International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2694-2700

Spoken Language Identification using Gaussian Mixture Model-Universal Background Model in Indian Context

[1]Sreedhar Potla
Associate Professor, Department of IT, SNIST, Hyderabad-501301, Telangana, India
E-mail: sreedharp@sreenidhi.edu.in

[2]Vishnu Vardhan. B
Professor, JNTUCE of Manthani, Department of CSE, Karimnagar-505501, Telangana, India.
E-mail: mailvishnuvardhan@gmail.com

Abstract

Speech processing is the area of analyzing the contents of speech signals by using the speech signal processing techniques. The speech signal acts as in input for man-machine interactive systems. The characteristics of a language are also influences the functionality of speech based systems. Spoken Language Identification (SLI) is one of the interesting research areas of speech processing. SLI determines the language that is been uttered from an unknown speaker. The objective of this work is to identify a specific language from a set of twenty three major Indian languages spoken by an unknown speaker. In this work, speech signals were divided into a sequence of short time segments and represented by Mel-Frequency Cepstral Coefficient (MFCC) vectors. A Universal Background Model (UBM) is constructed by selecting the minimal and equal number of speech samples from all languages with varying number of mixtures. The experiments were carried out with MFCC and MFCC-delta features on hypothesized Gaussian Mixture Model with Universal Background Model(GMM-UBM) and were compared against Gaussian Mixture Model(GMM). By using GMM-UBM model, the accuracy of spoken language identification in the Indian context has significantly improved when compared with an SLI using GMM classifier

Keywords: Spoken Language Identification, Gaussian Mixture Model, Universal Background Model, Mel- Frequency Cepstral Coefficients

INTRODUCTION

Speech processing is the analysis of speech signals by using the processing methods on various speech signals. The signals are usually processed in a digital representation and speech processing is regarded as a special example of digital signal processing which is applied to the speech signal. The speech processing is expanded as a major research domain which includes various sub domains as speech analysis, speech synthesis, speech coding, speech recognition, speaker recognition and language identification. Speech analysis deals with the methods of speech productions, which involves the extraction of needed information such as voiced and unvoiced segments for possible event detection. Speech synthesis is the process of converting the written text to a speech signal. Speech coding is a process of representation of a speech signal in a compact manner.

Speech Recognition (SR) identifies speech con- tent and represents it as text. SR depends on speech mode, type of the speaker, vocabulary size and speaking style. The mode of the speech is either isolated speech or continuous. A speaker is possibly independent, dependent or adaptive speaker. The vocabulary size of the data set conveys the number of speech samples and their time duration. Speaking style also influences the speech recognition which depends on the flow of the speech delivery as dictation or continuous. Speaker identification determines the person from the characteristics of a given speech input.

Language identification attempts to identify a language from a short speech sample uttered by an un- known speaker. The set of features used in SLI are divided into various levels corresponds to the human speech production system such as acoustic-phonetic information, phonotactic information, prosodic information, lexical and syntactic information [1]. The human speech production system produces possible sound units termed as phones. These units convey spectral information of the vocal tract. The acoustic information about a phone carries formant frequencies, articulatory patterns and spectral components. The acoustic-phonetic features discriminates a language from other languages as the number of phones varies based on the language. The phonotactic information specifies predefined rules that govern the possible combinations of different phonemes.

Every language has a unique set of phonemes and their sequences differentiate a language from other languages. Prosodic information corresponds to supra-segmental speech such as duration, stress, rhythm and intonation. Based on the combination of prosodic features, a language is identified. Phonological information of a language specifies rules to form words or morphemes using symbols. These rules are unique for each language which are used for language identification. The syntactic system specifies the rules to form phrases and utterances using words and morphemes of a language. Each language has unique syntactic system that is used for language identification.

SLI is an implicit system used in many speech processing applications such as automated call directing systems, language translation systems, multilingual dialog conversational systems, speech based interactive response systems, multilingual speech recognition system. It is also used for effective organization of speech documents in web domain and optimal spoken document retrieval system [2].

LITERATURE SURVEY

In speech processing, many researches were implicitly used language identification task as front-end application. Language Identification task is addressed by various researchers using the characteristics of the speech signal. Such characteristics include acoustic, phonotactic, prosodic, lexical and syntactic features or any combination such features is used to discriminate a language from other languages. In a detailed review by Muthusamy [3, 4] on language identification sys- tem which was used cues such as acoustic phonetics, prosodic, phonotactics and vocabulary entities. It is suggested that phonetic level information is sufficient to discriminate between languages with greater accuracy.

The phonetic information is represented as a distinctive feature vector or a set of segmental vectors. Cimarustri [5] demonstrated the usage of Linear Predictive Coding (LPC) feature using static feature extraction process which contains autocorrelation coefficients, cepstral coefficients, filter coefficients, log area ratios and formant frequencies. Navratil [6] presented a language identification system using binary tree structure on phonotactic features. Foil [7] demonstrated the extraction of prosodic features like rhythm and intonation from pitch and energy contour for language identification. Prosodic features used by Hazen [8] and Muthusamy [3, 4] and the other features like pitch and amplitude contour were used by Thyme- Gobbel [9]. It is proved that pitch and amplitude contour were efficient in discriminating one language from other. Bhaskararao

[10] has verified usage of all speech characteristics and suggested that variation in the phonotactics were prominent in comparison with the speech sounds for language identification. Schultz compared [11] language identification system based on phone level and word level with and without language model and demonstrated that word based system with trigram modeling of words superior than phone-based system. Kadambe [12] proposed a bottom-up approach through the lexical model that first identifies phones, followed by words, and subsequently a language. Thomas [13] proved that lexicons better discriminates a language from other and such lexicons were learned while training the system.

The spectral features are popular and efficient to represent variation acoustic-phonetic information of speech in several languages. Widely used spectral features for language identification are processed using Linear Predictive Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs), Shifted Delta Cepstrum (SDC) features and Linear Predictive Coding (LPC) features. Quatieri [14] has demonstrated that MFCCs are quite effective in most of the speech processing systems because they exploit auditory principles. MFCCs with delta represent the acceleration values using first order derivatives of MFCC vectors for capturing dynamics in a speech signal.

Zissman [15] has proved that the SLI performance is further enhanced by using spectral information which is represented as MFCCs and with a parallel Phone Recognition followed by Language Model (PRLM) based system. Pellegrino [16] explored about the influence of vowels that are essential in identifying a spoken language based on the MFCCs features and Gaussian Mixture Model. A different approach explored by Torres-Carrasquillo [17] by using phone tokenizers, mixture model order, and combination of n- gram classifier in conjunction PRLM. It is proved that GMM tokenization with language modeling achieves minimal error rate and efficient identification performance.

In the literature it found that most of SLI works were carried out on TIMIT corpus, Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus, SPEECHDAT corpus, NIST LRE (1994, 1996, 2003, 2005, 2007, 2009, 2011) corpus and on many other foreign languages corpus.

A subcontinent like India, designing an accurate SLI has several limitations related to features selection, availability of speech corpus and the models that are to simulate. In the Indian context, Mary [18] has proposed a new approach for prosodic features extraction and representation from the syllable based segmented speech signal. It is proved that language recognition task conducted on NIST LRE 2003 based on prosodic features is potential for discriminating languages. In India, there are several official languages and a majority of them are categorized into four language families, such as Indo-Aryan, Dravidian, Austro-Asiatic and Tibeto- Burman. Jothilakshmi [19] has prepared a database comprised with 10 major Indian languages and developed a hierarchical language identification model based on acoustic information represented as MFCCs and SDC features vectors with GMM. In her exploration these languages were categorized using hierarchical language identification. It is suggested in her work that SLI designed based on SDC features and GMM is efficient when com- pared with MFCC features and GMM. Maity [20] at- tempted to collect speech samples for 27 local official languages and represented them as spectral features in a language identification task modeled with the Gaussian Mixture Model. Further, the language identification method proposed by Ramu Reddy [21] evaluated on both spectral and prosodic characteristics that are captured at possible levels using IITKGP- MLILSC corpus [20]. Xu [22] proved that selection of acoustic scores along with language model scores improves efficiency of language identification activity.

In this paper the Universal Background Method is proposed which was used acoustic-phonetic information for discriminating language specific information. Speech acoustic information is represented as MFCCs and MFCCs with delta features. The speech corpus used in this work has twenty three major Indian languages collected from different news broad- cast archives. Balanced amount of speech samples for different language are taken into account for building a background model. A Universal Background Model is then adapted to all selected languages to build corresponding language models. Subsequently, performance of SLI is evaluated on both GMM and GMM- UBM systems by varying number of mixture components with different durations of speech samples.

The following sections of this paper are organized as follows. The development of SLI system with GMM is described in Section 3. Section 4 describes implementation of SLI using GMM-UBM. Section 5 shows the details of experimental evolution and result discussion that includes speech corpus specification and results observed on the corpus data using both GMM and GMM-UBM classifiers. Finally, Section 6 describes conclusions and directions for future works.

SLI USING GAUSSIAN MIXTURE MODEL (GMM)

In this paper, variations in the spectral features are used to differentiate a language from other. The speech signal spectral features are effectively represented as MFCC vectors. In this work, speech signals are divided into short time window segments of length 20msec with a shift of 10msec. Subsequently, spectral information of the segmented speech is captured through 26 mel scale filter bands to represent as MFCCs. Delta features are also captured using first order derivatives of MFCCs to represent dynamics in a speech signal.

The MFCC feature vector C is represented as a set of values mentioned as $c_1, c_2, c_3, c_x, \ldots \ldots . c_K$. In this C feature vector, x is the frame number that contains N dimensional MFCCs and is indentified as c_x . The K value corresponds to number of feature frames present for an input speech signal. The work flow of SLI using GMM classification method is demonstrated in Fig. 1.

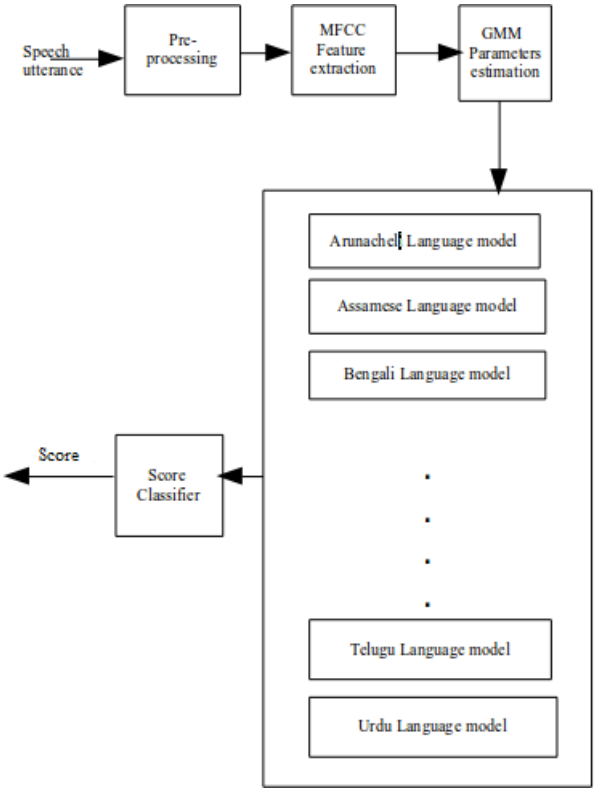

Figure 1. SLI using GMM Classification Training with GMM

The GMM classification method uses weighted set of multivariate Gaussian densities, through the probability density function of MFCCs feature vector C.

$$p\ C|\lambda = \sum_{i=0}^{m} p_i\ d_i\ C \quad (1)$$

In equation (1) p_i as the probability of mixture weight (as $\sum_{i=0}^{m} p_i = 1$), m as number of mixtures, i as the mixture index ($1 \leq i \leq m$) and λ is a language model specified as equation (2).

$$\lambda = p_i, \mu_i, \Sigma_i \quad (2)$$

and the d_i (c) shown in equation (3) multivariate Gaussian distribution function constructed with the help of associated means μ_i and set of diagonal co-variance matrices, represented as Σ_i.

$$d_i = \frac{1}{\sqrt{2\pi}^D} e^{\frac{c-\mu}{2}^2} \quad (3)$$

During the training activity the Expectation Maximization (EM) algorithm is used for better convergence of mixture attributes like mixture weights, means and variances. In this process twenty three language models (λ_i) were constructed for the designed corpus comprising major Indian languages.

Testing with GMM

The speech utterance of an unknown speaker is given for pre-processing followed by MFCC feature extraction and represented as C vector. The C observation is interpreted by the language model and calculates log likelihood specified as follows

$$p\ C|\lambda = \frac{1}{N} \sum_{n=0}^{N} logp\ C_n|\lambda \quad (4)$$

in equation (4) that λ corresponds to a language model. The classifier score from all language models are used to compute the maximum-likelihood, specified as follows

$$Score = argmax_{l=1}^{L} p\ C|\lambda_i \quad (5)$$

where $argmax$ function in equation (5) determines highest score value for a particular language index l out of all twenty three language models as shown in Fig. 1. The score is the final outcome that determines name of the spoken language. Based on the obtained score, the spoken language uttered by an unknown speaker is identified.

SLI USING GMM-UBM

The GMM-UBM is a specialized GMM which uses a large amount of training data in order to model the variability of these languages data. In most of the SLI systems, huge amount of training data is required to model the characteristics of languages. The UBM based SLI consumes very minimum amount of training data when compared with independent SLI language models. The training data used for UBM is classified into two groups namely algorithm parameters and data parameters. The algorithm parameters are corresponds to the amount of data used for adaptation process. The data parameters deals to various factors like corpus, the amount of speech data, number of speakers, amount of data used per speaker, method of selecting speakers, mode of feature vectors represented and data balancing according to language.

The training process of SLI based UBM is shown in Fig. 2. To build the UBM, 300sec of speech data was selected from 4 speakers of each of the twenty three Indian languages and represented as MFCCs. MFCCs features were processed through Expectation-Maximization algorithm for better converged UBM. Such an UBM represents behavior of all twenty three languages together. In this model, an input utterance of C contains a set of feature frames as $c_1, c_2, c_3, c_x, \ldots \ldots . c_t$. Thus, equation (6) explains the calculation of observation mixture attributes

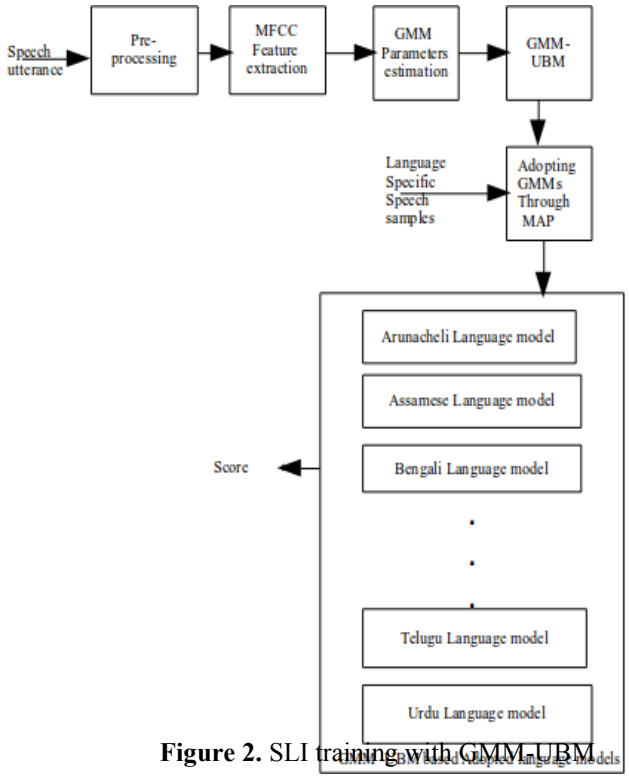

Figure 2. SLI training with GMM-UBM

$$p\,i|c_t = \frac{pidi(ct)}{\sum_{j=1}^{M} p_j\, d_j\,(c_t)} \quad (6)$$

Subsequently, twenty three language models (λ_i) were derived by adopting the Universal Background Model using 30% of training data from each of the twenty three languages. During the adaptation process 50% corpus data is used through a Maximum A. Posteriori (MAP) method described by Gauvain [23]. This results in building respective language models (λ_i) in optimal training process.

During testing activity shown in Fig. 3 an unknown speech sample (from remaining 20% of corpus data) is fed to the system, features are extracted and represented as MFCCs. The scores for the given input are determined against all twenty three language models. Collection of this score is fed to the score classifier where average log likelihood method determines the highest scored language model as an outcome. The language outcome of the test sample is assigned to that of the highest scored language model.

EXPERIMENTAL EVALUATION AND DISCUSSIONS

Speech Corpus in the Indian Context

The speech corpus used in this work, is prepared from different broadcasting channels for the twenty three major Indian languages such as Arunachali, Assamese, Bengali, Bhutiya, Chattisgarhi, Dogri, Gujarati, Hindi, Indian English, Kannada, Konkani, Ladkh, Lepcha, Malayalam, Manipuri, Marathi, Odiya, Panjabi, Rajasthani, Sanskrit, Tamil, Telugu and Urdu. All of the identified channels were recognized for broadcasting clean and noise free speech signals.

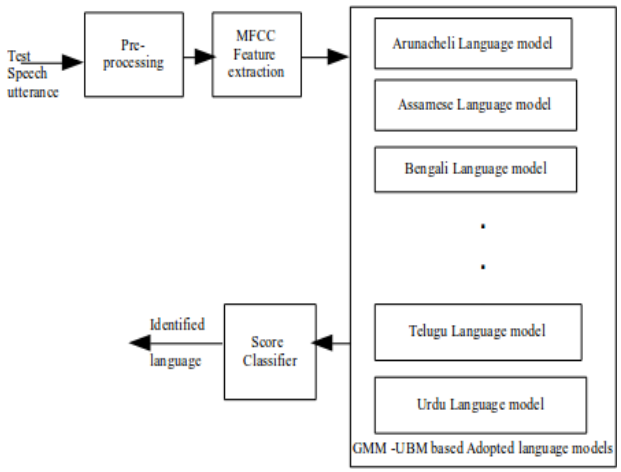

Figure 3. SLI testing with GMM-UBM

This collection has a total of 2815 minutes of speech data comprising all languages. The corpus has 182 male and 142 female native language speakers pertaining to all twenty three languages to avoid speaking style issues. The speech signals were captured at 16 kHz sampling rate and represented with 16 bits of data. Language wise speech samples details were described in Table 1. The corpus data collection was formulated from different archives that are available with major Indian broadcasting news channels. The first 90 seconds of each speech data was removed and the remaining voiced portion was retained to ensure clean and music free speech content. From this corpus data, speech data of 80% speakers is used to train the GMM based SLI and speech data of 20% speakers is used in testing phase. This work, also studied the performances of SLI using GMM method and SLI using GMM-UBM method by varying length of test speech samples at 5sec, 10sec and 15sec. Both systems are processed using 13 MFCC features vector and MFCC+Δ features vector. The subsequent sub sections describe more about both GMM based SLI and GMM-UBM based SLI implementations.

Table 1. Details of Speech corpus prepared for Indian context

Language name	Number of Speakers male	female	Duration (min)
Arunachali	4	2	85.39
Assamese	8	7	137.34
Bengali	9	4	120.49
Bhutia	10	2	65.28
Chattisgarhi	3	6	141.17
Dogri	8	9	112.10
Gujarati	12	4	126.24
Hindi	10	9	166.14
Indian English	4	10	105.09
Kannada	5	10	145.40
Konkani	8	6	140.17
Ladkh	4	8	92.56
Leptcha	7	5	59.24

Malayalam	11	5	152.02
Manipuri	9	6	172.41
Marathi	4	13	136.54
Odiya	6	5	93.25
Panjabi	12	4	169.10
Rajasthani	8	6	134.13
Sanskrit	9	4	59.28
Tamil	6	10	161.11
Telugu	10	5	136.53
Urdu	15	2	100.05

SLI Using Gaussian Mixture Model (GMM)

During the training phase, about 8 speech samples per language (4 male speakers and 4 female speakers) were used from the speech corpus. While at the testing phase, 2 speech samples were taken into consideration, which included distinct male and female speakers not present in the training data to nullify issues associated with the presence of speaker dependency in SLI system. The results of SLI using GMM method are given in Table 2.

Table 2. SLI using Gaussian Mixture Model

No. of mixtures	SLI Performance in %					
	MFCC			MFCC+Δ		
	Duration			Duration		
	5*sec*	10*sec*	15*sec*	5*sec*	10*sec*	15*sec*
32	34.62	34.87	35.43	35.61	35.79	36.06
64	35.64	35.92	36.37	36.11	36.85	37.14
128	36.01	36.67	36.81	36.23	37.13	37.88
256	37.58	36.79	38.64	38.67	39.48	**41.67**
512	36.23	37.97	37.04	36.55	37.18	37.93
1024	36.09	37.93	36.78	36.23	36.94	37.77

The GMM based SLI performance shown in Table 2 is the average percentage of 100 test cases. The number of mixture components used in GMM based SLI were specified at Column 1. Column 2 to 4 and Column 5 to 7 in order to represent SLI performances for 5sec, 10sec and 15sec speech samples using MFCC and MFCC+Δ features. From these results it was analyzed that SLI using GMM with MFCC+Δ features has the best performance of 41.67% for 15 sec speech with 256 mixture components and the worst performance 34.87% is observed in the case of MFCC features with 5 sec utterance with 32 mixtures. It was also observed that performances of MFCC + Δ features based SLI were effective compared with that of MFCC features based SLI. The increase in 4 - 5% of SLI performance is observed while increasing the number of mixture components from 32 to 256.

SLI Using Gaussian Mixture Model- Universal Background Model (GMM- UBM)

The UBM is constructed by using 5 mins of speech samples selected from each of the twenty three languages that consists both male and female speakers. The speech data of 600 secs from each of twenty three languages were used to adopt twenty three language models with 32 to 1024 mixtures. During the testing phase, 5 mins of speech samples comprising both male and female speakers were used. The average of 100 GMM-UBM based SLI test cases results over 5sec, 10sec, and 15sec were given in Table 3.

Table 3. SLI using Gaussian Mixture Model- Universal Background Model

No. of mixtures	SLI Performance in %					
	MFCC			MFCC+Δ		
	Duration			Duration		
	5*sec*	10*sec*	15*sec*	5*sec*	10*sec*	15*sec*
32	34.62	34.87	35.43	35.61	35.79	36.06
64	35.64	35.92	36.37	36.11	36.85	37.14
128	36.01	36.67	36.81	36.23	37.13	37.88
256	37.58	36.79	38.64	38.67	39.48	**41.67**
512	36.23	37.97	37.04	36.55	37.18	37.93
1024	36.09	37.93	36.78	36.23	36.94	37.77

Column 2 to 4 and Column 5 to 7 represents GMM-UBM based SLI performances for 5 sec, 10 sec and 15 sec speech samples using MFCC and MFCC+Δ features. From these results it was analyzed that SLI performance with MFCC+Δ features has best performance as 46.74% for 10 sec speech with 512 mixture components and the worst performance 36.11% is observed in the case of MFCC features with 5 sec utterance with 32 mixtures. It is also observed that the best SLI performance is shown in Table 3 is 5% superior than the best performance achieved by the GMM based SLI as shown in Table 2.

The best SLI performances observed for GMM and GMM-UBM are presented in Table 4. Results were observed that 15 sec samples using MFCC+Δ features with 512 mixtures of GMM-UBM resulted best language identification performance as 56.51% and for GMM using MFCC+Δ features with 256 mixtures resulted best SLI performance as 51.10%. In both cases, Tamil language has accomplished highest identification performance. It was also observed that the least performances for SLI are accomplished in GMM as 31.88% for 15sec duration using MFCC+Δ features with 256 mixtures and in GMM-UBM as 37.56% for 10sec samples using MFCC + Δ features with 512 mixtures. In both cases least SLI performances were corresponding to Ladkh language. It is also observed that averages of best case SLI performance found in both GMM for 15sec with 256 mixtures and GMM-UBM for 10sec with 512 mixtures as 41.67% and 46.74% respectively. This observation recorded an improvement in SLI performance of 5% to 6% with less training data and more number of mixture components.

Table 4. Language wise Best Identification Performance using GMM and GMM-UBM

Language	Identification Performance in %	
	GMM	GMM-UBM
	15sec	10sec
	256 mixtures	512 mixtures
Arunachali	34.22	39.39
Assamese	41.35	46.34
Bengali	46.07	51.49
Bhutia	39.46	45.28
Chattisgarhi	34.17	38.17
Dogri	40.21	44.10

Gujarati	48.76	52.24
Hindi	49.47	56.14
Indian English	50.26	55.09
Kannada	43.38	48.40
Konkani	34.30	39.87
Ladkh	31.88	37.56
Leptcha	33.61	39.24
Malayalam	47.20	52.02
Manipuri	43.94	47.41
Marathi	48.37	53.54
Odiya	40.14	45.25
Panjabi	40.35	47.10
Rajasthani	34.68	39.13
Sanskrit	41.62	46.28
Tamil	51.10	56.51
Telugu	47.57	52.43
Urdu	36.22	40.05
Average	41.67	46.74

SUMMARY AND FUTURE SCOPE

In this work, 23 major Indian languages speech data is used for SLI task. An UBM is created using 115 mins speech samples from all twenty three languages. The SLI performance has been carried out using both GMM method and GMM-UBM method. The GMM-UBM method has resulted in the highest SLI performance percentage for 10 sec duration with MFCC+Δ features as 46.74%. This method also improved average SLI performance by 5.07% with comparison to SLI using GMM method consuming only 4.67% amount of actual training speech data and with more number of mixture components.

The performance of SLI task may be increased further by combining language variable prosodic features along temporal characteristics. In the Indian Context the classification performance may be further improved by considering the family of the spoken language and language specific features in depth.

REFERENCES

[1] Ambikairajah, E., Li, H., Wang, L., Yin, B., and Sethu, V.: Language identification: a tutorial. Circuits and Systems Magazine, IEEE, 11(2), 2011, pp. 82–108.

[2] Chelba, C., Silva, J., and Acero, A.: Soft indexing of speech content for search in spoken documents. Computer Speech & Language, 21(3), 2007, pp. 458–478.

[3] Muthusamy, Y. K., Barnard, E., and Cole, R.: Reviewing automatic language identification. Signal Processing Magazine, IEEE, 11(4), 1994, pp. 33–41.

[4] Muthusamy, Y.K.: A Segmental Approach to Automatic Language Identification. Ph.D. thesis, Oregon Graduate Institute of Science and Technology, October 1993.

[5] Cimarusti, D., and Ives, R.B.: Development of an Automatic Identification System of Spoken Languages: Phase 1. Proc. ICASSP82, Vol. 7, May 1982, pp. 1661–1664.

[6] Navrtil, J.: Spoken Language Recognition A step Toward Multi-linguality in Speech Processing, IEEE Trans. Speech Audio Processing, Vol. 9, September 2001, pp. 678–685.

[7] Foil, J.T.: Language Identification Using Noisy Speech, Proc. ICASSP86, April 1986, pp. 861– 864.

[8] Hazen, T.: Automatic language identification using a segmented-based approach, Ph.D. Thesis, MIT, 1993.

[9] Thyme-Gobbel, A.E., Hutchins, S.E.: On using prosodic cues in automatic language identification, International Conference on Spoken Language Processing, Vol. 3, 1996, pp. 1768–1772.

[10] Bhaskararao, P.: Salient phonetic features of Indian languages in speech technology. Sadhana, 36(5), 2011, pp. 587–599.

[11] Schultz, T., Rogina, I., and Waibel, A.: LVCSR- Based Language Identification. Proc. ICASSP96, May 1996, pp. 781–784.

[12] Kadambe, S., Hieronymus, J.: Language identification with phonological and lexical models. Proc. ICASSP95, Vol. 5, 1995, pp. 3507-3511.

[13] Thomas, H.L., Parris, E.S., and Write, J.H.: Re- current substrings and data fusion for language recognition. International Conference on Spoken Language Processing, Vol. 2, 1998, pp. 169- 173.

[14] Quatieri, T. F.: Discrete-Time Speech Signal Processing: Principles and Practice. Engle- wood Cliffs, NJ, USA: Prentice-Hall, 2002.

[15] Zissman, M. A.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing, 4(1), 1996, 31.

[16] Pellegrino, F., and Andr-Obrecht, R.: Automatic language identification: an alternative approach to phonetic modelling. Signal Processing, 80(7), 2000, pp. 1231–1244.

[17] Torres-Carrasquillo, P. A., Reynolds, D., and Deller Jr, J. R.: Language identification using Gaussian mixture model tokenization. In Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE May 2002, pp. I–757.

[18] Mary, L., and Yegnanarayana, B.: Extraction and representation of prosodic features for language and speaker recognition. Speech communication, 50(10), 2008, pp. 782–796.

[19] Jothilakshmi, S., Ramalingam, V., and Palanivel, S.: A hierarchical language identification system for Indian languages. Digital Signal Processing, 22(3), 2012, pp. 544–553.

[20] Maity, S., Vuppala, A. K., Rao, K. S., and Nandi, D.: IITKGP-MLILSC speech database for language identification. In Communications (NCC), February 2012, pp. 1–5.

[21] Reddy, V. R., Maity, S., and Rao, K. S.: Identification of Indian languages using multi- level spectral and prosodic features. International Journal of Speech Technology, 16(4), 2013, pp. 489–511.

[22] Xu, Y., Yang, J., and Chen, J.: Methods to improve Gaussian mixture model for language identification. In Measuring Technology and Mechatronics Automation (ICMTMA), Vol. 2, IEEE March 2010, pp. 656–659.

[23] Gauvain, J. L., and Lee, C. H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE transactions on Speech and audio processing, 2(2), 1994, pp. 291–298.