the identified language. To reduce the overhead in feature extraction part, the same features has been used for both language identification as well as speaker verification.

II. ARUNACHALI LANGUAGE SPEECH DATABASE

The main purpose of building speech corpora containing recording from a large number of speakers from phonologically distinct dialect areas. Data should be collected from each speaker in different languages for cross-lingual, multilingual and multichannel experiments like speaker recognitions and language identifications. Recently, linguistic data consortium for Indian language (LDC-IL) has taken up an initiative to build speech corpora for all Indian language in the area of speech recognition [26]. To make efficient research in future and to improve the ASR system's capability, it is needed to work including in Indian context and collect multilingual or multi-dialect and multichannel corpora.IIT Guwahati has already taken initiative in building a multi-device, multi-lingual and multi-environment speech database for speaker recognition tasks by covering more than 10 Indian major languages, which is referred to as the IITG Multi-Variability (IITG-MV) speaker recognition database [27]. Similar approach has been made by IIT Kharagpur in order to develop a multilingual Indian language speech corpus namely Indian Institute of Technology Kharagpur-MutliLingual Indian Speech Corpus (IITKGP-MLILSC) that covers 27 Indian languages and dialects of all over India [28]. Arunachal Pradesh which is one of the most linguistically rich state in North-eastern India having different sub languages belonging to two important language family namely Tibeto-Burman and Indo-European which is still not included in any speech database. So, in this work, we developed a multilingual speech database for the Arunachali languages including Hindi and English language.

In this section a multilingual and multichannel speech database has been described namely Arunachali Language Speech Database (ALS-DB) that recently collected in Arunachal Pradesh [24][25]. To study the impact of language variability, impact of channel variability on speaker verification and language identification in multilingual environments, ALS-DB is collected in multilingual and multichannel environments. Each speaker is recorded for three different languages – English, Hindi and a local language, which belongs to any one of the four major Arunachali languages - Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration.

The speakers are recorded for reading style of conversation. The speech data collection was done in laboratory environment with air conditioner, server and other equipment's switched on. The speech data was contributed by 100 male and 100 female informants chosen from the age group 20-50 years. During recording, the subject was asked to read a story from the school book of duration 4-5 minutes in each language for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

III. CEPSTRAL SYSTEM AND FRONTETHODOLOGY

In a robust Speaker Verification system, it has been observed that the features that depends not only the speaker specific but also that of the specific language has enhanced its performance. It is the most essential task to find the language dependent features in speaker verification system in multilingual environments.

A. Spectral Feature

In the earlier Language Identification (LID) system, researchers heavily focus on the spectral contents of the language. The basic idea behind it is that different languages contain different phonemes and phones. A set of short-term spectra is obtained from the training utterances of each language and these prototypes are compared to the ones obtained from the test speech.

B. Language Dependent Prosody Features

The general conception that prosodic features do not help in language identification (LID) is often the consequence of an oversimplified implementation in feature extraction. Muthusamy indicates the feasibility for prosodic features to contribute to LID [22]. It is shown in recent studies by Mary and Yegnanarayana that prosodic features alone can help in an LID task [23].

Pitch frequency (fundamental frequency) of speech is defined as the frequency at which the vocal cords vibrate

during a voiced sound. Fundamental frequency (f_0) is usually processed on a logarithmic scale rather than a linear one in order to match the resolution of human auditory system. Normally, f_0 is in between the range from 50 Hz to 500 Hz for voiced speech. For unvoiced speech f_0 is undefined and by convention, it is zero in log scale.

Since the fundamental frequency implies the characteristics of the speaker, it does not give global information about the language or the utterance. The slope of the pitch frequency, however, gives some clues about the prosody and the stress on the utterance, which might differ from language to language.

The cepstral-based system used a common set of speech activity detection marks from a GMM- based voice activity detection (VAD) system. Two sets of features are used for verification – MFCCs and Prosodic features. For MFCCs, 13 cepstral coefficients and deltas as well as double deltas were computed to produce a 39 dimensional feature vectors. The feature vectors stream is processed through VAD to eliminate non-speech as well as low energy based vectors. CMS, CVN are then applied to the feature stream.

For Prosodic based feature processing, a total 6 dimensional features vector consist of pitch, short time energy and its first and second order derivatives (Δpitch, Δenergy, $\Delta\Delta$pitch and $\Delta\Delta$energy). For Prosodic based feature extraction, windowing, frame size, frame rate, VAD, CMS and CVN performed in the same manner as for MFCCs.

Finally, we have concatenated 39 dimensional MFCCs with 6 dimensional Prosodic features to get 45 dimensional language dependent feature vectors.

IV. SCORE-NORMALIZATION TECHNIQUES

Score normalization techniques have been mainly derived from the study of Li and Porter [21].It is observed that a large variances from both distributions of client scores (intra-speaker scores) and impostor scores (inter-speaker scores) during speaker verification tests. Based on these observations, a solution has been proposed that based on impostor score distribution in order to reduce the overall score distribution variance (both client and impostor distributions) of the speaker verification system. The basic of the normalization technique is to center the impostor score distribution by applying normalization on each score generated by the speaker verification system.

Let $S_\lambda(X)$ denote the score for speech signal X and speaker model λ. The normalized score $N_\lambda(X)$ is then given as follows:

$$N_\lambda(X) = (S_\lambda(X) - \mu_\lambda)/\sigma_\lambda$$

Where μ_λ and σ_λ are the normalization parameters for speaker λ. Those parameters need to be estimated from the imposter distribution. Various kinds of score normalization techniques have been proposed in the literature. In this case we only use two techniques namely Test Normalization (T-Norm) at score level and Distance Normalization (D-Norm) at model level.

A. *Test-Normalization (T-Norm)*

T-Norm is one of the most popular score normalization technique. It is done by comparing the incoming test signal with claimed speaker model as well as with a set of impostor models to estimate impostor score distribution and normalization parameters are estimated from these scores. If Zero Normalization (Z-Norm) is considered as a speaker-dependent normalization technique, T-Norm is a test-dependent one. As the same test utterance is used during both testing and normalization parameter estimate, T-Norm avoids a possible issue of Z-Norm based on a possible mismatch between test and normalization utterances. Conversely, T-Norm has to be performed online during testing.

B. *Model Distance Normalization (D-Norm)*

Speaker modeling distance normalization (D-Norm) is one of the important model level normalization techniques in speaker verification (SV) system which is based on the Kullback-Leibler **(**KL)-divergence that commonly used in statistics as a measure of similarity between two density distributions. For D-Norm implementation, it doesn't need any additional speech data or external speaker population. It is because D-Norm has reduced computational time which is one of the essential advantages of D-Norm.

.

V. GMM-UBM AS SPEAKER MODELING TECHNIQUE

Over the last decade, the Gaussian Mixture model GMM [11] has become established as the standard classifier for text-independent speaker recognition. Gaussian Mixture model (GMM) often to be used to the speaker verification because this mode has good ability of recognition [12]. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped distributions [13].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 [14].

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 [13].

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

$$P(x|\lambda) = \sum_{i=1}^{M} w_i\, b_i(x) \qquad (2)$$

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

$$b_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\sum_i^{-1}(x-\mu_i)\right\} \quad (3)$$

With mean vector $\mu_i$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:

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

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:

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

Once a model is trained then (3) can be used to compute the log-likelihood of model λ for an input test set of feature vector X can be defined as

$$log\ p(X|\lambda) = \sum_{i=1}^{T} log\ p(x_i|\lambda) \quad (7)$$

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.

A. Maximum Likelihood Parameter Estimation

For a given training vectors and a GMM configuration, we have to estimate the parameters of the GMM, λ that 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

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda). \quad (8)$$

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 $P(X|\lambda') \geq P(X|\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached [13].

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

Mixture Weights :

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

Means :

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

Variance (diagonal covariance):

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

The a posteriori probability for component *i* is given by

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

B. 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 the probabilistic alignment of the training vectors into the prior mixture components are computed first. For mixture *i* in the prior model, $Pr(i|x_t,\lambda_{prior})$is computed as in Equation (13). Then the sufficient statistics for the

weight, mean and variance parameters are computed as follows.

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

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

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

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

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

Adapted mixture variance, $\sigma_i^{'2} = a_i^v E_i(x^2) + (1-a_i^v)(\sigma_i^2+\mu_i^2) - \mu_i^{'2}$ (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 $a_i^{\tilde{n}}$, $\tilde{n} \in \{w,m,v\}$,is used in the above equation defined as

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

Where $r^{\tilde{n}}$is a fixed "relevance" factor for parameter .

It is common in speaker recognition application to use one adaptation coefficient for all parameters $(a_i^w = a_i^m = a_i^v = n_i/(n_i + r))$ and to adapt only certain GMM parameter such as the mean vectors. There are lots of reasons to consider in contrasting one of the standard MAP approaches to its iterative form. The standard MAP technique is simply a single iteration while EM based result is iterative. A single iteration assumes that the mixture mean components vary in a completely independent manner [17], and consequently, only a single iteration would be required to solve the MAP solution.

Since the environment and even the speaker's voice characteristics may change over time, one can adapt the model for *P***,** when it is sure that the current speaker is *P*. Maximum a posteriori probability (MAP) adaptation combined with confidence weighting improved authentication performance under channel mismatch conditions by 61%, despite impostor attacks [18].

C. Universal Background Models (UBM)

A Universal Background Model (UBM) is one of the most popular models used in a biometric verification system [19]. 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. In speaker recognition 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 act as a prior model in the training of speaker-specific 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 [19].

$$LR(X) = \frac{p(X|\lambda_p)}{p(X|\lambda_{p'})} \qquad (20)$$

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

$$p(X|\lambda p') = f(p(X|\lambda_1)\, p(X|\lambda_2) \ldots\ldots p(X|\lambda_N)) \qquad (21)$$

Where *f(.)* is some function, such as average or maximum, of the likelihood values from the background speaker set.

In GMM-UBM system we use a single, speaker-independent background model to represent $p(X|\lambda p')$.That is, for a single feature vector observation, the statistics for the two speaker classes: the target and non-target are specified by the models λ_{target} and λ_{ubm} respectively.

For a *T* independent and identically distributed observations $X = \{x_1, x_2, x_3, \ldots\ldots x_T\}$.like frame-based observation, the joint likelihood ratio may be determined as follows:

$$E[LLR(x)] = E[\log p(x|\lambda_{target}) - \log p(x|\lambda_{ubm})] \qquad (22)$$

$$E[LLR(x)] = \frac{1}{T} \sum_{t=1}^{T} (log\ p(x_t|\lambda_{target}) - log\ p(x_t|\lambda_{ubm})) \qquad (23)$$

The UBM is a large GMM (1024 mixtures) trained to represent the speaker-independent distribution of features. To train a UBM, the simplest approach is to merely pool all the data and use it to train the UBM via the EM algorithm.

D. Results of GMM-UBM based MSV system

The following Figure 2. Shows the performance of the language dependent speaker verification system using

MFCC with the same Prosodic features with the combined normalization techniques namely T-Norm and D-Norm.

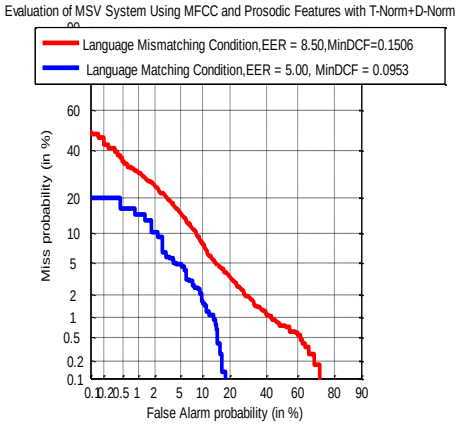

Figure 2.DET curves for the multilingual speaker verification system using MFCC and Prosodic Features with T-Norm + D-Norm for language matching and mismatching conditions.

Table 1. ERR values and MinDCF values of SV System Using MFCC and Prosodics as Feature Vectors

Training and Testing Language	***Feature Vectors***	***Normalization Techniques***	***EER%***	***Recognition Rate%***	***Minimum DCF Values***
Language Matching Condition	MFCC+ Prosodic Features	T-Norm+D-Norm	5.00	95.00	0.0953
Language Mismatching Condition	MFCC+ Prosodic Features	T-Norm+D-Norm	8.50	91.50	0.1506

From the above experiments, it has been observed that when T-Norm and D-Norm applied together, the performance of the SV system improves for both language matching and mismatching conditions. The performance of the baseline system is improved approximately **3.5%** in case of language matching condition that of language mismatching conditions as its EER value reduced from **8.50** to **5.00**.

VI. BASELINE SYSTEM FOR LANGUAGE IDENTIFICATION

The Language Identification (LID) system has been developed using Gaussian Mixture Model based modeling approach utilizing the same cepstral systems discussed in the Section 3.

The Gaussian mixture model with 1024 Gaussian components has been used for constructing language models. The individual language models were trained using the Expectation Maximization (EM) algorithm. Each language model has been trained with equal number of male and female speakers' data. Total 50 speakers' data has been considered for construction of a language model. Data from a single session has been considered for model construction and rest of the sessions have been considered for generation the test segments.

The detection error trade-off (DET) curve has been plotted using log likelihood ratio. The equal error rate (EER) obtained from the DET curve has been used as a measure for the performance of the LID system.

A. Experiments on LID System

In this case, preliminary language identification system has been developed to evaluate the performance of LID system using single language training and testing.
The performance of the baseline system for LID system has been given below in terms of EER values.

B. Results of GMM based LID System

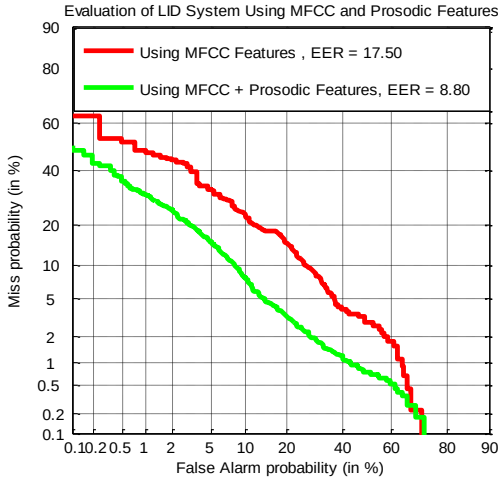

Figure 3. DET curves for the language identification (LID) system using MFCC and Prosodic Features.

Table 2. ERR values of Language Identification Using MFCC and Prosodic as Features Vectors.

Feature Vectors of LID system	***EER%***	***Recognition Rate***
MFCC	17.50	82.50
MFCC + Prosodic	8.80	91.20

The main observation that has been drawn from the experiment part is that prosodic features can significantly improve the performance of a language identification system. Combining prosodic features pitch and intensity along with MFCC features, it has been observed that there is a sharp improvement of performance by **11%**. Important language dependent feature duration has not been considered in the present study as a limitation to the feature combination.

However, there is scope of further improvement in the performance of the LID system by considering all the prosodic features. To overcome the problem of feature combination, model combination approach can be used.

VII. LANGUAGE DEPENDENT SPEAKER VERIFICATION SYSTEM

The proposed structure of the baseline system for the speaker verification system can be divided into phases: First phase is responsible for identifying the particular language of that particular speaker speaking and the second phase is the speaker verification phase based on that LID module and final take decision whether the particular speaker is accepted or rejected. The baseline system has been developed using Gaussian Mixture Model with Universal Background Model (GMM-UBM) modeling approach. Here we first constructed acoustic-prosodic features based language model for each language using language specific speech data using GMM modeling techniques by applying Expectation Maximization (EM) algorithm. A set of speaker models – one for each language would be used for each speaker. The language identification (LID) module is used to identify the language of the test utterance. The same feature vectors both MFCC and Prosodic has been utilized for the evaluation of the language identification system and accordingly the speaker verification phase have been done using particular language dependent speaker model for both language matching and mismatching conditions.

Training and Testing Language	***Feature Vectors***	***EER%***	***Recognition Rate%***	***Min. DCF Values***
Languages Matching Condition	MFCC and Prosodic	4.18	95.82	0.0725
Languages Mismatching Condition	MFCC and Prosodic	4.80	95.20	0.0863

A. Experiment for the language dependent speaker verification system

In this experiment for Language dependent Speaker Verification system using LID module as the initial phase of the system to identify the language first and then in the second phase used LID module to SV system. Also different normalization techniques and combined them to get better performance of the baseline system. In this case, we apply the T-Norm and D-Norm which gives the best performance.

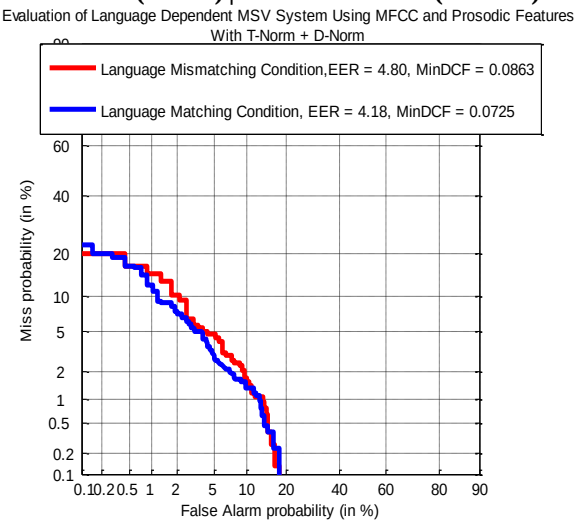

Figure 5. DET curves for the Language Dependent Speaker Verification System using MFCC and Prosodic Features with T-Norm + D-Norm for language matching and mismatching conditions.

Table 3. ERR values and MinDCF values for Language Dependent SV System with T-Norm and D-Norm Using MFCC and Prosodic Features.

From the above experiments, it has been observed that the performance of the language dependent speaker verification system improves up to **4.18%** and **4.80%** of EER levels for language matching and mismatching conditions respectively.

VIII. CONCLUSIONS

In this case approximately **1.00%** and **3.70%** improvements has been observed by applying the LID module with T-Norm and D-Norm in the **SV** system than that of the traditional without deploying LID module for language matching and mismatching conditions respectively. Although score normalization and model level normalization techniques improves the performance of SV system as usual. The improvement of the SV system applying the new approach by hybridization of LID modeling technique has been remarkable. Furthermore another important observation based on the above experiment that the performance of the proposed approach shows better improvement in case of language mismatching conditions than that of matching situations of languages.

REFERENCES

[1] D.A. Reynolds, et al., "Speaker Verification using Adapted Gaussian Mixture Models", Digital Signal Processing, pp.19-41, 2000.

[2] L. P. Heck, "Integrating high-level information for robust speaker recognition", John Hopkins University workshop on SuperSID, Baltimore, Maryland, http: www.cslp.jhu.edu/ws2002/groups/supersid, July 2002.

[3] L. Mary, and B.Yegnanarayana, "Prosodic Features for Speaker Verification", INTERSPEECH ,ICSLP, Pittsburgh, pp.917-920, September 17-21, 2006.

[4] W.M.Campbell et al., "Fusing discriminative and generative methods for speaker recognition :Experiments on Switchboard and NFI/TNO field data", in Proc. Odyssey Workshop, pp.41-44, 2004.

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

[6] W.M.Campbell et al., "The MIT-LL/IBM 2006 speaker recognition system.High –Performance reduced –complexity recognition" ,2006.

[7] A.G. Adami, et al., "Modeling Prosodic Dynamics For Speaker Recognition", in Proc. ICASSP, Hong Kong, pp.788–791, 2003.

[8] L. Ferrer, et al., "Modeling Duration Patterns For Speaker Recognition.in Eurospeech", Geneva, pp.2017–2020, 2003.

[9] N.Dehak, et al., "Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.pp.1-9.

[10] H.Mixdorff, "Speech Technology, ToBI, and Making Sense of Prosody. In Bel, Bernard &Marlien, Isabelle (Eds.) Speech Prosody" In Proceedings, Aix-en-Provence, France.2002.

[11] D.A. Reynolds, and R.Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", In IEEE Trans. on Speech and Audio Processing 3 , pp. 72–83,January 1995.

[12] N. Malayath, et al., "Data –driven temporal filters and alternatives to GMM in speaker verification", In Digital Signal Processing, pp.55-74, 2000.

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

[14] A.Fazel, and S.Chakrabartty, "An overview of Statistical Pattern Recognition Techniques for Speaker Verification", In IEEE CIRCUITS AND SYSTEMS MAGAZINE. 1531-636X/11/©2011 IEEE, 2011.

[15] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification" In Proceeding of EUROSPEECH '97, Rhodes, Greece, Vol.2, pp.963-966, 1997.

[16] D.A. Reynolds, et al., "Speaker verification using adapted Gaussian mixture models", In Digital Signal Processing, Vol.10, pp.19-41, 2000.

[17] J. Pelecanos, et al., "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.

[18] L.Heck, and N.Mirghafori, "On-Line Unsupervised Adaptation in Speaker Verification", In Proc. ICSLP-2000, Beijing, China, Vol. 2, pp.454-457 Oct. 2000.

[19] D.A Reynolds, "Universal Background Models. In Encyclopedia of Biometric Recognition", Springer, Journal Article, February 2008.

[20] Arunachal Pradesh, http://en.wikipedia.org/wiki/Arunachal_Pradesh.

[21] K. P. Li, and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring", in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing(ICASSP '88), vol. 1, pp.595–598, New York, NY, USA, April 1988.

[22] Y.K.Muthusamy, et al., "Perceptual benchmarks for automatic language identification", In Proceedings IEEE International Conference on Acoustic,Speech and Signal Processing 94, Adelaide, Australia, April 1994.

[23] L.Mary, and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition", Speech Communication, Vol.50, Issue 10,pp.782-796,2008.

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

[25] U.Bhattacharjee, and K.Sarmah, "Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Vol.2, Issue-6, pp.443-446. January 2013.

[26] Linguistic Data Consortium for Indian Languages. http://www.ldcil.org. 2008.

[27] B.C. Haris, et al., "Multi-variability Speech Database for Robust Speaker Recognition", In Proc. NCC, pp.1-5, 2011.

[28] S.Maity, et al., "IITKGP-MLILSC Speech Database for Language Identification" IEEE proceedings, 2012.

Authors Profile

Dr. Kshirod Sarmah received his Master of Science (M.Sc.) in Computer Science from Gauhati University, and Ph.D. in Computer Science and Engineering from Rajiv Gandhi University, India in the year 2004 and 2015 respectively. Currently he is working as an Assistant Professor and HOD in the department of Computer Science of Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Amjonga, Goalpara under Gauhati University, Assam, India. His research interest is in the field of Speech Processing and Robust Speaker Recognition and Machine Learning.

Dr. Swapnanil Gogoi received his Master of Science (M.Sc.) in Computer Science from Gauhati University, and Ph.D. in Computer Science and Engineering from Rajiv Gandhi University, India in the year 2004 and 2017 respectively. Currently he is working as an Assistant Professor and in the department of Computer Science and IT of Gauhati University, IDOL, Assam, India. His research interest is in the field of Speech Processing and Robust Speaker Recognition.

Prof. Utpal Bhattacharjee received his Master of Computer Application (MCA) from Dibrugarh University, India and Ph.D. from Gauhati University, India in the year 1999 and 2008 respectively. Currently he is working as a Professor in the department of Computer Science and Engineering of Rajiv Gandhi University, Arunachal Pradesh, India. His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.