International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

GMM-UBM Based Modeling for Language Identification using New Feature Vectors

A. Nagesh, M. Sadanandam

Abstract: ***The most of the existing LID systems based on the Gaussian Mixture model. The main requirement of the GMM based LID system is it require large amount of speech data to train the GMM model. Most of the Indian languages have the similarity because they are derived from Devanagari. Even though common phonemes exists in phoneme sets across the Indian languages, each language contain its unique phonotactic constraints imposed by the language. Any modeling technique capable of capturing all these slight variations imposed by the language is one of the important language identification cue. To model the GMM based LID system which captures above variations it require large number of mixture components.To model the large number of mixture components using Gaussian Mixture Model (GMM), the technique requires a large number of training data for each language class, which is very difficult to get for Indian languages. The main objective of GMM-UBM based LID system is it require less amount of training data to train(model) the system. In this paper, the importance of GMM-UBM modeling for language identification (LID) task for Indian languages are explored using new set of feature vectors. In GMM-UBM LID system based on the new feature vectors, the phonotactic variations imparted by different Indian languages are modeled using Gaussian Mixture model and Universal Background Model (GMM-UBM) technique. In this type of modeling, some amount of data from each class of language is pooled to create a universal background model. From this UBM model each model class is adapted. In this study, it is found that the performance of new feature vectors GMM-UBM based LID system is superior when compared to conventional new feature vectors based GMM LID system.***

Keywords: Universal Background Model(UBM), Gaussian Mixture Model(GMM), Language Identification (LID)

I. INTRODUCTION

The speech utterance not only contains message being conveyed, but also contains information related to the speaker, language and speaker emotional state related information in it. Language identification (LID) is the task of identifying the language being is spoken from short speech duration. It is an important enabling technology for spoken document retrieval and multilingual speech recognition.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Dr. A. Nagesh*, Professor, Department of CSE, MGIT, Hyderabad, India.
Dr. M. Sadanandam, Assistant Professor & BOS, Department of Computer Science & Engineering at Kakatiya University, Warangal, India.

The performance of LID system mainly depends on amount of language specific discriminative information incorporated into the system and type of modeling techniques that are employed for LID task [1].

There is a lot of similarity among Indian languages, this will have influence on phonotactic constraints are impaired on that languages. Owning to this similarity, to develop a LID system with large number of mixture components and huge amount of training data of a language is a prerequisite and in practice, we often suffer from insufficient training data to build a GMM from scratch. The Gaussian mixture model (GMM) – Universal Background Model (UBM) training process offers a solution to overcome such a problem. The most of the conventional GMM based LID systems require large amount of training data to model the characteristics of the languages. The GMM-UBM based LID system requires minimum amount of training data when compared to the conventional GMM based LID system. In this work the universal background method (UBM) is proposed which uses acoustic-phonetic information using new feature vectors for discriminating language specific information [2].

In this paper, a GMM LID system and GMM-UBM LID system is developed based on new feature vectors. First the new feature extraction method based GMM LID system is developed. Followed by the GMM-UBM LID system model is introduced as an alternate to GMM LID system model. To build a GMM-UBM LID system, the speech corpus of different language are taken into account for building a background model. Later, the UBM was used as an initial model for the enrollment phase in the GMM-UBM based LID system. In this approach, a language-specific GMM is adapted or derived from the UBM using Bayesian adaptation process. Instead of performing maximum likelihood training of the GMM for each language, this model is adapted from the well trained UBM parameters. Subsequently, performance of language identification (LID) performance is evaluated on both GMM and GMM- UBM LID systems by varying number of mixture components with different test durations of speech. This GMM-UBM model based LID system gives better performance than the GMM based LID system.

This paper is organized as follows. First the new feature extraction method based on frequency of occurrence phonemes is different between languages is described. Followed by GMM based modeling and GMM-UBM based modeling approaches for LID is discussed in section three.

Finally GMM based LID system, GMM-UBM based LID system and its identification performance is analyzed.

II. FEATURE EXTACTION

The any LID task basically divided into three phases, namely feature extraction, training and testing.

A. Exploring New Feature Vectors for GMM-UBM LID

The feature extraction is based on the frequency of occurrence of phonemes, also called probability of each feature vector in the acoustic class. Each Gaussian represents one acoustic class (cluster).This new type of feature vectors capture variations in the frequency of occurrence of phonemes across the languages effectively. In traditional MFCC feature vector based LID system, the phonemic differences is represented as a scalar (probability) value and this is given to the classifier for identification. Here, the phonemic differences is represented as a vector instead of scalar value and this is given to the GMM-UBM classifier for language identification [3].

In the first phase, from the speech signal language L_i , a 12 dimensional MFCC feature vectors are extracted. Using these MFCC feature vectors R Gaussians mixtures (clusters) are formed.

Once R clusters(Gaussians) are formed, the feature vector $X = \{x_1, x_2, , x_n\}$ is passed through a each Guassian G_i by producing the probability P_1 using probability density function. For each feature vector the probability $P_1$is calculated against Gaussian using probability density function. This P_1 represent the first coefficient in the new feature vector. In this way the feature vector pass through R Gaussians by creating R coefficients namely $P_1,P_2,\ldots,P_R$ in the feature vector as shown Fig.1. The new feature vector with dimension is R.

In same way, all the feature vectors passed through 'R' Gaussians $G_1,G_2,\ldots,G_R$ generating R dimensional feature vector $P_1,P_2,\ldots,P_R$. That means the 12 dimensional MFCC feature of size n are transformed into R dimensional new feature vector of size n. The 12 dimensional MFCC feature vector is represented into R dimensional new feature vector. In the new feature vector, each Gaussian probability density represents one coefficient. When the number of coefficients are 15, the good identification performance is achieved. The 12 dimensional MFCC feature vector is represented as a 15 dimensional feature vector as shown in Fig.2. The newly formed feature vector is given to GMM-UBM based LID classifier for language identification purpose.

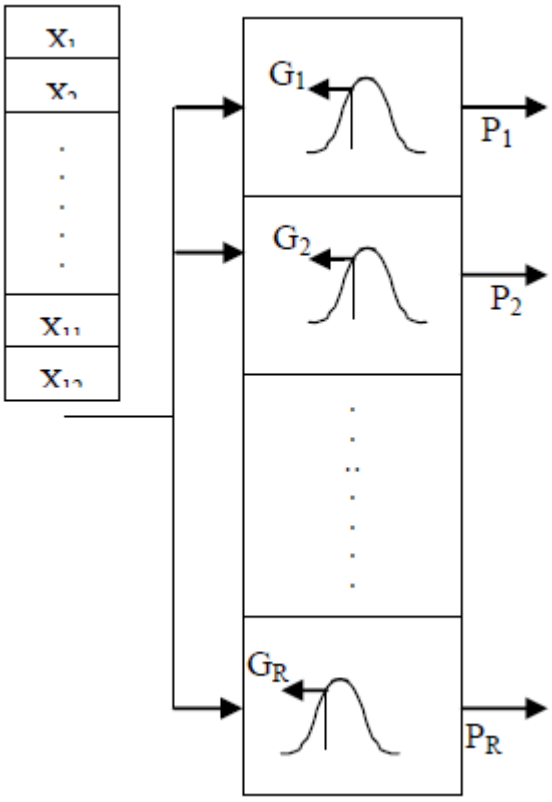

Fig.1.: Parameter estimation for new feature vector P.

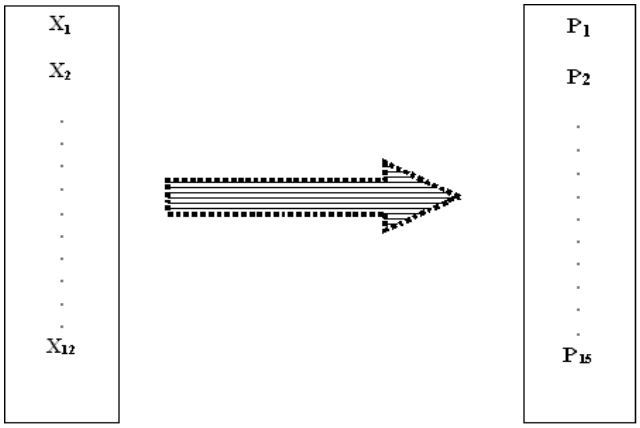

Fig.2: Transforming from 12 dimensional MFCC feature vector to 15 dimensional feature vector

III. NEW FEATURE VECTORS BASED LID SYSTEM

Using new formed feature vectors the language models are created. For the GMM LID task, the language models are created by modeling the new feature vectors using GMM as explained in the next section. Next GMM-UBM LID task, language models are created by modeling the new feature vectors using GMM parameters are adopted from UBM as explained in the next section[4] [5].

A. GMM Based LID System

The $X = \{x_1, x_2, , x_n\}$ be a new feature vector in D dimensional observation data.

The probability distribution of observation data is given by

$$p(x) = \sum_{i=1}^{M} p_i b_i(x) \quad (1)$$

Where M indicate the number of mixture components (Gaussian), the p_i indicate the weights of mixture component for i = 1, .., M and The mixture component weights satisfy the condition $\sum_{i=1}^{M} p_i = 1$. The $b_i(x)$ represent the mixture component density and given by

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} e^{\left\{-\frac{1}{2}(x-\mu)^T \sum_i^{-1}(x-\mu)\right\}} \quad (2)$$

These parameters can be collectively represented as $\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\}$ for i = 1, … M . Each language in a language identification system can be represented by one distinct GMM and is referred by the language models λ_i, for i= 1, 2, 3,…..K, where K is the number of languages under consideration.

During GMM training, clusters are formed within the training data. The main goal is to estimate the GMM parameters for the new feature vectors. The most common approach to estimate the GMM parameters is the maximum likelihood estimation. The main objective of maximum likelihood is to estimate the model parameters which maximize the likelihood of the GMM for the given new feature vectors. For the sequence of n new feature vectors GMM likelihood can be defined as

$$p(X/\lambda) = \prod_{t=1}^{T} p(x_t/\lambda) \quad (3)$$

The maximum likelihood GMM parameters are estimated using Expectation-maximization (EM) iterative algorithm. The basic idea of EM process is to start with a initial model λ and estimate a new model such a way that $p(X/\lambda) < p(X/\hat{\lambda})$. In the next iteration the new model $\hat{\lambda}$ model becomes the initial model and the process is repeated until the a convergence is reached. For each EM iteration, to re-estimate the GMM parameters the following formulas are used which indicate the monotonic increase in the GMM likelihood value.

Mixture weight $p_i = \frac{1}{T}\sum_{t=1}^{T} pr(i/x_t, \lambda)$ (4)

Means: $\mu_i = \frac{\sum_{t=1}^{T} pr(i/x_t,\lambda)x_t}{\sum_{t=1}^{T} pr(i/x_t,\lambda)}$ (5)

Variance: $\Sigma_i = \frac{\sum_{t=1}^{T} pr(i/x_t,\lambda)x_t^2}{\sum_{t=1}^{T} pr(i/x_t,\lambda)} - \mu_i^2$ (6)

a. New Feature Vectors Based LID using GMM

During feature extraction in GMM based LID system, using new feature extraction method new feature vectors are extracted. First from the speech corpus of each language L_i, 12 dimensional MFCC feature vectors are extracted. The extracted 12 dimensional feature vectors are transformed into 15 dimensional new feature vectors. This is repeated for all the languages under consideration and separate set of new feature vectors are obtained for each language under consideration.

Training the Model

During training, the new feature vectors of the language L_i are clustered using K-means clustering algorithm. Using EM algorithm GMM parameters are re-estimated for the language the L_i. In this way GMM is created λ_i as shown in the Figure.3. In this Now the GMM parameters are re-estimated using EM algorithm. For each language L_i , one GMM is created. This procedure is repeated for all the languages under consideration and separate GMMs are created for each language. The steps involved in the GMM training as follows.

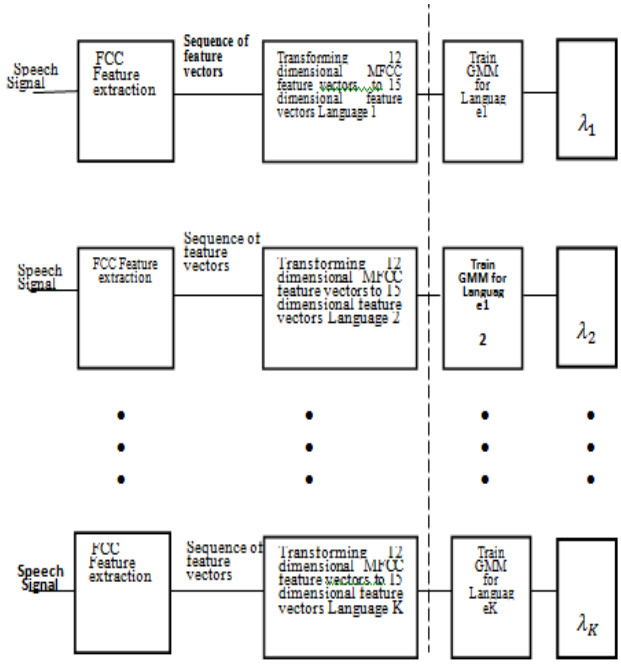

Fig.3: GMM training for language identification system

Training Phase:

for each language L_i $(1 \le i \le M)$ from language L_1 to L_M do

i) for each language L_i extract 12 dimensional MFCC features

ii)create 15 Gaussians using MFCC feature vectors of language L_i

iii) for each MFCC feature vector X_k $(1 \le k \le N)$ of language L_i do

X_k passing through 15 Gaussians to create new feature vector P_k

end

Initialize GMM parameters $\lambda_i=(p_i, \mu_i, \sum_i)$ using K-means clustering for language L_i

train λ_i for optimal solution using EM algorithm.

end.

During testing, for the given test speech observation $X=(x_1,x_2,\ldots,x_t)$, the 12 dimensional MFCC feature vectors are extracted. This 12 dimensional feature vectors are transformed into 15 dimensional feature vectors. Using these new test feature vectors, the likelihood value is calculated against each GMM model. The model which gives maximum likelihood value is declared as the identified language as shown in Fig.4. The steps involved in the GMM testing as follows.

Testing Phase:

// The procedure of generating new feature vector with 15 elements is identical as in the training phase

i)from test speech utterance $X=(X_1,X_2,\ldots..,X_t)$ 12 dimensional extract MFCC features

ii)X passing through 15 Gaussians to create new feature vectors P

iii)for each model $\lambda_1, \lambda_1,\ldots. \lambda_M$ do

p(P|λ) is calculated, where p(P| λ_i) is the probability of the observation sequence of new feature vectors P(P_1, P_2 ,…. P_t).

end

iv)Calculate 1-max result for a given testing speech utterance using

$$\arg \max_{1 \le i \le M} p(P/\lambda_i)$$

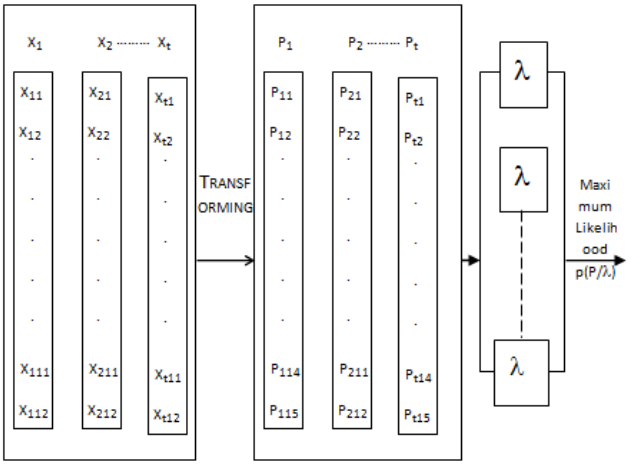

Fig. 4: GMM testing for language identification system.

b. Database Used for the Present Study

For this study IITKGP-MLILSC multi lingual Indian language speech data base is used. This speech database consists of 27 regional languages. For each language ten speakers including both male and female are present. For each language one hour speech data is available in the data base.

The performance of the GMM LID system not only depends on the feature vectors newly derived feature vectors it also depends on the GMM parameters. The GMM parameters consists of dimension of the new feature vectors, total number of feature vectors (size) and number of mixture components. Here for language identification purpose five languages are considered. For training and testing purpose different speaker data are used. For training purpose eight speakers and testing purpose two speakers data are considered. For the training 25-30 minutes of speech data for each language is used under consideration. For testing purpose, different duration of speech such as 3ses, 5sec and 10 sec are considered and 100 samples are used for the testing purpose.

B. UBM-GMM Based LID System

In language identification, adaptation of the acoustic models to new operating conditions is very important because of variability of data due to different environment, speaking styles, speakers and so on. As in the conventional GMM based LID, the performance of the system suffers due to variability of speaker, channel and environment present in the speech data. But in GMM-UBM based LID system, it is vulnerable to undesired variability due to non language effects, such as speaker and channel present in the speech data[6]. In this, the language specific GMM models are created by adapting from the UBM based language model. The UBM is language model which represents a general language model, which is independent of language, speaker and channel.

a. Maximum A Posteriori (MAP) Parameter Estimation

Like in EM iterative algorithm, the adaptation method is a two-step estimation process. The EM algorithm is used to train the language models, same way in GMM-UBM adaptation based models are used to train using maximum a posteriori estimation (MAP) approach.

The adaptation MAP process first, align target training data into UBM mixture components using equation (7). Next calculate the sufficient statistics, which are known as Baum-Welch statistics from the new feature vectors using following equations (8). It is possible to adopt all parameters or some of them from the UBM model. But it is found that, adopting the means only will work well in practice using equation (9). Then update the target language model parameters using Baum-Welch statistics and adapt parameter() using equation (4). Here r indicate relevance factor, which controls the rate of adaptation.

$$pr(i|x) = \frac{p_i b_i(x)}{\sum_{j=1}^{M} p_i b_i(x)} \quad (7)$$

$$n_i = \sum_{n=1}^{N} pr(i|x_n) \quad (8)$$

$$E_i(x) = \sum_{n=1}^{N} pr(i|x_n)\, x_n \quad (9)$$

$$\alpha_i = \frac{n_i}{n_i + r} \quad (10)$$

$$\mu_i = \alpha_i E_i(x) + (1 - \alpha_i)\mu_i^{ubm} \quad (11)$$

a. New Feature Vectors Based LID using GMM-UBM

During feature extraction in GMM-UBM based LID system, from speech database new feature vectors are extracted. For this first from the speech corpus of each language L_i, 12 dimensional MFCC feature vectors are extracted. The extracted 12 dimensional feature vectors are transformed into 15 dimensional new feature vectors using new feature extraction method explained in the previous section. This is repeated for all the languages under consideration and separate set of new feature vectors are obtained for each language under consideration.

Training the Model

In this new feature vectors ten speaker data of each language data is present in the database. From these three speaker speech data of the all the languages are polled to create a universal background model. The other five speaker speech data is used to adapt the universal background model to develop the corresponding language model. Here the five language GMM-UBM LID system is developed with varying the number of mixture components such as 128,256,512 and 1024.

During the training, first to create the UBM, 120-150 minutes of speech data of all languages are used. This universal background model is used to adopt all the classes to develop the corresponding language models. During this adaptation process, 18sec data from all the five speakers of a language i.e total 90 sec of speech data per language is used to develop the corresponding language models.

IV. RESULT AND DISCUSSION

a. New Feature Based GMM LID System

For GMM based LID task five languages are considered namely Hindi, Kanda, Telugu, Tamil and Indian English. The language identification performance is analyzed varying number of mixture components and test duration is shown in Fig.5. For each language multiple LID systems are developed with varying the number of mixture components from 16 to 128 as shown in Fig.4. The average performance of the three test cases of five languages LID system is shown in Table.1. The table it is noticed that the performance of LID system is increased the number of mixture components are increased. When the number of mixture components is 128, the LID system performance is high. The LID performance is also increased when the test speech duration is increased from 3sec to 10sec.

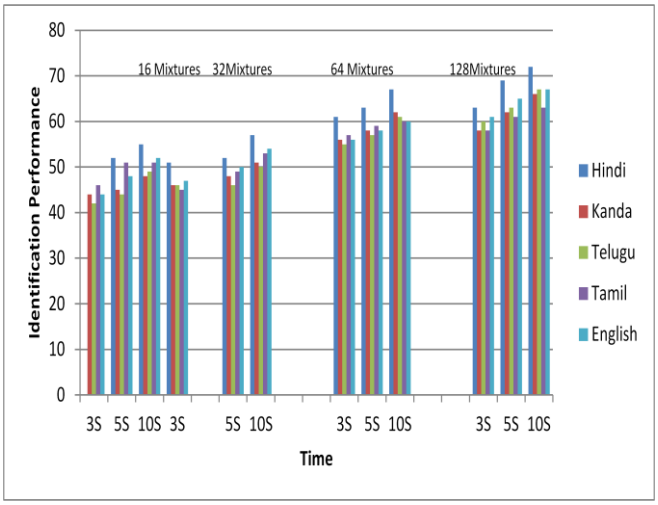

Fig.5: The five language GMM based LID performance for varying length of mixture components and test duration.

Table.1: The five language GMM based LID performance for varying length of mixture components and test duration.

Language	16mixtures			32 mixtures			64 mixtures			128 mixtures		
	Time in Sec			Time in Sec			Time in Sec			Time in Sec		
	3	5	10	3	5	10	3	5	10	3	5	10
Hindi	48	52	55	51	52	57	61	63	67	63	69	72
Kanda	44	45	48	46	48	51	56	58	62	58	62	66
Telugu	42	44	49	46	46	50	55	57	61	60	63	67
Tamil	46	51	51	45	49	53	57	59	60	58	61	63
English	45	48	52	47	50	54	56	58	60	61	65	67
Avg	44	48	51	47	49	53	57	59	62	60	64	67

b. New Feature Based GMM-UBM LID System

The language identification performance of five languages Hindi, Kanda, Telugu, Tamil and Indian English for varying number of mixture components and test duration. For each language multiple LID systems are developed with varying the number of mixture components from 64 to 512 as shown in Fig.6. The lid performance is increased when the number of mixture components is also increased. The average performance of the all the test cases of five languages GMM-UBM based LID system is shown in table.2. From the table.2 it is clearly evident that the performance of present GMM-UBM based LID system is superior when compared to base line GMM based LID system. The improvement is due to the usage of a large number of mixture components in the GMM-UBM LID system. It is also noticed in the training, GMM-UBM based LID system require very less amount of training data when compared to conventional GMM based LID system.

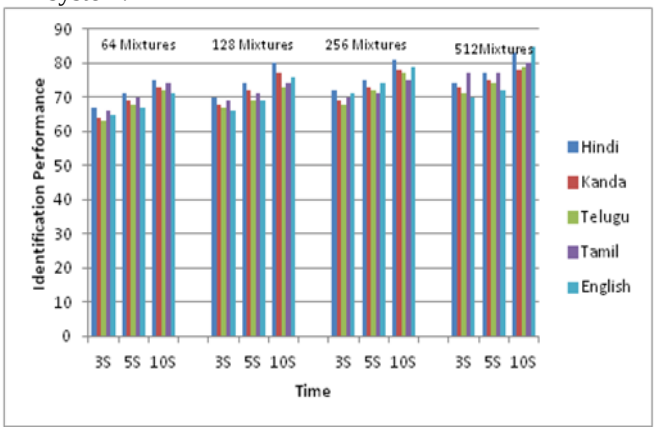

Fig.6:The five language GMM-UBM based LID performance for varying length of Gaussian components and test duration

Table.2:The five language GMM-UBM based LID performance for varying length of Gaussian components and test duration

Language	64mixtures			128 mixtures			256mixtures			512mixtures		
	Time in Sec			Time in Sec			Time in Sec			Time in Sec		
	3	5	10	3	5	10	3	5	10	3	5	10
Hindi	67	71	75	70	74	80	72	75	81	74	77	83
Kanda	64	69	73	68	72	77	69	73	78	73	75	78
Telugu	63	68	72	67	69	73	68	72	77	71	74	79
Tamil	66	70	74	69	71	74	70	71	75	77	77	80
English	65	67	71	66	69	76	71	74	79	70	72	85
Avg	65	69	73	68	71	76	70	73	78	71	75	81

V. CONCLUSION

In the present work the importance of GMM-UBM modeling for language identification (LID) task for Indian languages are explored using new set of feature vectors. The slight variations in phonotactic information exist among Indian languages. The variations effectively captured using Gaussian. The conventional GMM based LID system require more number of mixture components to capture slight variation in phonotactic information imparted by the language for training. To train GMM LID system with more number of mixture components requires large amount of speech training data.

To get this large amount of training data for Indian languages is tedious task. In GMM–UBM based LID system, to training model offers a solution to overcome such a problem. In this approach certain amount of data is pooled to develop the universal background model (UBM) with large number of mixture components and from this UBM model the GMM language specific models are created. The conventional GMM based LID system for five languages are developed with varying number of mixture components such as 16, 32, 64 and 128 and the language identification performance is analyzed for different test duration. The five language GMM-UBM LID system is also developed with varying the number of mixture components such as 128,256,512 and 1024. The Here GMM-UBM LID system with more number of mixture components developed with minimum amount of training data. The LID performance of proposed GMM-UBM based LID system is superior compared to GMM based LID system. This is due to usage of GMM-UBM based modeling technique.

REFERNCES

1. Li, H., Ma, B., & Lee, K. (2013). Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE, 101 (5), 1136–1159.
2. E. Wong, "Automatic spoken language identification utilizing acoustic and phonetic speech information," Ph.D. dissertation, Speech and Audio Research Laboratory, Queensland Univ. Technol., 2004.
3. L. Mary and B. Yegnanarayana, Extraction and Representation of Prosodic Features for Language and Speaker Recognition, Speech Communication, vol. 50(10), pp. 782–796, (2008)
4. M.A.Zissman "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech" lEEE Transactions on Speech and Audio Processing Vol 4, NO. 1, p.p 31-44 January 1996.
5. E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in Proc. Int. Conf. Spoken Language Processing (ICSLP-2002), 2002, pp. 93–96.
6. T. Hasan, Y. Lei, A. Chandrasekaran, and J. H. L. Hansen, "A novel feature sub-sampling method for efficient universal background model training in speaker verification," in *Proc. IEEE ICASSP'10*, Dallas, TX, Mar. 2010, pp. 4494–4497.
7. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process*, vol.10, no. 1–3, pp. 19–41, 2000.
8. Q. Dan and W. Bingxi, "Discriminative Training of GMM for Language Identification", *in ISCA* & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. MAP8, Apr. 2003.

AUTHORS PROFILE

Dr A. Nagesh, is currently working as a professor in CSE at MGIT, Hyderabad. He completed B.E and M.Tech from Osmania University, Hyderabad in 1996 and 2002 respectively. He did Ph.D in CSE from JNTUH, Hyderabad in the year 2012. He is having total 22 years of teaching experience. At present he his supervising five Ph.D students. Total he is having 40 publications in national & international journals. His research areas includes pattern recognition , speech processing and data mining.

Dr. M. Sadanandam is currently working as a Assistant professor & BOS in department of computer science & Engineering at Kakatiya University, Warangal. He completed B.E Kakatiya University, Warangal and M.Tech from JNTUH. He did Ph.D in CSE from JNTUH, Hyderabad in the year 2016. He is having total 15 years of teaching experience. At present he his supervising 4 Ph.D students. Total he is having 30 publications in national & international journals. His research areas includes pattern recognition , speech processing and data mining.