IITKGP-MLILSC Speech Database for Language Identification

Sudhamay Maity[1], Anil Kumar Vuppala[2], K. Sreenivasa Rao[1], and Dipanjan Nandi[1]
[1]School of Information Technology
[2]G.S.Sanyal School of Technology, I.I.T Kharagpur, 721302, India
E-mail: friendsudha@gmail.com, anil.vuppala@gmail.com, ksrao@sit.iitkgp.ernet.in and dipanjanconnect.08@gmail.com

***Abstract*—In this paper, we are introducing speech database consists of 27 Indian languages for analyzing language specific information present in speech. In the context of Indian languages, systematic analysis of various speech features and classification models in view of automatic language identification has not performed, because of the lack of proper speech corpus covering majority of the Indian languages. With this motivation, we have initiated the task of developing multilingual speech corpus in Indian languages. In this paper spectral features are explored for investigating the presence of language specific information. Mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) are used for representing the spectral information. Gaussian mixture models (GMMs) are developed to capture the language specific information present in spectral features. The performance of language identification system is analyzed in view of speaker dependent and independent cases. The recognition performance is observed to be 96% and 45% respectively, for speaker dependent and independent environments.**

***Index Terms*—Language Identification, Indian Language Database, Mel-frequency cepstral coefficients (MFCCs), Gaussian mixture models (GMMs), Linear prediction cepstral coefficients (LPCCs).**

I. INTRODUCTION

The basic goal of the language identification (LI) system is to accurately identify the language from the given speech sample. Language identification has numerous practical applications such as: front end for automatic speech recognition, speech to speech translation, speech activated automated systems, and information retrieval from databases [1] [8] [10] [13]. In multilingual automatic speech recognition (ASR) system, LI system acts as a front end for recognizing the language of the input speech signal. After recognizing the language, speech signal will be passed to corresponding language ASR system. Similarly, in speech to speech translation system source language will be identified by using LI system, and then the message is translated into the target language, and the speech is generated in target language. One of the voice activated applications such as automatic call routing use LI system to recognize the language from the incoming call, and it will be routed to the desired application. For providing the information retrieval through queries in native language, LI system is needed to transform the speech queries from the native language to the formal language.

Language identification systems are broadly classified into two types [10], namely explicit and implicit LI systems. In explicit language identification system, phoneme sequence is derived from the given speech sample in the first step, later, based on the obtained phoneme sequence language is identified. In implicit LI system, language identity is determined directly, using language specific speech features. There is no need to determine the phoneme sequence prior to identify the language. Performance of explicit LI systems is high compared to their implicit counterparts, but implicit LI systems are less complex compared to explicit LI systems. In this work, implicit LI systems are used for analyzing language specific information present in speech.

A detailed survey of the state of the art language identification in view of speech features and models is given in [1]. Most of the approaches explored spectral and prosodic features for language identification task [2]–[8]. In the context of Indian languages, very few attempts are reported in the area of language identification. Jyotsna *et. al.,* [9] have first attempted to identify four Indian languages. They have used mel frequency cepstral coefficients (MFCC) as features and vector quantization models for language recognition task. Nagarajan *et. al.,* [10] have explored different code book methods for building LI system. Later, using automated segmentation of speech into syllable like units and parallel syllable like unit recognition are used to build implicit language identification system. Leena *et. al.,* [11]–[13] have explored spectral features with auto associative neural network models for language identification with varying durations of test speech samples. In their later work, they focused on prosodic (intonation and rhythm) and syllabic features for language recognition.

Building an efficient LI system for Indian languages is lacking due to unavailability of proper database covering majority of languages. In the context of Indian languages most of the works related to LI systems are limited to 4 languages. So to overcome this deficiency, we have initiated the task of developing multilingual speech corpus in Indian languages. We have collected speech corpus for 27 Indian languages and details of the database are provided in the following section. The language specific characteristics of the speech can be attributed to the characteristics of vocal tract system, behavioral characters of speaker, excitation source characteristics and suprasegmental characteristics. Language specific vocal tract information is mainly represented by spectral features such as Mel-frequency cepstral coefficients (MFCC) or linear predictive cepstral coefficients (LPCC). The parameters like pitch, duration and energy can be used as

978-1-4673-0816-8/12/$31.00 ©2012 IEEE

basic prosodic features to represent language specific prosodic information. In this work, only spectral features are analyzed for evaluating the performance of LI system developed using the proposed database. Gaussian mixture models (GMMs) are used to capture the language specific distributions from the derived spectral features.

Rest of the paper is organized as follows. The details of proposed IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus) are discussed in Section II. Section III, discusses about development of various LI systems used in this work. Evaluation of the developed LI systems is discussed in Section IV. Summary of the present work and the future directions to extend present work are provided in the final section of the paper.

II. IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus)

The proposed database is recorded from broadcast television channels using DISHTV direct to home connection. For some languages, where TV broadcast channels are not available, broadcast radio channels are used for collecting the speech. The speech corpus is recorded from news bulletins, interviews, live shows and talk shows. Around 80% of the database is collected from news bulletins. Database covers 27 Indian languages, out of which 16 popular (widely spoken) languages: Assamese, Bengali, Bhojpuri, Gujarati, Hindi, Indian English, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Tamil, Telugu and Urdu are considered in this paper for language identification task. From each speaker, about 5-10 minutes of speech is collected. On the whole, each language contains minimum of 1 hour speech data. Details of the database are described in Table I. Pixelview TV tuner card is used to access the required TV channels through the computer, using VentiTV software. Audacity software is used to record the speech from TV channels. Speech data from broadcast radio channels are collected from archives of Prasar Bharati, All India Radio. The main reasons for choosing TV and Radio channels for collecting the proposed database are: (1) It is difficult to get sufficient number of native speakers for each of the above mentioned languages, in a single place, (2) It is highly time consuming to record the required speech data from different speakers at different places, (3) The quality of speech data used in broadcast TV channels is observed to be clean and noise free and (4) Speakers from TV channels are more professional and matured. Since, most of the speakers in the database are well matured and professional, hence the collected speech corpus ensures standard quality in terms of articulation, speaking rate and pronunciation. The speech signal is recorded with the sampling rate of 16 kHz, and each sample is stored as 16 bit number. The speech recorded through TV channels may contain some inherent problems such as (1) News bulletins contain background music during headlines, background videos and commercials, and (2) During talk shows and interviews, there are chances of overlapping of speech of different speakers. Therefore, while recording and editing the database proper care has been taken to minimize the above mentioned problems.

TABLE I
DESCRIPTION OF INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR - MULTI LINGUAL INDIAN LANGUAGE SPEECH CORPUS (IITKGP-MLILSC)

Language	Region	Speaking Population @2001 census (Mil)	#Speakers		Duration (Minutes)
			F	M	
Arunachali	Arunachal Pradesh	0.41	6	15	72.00
Assamese	Assam	13.17	6	8	67.33
Bengali	West Bengal	83.37	14	10	69.78
Bhojpuri	Bihar	38.55	5	7	59.82
Chhattisgarhi	Chhattisgarh	11.50	9	11	70.00
Dogri	Jammu and Kashmir	2.28	8	12	70.00
Gojri	Jammu and Kashmir	20.00	3	12	44.00
Gujrati	Gujarat	46.09	7	6	48.96
Hindi	Uttar Pradesh	422.05	14	24	134.70
Indian English	All over India	125.23	12	13	81.66
Kannada	Karnataka	37.92	4	8	69.33
Kashmiri	Jammu and Kashmir	5.53	2	19	59.64
Konkani	Goa and Karnataka	2.49	5	15	50.00
Manipuri	Manipur	1.47	11	11	64.00
Mizo	Mizoram	0.67	3	8	48.00
Malyalam	Kerala	33.07	7	12	81.09
Marathi	Maharashtra	71.94	7	9	74.33
Nagamese	Nagaland	0.03	11	9	60.00
Neplai	West Bengal	2.87	7	6	54.19
Oriya	Orissa	33.02	10	4	59.87
Punjabi	Punjab	29.10	7	10	80.91
Rajasthani	Rajasthan	50.00	10	10	60.00
Sanskrit	Uttar Pradesh (UP)	0.014	0	20	70.00
Sindhi	Gujarat and Maharashtra	2.54	14	6	50.00
Tamil	Tamil Nadu	60.79	7	10	70.96
Telugu	Andhra Pradesh (AP)	74.00	7	8	73.72
Urdu	UP and AP	51.54	5	16	86.49

III. Development of Language Models

In this work, GMMs are used for developing the language identification (LI) systems using spectral features. GMMs are well known to capture the distribution of data in the feature space [8]. The accuracy in capturing the true distribution of data depends on various parameters such as dimension of feature vectors, number of feature vectors and number of mixture components. In this work, GMMs are assumed to capture the language specific information from the given spectral features.

In this work, we have explored two popular spectral features namely, linear prediction cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC). 13 MFCC and 13 LPCC features are derived from a speech frame of 20 ms with a frame shift of 10 ms. For deriving the MFCCs, 24 filter bands are used. LPCCs are derived using 16^{th} order LP filter [8]. For developing the language identification system using GMMs, we need to develop a specific GMM for each of the language. These GMMs are developed using the spectral vectors derived from the speech corresponding to the languages considered. In this work, 16 Indian languages are considered for analyzing the language identification performance using spectral features. Therefore, the proposed LI system consists of 16 GMMs (language models) developed using speech corresponding to 16 languages. The block diagram of the proposed language identification system is shown in Fig.1 For evaluating the developed LI system, feature vectors derived from test speech samples are given to all the language models. The evidence in response to test speech samples from all the models are analyzed, and the highest evidence among the models is hypothesized as the language identity corresponding to the given speech sample.

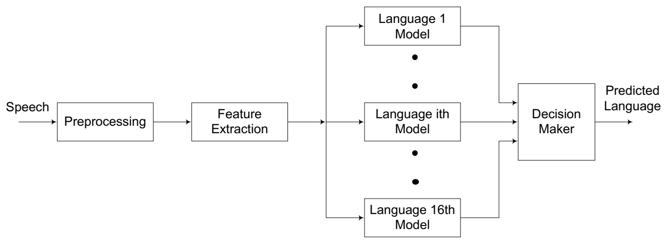

Fig. 1. Language Identification System

IV. Evaluation of Language Models

In this work, about 80% of the data is used for training the GMMs (developing the language models), and the rest 20% data is used for testing or evaluating the performance of developed language models. Initially, we have developed LI systems separately, using LPCC and MFCC features. In this study, LI systems are developed in speaker dependent and speaker independent environments. In speaker dependent LI system, speech from all the speakers of a given language is used for developing and evaluating the models. Whereas for speaker independent LI system, speaker set used for developing the models or LI system is different from the set of speakers used for evaluating the models. The following subsections discuss more details about speaker dependent/independent LI systems.

A. Speaker dependent LI system

In this system, 80% of speech from all the speakers of a given language is used for developing the specific language model. For evaluating the developed speaker dependent LI system, 20% of speech from each speaker (which is not used for developing the models) is used. For analyzing the effect of number of mixture components of GMMs on the identification performance, various LI systems are developed by varying the number of mixture components of GMMs from 32 to 512. Similarly, for analyzing the effect of length of test speech utterance on identification performance, three different lengths (5, 10 and 20 secs) of test speech utterances are analyzed. The performance of speaker dependent LI system with different spectral features, different mixture components and different lengths of test speech utterances is given in Table II. In Table II, first column shows the number of mixture components used for developing the LI system. Columns 2-4, indicate the performance of LI system developed using MFCC features for different durations of test speech utterances. Columns 5-7, indicate the performance of LI system developed using LPCC features for different durations of test speech utterances. From the results, it is observed that the identification performance has improved significantly, for both LPCC and MFCC based systems while increasing the number of mixture components from 32 to 256, and beyond 256 mixture components a slight improvement is observed. The improvement in identification rate by increasing the number of mixtures from 256 to 512 is not worth in view of time complexity associated to building the models. Language recognition performance using MFCC features seems to be slightly better compared to LPCC features. But, it is observed that LPCC features seems to perform better compared to MFCC features for the models built with lower number of mixture components (32 to 256). In view of test utterance duration, the performance seems to be better at 10 secs, compared to 5 and 20 secs.

TABLE II
Performance of Speaker Dependant LI System

No of Components	Average Recognition Performance (%)					
	MFCC			LPCC		
	5 sec	10 sec	20 sec	5 sec	10 sec	20 sec
32	73.19	77.34	78.12	79.37	82.46	81.67
64	80.35	85.23	84.92	86.79	89.12	88.76
128	87.29	90.33	90.04	89.96	91.32	90.97
256	90.32	94.72	93.89	91.62	**94.37**	93.92
512	91.78	**95.83**	94.65	88.64	94.22	94.16

B. Speaker independent LI system

In this system, one male and one female speakers' speech data in each language is omitted during development of language models. In speaker dependent LI system, all the speakers present during training stage are also present during testing stage. Hence, we may observed the better recognition performance. But, in practice LI system should perform well even under speaker independent environment. Therefore, to analyze the LI performance in speaker independent environment, we have developed and evaluated the LI system with nonoverlapping speaker sets. For testing the speaker independent LI system, speech utterances of 2 (one male and one female) speakers (who are not involved during training) from each language are used. The performance of speaker independent LI system is given in Table III. From the results presented in Table III, it is observed that the performance of LI system in speaker independent environment is drastically reduced compared to speaker dependent environment (see Tables II and III). From this observation, it may be noted that LI system in speaker dependent environment has captured speaker specific information in addition to language specific information. In speaker independent environment, LI system with MFCC features has achieved the highest recognition performance of 37% with 10 secs test utterance speech duration and 128 Gaussian mixture components. LI system with LPCC features seems to perform better compared to MFCC features. The language recognition performance with LPCC features is observed to be about 41% with 32 mixture components and test utterances with 10 secs duration. From the results present in Tables II and III, it is observed that there exists a strong speaker bias in the developed language models. Therefore, the identification performance is superior (about 96%) in the case of speaker dependent environment, and inferior (about 41%) in the case of speaker independent environment.

TABLE III
RECOGNITION PERFORMANCE OF SPEAKER INDEPENDENT LI SYSTEM

No of Components	Average Recognition Performance (%)					
	MFCC			LPCC		
	5 Sec	10 Sec	20 Sec	5 Sec	10 Sec	20 Sec
2	28.75	30.54	31.53	21.94	23.37	21.16
4	26.70	26.52	26.52	22.13	25.05	22.93
8	27.83	29.09	29.25	31.62	35.63	36.00
16	33.46	34.22	33.70	30.05	36.46	33.69
32	33.12	35.24	35.22	**34.83**	**40.68**	**38.41**
64	**35.62**	36.26	36.62	32.37	40.00	38.13
128	34.15	**36.85**	**36.93**	30.93	38.66	36.09
256	33.86	35.68	36.28	33.03	38.65	37.67
512	33.07	34.76	36.59	34.62	40.17	37.89

C. Speaker independent LI system with speaker specific language models

For improving the performance of LI system in speaker independent environment, we have proposed speaker specific language models in this work. In this framework, each language model is represented by set of speaker models associated to that particular language. The block diagram of the LI system with proposed speaker specific language models is given in Fig. 2. For analyzing the language identification performance in speaker independent environment, language models are developed with multiple speaker models by omitting two speakers (one male and one female) from each language. The process of building language models is similar to speaker independent LI system discussed above.

The basic idea behind the proposed speaker specific language models is that, the multiple evidence from each language model (see Fig. 2) can be exploited in a better way compared single evidence from each language model (see Fig.1). In speaker independent case, it is known that speakers during training and testing are mutually exclusive, and the common information may present in both training and testing is only language specific information. Hence, our hypothesis is that the evidences from the matched language is high compared to non-matched languages. The proposed speaker specific language models are analyzed with respect to various spectral features and varying Gaussian mixture components. In this study, the length of test speech utterances is chosen as 10 secs. The evidence from multiple speaker models of each language are explored in various ways such as (a) maximum of the evidence, (b) mean of the evidence, (c) sum of best 2, 3, 4 and 5 evidences and (d) maximum plus mean.

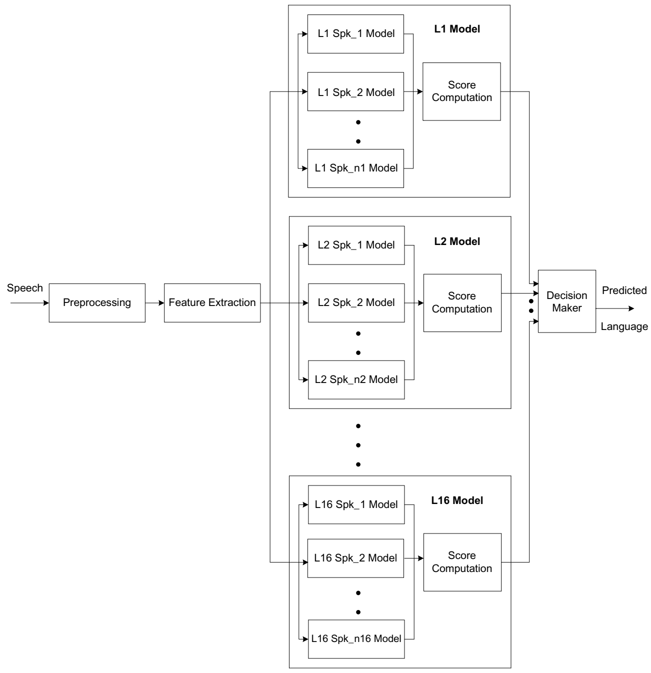

Fig. 2. Language identification system with speaker specific models

The identification performance of the speaker independent LI system developed using speaker specific language models is given in Table IV. Column 1 indicates the number of Gaussian mixture components used for developing various LI systems. Columns 2-8, indicate the % of recognition by exploring the available multiple evidence from each language model in various ways. From the results, it is observed that the identification performance of the speaker independent LI system using the proposed speaker specific models with MFCC and LPCC features is about 45% and 36% respectively. Among the various strategies to exploit the multiple evidences, maximum evidence from multiple speaker models seems to be out performed compared to other strategies.

For more detailed analysis of the recognition results, we have also examined the recognition performance of individual languages, for the best performed LI systems (32 component MFCC system and 64 component LPCC system (see Table IV)). Table V shows the recognition performance of the 16 Indian languages, individually. We have also examined the recognition performance by considering the desired language in either rank 1 or rank 2 or rank 3 position. Column 2, 3 and 4 indicate the recognition performance of individual languages within rank 1, rank 2 and rank 3 positions respectively. From the results, it is observed that the average recognition performance has been improved from 45% to 62% by considering the top 3 ranks. Among the 16 languages, Telugu, Urdu, Marathi and Malayalam have achieved the recognition performance less than 20%, Hindi, Assamese and Gujarati have achieved the recognition performance in between 25-45% and rest have achieved more than 50% recognition performance based on rank 1 (see column 2). From column 4

TABLE IV
RECOGNITION PERFORMANCE OF SPEAKER INDEPENDENT LI SYSTEM DEVELOPED USING SPEAKER SPECIFIC MODELS

No of Components	Average Recognition Performance (%)						
	Max	Mean	Top 2	Top 3	Top 4	Top 5	Max + Mean
MFCC							
2	37.66	29.39	37.68	32.87	29.60	23.17	32.78
4	37.57	32.03	39.74	35.03	29.07	24.50	34.09
8	42.78	31.13	41.57	37.30	32.99	27.70	36.37
16	44.16	29.78	40.83	37.93	34.51	27.94	34.44
32	**44.92**	28.65	41.35	35.56	34.82	26.03	35.26
64	41.27	29.10	41.36	35.74	32.81	26.09	33.63
128	39.77	23.22	37.66	34.58	29.39	24.69	31.73
LPCC							
2	29.11	21.36	28.69	29.38	25.55	22.22	24.49
4	30.57	23.37	30.67	28.83	23.52	23.59	31.21
8	34.66	25.48	32.48	29.82	23.19	21.77	34.37
16	35.21	29.30	32.04	29.46	23.27	21.19	35.88
32	34.55	31.53	31.63	27.35	25.81	22.86	34.28
64	**35.96**	30.32	31.79	28.51	24.68	21.76	33.68
128	34.09	30.99	33.60	29.07	25.58	20.92	32.67

of Table V, it can be noted that the recognition performance of Telugu, Urdu and Malayalam has still less than 30% even after considering top 3 choices. The cause for the poor performance of the above mentioned 3 languages has to be thoroughly investigated by examining the quality of test speech samples and by using other speaker sets.

TABLE V
ANALYSIS OF RECOGNITION PERFORMANCE OF SPEAKER INDEPENDENT LI SYSTEM FOR BEST 3 CHOICES

Language	k-Best Performance					
	MFCC			LPCC		
	k=1	k=2	k=3	k=1	k=2	k=3
Assamese	31.82	31.82	45.45	100.00	100.00	100.00
Bengali	50.00	54.55	68.18	25.00	40.00	40.00
Bhojpuri	77.14	82.86	88.57	27.27	30.30	27.27
Gujarati	41.67	66.67	66.67	75.00	75.00	75.00
Hindi	27.14	44.29	47.14	6.15	13.85	35.39
Indian English	52.71	58.21	65.70	40.12	48.51	60.15
Kannada	72.00	76.00	76.00	76.00	80.00	80.00
Kashmiri	54.84	74.19	74.19	46.15	57.69	65.39
Malayalam	18.75	25.00	28.13	3.03	24.24	39.39
Marathi	7.14	50.00	71.43	0.00	0.00	29.63
Nepali	96.88	100.00	100.00	40.74	66.67	77.78
Oriya	61.70	61.70	61.70	60.00	62.50	92.50
Punjabi	56.52	65.22	78.26	32.00	56.00	92.00
Tamil	53.19	68.09	85.11	42.22	64.44	73.33
Telugu	6.67	13.33	13.33	0.00	0.00	0.00
Urdu	18.37	20.41	22.45	4.65	30.23	27.91
Average	44.92	55.61	61.77	35.96	46.73	57.04

V. SUMMARY AND CONCLUSIONS

In this paper we have proposed IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus) speech database for language identification task in the context of Indian languages. The proposed speech database consists of 27 Indian languages, 16 of them which are widely spoken were considered in this study for analyzing the language identification performance. The language identification system is analyzed with speaker dependent and independent environments. The performance of LI system in speaker independent environment is observed to be very low compared to speaker dependent environment. The dominant performance in speaker dependent case is mainly due to speaker bias in the developed language models. The performance of speaker independent LI system has been improved by proposing multiple speaker models for each language.

The obtained recognition performance of the developed LI system is slightly inferior compared to state of the art. The main reason may be that all Indian languages are derived from Sanskrit, and hence there exists lot of overlapping of linguistic characteristics and it leads to ambiguity in discriminating them. In this paper, we have just explored GMMs and spectral features for identifying the languages. In addition to spectral features, one can explore language specific prosodic and source features for identifying the languages. The recognition performance may be further improved by combining spectral, prosodic and source information effectively. Other nonlinear models such as neural networks, support vector machines and hidden Markov models can be explored in future for further improvement in the performance.

REFERENCES

[1] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin and Vidhyasaharan Sethu, "Language Identification: A Tutorial", *IEEE Circuits and Systems Magazine*, 2011, Vol.11, Issue 2, pages 82-108.

[2] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds and J.R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features", *Proc. ICSLP*, 2002, pages 89-92.

[3] E.Wong, " Automatic Spoken Language Identification utilizing Acoustic and Phonetic Speech Information", *Ph.D Dissertation, Speech and Audio Research Laboratory, Queensland Univ. Technol.*, 2004.

[4] F. Allen, "Automatic Language Identification", *Thesis, School of Electrical Engineering and Telecommunication, The Univ. New South Wales*, 2005.

[5] Chi-Yueh Lin and Hsiao-Chuan Wang, "Language Identification using Pitch Contour Information in the Ergodic Markov Model", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pages 193-196.

[6] Haizhou Li, Bin Ma and Chin-Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification", *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, Vol.15, pages 271-284.

[7] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro and C. Vair ,"Language Identification using Acoustic Models and Speaker Compensated Cepstral-Time Matrices", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pages 1013-1016.

[8] J. Benesty, M. Sondhi and Y. Huang, "Springer Handbook of Speech Processing", *Springer-Verlag*, 2007, New York.

[9] Jyotsna Balleda, Hema A. Murthy, and T. Nagarajan, "Language identification from short segment of speech", *Proc. ICSLP-2000*, 2000, Vol.III, pages 1033-1036.

[10] T. Nagarajan and Hema A. Murthy, "Language identification using spectral vector distribution across languages", *Proc. International Conference on Natural Language Processing*, 2002, pages 327-335.

[11] Leena Mary and B. Yegnanarayana, "Autoassociative neural network models for language identification", *Proc. Int. Conf. Intelligent Sensing and Information Processing*, 2004, Chennai, India, pages. 317-320.

[12] Leena Mary, K. Sreenivasa Rao and B. Yegnanarayana, "Neural Network Classifiers for Language Identification using Syntactic and Prosodic features", *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing*, 2005, Chennai, India, January 4-7, pages 404-408.

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