Parametric representation of excitation source information for language identification

Article *in* Computer Speech & Language · June 2016

DOI: 10.1016/j.csl.2016.05.001

Dipanjan Nandi
University of Alberta
17 PUBLICATIONS 95 CITATIONS

Debadatta Pati
National Institute of Technology Nagaland
22 PUBLICATIONS 159 CITATIONS

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 88–115

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Parametric representation of excitation source information for language identification

Dipanjan Nandi [a,*], Debadatta Pati [b], K. Sreenivasa Rao [a]

[a] School of Information Technology, Indian Institute of Technology, Kharagpur 721302, West Bengal, India

[b] Department of Electronics and Communication Engineering, National Institute of Technology, Nagaland 797103, India

Received 11 August 2015; received in revised form 5 May 2016; accepted 12 May 2016

Available online 16 June 2016

Abstract

In this work, the linear prediction (LP) residual signal has been parameterized to capture the excitation source information for language identification (LID) study. LP residual signal has been processed at three different levels: sub-segmental, segmental and supra-segmental levels to demonstrate different aspects of language-specific excitation source information. Proposed excitation source features have been evaluated on 27 Indian languages from Indian Institute of Technology Kharagpur-Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC), Oregon Graduate Institute Multi-Language Telephone-based Speech (OGI-MLTS) and National Institute of Standards and Technology Language Recognition Evaluation (NIST LRE) 2011 corpora. LID systems were developed using Gaussian mixture model (GMM) and *i*-vector based approaches. Experimental results have shown that segmental level parametric features provide better identification accuracy (62%), compared to sub-segmental (40%) and supra-segmental level (34%) features. Excitation source features obtained from three levels show distinct language-specific evidence. Therefore, the scores from all three levels are combined to obtain the complete excitation source information for the LID task. LID performances achieved from both the excitation source and vocal tract system are compared. Finally, the scores obtained by processing the vocal tract and excitation source features are combined to achieve better improvement in LID accuracy. The best recognition accuracies obtained from stage-IV integrated LID systems I, II and III are 69%, 70% and 72% respectively.

Keywords: Excitation source information; Language identification (LID); LP residual; GFD parameters; RMFCC; MPDSS; Pitch contour; Epoch strength contour; MFCC; *i*-Vector; IITKGP-MLILSC; OGI-MLTS; NIST LRE; C_{avg}

1. Introduction

Human speech is intended to convey messages. Speech signal carries not only the message information but also the information about the speaker, language and emotion. The primary objective of a language identification (LID) task is to determine the identity of language from the uttered speech. Due to several real-life applications of automatic LID systems such as speech to speech translation systems, information retrieval from multilingual audio databases and multilingual speech recognition systems, it has become an active research problem. Indian languages belong to several language groups and sub-groups. The major two language groups are the Indo-Aryan languages spoken by 76.86% of Indian citizens and the Dravidian languages spoken by 20.82% Indians (Vanishree, 2011). Most of the

[*] Corresponding author at: School of Information Technology, Indian Institute of Technology, Kharagpur 721302, West Bengal, India. Tel.: +91-3222-282336; fax: +91-3222-282206.

E-mail address: dipanjanconnect.08@gmail.com (D. Nandi).

http://dx.doi.org/10.1016/j.csl.2016.05.001

0885-2308/© 2016 Elsevier Ltd. All rights reserved.

languages in India have a common set of phonemes and also follow similar grammatical structure. To develop a language identification system, it is necessary to derive non-overlapping language-specific information for each language. Therefore, building an automatic LID system in the Indian context is quite a challenging task.

Human speech production system has two major components: vocal tract system and source of excitation. The constriction caused by expiration of air acts as an excitation source during the production of speech. The quasi-periodic air pulses generated by the vocal folds vibration act as the primary source of excitation to vocal tract resonator during voiced speech production. During the production of unvoiced speech, the expiration of air is constrained either completely (e.g. unvoiced stops) or partially (e.g. fricatives). In speech production, the majority of excitation takes place during the production of voiced speech. This excitation source information can be captured by passing the speech signal through the inverse filter (Makhoul, 1975). We have used 10th order linear prediction (LP) analysis followed by inverse filtering the speech signal (sampled at 8 kHz) for estimating LP residual signal. Both the vocal tract system and excitation source have significant contribution in the production of voiced speech. Most of the contemporary works exploited the vocal tract system characteristics to determine the language-specific cues from the speech signal. The dynamics of vocal tract system are captured by spectral analysis of speech, which can be represented by Mel-frequency cepstral coefficients (MFCCs) (Balleda et al., 2000), linear prediction coefficients (LPCs) and linear prediction cepstral coefficients (LPCCs) (Sugiyama, 1991). Language related prosodic features extracted from syllable, word and sentence levels have also been explored in recent works (Reddy et al., 2013). However, the source characteristics have not been investigated to obtain language-specific information present in a speech signal.

In this work, the excitation source information has been studied for developing automatic LID system. Excitation source features are extracted from the linear prediction residual (LPR) signal (Makhoul, 1975). In the present study, LP residual signal has been processed at three different levels: within a glottal cycle known as sub-segmental (*sub*) level information, within 2–3 glottal cycles known as segmental (*seg*) level information and across 50 glottal cycles known as supra-segmental (*supra*) level information. The glottal flow derivative (GFD) parameters are extracted from LP residual signal to capture the *sub* level excitation source information. The energy and periodicity of excitation source are captured by parameterizing the LPR signal at *seg* level. The pitch and epoch strength contour information are obtained by processing the LPR signal at *supra* level. The excitation source parameters derived from different levels may capture some non-overlapping language-specific information. Therefore, scores from different levels are combined to obtain the complete parametric excitation source information. The LID accuracies achieved by proposed excitation source features are also compared with the LID performance obtained by processing vocal tract information represented by MFCCs. The vocal tract system and source for exciting the vocal tract are two different components of speech production system. Hence, non-overlapping language-specific information may be present in these two components. Therefore, the combination of these two information sources may provide improved LID accuracy. For developing the systems GMM and *i*-vector based approaches are explored. It is observed that *i*-vector based systems work better, compared to GMM based system. The significance of excitation source features is also examined by investigating five different dialects of Hindi language. Excitation source features are also evaluated on OGI-MLTS and NIST LRE 2011 corpora.

The rest of the paper is organized as follows: In Section 2, earlier works on LID are discussed. Section 3 presents the motivation for the present work. A brief description of the language databases used in this work has been provided in Section 4. Proposed parametric representation of excitation source is described in Section 5. In Section 6, experimental setup and methodology have been explained. Description of LID systems developed using the proposed features is laid out in Section 7. Evaluation of LID systems using proposed excitation source features is given in Section 8. Section 9 concludes the present work.

2. Previous works

This section presents a brief overview of existing works on language identification systems. Sugiyama (Sugiyama, 1991) has explored linear prediction coefficients (LPCs) and cepstral coefficients (LPCCs) for language recognition. Morgan et al. (1992) and Zissman (Zissman, 1993) have proposed the Gaussian mixture models (GMMs) (Reynolds and Rose, 1995) for language identification study. In the Indian context, Balleda et al. (2000) have first attempted to identify Indian languages automatically using speech signal. In their work, vector quantization (VQ) has been used for classification purpose and MFCC features have been used to represent the language-specific information. Vector quantization is used to represent large number of multi-dimensional feature vectors into few representative vectors known to be code vectors. These code vectors are derived from sufficiently large number of vectors using clustering methods. These code vectors usually represent the centroids of the clusters. The set of all the code vectors is known as code book. In case of language identification, a code book is prepared for each language and the testing is performed by analyzing the distances offered by various code books. Detailed explanation about VQ can be found in Rabiner and Juang (1993). In 2004, Leena Mary and B. Yegnanarayana have explored the auto-associative neural networks (AANNs) for capturing language-specific features for developing LID system (Mary and Yegnanarayana, 2004). They have also explored prosodic features for capturing the language-specific information (Mary, 2006). Rao et al. (2013) have explored spectral features using block processing (20 ms block size), pitch synchronous and glottal closure region (GCR) based approaches for discriminating 27 Indian languages. The language-specific prosodic features have also been explored by Reddy et al. (2013) for discriminating 27 Indian languages. In this work, prosodic features are extracted from syllable, word and sentence levels to capture language-specific prosodic knowledge. Jothilakshmi et al. (2012) have explored a hierarchical approach for identifying the Indian languages. This method first identifies the language group of a given test utterance and then identifies the particular language within that group. They have carried out the LID task by using different acoustic features such as MFCC, MFCC with velocity and acceleration coefficients, and shifted delta cepstrum (SDC) features.

Singh et al. (2013) explored *i*-vector based approach to classify sparse language data. For comparison purpose another LID system was also developed using GMM mean shifted supervectors from all training data. LID studies were carried out on NIST LRE 2007 dataset. Li and Narayanan (2014) proposed simplified and supervised *i*-vector based approach with applications to language identification and speaker verification. The level-regularized supervised *i*-vectors are created from traditional *i*-vectors by concatenating the label vector and the linear regression matrix at the end of *i*-vector factor loading matrix. The two LID systems, simplified and supervised *i*-vector based systems, were fused and the hybrid system outperformed the simplified *i*-vector based LID system both in robustness and efficiency aspects. Furthermore, Gammatone frequency cepstral coefficients (GFCC) and Gabor features have been incorporated with the traditional MFCC and SDC features based system, which significantly improved the efficiency of the LID system. Travadi et al. (2014) proposed novel techniques to handle the problems of *i*-vector variability for short duration segments. The prior distributions of *i*-vectors are modified. Dehak et al., (2011a) explored factor analysis for speaker verification task. Dehak et al. (2011b) proposed total variability subspace approach for language identification task. Moreno et al. (2014) developed LID system using deep neural networks (DNNs) and proposed approach was compared with the *i*-vector based LID system. Google 5M LID corpus and NIST LRE 2009 were used for LID study. They achieved relative improvement up to 70% of average cost (C_{avg}) metric as defined by NIST LRE 2009 evaluation plan. Dominguez et al. (2014) explored Long Short-Term Memory (LSTM) recurrent neural networks (RNNs). The proposed approach is compared to the baseline *i*-vector and feed forward Deep Neural Network (DNN). NIST LRE 2009 dataset was used for LID study. Furthermore, the fusion of different systems provides LID performance improvement up to 28%.

From the previous studies, it is observed that most of the LID studies have explored only spectral and prosodic features, and the excitation source component of speech has not been explored. However, the excitation source features have been explored for robust speaker recognition (Gupta et al., 2002; Pati and Prasanna, 2011, 2013; Yegnanarayana et al., 2005), audio clip classification (Bajpai and Yegnanarayana, 2004), speech enhancement (Yegnanarayana et al., 1997) and emotion recognition (Rao and Koolagudi, 2013) tasks. In the present study, we have parameterized the LP residual signal at sub-segmental, segmental and supra-segmental levels to capture different aspects of excitation source for language discrimination task.

3. Motivation

State-of-the-art LID systems mostly use the vocal tract information for discriminating the languages. However, the characteristics of the excitation source and articulatory constraints are distinct for each sound unit. Although there is a significant overlap in the set of sound units in different languages, the characteristics of the same sound unit may differ across different languages due to the co-articulation effects and phonotactic constraints. Hence, we conjecture that the characteristics of excitation source may contain some language related information, and this source information has not been explored so far for identifying the languages. The motivation for using excitation source features for LID task can also be observed from Table 1. In this section, the significance of the excitation source information for language identification task is shown by their respective correlation coefficients for within and between

Table 1
Correlation coefficients across the languages derived using glottal flow derivative (GFD), residual mel frequency cepstral coefficient (RMFCC), mel power differences of spectrum in sub-bands (MPDSS), pitch contour (PC) and epoch strength contour (ESC) features.

Languages	Correlation coefficients									
	sub		seg				supra			
	GFD		RMFCC		MPDSS		PC		ESC	
	WL	BL	WL	BL	WL	BL	WL	BL	WL	BL
Arunachali	0.85	0.24	0.15	0.06	0.27	0.16	0.37	0.26	0.81	0.69
Assamese	0.43	0.14	0.09	0.06	0.92	0.35	0.37	0.18	0.65	0.29
Bengali	0.66	0.28	0.12	0.06	0.53	0.27	0.58	0.31	2.21	0.92
Bhojpuri	0.74	0.42	0.08	0.06	0.91	0.55	0.21	0.09	2.17	1.38
Chattisgarhi	0.63	0.16	0.10	0.02	0.17	0.05	0.26	0.12	0.96	0.29
Dogri	0.71	0.35	0.28	0.10	0.47	0.19	0.20	0.13	2.92	0.91
Gojri	0.52	0.13	0.03	0.02	0.82	0.17	0.36	0.18	0.71	0.33
Gujarati	0.59	0.25	0.05	0.05	0.28	0.08	0.28	0.15	0.45	0.21
Hindi	0.39	0.22	0.04	0.04	0.70	0.19	0.43	0.16	0.77	0.34
Indian English	0.45	0.13	0.12	0.07	0.56	0.14	0.93	0.14	0.98	0.43
Kannada	1.21	0.77	0.09	0.06	0.25	0.09	1.21	0.79	0.57	0.25
Kashmiri	1.08	0.73	0.08	0.06	0.51	0.25	0.72	0.22	0.99	0.47
Konkani	0.82	0.41	0.01	0.02	0.46	0.12	0.81	0.15	0.61	0.41
Malayalam	0.73	0.31	0.12	0.08	0.71	0.19	0.41	0.20	0.41	0.21
Manipuri	0.91	0.26	0.11	0.07	0.16	0.09	0.26	0.08	0.65	0.41
Marathi	0.41	0.24	0.09	0.07	0.75	0.36	0.45	0.16	1.29	0.46
Mizo	0.53	0.14	0.36	0.11	0.43	0.14	0.23	0.12	0.48	0.19
Nagamese	0.85	0.35	0.34	0.11	0.72	0.28	0.33	0.25	0.70	0.23
Nepali	1.12	0.74	0.01	0.02	0.35	0.08	0.75	0.11	1.30	0.46
Oriya	1.17	0.35	0.15	0.08	0.67	0.26	0.30	0.13	0.32	0.11
Punjabi	1.10	0.44	0.05	0.03	0.14	0.06	0.39	0.27	0.44	0.12
Rajasthani	0.59	0.27	0.05	0.04	0.76	0.31	0.27	0.06	0.61	0.26
Sanskrit	1.34	0.81	0.22	0.08	0.68	0.36	0.25	0.19	1.23	0.76
Sindhi	1.01	0.70	0.32	0.13	0.64	0.19	0.29	0.16	0.97	0.35
Tamil	0.84	0.44	0.06	0.05	0.58	0.22	0.81	0.46	0.75	0.34
Telugu	0.52	0.19	0.01	0.01	0.39	0.24	0.38	0.19	0.57	0.15
Urdu	0.94	0.51	0.17	0.08	0.49	0.28	0.34	0.15	0.88	0.52

languages. Correlation determines the degree of similarity between two signals. Suppose that we have two real signal sequences $x(n)$ and $y(n)$ each of which has finite energy. The *cross-correlation* of $x(n)$ and $y(n)$ is a sequence $r_{xy}(l)$, which is defined as follows:

$$r_{xy}(l)=\sum_{n=1}^{p} x(n)y(n-l), \quad l=0,\pm 1,\pm 2,\ldots \tag{1}$$

where the l is the time shift parameter and p is the total number of samples. The x and y are the two signals being correlated. If the signals are identical, then the correlation coefficient is maximum and if they are orthogonal then the correlation coefficient is minimum. When $x(n)=y(n)$, then their correlation is known as *autocorrelation* of $x(n)$.

From each language database, one male speaker's data of 5 minutes duration is considered, and the LP residual has been extracted. The LP residual signal is processed to estimate the *seg* level energy information represented by Residual Mel Filter Cepstral Coefficients (RMFCC). To normalize the speaker variability across languages, the mean subtraction is imposed on all the feature vectors across all languages. Then the RMFCC feature vectors are modeled with 128 Gaussian mixtures for each language. The average of 128 *mean* vectors is considered as the signal for a particular language to compute the correlation coefficients. The values in the 4th column of Table 1 indicate the correlation coefficients within the languages (abbreviated as WL). The correlation coefficient within a language has been computed from two different speech utterances spoken by a speaker. The first element of the 4th column indicates the correlation coefficient of Arunachali language. The first element of the 5th column indicates the average of cross-correlation coefficients of Arunachali language with respect to other 26 languages. The lower average cross-correlation coefficient value between the languages (abbreviated as BL) indicates more dissimilarity between Arunachali and the rest of the languages. This average cross-correlation coefficient value of Arunachali language with respect to other 26 languages (0.06) is less than the correlation coefficient value of Arunachali language (0.15). This portrays that the *seg* level excitation source feature has significant language discriminative capability. Similar characteristics can be observed in other languages also. The correlation coefficients computed from *sub* and *supra* levels' excitation source features also show significant language discrimination capability.

The correlation coefficients are also computed using *sub* level Glottal flow derivative (GFD) feature, *seg* level Mel power differences of spectrum in sub-bands (MPDSS) feature, *supra* level pitch contour (PC) and epoch strength contour (ESC) features (see Table 1). The first element of the 2nd column of Table 1 indicates the auto-correlation coefficient of Arunachali language computed by using *sub* level GFD feature. The first element of the 3rd column indicates the average of cross-correlation coefficients of Arunachali language with respect to other 26 languages. Lower average cross-correlation coefficient value between the languages indicate more dissimilarity between Arunachali and the rest of the languages. This average cross-correlation coefficient value of Arunachali language with respect to other 26 languages (0.24) is less than the auto-correlation coefficient value of Arunachali language (0.85). This shows that the *sub* level GFD feature has significant language discriminative capability. Similar characteristics can be observed for other languages also. The correlation coefficients computed from MPDSS, PC and ESC features also show significant language discrimination capability (see Table 1). This theoretical discussion elicits the significance of the excitation source features in language identification task, which is the motivation for the present work.

4. Speech corpora

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

In this work, LID study has been carried out on Indian Institute of Technology Kharagpur-Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) (Maity et al., 2012). This database contains 27 Indian regional languages. Sixteen languages are collected from news bulletins of broadcasted radio channels and the remaining are recorded from broadcasted TV talk shows, live shows, interviews and news bulletins. The broadcasted television channels are accessed using VentiTV software (VentiTV software) and the Pixelview TV tuner card. Audacity software (Audacity software) is used for recording the speech data from TV channels. The language data of broadcasted Radio channels are collected from the archives of Prasar Bharati, All India Radio (AIR) website (All India Radio, 2014). The detailed description of the database is given in Table 2. We are preparing this database to publish it for academic research purposes.

4.2. Oregon Graduate Institute Multi-Language Telephone-based Speech (OGI-MLTS)

Oregon Graduate Institute (OGI) Multi-Language Telephone-based Speech (MLTS) consists of 11 languages. Muthusamy et al. (1992) have collected the following 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese. Later Hindi language data have been added. During collection of this database, each speaker was asked a series of questions designed to elicit:

- Fixed vocabulary speech
- Domain-specific vocabulary speech
- Unrestricted vocabulary speech

The "unrestricted vocabulary speech" was obtained by asking the callers to speak on any topic of their choice. The "unrestricted vocabulary speech" of each speaker consists of two separate utterances. The duration of one utterance is 50 s and duration of another utterance is 10 s. In our work, we have used only the utterances of 50 s from each speaker for LID study. From each language we have considered calls both from male and female speakers.

4.3. NIST 2011 database

The National Institute of Science and Technology (NIST) 2011 Language Recognition Evaluation (LRE) corpus is composed of 24 languages collected over telephone conversations and narrowband recordings. These 24 languages

Table 2
Description of the IITKGP-MLILSC Language Database.

Languages	Region	Speaking Population at 2001 census (Mil)	No. of speakers F	No. of speakers M	Duration (minutes)
Arunachali	Arunachal Pradesh	0.41	6	6	175.84
Assamese	Assam	13.17	6	7	43.41
Bengali	West Bengal	83.37	13	9	68.63
Bhojpuri	Bihar	38.55	2	6	26.75
Chhattisgarhi	Chhattisgarh	11.50	6	6	54.58
Dogri	Jammu and Kashmir	2.28	5	4	101.81
Gojri	Jammu and Kashmir	20.00	2	6	70.03
Gujarati	Gujarat	46.09	4	7	76.72
Hindi	Uttar Pradesh	422.05	6	12	85.56
Indian English	All over India	125.23	6	6	63.01
Kannada	Karnataka	37.92	1	5	29.43
Kashmiri	Jammu and Kashmir	5.53	1	14	47.67
Konkani	Goa and Karnataka	2.49	5	4	66.92
Malayalam	Kerala	33.07	7	7	60.49
Manipuri	Manipur	1.47	7	7	107.54
Marathi	Maharashtra	71.94	5	6	41.34
Mizo	Mizoram	0.67	6	6	110.13
Nagamese	Nagaland	0.03	5	6	104.31
Nepali	West Bengal	2.87	6	6	54.19
Oriya	Orissa	33.02	7	4	53.86
Punjabi	Punjab	29.10	0	8	64.01
Rajasthani	Rajasthan	50.00	5	6	69.43
Sanskrit	Uttar Pradesh (UP)	0.014	0	12	58.2
Sindhi	Gujarat and Maharashtra	2.54	6	6	109.83
Tamil	Tamilnadu	60.79	8	10	75.38
Telugu	Andhra Pradesh (AP)	74.00	0	9	99.08
Urdu	UP and AP	51.54	1	8	54.03

are: Arabic Iraqi, English Indian, Russian, Arabic Levantine, Farsi/Persian, Slovak, Arabic Maghrebi, Hindi, Spanish, Arabic MSA, Lao, Tamil, Bengali, Mandarin, Thai, Czech, Panjabi, Turkish, Dari, Pashto, Ukrainian, English American, Polish, Urdu. For training purpose we have used the data from the following sources: The Voice of America (VOA) corpus from NIST LRE 2009, Corpora from NIST 1996, 2003, 2005 and 2007 LRE, LDC corpora. The evaluation set consists of speech segments of durations 30 s, 10 s and 3 s. In this work, we have used only the segments of 10 s durations during evaluation.

5. Excitation source information

The excitation source information mostly represents the dynamic nature of the vocal folds vibration during voiced sound production. This information can be captured by processing the LP residual signal (Makhoul, 1975). LP residual signal can be analyzed at three different regions: within a glottal cycle known as sub-segmental level information (2–5 ms), within 2–3 consecutive glottal cycles (10–20 ms) known as segmental level information and across several glottal cycles (50 cycles) known as supra-segmental level (100–300 ms) information. In Section 5.1, Section 5.2 and Section 5.3, parametric approaches have been proposed to capture the sub-segmental, segmental and supra-segmental level excitation source information, respectively.

5.1. Parametric representation of sub-segmental level excitation source information

5.1.1. LF model of glottal flow derivative

The glottal air pulses are the primary source to provoke the vocal tract resonator during the production of voiced speech. The vibration of the vocal folds during the production of voiced speech generates the glottal flow that is

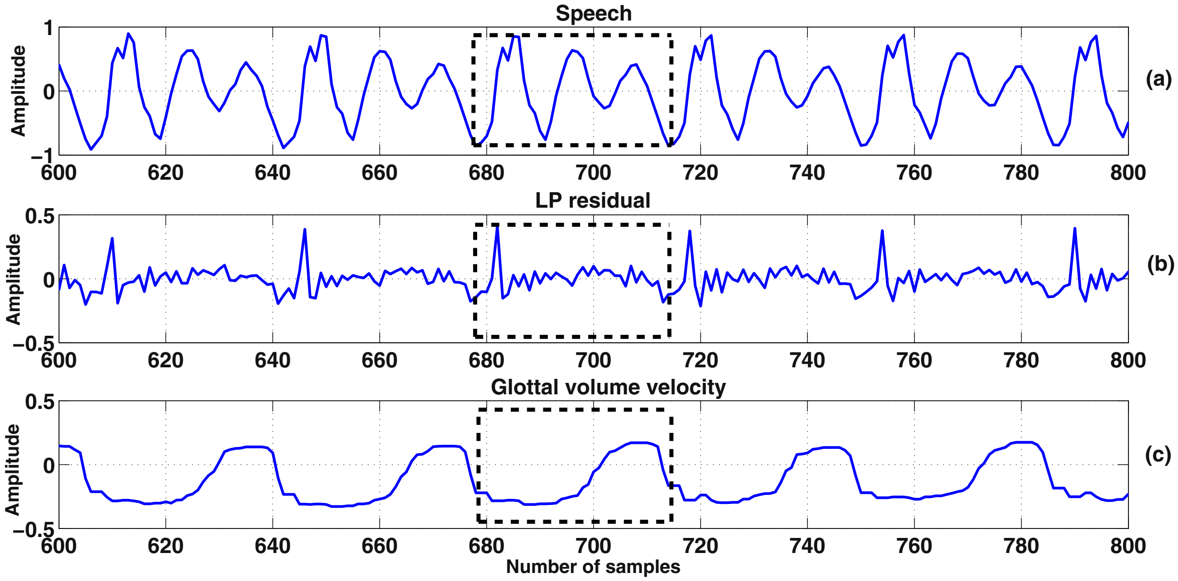

Fig. 1. (a) Speech signal, (b) corresponding LP residual signal and (c) glottal volume velocity marked by bounding boxes represent sub-segmental level frame (5 ms).

quasi-periodic. A segment of voiced speech and corresponding LP residual signal have been shown in Fig. 1(a) and (b). Fig. 1(c) portrays the glottal volume velocity (GVV) waveform corresponding to the speech signal. In the sub-segmental level, LP residual signal is analyzed in a block size of 5 ms with a shift of 2.5 ms to capture the characteristics of one glottal cycle. One glottal pulse and the corresponding glottal flow derivative (GFD) are shown in Fig. 2(a) and (b), respectively. Glottal pulse is defined as a single period of the glottal air flow. Each glottal flow cycle is divided into three phases: *closed-phase*, *open-phase* and *return-phase*. The characteristics of vocal folds vibration is unique for each sound unit and the same sound unit may vary across the languages due to phonotactic characteristics. Hence, the shape of the glottal pulse and the corresponding GFD waveform during the *open-phase* interval is unique across different sound units and may provide language discriminative information. We presume that the *open-phase* of the GFD cycle may contain significant language-specific knowledge. In *return-phase* (Fig. 2(a)), vocal folds tend to close. High frequency energy is generated due to the sudden closing of vocal folds which is also reflected in the *return-phase* characteristics. The duration of this phase and the slope of GFD waveform corresponding to the *return-phase* depends on the swiftness of vocal folds closing. The swiftness of closing of vocal folds reduces the duration of the *return-phase*, which yields more high frequency energy. These characteristics can be observed from the exponential nature of the GFD during the *return-phase* shown in Fig. 2(b). The sudden change of glottal pulse in *return-phase* may be context dependent. The rapidness of vocal folds closing for one phoneme may vary from one language to another which reflects the uniqueness of a language. Hence, we conjecture that the *return-phase* may contain language-specific information.

The LF model (Ananthapadmanabha and Fant, 1982) of the glottal air flow illustrates the GFD waveform in terms of exponentially growing sinusoid in the *open-phase* and a decaying exponential in the *return-phase* (Plumpe et al., 1999). The LF model parameters of a GFD cycle shown in Fig. 2(b) are listed below.

1. T_c: Time instant of glottal closure (GCI).
2. T_o: Time instant of glottal opening (GOI).
3. E_e: Absolute value of the maximum rate of glottal flow decrease.
4. T_e: The time instant of maximum rate of glottal flow decrease.
5. α: The growth factor defined as the ratio of E_e to maximum rate of glottal flow increase.
6. β: Exponential time constant that determines how quickly glottal flow derivative returns to zero after time T_e.
7. ω_z: Frequency that determines flow derivative curvature to the left of the first GFD zero crossing (T_z), $\omega_z = \dfrac{\pi}{T_z - T_o}$.

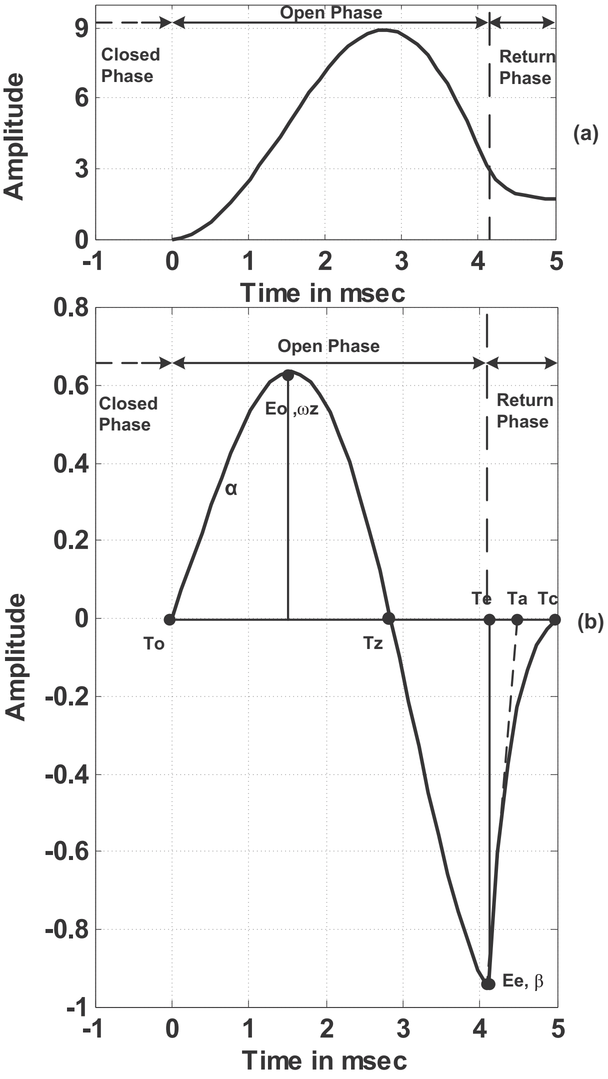

Fig. 2. (a) Glottal pulse and (b) LF model of GFD.

These seven LF model parameters are estimated from LP residual signal. The three phases of a single GFD cycle denoted as $e_{LF}(t)$ can be expressed mathematically by the following way (Plumpe et al., 1999; Qi and Bi, 1994).

$$\begin{aligned} e_{LF}(t) &= 0, && 0 \le t < T_o \\ &= E_0 e^{\alpha(t-T_0)} \sin[w_z(t-T_0)], && T_0 \le t < T_e \\ &= -\frac{E_e}{\beta T_a}\left[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}\right], && T_e \le t < T_c \end{aligned} \tag{2}$$

where E_o is a gain constant and T_a is (time constant of the return phase) the time instant where the slope of *return-phase* crosses the time axis. In the LF model analysis it is generally assumed that there is no air flow during the *closed-phase*. Hence, $e_{LF}(t) = 0$ is assumed at *closed-phase*. From Fig. 2(b) it can be observed that, $T_o, T_e, E_e, \alpha, \omega_z$ and E_o characterize the *open-phase* and E_e, β and T_c characterize the *return-phase*. All the seven parameters estimated from the LP residual signal are used as feature vector to capture the sub-segmental level excitation source information for LID task.

5.1.2. *Proposed method for computation of the LF model parameters*

The LF model parameters are computed from each glottal cycle (5 ms window) to capture the language-specific information present within one glottal cycle. The proposed method for calculating seven LF parameters is discussed below.

1. **Computation of T_c:**

T_c Indicates the glottal closing instant (GCI). In our work, we have used *zero-frequency filtering* (Murty and Yegnanarayana, 2008) method to calculate the GCI locations. In this method, the speech signal is passed through the *zero-frequency resonator* twice. The reason for passing the speech signal twice through the *zero-frequency resonator* is to mitigate the effects of high frequency resonances. This yields a filtered output that varies as a polynomial function of time which is known as zero-frequency filtered signal (ZFFS). The positive zero crossings in the ZFFS indicates the GCI locations (Murty and Yegnanarayana, 2008).

2. **Computation of T_o:**

The steps to determine the GOIs are described below:

- Compute the pitch period (P_g) of the g^{th} glottal cycle as $P_g = T_c(g) - T_c(g-1)$, where $T_c(g)$ and $T_c(g-1)$ are the closing instants of the g^{th} and its immediate previous glottal cycle respectively.
- Compute the average pitch period P_{avg}.
- Compute the opening instant of the $(g+1)^{th}$ glottal cycle $\left(T_{o(g+1)}\right)$ by using equation (3) (Naylor et al., 2007).

$$T_{o(g+1)} = T_{cg} + 0.3 \times min\left[T_{c(g+1)} - T_{c(g)}, P_{avg}\right] \tag{3}$$

3. **Computation of E_e:**

E_e indicates the absolute value of the maximum rate of glottal flow decrease. The magnitude of the negative peak of the GFD waveform shown in Fig. 2(b) indicates the maximum rate of glottal flow decrease. This can be calculated by finding the absolute maximum value from LP residual samples within one glottal cycle.

4. **Computation of T_e:**

Once the E_e is computed, the corresponding time instant T_e can also be obtained, which denotes the time instant corresponding to the maximum rate of glottal flow decrease.

5. **Computation of α and β:**

The amplitude of the GFD cycle is zero during the *closed-phase*. Therefore, there is no significant information present in *closed-phase*. Hence, we can consider that the effective GFD cycle is from the instant of the glottal opening (T_o) to the instant of the glottal closing (T_c). By assuming $T_o = 0$ (i.e. the glottal opening instant starts from $t = 0$), equation (2) becomes:

$$e_{LF}(t) = E_o e^{\alpha(t)} \sin\left[w_z(t)\right], \quad 0 \le t < T_e \tag{4}$$

$$= -\frac{E_e}{\beta T_a}\left[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}\right], \quad T_e \le t < T_c \tag{5}$$

Now, we can observe from the LF model of GFD cycle shown in Fig. 2(b) that the value of the negative peak of GFD waveform is $-E_e$ at time instant $t = T_e$ (i.e. $\left[e_{LF}(t)\right]_{t=T_e} = -E_e$). Substituting $t = T_e$ in equations (4) and (5), we get:

$$[e_{LF}(t)]_{t=T_e} = E_o e^{\alpha T_e} \sin[w_z(T_e)] \tag{6}$$

$$[e_{LF}(t)]_{t=T_e} = -\frac{E_e}{\beta T_a}\left[1 - e^{-\beta(T_c - T_e)}\right] \tag{7}$$

Therefore, we get the following relation:

$$E_o e^{\alpha T_e} \sin(\omega_z T_e) = -E_e \quad \Rightarrow \alpha = \frac{1}{T_e} \ln\left[-\frac{E_e}{E_o \sin(\omega_z T_e)}\right] \tag{8}$$

Again, comparing equation (7) with the equation $[e_{LF}(t)]_{t=T_e} = -E_e$, we get:

$$-\frac{E_e}{\beta T_a}\left[1 - e^{-\beta(T_c - T_e)}\right] = -E_e \quad \Rightarrow 1 - e^{-\beta(T_c - T_e)} = \beta T_a \tag{9}$$

In Qi and Bi (1994), it is assumed that the return flow is relatively faster. So, the assumption $\beta(T_c - T_e) \gg 1$ has been taken. With this assumption, equation (9) gets reduced to equation (11) as follows:

$$\beta(T_c - T_e) \gg 1 \quad \Rightarrow e^{-\beta(T_c - T_e)} \simeq 0 \tag{10}$$

Therefore,

$$\beta T_a = 1 \tag{11}$$

A constraint is imposed with the above assumption that the GFD cycle returns to zero at the end of each cycle. This implies that the area under the GFD curve is 0.

$$\int_0^t e_{LF}(t)dt = 0 \tag{12}$$

To compute β, we need to solve equation (12).

$$\begin{aligned} &\int_0^t e_{LF}(t)dt = 0 \\ &\quad \Rightarrow \int_0^{T_e} e_{LF}(t)dt + \int_{T_e}^{T_c} e_{LF}(t)dt = 0 \\ &\quad \Rightarrow \int_0^{T_e} E_o e^{\alpha(t)} \sin(\omega_z t)dt + \int_{T_e}^{T_c} -\frac{E_e}{\beta T_a}\left[e^{-\beta(t - T_e)} - e^{-\beta(T_c - T_e)}\right]dt = 0 \end{aligned} \tag{13}$$

Solving the above equation, we obtain an expression for β as follows:

$$\beta = \frac{E_e(\alpha^2 + \omega_z^2)}{E_o\{e^{\alpha T_e}[\alpha sin(\omega_z T_e) - \omega_o cos(\omega_z T_e)] + \omega_z\}} \tag{14}$$

The α and β are modified iteratively until equation (11) is satisfied. The number of iterations are fixed based on experimental studies. To terminate the computation procedure it is bounded by 10 iterations which provides the best LID performance reported in this work.

6. **Computation of ω_z:**

As the LP residual signal is a noise like signal, it is difficult to find out the T_z and E_o accurately. In our work, we have proposed an iterative procedure to find out these two parameters. The parameters T_z and E_o are related to the *open-phase* of the glottal cycle which is generally larger than the *return-phase*. Initially, we assume that T_z (the first zero crossing of GFD cycle) is 50% of the total glottal cycle duration. With this assumption, E_o is measured as the absolute maximum value of the glottal cycle up to T_z. Now, we can easily compute the α and β by equation (8) and equation (14) respectively. Thus, to verify the accuracy of the initial estimation of the parameters, the constraint of equation (11) has been imposed. In every iteration, the T_z value is increased by 5% of the glottal cycle. The reason for increasing the T_z value is due to the larger duration of the open phase. In our work, we have used 10 iterations which are fixed by experimental observations.

We also have explored the dynamic nature of the LF parameters to capture the fine variations of the glottal activities. The variations in the LF parameters from one glottal cycle to the other may be attributed to the fine variations in the glottal cycles. The dynamic nature can be captured by velocity and acceleration coefficients. The Delta–Delta (acceleration) coefficients are computed by performing time derivative over the Delta (velocity) coefficients. To minimize the speaker variability in speech corpus, we have explored speaker normalization using a mean subtraction (MS) method. The LID system is developed by imposing the MS on GFD concatenated with dynamic coefficients.

5.2. *Parametric representation of segmental level excitation source information*

Parameters extracted from the LP residual signal consisting of (2–3) consecutive glottal cycles may carry segmental level information. At segmental level, the LP residual signal is processed in block size of 20 ms with a shift of 2.5 ms. The LP residual signal is parameterized at segmental level to capture the language-specific energy and periodicity information present within (2–3) glottal cycles. A segment of voiced speech and corresponding LP residual signal is shown in Fig. 3(a) and (b) respectively. The bounding boxes in the Fig. 3(a) and (b) represent the 20 ms segmental level frame. In this work, the LP residual signal is processed in spectral and cepstral domains to capture the segmental level energy and periodicity information respectively. Sections 5.2.1 and 5.2.2 describe the processing of LP residual signal in cepstral and spectral domains to capture the energy and periodicity information of excitation source respectively.

5.2.1. *Language-specific energy information from cepstral analysis of LP residual spectrum*

Cepstral analysis is performed on LP residual signal to capture the energy of excitation source. This process is similar to the computation of conventional MFCC features except that the input is LP residual signal instead of speech signal. The cepstral coefficients computed from Mel sub-band spectra are termed as Residual Mel Filter Cepstral

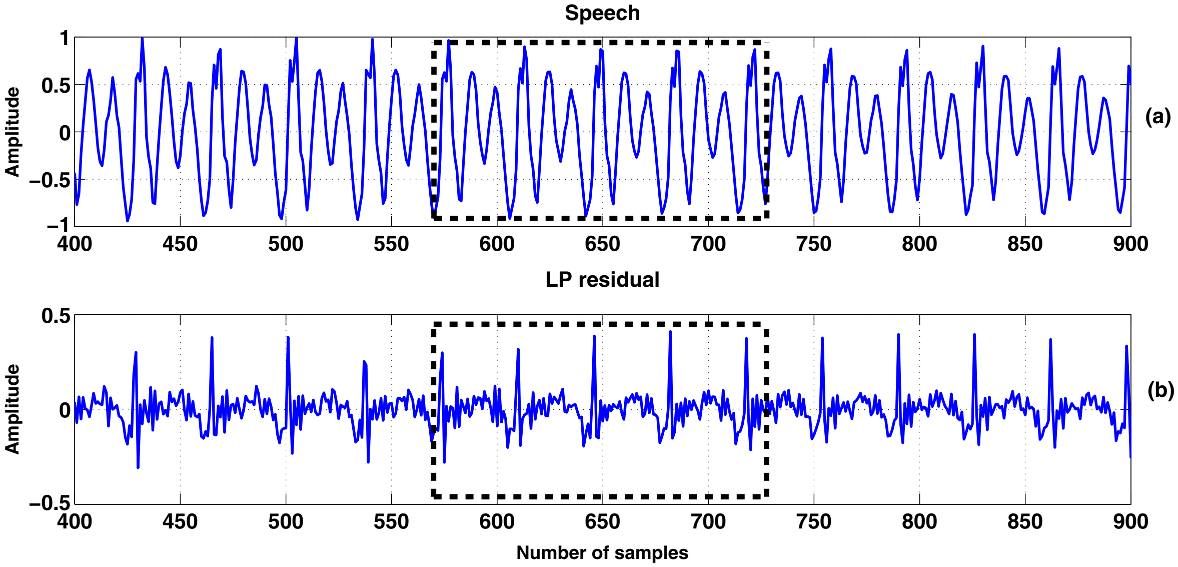

Fig. 3. (a) Speech signal and (b) corresponding LP residual signal marked by bounding boxes represent the segmental level frame (20 ms).

Coefficients (RMFCC). In this work, only the first 13 cepstral coefficients are used as a feature vector and are modeled to capture the spectral energy information of excitation source.

There is significant overlap in the set of sound units for different languages in India. However, each language has unique phonotactic constraints based on a sequence of occurrence of sound units. Therefore, the production characteristics of a particular sound unit may vary from one language to another due to language-specific co-articulation effects. Hence, the characteristics of vocal tract system and vocal folds vibration also contain specific information due to language-specific phonotactics which is unique for individual languages. In this work, we have focused only on the characteristics of the vocal folds vibration. As we know from literature, the excitation source information can be captured by processing the LP residual signal. Hence, we have proposed a method to capture the energy of excitation source through RMFCCs extracted from a segment of LP residual. We hypothesize that the segmental level energy information may contain language-specific phonotactic knowledge.

5.2.2. Language-specific periodicity information from spectral analysis of LP residual signal

Each sound unit corresponds to a unique articulatory configuration of both the vocal tract and vocal folds. Hence, the characteristics of vocal tract shapes and vocal folds vibration are also unique for a particular sound unit. The rate of the vocal folds vibration varies from one sound unit to another. Thus, the dynamic range of the residual spectrum varies from one sound unit to another. The variation in the dynamic range depends on the periodicity nature of the spectrum. Therefore, the periodicity information or the harmonic structure of the excitation source also varies from one sound unit to another. To capture the periodicity information of excitation source, spectral analysis has been performed. The periodicity of the excitation source is commensurable with the difference between peaks and dips of the LP residual spectrum.

The harmonic structure information can be obtained by analyzing the power spectrum $p(k)=[R(k)]^2$ of the LP residual signal. This periodicity information can be represented by Power Differences of Spectrum in Sub-bands (PDSS) features (Hayakawa et al., 1997). The PDSS feature represents the periodicity nature of the excitation source (Gray and Markel, 1974). PDSS can be represented as a spectral flatness (SF) measure of the power spectrum in sub-bands. SF can be measured by the ratio of geometric mean (GM) to arithmetic mean (AM) of the power spectrum. PDSS of residual sub-band spectra is defined as follows (Hayakawa et al., 1997):

$$V(i)=1.0-\frac{\left[\prod_{k=L_i}^{H_i} p(k)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p(k)} \tag{15}$$

where $N_i=h_i-l_i+1$ is the sample number of frequency points in the i^{th} filter and l_i and h_i are the lower and upper limits of frequency in i^{th} sub-band, respectively.

Since, $0 \le SF \le 1$, the values of PDSS also vary from 0 to 1. High SF value indicates that the power of the spectrum is distributed throughout the spectrum and the power spectrum would appear relatively smooth. Low SF value illustrates that the spectral power is concentrated only within smaller regions of the power spectrum. The spectral flatness can also be measured within a specified sub-band, rather than across the whole spectrum. In the present work, the spectral flatness has been measured in sub-bands. If the power spectrum has less dynamic range, for example nearly flat, then $\mathrm{GM} \simeq \mathrm{AM}$ and the PDSS value will be less than one. Alternatively, if PDSS is low, the spectrum is less periodic. If the spectrum has peaks and dips, for example, the dynamic range is more, then GM is less than AM and PDSS value is close to one. In this case the spectrum is more periodic. Therefore, PDSS measure gives information about the periodicity nature of a spectrum. Sub-band spectra are obtained by multiplying the residual power spectrum with a filterbank and PDSS values are computed from each sub-band using equation (15). In this work, the Mel filter bank is used for computing the PDSS from Mel sub-bands. The motivation for using Mel filters is from the property of the Mel filterbank that provides less spectral samples to lower bands and more to higher bands (beyond 1 kHz). The dominant information about the excitation source is manifested in the higher frequency range. Therefore, it is expected that PDSS computed from Mel sub-bands may provide better identification accuracy. Hence, in this work PDSS values have been calculated using Mel filter-bank which is termed as Mel PDSS (MPDSS) features.

The pitch of a particular sound unit is context-dependent. Since each language has unique phonotactics, even though sound units may be mostly common across the languages, the co-articulation effect (due to context) will be unique for each of the languages. Due to the pitch difference between two languages, the periodicity of voiced sound units may also vary. Therefore, we can presume that the periodicity information of excitation source captured by MPDSS feature may provide language discriminative information.

5.3. Parametric representation of supra-segmental level excitation source information

In this section, the procedure to parameterize the LPR signal at the supra-segmental level has been discussed. The excitation source characteristics vary with time. At the supra-segmental level, these temporal variations are captured by processing larger segment of LPR signal. The tension of the vocal folds and the sub-glottal air pressure changes continuously during speech production (Atal, 1972; Wolf, 1972). As a consequence, the average rate of vocal folds vibration or pitch and the strength of excitation at glottal closing instants also vary with time. Therefore, in this work, the LPR signal has been parameterized for 100 ms frame at supra-segmental level to capture the language-specific excitation source information. The pitch and epoch strength contours are captured at supra-segmental level. To estimate the pitch, zero-frequency filtering approach has been used in this work. The epoch strength can also be measured from the zero-frequency filtered signal (ZFFS). The pitch and epoch strength contours are processed separately to capture the language-specific information. Finally, the scores are combined to capture the complete excitation source information at supra-segmental level. The zero-frequency filtering method estimates the pitch and epoch strength by locating epochs or glottal closure instants (GCIs) (Murthy and Yegnanarayana, 2008, 2009; Yegnenarayana and Murthy, 2009). The speech signal is given to the input of zero-frequency filter (ZFF). The output signal of ZFF is known as ZFFS. The positive zero crossings of ZFFS signal are the glottal closing instants (GCIs) (Murthy and Yegnanarayana, 2008). The interval between two GCIs is the instantaneous pitch period T_0. The reciprocal of the instantaneous period is instantaneous pitch $\left(P_0 = \frac{1}{T_0}\right)$ (Yegnenarayana and Murthy, 2009). The slope of the ZFFS around the positive zero crossings corresponding to the locations of GCIs gives the measure of epoch strength (A_0) (Murthy and Yegnanarayana, 2009). The slope around the epochs or GCI locations can be computed as the absolute difference between the preceding and succeeding sample amplitudes of ZFFS around the GCIs (Murthy and Yegnanarayana, 2009).

6. Experimental setup and methodology

In this work, we have considered 27 Indian languages for carrying out LID studies. From each language database, 45 minutes of speech data are taken for training the language models. All the LID systems are evaluated using leave-two-speaker-out approach. In each iteration, $(n-2)$ speakers (where n is the total number of speakers present in each language database) from each language database are used to develop language models, and two other speakers of each language, who have not participated during training phase, are considered for evaluation. The LID performances reported in this paper are obtained from the average of all iterations. 20 test utterances from each language, each of 10 s duration, have been considered during evaluation. In this work LID systems are developed using GMMs. We have also compared the performance of GMM based LID systems with *i*-vector based approach. The details of building LID systems using GMM and *i*-vector based approaches are given in the following subsections.

6.1. GMM based approach

6.1.1. Training

Gaussian Mixture Model (GMM) can capture the language-specific excitation source information in terms of mixture parameters. Parametric excitation source features discussed in Section 5 are modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix. Therefore, a mixture of single densities, i.e., a Gaussian Mixture Model (GMM), is used to model the complex structure of probability density. For a D-dimensional feature vector denoted as x_t, the mixture density for a language Ω is defined as a weighted sum of M component Gaussian densities as given by the following expression (Reynolds and Rose, 1995):

$$P(x_t|\Omega) = \sum_{i=1}^{M} w_i P_i(x_t) \tag{16}$$

where w_i are the weights and $P_i(x_t)$ are the component densities. The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities. These parameters are collectively represented by

$$\Omega = \{w_i, \mu_i, \Sigma_i\}; \quad i = 1, 2, \dots M. \tag{17}$$

To determine the model parameters of a GMM for a particular language, the model has to be trained. The maximum likelihood (ML) (Reynolds and Rose, 1995) technique was used for estimating the optimal parameters. Maximization was carried out iteratively using an Expectation Maximization (EM) algorithm (Dempster et al., 1977). The performance of EM algorithm depends on the initialization. In the present work, GMM parameters were initialized using k-means clustering and $M = 16$ was empirically found to perform the best. In each iteration, the posterior probability for the i^{th} mixture was computed as given by the following equation (Reynolds and Rose, 1995):

$$Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)} \tag{18}$$

The model parameters were updated. Empirically we found that after 50 iterations of the EM algorithm the model parameters reached their optimum values with negligible changes in any subsequent iteration.

6.1.2. Testing

In identification phase, mixture densities are calculated for every feature vector for all languages. The language with maximum likelihood is selected as identified language among all languages. For example, if S language models $\{\Omega_1, \Omega_2, \dots, \Omega_S\}$ are available after the training, language identification can be carried out using test speech dataset. First, the sequence of feature vectors $X = \{x_1, x_2, \dots, x_T\}$ is calculated. Then the language model $\hat{s}$ is determined which maximizes the a posteriori probability $P(\Omega_S|X)$. The posterior probability associated to each language is computed using the Bayes rule (Reynolds and Rose, 1995).

Assuming equal probability of all languages and the statistical independence of the observations, the decision rule for the most probable language can be redefined as:

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log P(x_t|\Omega_s) \tag{19}$$

with T the number of feature vectors of the speech data set under test and $P(x_t|\Omega_s)$ given by Equation (16). The identification accuracy of each language is defined as the percentage of the ratio of number of test utterances correctly identified to the total number of utterances used per language. The average of the individual language accuracies is coined as average LID accuracy.

6.1.3. Fusion of scores

In this work, adaptive weighted combination scheme (Reddy et al., 2013) has been used for combining various LID systems. We have considered 27 Indian languages and 20 test utterances from each language (total 540 test samples) for the evaluation. The scores of a particular test utterance obtained from different sub-systems are combined by adaptive weighted scheme. The combined score C is given by:

$$C = \frac{1}{k} \sum_{i=1}^{k} w_i c_i \tag{20}$$

where w_i and c_i denotes weighting factor and confidence score of the i^{th} model and k denotes the number of sub-systems considered for fusion. The Weighting factor w_i varies from 0 to 1 with a step size of 0.01 and sum up to 1 (i.e. $\sum_{i=1}^{k} w_i = 1$). In this work, we have explored 4753 and 98 different sets of weighting factors for three and two sub-systems, respectively. For a particular set of weighting factors average performance of 27 languages has been calculated. Out of 98 different average performance values while combining two sub-systems (or 4753 for three sub-systems), best average accuracy and corresponding weighting factor set are considered as the optimum one and has been reported for the combined systems.

In this work, LP residual signal has been parameterized at three different levels (*sub*, *seg* and *supra*) to capture distinct aspects of excitation source. *Sub* level information is captured by GFD features for LID task. At *seg* level, RMFCC and MPDSS features are extracted to model the energy and periodicity information present within 2–3 glottal cycles. At *supra* level, variation of pitch and epoch strength with respect to time are modeled to capture language-specific information present in 50 glottal cycles. Hence, we have combined the scores from three different levels to obtain complete parametric excitation source information. Finally, the confidence scores from excitation source and vocal tract features are combined to investigate the complementary nature of these two features in language identification context. All these different combinations are discussed in Section 7.

6.2. i-*Vectors based approach*

i-Vector is one kind of subspace modeling approach which uses the dimensionality reduction of training data and applies the classifier to determine the language identity of a test utterance. The main purpose of reducing the dimensions is to separate the common patterns of the data across all languages from the unique information between utterances. Dimensionality reduction reduces the computational load at training stage, and thus more amount of training data can be used. Extracted *i*-vectors highlight both the class and channel-specific information of a particular utterance.

6.2.1. i-*Vector extraction*

In this work, language and channel dependent information has been represented by GMM based supervectors created by concatenating GMM mean vectors. These GMM supervectors can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{21}$$

where **m** is language and channel independent supervector of concatenated UBM means. **T** is the subspace matrix covering the major variabilities (both language- and session-specific) in the supervector space. **w** is a normally distributed latent variable. *i*-Vector for each utterance is nothing but the maximum a posteriori (MAP) point estimate of the latent variable **w**. Detailed procedure of *i*-vector extraction is given in Dehak et al. (2011a). The rank of T-matrix 400 was chosen and the UBM size of 2048 was chosen empirically in our work.

6.2.2. *Classification*

A linear generative classifier is trained using the extracted *i*-vectors (Martinez et al., 2011). The *i*-vectors obtained from individual languages are modeled by Gaussian distributions with a single within a class (WC) full covariance matrix. The log-likelihood is computed with an *i*-vector corresponding to a test utterance by the following formula.

$$\ln p(\mathbf{w}|l) = -\frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} + \mathbf{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l - \frac{1}{2}\boldsymbol{\mu}_l^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l + C. \tag{22}$$

where $\boldsymbol{\mu}_\mathbf{l}$ is the mean vector for language l. $\boldsymbol{\Sigma}$ is the common covariance matrix and C is language and *i*-vector independent constant. The fusion of scores is performed using the adaptive weighting mechanism described in Section 6.1.

7. Development of LID systems using proposed excitation source features

In this work, LID systems are developed at four stages by processing the proposed excitation source features. Seven LF parameters concatenating with Delta and Delta–Delta coefficients (21 dimensions) from GFD feature vector is used to capture *sub* level parametric source information. Thirteen RMFCC coefficients with velocity and acceleration coefficients (39 dimensions) are used as *seg* level parametric features. Ten successive pitch and epoch strength values were used to construct *supra* level feature vector. The above-mentioned feature extraction programs used in

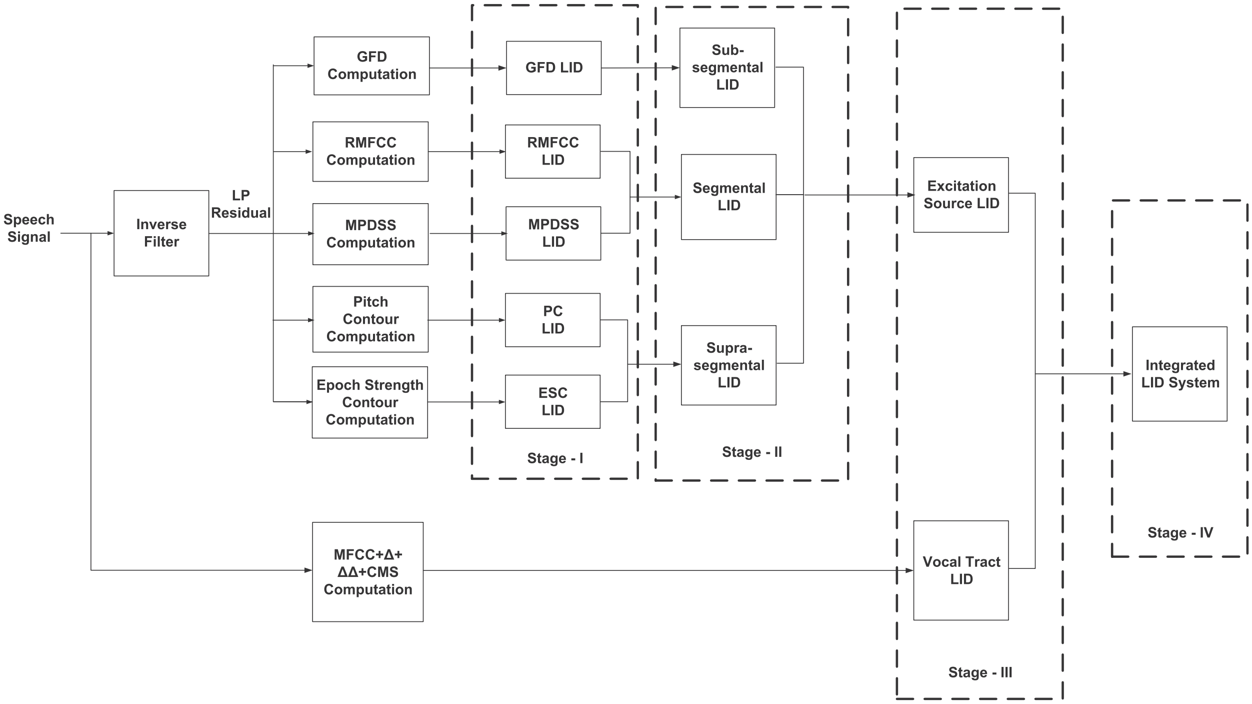

Fig. 4. Development of LID systems using excitation source and vocal tract information.

this work are developed using MATLAB functions (Matlab and Statistics Toolbox Release, 2012b). The block diagram of all LID systems developed in various stages is shown in Fig. 4.

7.1. Stage-I

The primary objective of stage-I is to develop LID systems using individual excitation source features of three different levels. Each feature carries distinct language-specific excitation source information. Hence, we have developed five different LID systems using five different excitation source features. At the stage-I, five different LID systems are developed:

- First LID system is developed by processing the GFD feature at *sub* level which is denoted as GFD LID in Fig. 4.
- Second LID system is developed by processing the RMFCC feature at *seg* level which is denoted as RMFCC LID in Fig. 4.
- Third LID system is developed by processing the MPDSS feature at *seg* level which is denoted as MPDSS LID in Fig. 4.
- Fourth LID system is developed by processing the pitch contour at *supra* level which is denoted as PC LID in Fig. 4.
- Fifth LID system is developed by processing the epoch strength contour at *supra* level which is denoted as ESC LID in Fig. 4.

7.2. Stage-II

At *seg* level we have explored two different excitation source features: RMFCC and MPDSS for capturing the instantaneous energy and periodicity information present at *seg* level. To represent complete *seg* level information we have combined the scores from RMFCC and MPDSS features. *Supra* level PC and ESC features individually represent the variation of pitch and energy over 50 glottal cycles. So, to capture the complete *supra* level excitation source information we have combined the scores obtained from LID systems based on PC and ESC features. Only GFD feature has been explored at *sub* level. Therefore, subsegmental LID of stage-II is same as GFD LID at stage-I.

- The segmental LID system is developed by combining the scores obtained from RMFCC and MPDSS LID systems of stage-I to capture the complete excitation source information at *seg* level.
- The supra-segmental LID system is developed by combining the scores obtained from PC and ESC LID systems to derive the full excitation source information at *supra* level.

7.3. Stage-III

Excitation source LID at stage-III is developed by combining the scores of sub-segmental, segmental and supra-segmental LID systems from stage-II. Vocal tract LID system is designed at this stage using MFCC + Δ + $\Delta\Delta$ + CMS feature. The primary objective of this phase is to develop LID system using overall excitation source information and examine the complementary nature of vocal tract and excitation source features from language discrimination perspective.

- Excitation source LID at this stage is developed by combining the scores from *sub*, *seg* and *supra* LID systems to represent the complete excitation source information.
- The vocal tract LID has been developed by imposing the cepstral mean subtraction (CMS) on MFCCs concatenated with dynamic coefficients.

7.4. Stage-IV

The primary goal of this stage is to develop the integrated LID system by fusing the vocal tract and excitation source based LID systems of stage-III.

8. Evaluation of language identification systems using proposed excitation source features

In this work, we have carried out the evaluation of the language models using leave-two-speaker-out approach. As a first step preprocessing is performed to remove the silence portions from a continuous speech. The silence removed speech signal is divided into segments of 10 seconds in a non-overlapping manner. The Gaussian mixture models (Reynolds and Rose, 1995) are used to train the language models. Different Gaussian mixtures (32, 64, 128 and 256) have been explored for modeling the language-specific excitation source information.

8.1. Stage-I

At stage-I, five different LID systems are developed by parameterizing the LP residual signal at three different levels. Table 3 portrays the identification performances of individual languages as well as the average performances of LID systems. The first column represents the 27 Indian languages used in this LID study. The 2nd, 3rd, 4th, 5th and 6th columns of Table 3 represent the LID performances obtained by processing the *seg* level RMFCC and MPDSS features, *supra* level pitch and epoch strength contours and *sub* level GFD features, respectively. The average recognition accuracies obtained from the corresponding systems are 61%, 51%, 33%, 18% and 40%, respectively. The energy of the excitation source is represented by *seg* level RMFCC feature which provides better LID accuracy compared to other features. To understand the confusion pattern among the languages, confusion matrix is shown in Table 4. This confusion matrix is obtained by evaluating 27 languages using *seg* level RMFCC features. Comparison of individual language performances obtained from the RMFCC and MPDSS features portray that these two features contribute distinct language-specific information. Hence, the scores derived from these two features are combined at stage-II to capture the complete *seg* level excitation source information. Similarly, the LID accuracies obtained from pitch and epoch strength contours show distinctness between them from language identification perspective. Hence, scores from these two features are also combined at stage-II to capture the overall *supra* level parametric source information.

8.2. Stage-II

At stage-II, three different LID systems are developed (see Fig. 4). The 6th column of Table 3 shows the LID performance obtained by processing the sub-segmental level GFD parameters. An average accuracy of 40% is achieved

Table 3
LID performances using excitation source and vocal tract features on IITKGP-MLILSC.

Languages	Average Recognition Performances (%)													
	Stage-I					Stage-II		Stage-III				Stage-IV		
	MFCC	MPDSS	Pitch contour	Epoch strength contour	GFD (*sub*)	*seg*	*supra*	*src*	MFCC	MFCC + Δ + ΔΔ	MFCC + Δ + ΔΔ + CMS	*src* + MFCC	*src* + MFCC + Δ + ΔΔ	*src* + MFCC + Δ + ΔΔ + CMS
Arunachali	50	50	0	10	0	50	0	50	100	100	100	50	65	100
Assamese	20	0	35	5	35	10	35	30	5	0	5	25	10	15
Bengali	65	55	0	0	25	65	10	65	55	65	70	65	65	65
Bhojpuri	45	40	80	25	40	45	80	50	35	50	35	55	50	55
Chhattisgarhi	100	100	45	40	85	100	70	100	100	100	100	100	100	100
Dogri	95	50	5	15	30	95	5	95	65	100	100	90	100	100
Gojri	100	100	60	40	60	100	60	100	100	100	100	100	100	100
Gujarati	35	0	10	5	0	50	15	50	50	50	55	50	50	55
Hindi	10	20	0	0	0	10	0	10	10	20	25	10	15	25
Indian English	100	100	45	15	100	100	30	100	100	100	100	100	100	100
Kannada	40	35	35	5	55	40	40	35	50	0	0	45	25	10
Kashmiri	70	70	65	5	70	70	65	90	75	70	90	90	80	90
Konkani	20	0	20	35	50	20	10	35	60	50	50	65	55	50
Malayalam	0	0	45	15	45	0	45	30	30	35	35	45	40	40
Manipuri	100	75	0	0	0	100	0	90	100	100	100	100	100	100
Marathi	40	20	0	5	0	40	0	45	45	60	35	50	60	45
Mizo	60	40	10	0	50	55	10	50	15	50	70	30	75	95
Nagamese	90	50	40	45	45	85	50	100	100	100	100	100	100	100
Nepali	95	45	10	5	15	95	10	90	90	100	100	100	100	100
Oriya	40	40	35	25	20	40	35	40	40	40	40	40	40	40
Punjabi	40	50	95	45	90	45	80	50	65	50	55	55	50	55
Rajasthani	100	100	45	70	55	100	50	100	100	100	100	100	100	100
Sanskrit	75	95	25	0	25	95	30	85	5	65	100	75	95	100
Sindhi	50	35	0	45	40	50	0	70	95	85	100	90	90	90
Tamil	50	25	20	10	45	45	25	50	25	45	35	50	50	25
Telugu	100	100	100	20	100	100	100	100	100	100	100	100	100	100
Urdu	65	90	60	10	5	75	65	80	70	60	95	75	75	80
Average Performances	61	51	33	18	40	62	34	66	62	66	70	69	70	72

Table 4
LID performances (in confusion matrix form) obtained from segmental level RMFCC features. Here, the numerical values in the table indicate number of test utterances classified under the particular class. The total number of test utterances per language used for evaluation is 20.

Languages	Aru	Ass	Ben	Bho	Cha	Dog	Goj	Guj	Hin	Ind	Kan	Kas	Kon	Mal	Man	Mar	Miz	Nag	Nep	Ori	Pun	Raj	San	Sin	Tam	Tel	Urd
Aru	10	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
Ass	0	4	4	0	0	0	0	10	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Ben	0	0	13	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bho	0	1	8	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
Cha	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dog	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Goj	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Guj	0	0	0	0	0	2	0	7	0	0	0	0	4	0	0	0	0	0	0	0	1	0	2	0	0	0	4
Hin	0	1	3	0	0	0	0	0	2	0	0	4	0	2	0	2	0	0	1	4	0	0	0	0	1	0	0
Ind	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kan	0	0	1	0	0	0	0	0	0	0	8	2	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0
Kas	0	1	0	0	0	0	0	1	4	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kon	3	0	0	0	0	0	8	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	5	0	0	0
Mal	0	1	5	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	10	0	0
Man	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
Mar	0	0	7	0	0	0	0	0	0	0	0	0	0	4	0	8	0	0	0	0	0	0	0	0	1	0	0
Miz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	8	0	0	0
Nag	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	2	0	0	0
Nep	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
Ori	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	6	8	0	0	0	0	0	0	0
Pun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	8	0	0	0	0	0	10
Raj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
San	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	3
Sin	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	6
Tam	0	0	0	0	0	0	0	0	3	0	0	6	0	1	0	0	0	0	0	0	0	0	0	0	10	0	0
Tel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
Urd	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3	0	0	0	13

from sub-segmental LID at stage-II. The segmental LID system at stage-II has been developed by combining the scores obtained from RMFCC and MPDSS features to represent the *seg* level excitation source information. An LID accuracy of 62% is obtained by processing the *seg* level excitation source information, which is shown in the 7th column of Table 3. The supra-segmental LID system of stage-II has been developed by combining the scores from pitch and epoch strength contours. The LID performance obtained from the supra-segmental LID system of stage-II is 34% as shown in the 8th column of Table 3.

The comparison has been carried out among the individual language performances obtained from *sub*, *seg* and *supra* levels shown in the 6th, 7th and 8th columns of Table 3. The LID performances for Arunachali, Gujarati, Hindi, Manipuri and Marathi languages obtained by processing *seg* level excitation source features are 50%, 50%, 10%, 100% and 40% respectively. However, no test utterances have been correctly identified using *sub* level GFD feature for the corresponding languages. Similarly, the *supra* level excitation source features were not able to identify any test utterances belonging to Arunachali, Hindi, Manipuri and Marathi languages. For Malayalam language, the excitation source information at *sub* and *supra* levels provide 45% LID accuracy. However, the *seg* level excitation source features were not able to identify any test utterance for Malayalam. The excitation source information contributes distinct language-specific information at three different levels. This empirical analysis shows the disparate nature of excitation source information extracted from *sub*, *seg* and *supra* levels in the context of language identification.

8.3. Stage-III

The scores from *sub*, *seg* and *supra* levels are combined to capture the complete excitation source information for language discrimination task. The excitation source LID system at stage-III has been developed by processing the complete excitation source information. The most outstanding LID accuracy obtained from complete excitation source information (denoted as *src*) is 66% (see 9th column of Table 3). The main purpose of stage-III is to develop LID systems based on the overall excitation source features and spectral features. The identification accuracy obtained from MFCC feature based system is 62% which is shown in the 10th column of Table 3. The LID performance obtained by processing the MFCCs concatenated with the velocity and acceleration coefficients is 66% which has been shown in the 11th column of Table 3. The LID performance obtained from MFCC + Δ + $\Delta\Delta$ + CMS feature based system is 70% (12th column of Table 3). The individual language performances achieved from the proposed excitation source features are compared with the state-of-the-art vocal tract features represented by MFCCs, which show the discrepancies between these two features from language identification point of view. Similarly, the distinct nature can also be observed by comparing the individual language performances of the 9th column with the 11th and 12th columns shown in Table 3. The LID system based on MFCC + Δ + $\Delta\Delta$ + CMS feature provides better accuracy, compared to others. Hence, we have displayed only one vocal tract feature based system at stage-III of Fig. 4 based on this feature.

8.4. Stage-IV

By analyzing the recognition accuracy values achieved from stage-III LID systems, it has been observed that the vocal tract and excitation source features provide complementary information for language discrimination task. Hence, to improve the LID accuracies, we have combined the scores from these two complementary features at stage-IV. We have developed three integrated LID systems by combining *src* with MFCC, MFCC + Δ + $\Delta\Delta$ and MFCC + Δ + $\Delta\Delta$ + CMS features. The LID accuracies obtained from the integrated LID systems have been shown in the 13th, 14th and 15th columns of Table 3. The average LID performances achieved from these integrated LID systems are 69%, 70% and 72%, respectively, which are better compared to individual features. Out of these three integrated LID systems, *src* + MFCC + Δ + $\Delta\Delta$ + CMS features provide the best average LID performance of 72%. Hence, at stage-IV of Fig. 4, we have only shown the integrated LID system based on *src* + MFCC + Δ + $\Delta\Delta$ + CMS feature.

In this work, we have analyzed the statistical significance of variation in identification accuracy using confidence intervals (CIs). Here, we have determined the CIs for 95% significance level. In this study, we computed confidence intervals for 12 pairs of systems. Out of 12 pairs of systems, the difference in the recognition accuracy is statistically significant for 11 pairs. If the confidence interval includes zero, then it indicates that the difference of identification

Table 5
Confidence intervals of the pairs of LID systems developed using different features.

Features	Confidence Intervals
Sub-Seg	{19.9810, 24.0930}
Seg-Supra	{26.0921, 30.2041}
Sub-Supra	{4.0551, 8.1671}
Src-Sub	{24.0551, 28.1671}
Src-Seg	{2.0181, 6.1301}
Src-Supra	{30.1662, 34.2782}
Src-MFCC	{1.8329, 5.9449}
Src-MFCC + Δ + $\Delta\Delta$	{−1.8708, 2.2412}
Src-MFCC + Δ + $\Delta\Delta$ + CMS	{1.8329, 5.9449}
Src-Src + MFCC	{0.3514, 4.4634}
Src-Src + MFCC + Δ + $\Delta\Delta$	{1.6477, 5.7597}
Src-Src + MFCC + Δ + $\Delta\Delta$ + CMS	{3.3144, 7.4264}

accuracies between the pair of systems is not statistically significant. Table 5 shows the CIs for 12 pairs of LID systems. Column 1 indicates the pairs of LID systems considered for this study. Column 2 shows the confidence intervals for various pairs of LID systems. Among 12 pairs of LID systems, a pair of LID systems consisting of *Src* and MFCC + Δ + $\Delta\Delta$ has CI (−1.8708, 2.2412), which includes zero, indicating that the differences among LID accuracies are not significant. This observation also coincides with their individual performances (see columns 9 to 11 of Table 3). It is noteworthy that CIs are far from zero for the set of features providing acutely distinct LID performances. For example, CIs are far from zero for *sub-seg*, *seg-supra*, *src-sub* and *src-supra* systems, because the variations in LID performances are large for these set of features. However, CIs are closer to zero for *sub-supra*, *src-seg*, *src*-MFCC, *src*-MFCC + Δ + $\Delta\Delta$ + CMS, *src-src* + MFCC, *src-src* + MFCC + Δ + $\Delta\Delta$ and *src-src* + MFCC + Δ + $\Delta\Delta$ + CMS because the LID performances are closer for these set of features.

To show the significance of proposed excitation source features for identification of similar kind of languages or dialects, we have carried out experiments on five major dialects of Hindi language. There are five major dialect groups in Hindi language: Western Hindi group, Eastern Hindi group, Rajasthani group, Pahari group and Bihari group. We have used the proposed excitation source features to identify the major five dialects corresponding to the five groups (Khariboli (Western Hindi group), Chattisgarhi (Eastern Hindi group), Marwari (Rajasthani group), Eastern Pahari (Pahari group) and Bhojpuri (Bihari group)). These five dialect groups are abbreviated in Tables 6 and 7 as “Kha”, “Cha”, “Mar”, “EasP” and “Bho”, respectively. Performance of the dialect identification (DI) system using different excitation source and spectral features are given in a confusion matrix form (see Table 6 and Table 7) to understand the confusion pattern. Average performance of the DI system using GFD, RMFCC, MPDSS, PC and ESC features are 51%, 62%, 59%, 55% and 48%, respectively. The average performance is 66% and 64% observed for the DI systems developed using overall excitation source and MFCC + Δ + $\Delta\Delta$ + CMS feature. The combined vocal tract and excitation source information provide 70% accuracy, which is better than individual features. This empirical observation indicates that the proposed excitation source features are also able to identify different dialects within a language as well.

In this work, we have also developed LID systems using *i*-vector based approach. The performance of LID systems developed using GMM and *i*-vector based methods is given in Table 8. From the results, it is observed that the performance of LID systems is slightly better using *i*-vector based approach compared to GMM based approach. About 2–3% improvement is observed in case of *i*-vector based systems. In Table 9, the individual language performances of *i*-vector based systems are shown. It is observed that the individual LID accuracy values obtained from RMFCC and MPDSS features are distinct. For example, RMFCC feature works better than MPDSS feature for Dogri, Gujarati, Malayalam, Nagamese and Nepali languages. Similar observation can be made for PC and ESC features based systems. Therefore, we have combined RMFCC and MPDSS features to capture the *seg* level language-specific information. PC and ESC features are also combined to acquire *supra* level excitation source information. Finally, scores obtained from *sub*, *seg* and *supra* levels are combined to capture overall excitation source information for language discrimination task. LID accuracy of 68% is obtained from overall excitation source feature and *i*-vector based modeling approach. To understand the confusion pattern of *i*-vector based system, confusion matrix of RMFCC feature is shown in Table 10.

Table 6
Dialect identification results (in confusion matrix form) obtained from GFD, RMFCC, MPDSS, PC and ESC features. Khariboli, Chattisgarhi, Marwari, Eastern Pahari and Bhojpuri are abbreviated as “Kha”, “Cha”, “Mar”, “EasP” and “Bho”, respectively.

Dialects	Performances (in %) of DI recognition systems																								
	GFD					RMFCC					MPDSS					PC					ESC				
	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho
Kha	55	10	20	5	10	70	10	5	5	10	65	5	10	10	10	55	20	10	5	10	60	20	10	5	5
Cha	5	45	20	15	15	10	75	5	5	5	25	40	10	10	15	15	50	15	10	10	15	50	15	10	10
Mar	20	25	40	10	5	5	5	50	25	15	15	10	55	15	5	10	15	60	5	10	20	15	40	15	10
EasP	10	20	5	50	15	15	5	10	55	15	5	10	10	60	15	20	10	10	45	15	30	10	20	35	5
Bho	10	10	10	5	65	10	5	15	10	60	10	5	5	5	75	15	5	5	10	65	15	10	10	10	55

Table 7
Dialect identification results (in confusion matrix form) obtained from src, MFCC + Δ + ΔΔ + CMS and src + MFCC + Δ + ΔΔ + CMS features. Khariboli, Chattisgarhi, Marwari, Eastern Pahari and Bhojpuri are abbreviated as "Kha", "Cha", "Mar", "EasP" and "Bho", respectively.

Dialects	Performances (in %) of DI recognition systems														
	src					MFCC + Δ + ΔΔ + CMS					src + MFCC + Δ + ΔΔ + CMS				
	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho	Kha	Cha	Mar	EasP	Bho
Kha	65	10	5	10	10	60	5	20	5	10	70	5	10	5	10
Cha	5	60	10	5	20	10	70	10	5	5	10	70	10	5	5
Mar	5	10	75	5	5	5	25	50	10	10	5	15	65	5	10
EasP	5	5	5	80	5	5	5	5	75	10	5	5	5	80	5
Bho	5	15	10	20	50	15	5	10	5	65	15	5	10	5	65

Table 8
LID Performances obtained from *i*-vector and GMM based approaches.

Classification approaches	Features							
	GFD (*sub*)	RMFCC	MPDSS	PC	ESC	*seg*	*supra*	*src*
GMM	40	61	51	33	18	62	34	66
i-vector	43	63	53	35	21	65	37	68

Table 9
LID performances obtained from *i*-vector based systems.

Languages	Average recognition performances (%)							
	RMFCC	MPDSS	Pitch contour	Epoch strength contour	GFD (*sub*)	*seg*	*supra*	*src*
Arunachali	55	50	10	10	15	55	10	50
Assamese	30	15	35	15	35	10	35	30
Bengali	60	55	5	10	25	65	20	65
Bhojpuri	55	40	80	30	40	50	80	50
Chattisgarhi	100	100	45	40	85	100	70	100
Dogri	95	50	10	20	30	100	20	100
Gojri	100	100	60	40	60	100	60	100
Gujarati	50	5	10	20	10	50	25	50
Hindi	20	20	10	0	0	30	15	20
Indian English	100	100	45	15	100	100	30	100
Kannada	40	40	35	5	55	40	40	40
Kashmiri	70	70	65	25	70	70	60	90
Konkani	20	10	20	35	50	30	10	45
Malayalam	10	0	45	15	55	10	45	40
Manipuri	100	80	0	10	0	100	0	90
Marathi	40	35	0	10	0	50	0	50
Mizo	60	40	30	0	50	55	25	50
Nagamese	90	50	40	45	45	85	50	100
Nepali	95	45	15	5	35	100	10	90
Oriya	40	40	35	25	30	50	40	50
Punjabi	40	50	95	45	90	45	80	50
Rajasthani	100	100	45	70	55	100	50	100
Sanskrit	75	95	25	0	30	95	30	85
Sindhi	50	35	0	45	40	50	0	70
Tamil	50	25	20	10	45	45	30	50
Telugu	100	100	100	20	100	100	100	100
Urdu	65	90	60	10	15	75	65	80
Average performances	63	53	35	21	43	65	37	68

Table 10
LID performances (in confusion matrix form) obtained from segmental level RMFCC features. Here, the numerical values in the table indicate number of test utterances classified under the particular class. The total number of test utterances per language used for evaluation is 20.

Languages	Aru	Ass	Ben	Bho	Cha	Dog	Goj	Guj	Hin	Ind	Kan	Kas	Kon	Mal	Man	Mar	Miz	Nag	Nep	Ori	Pun	Raj	San	Sin	Tam	Tel	Urd
Aru	11	0	0	2	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Ass	0	6	4	0	2	0	0	4	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
Ben	0	0	12	0	0	2	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Bho	0	1	8	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cha	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dog	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Goj	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Guj	0	0	0	2	0	4	0	10	0	0	0	0	1	0	2	0	0	1	0	0	0	0	0	0	0	0	0
Hin	0	1	2	0	0	2	0	0	4	0	0	1	0	5	0	2	0	0	3	0	0	0	0	0	0	0	0
Ind	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kan	0	0	1	0	3	0	0	0	0	0	8	2	0	4	0	0	0	0	0	0	0	0	0	0	2	0	0
Kas	0	1	0	0	0	0	0	1	4	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kon	3	0	0	0	0	0	7	0	0	0	0	0	4	0	0	1	0	0	0	0	0	0	0	5	0	0	0
Mal	0	1	5	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	2	0	0	4	0	4	0	0
Man	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
Mar	0	0	7	0	0	0	0	0	0	0	0	0	0	3	0	8	0	0	0	0	0	0	0	0	2	0	0
Miz	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	8	0	0	0
Nag	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	2	0	0	0
Nep	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
Ori	0	0	0	0	0	0	0	2	0	0	0	6	0	0	0	0	0	0	4	8	0	0	0	0	0	0	0
Pun	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	2	0	0	0	8	0	0	0	0	0	5
Raj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
San	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	4
Sin	0	0	0	0	0	5	0	0	0	0	3	2	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
Tam	0	0	0	0	0	2	0	0	3	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	10	0	0
Tel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
Urd	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	1	0	0	0	13

Table 11
Evaluation of parametric excitation source features on OGI-MLTS database.

Languages	Average recognition performances (%)									
	GFD (*sub*)	RMFCC	MPDSS	Pitch contour	Epoch strength contour	*seg*	*supra*	*src*	MFCC	*src* + MFCC
English	77	66	66	33	88	77	33	77	100	100
Farsi	66	77	100	55	100	77	100	77	77	55
French	100	88	100	22	11	88	11	77	77	88
German	55	66	77	55	11	77	44	77	100	100
Hindi	22	77	100	100	44	88	100	88	55	66
Japanese	22	33	88	44	55	33	55	33	33	0
Korean	11	88	22	66	11	88	33	88	100	100
Mandarin	55	66	33	33	100	66	88	88	100	100
Spanish	77	44	44	77	55	33	55	55	66	77
Tamil	33	100	66	100	77	88	100	100	88	100
Vietnam	55	77	22	11	11	77	11	88	44	88
Average performances	52	71	65	54	51	72	57	77	76	79

Table 12
Average LID performances (in %) of parametric excitation source features obtained from NIST 2011 database.

RMFCC	MPDSS	Pitch contour	Epoch strength contour	GFD (*sub*)	*seg*	*supra*	*src*	MFCC	MFCC + *src*
60	55	23	14	25	64	28	67	70	76

8.5. Evaluation of proposed excitation source features on OGI-MLTS and NIST LRE 2011 databases

The effectiveness of the proposed excitation source features for language identification task has been analyzed on OGI-MLTS database (Muthusamy et al., 1992). 11 languages from 10 different countries are used for LID task. From each language around 1 hour of data are used for building the language models. Three speakers from each language who had not participated during training stage are considered during evaluation. From each speaker, three test utterances, each of 10 s duration are considered for evaluation. LID systems have been developed using proposed excitation source information at the *sub*, *seg* and *supra* levels. Identification accuracies obtained from the empirical analysis portray the similar characteristics of excitation source features as they have been observed for Indian languages. The experimental results obtained from OGI-MLTS database are shown in Table 11. Segmental level features provide better accuracy (72%) compared to the *sub* (52%) and *supra* (57%) levels. The important observation is LID accuracies of individual features are better in OGI-MLTS database compared to IITKGP-MLILSC database. The reason is IITKGP-MLILSC speech corpus consists of 27 Indian regional languages and most of the Indian languages originated from one or two root languages. Hence, there is a lot of similarity among 27 Indian languages. Most of the Indian languages have a common set of phonemes and also follow similar grammatical structure. However, the OGI-MLTS database contains languages across the globe. Most of the languages are from different countries and belong to different origins. So, the discrimination is implicitly present among the languages of OGI-MLTS database. To develop a language identification system, it is necessary to derive non-overlapping language-specific information for each language. Therefore, building a robust automatic language identification system in Indian context is really a challenging task. Hence, the LID accuracy values obtained from OGI-MLTS database using the proposed excitation source features are better compared to the accuracies obtained from Indian languages.

We have also performed LID studies on NIST 2011 database using proposed parametric features at sub-segmental, segmental and supra-segmental levels. The average LID performances are given in Table 12. It is observed that segmental level parametric features provide better accuracy, compared to sub-segmental and supra-segmental levels. We have combined the evidence from all three levels to acquire complete parametric source information (67% LID accuracy). LID performance obtained from vocal tract feature is 70%. Combined vocal tract and excitation source feature provide 76% LID accuracy. Similar trends of LID accuracies are observed on NIST database as

Table 13
$C_{avg} \times 100$ values obtained from parametric excitation source features modeled with *i*-vector based approach on NIST LRE 2011 database.

Features	$C_{avg} \times 100$
MFCC + Δ + ΔΔ + CMS	6.32
GFD	9.48
RMFCC	7.04
MPDSS	8.45
PC	10.14
Esc	12.08
src	5.91
src + MFCC + Δ + ΔΔ + CMS	4.28

observed in IITKGP-MLILSC and OGI-MLTS databases. The conclusion can be drawn from this LID study that proposed parametric excitation source features have potential capability in language discrimination task.

The experimental results on NIST database are also shown in terms of average cost ($C_{avg} \times 100$) in Table 13. Average cost of the overall system $C_{avg} \times 100$ as defined in NIST LRE 2009 Evaluation Plan was calculated for excitation source and MFCC feature based LID systems. Low $C_{avg} \times 100$ value indicates the system with better accuracy. The excitation source and MFCC feature based systems are fused to improve the accuracy. $C_{avg} \times 100$ values for MFCC and excitation source feature (*src*) based systems are 6.32 and 5.91, respectively. The $C_{avg} \times 100$ value of the fused system has improved to 4.28. This result shows the importance of excitation source features for language identification.

9. Summary and conclusions

In this work, a framework has been proposed for deriving the language-specific excitation source information. Parametric methods are proposed to estimate excitation source information at *sub*, *seg* and *supra* levels. Experimental analysis indicates that language-specific information present at *sub*, *seg* and *supra* levels are fundamentally distinct. Therefore, the scores obtained from three different levels are combined to achieve complete excitation source information. Segmental level information provides better LID accuracy (62%), compared to sub-segmental (40%) and supra-segmental (34%) levels. This indicates the importance of *seg* level excitation source information for language discrimination. The scores of excitation source information are further combined with the scores obtained from vocal tract features to examine the existence of non-overlapping language-specific information between these two features. The stage-IV integrated LID system I, II and III provide LID accuracies of 69%, 70% and 72% respectively. The proposed excitation source features are also evaluated on large and publicly available databases (OGI-MLTS and NIST-LRE-2011), and the identification performances show a similar trend as it is observed with Indian language database. The statistical significance of the LID accuracies obtained by various proposed features is analyzed using confidence intervals. The LID systems were also developed using *i*-vector based approach. The average LID accuracy of *i*-vector based systems at *sub*, *seg* and *supra* levels are 43%, 65% and 37%, respectively. The complete excitation source feature (*src*) provides 68% LID accuracy. To show the significance of excitation source features, five different dialects of Hindi language were also explored.

Acknowledgement

This research is funded by the Department of Information Technology (DIT), Government of India (11(6)/2011-HCC(TDIL)), through the project “Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages”.

References

Ananthapadmanabha, T.V., Fant, G., 1982. Calculation of true glottal flow and its components. Speech Commun. 1, 167–184. Elsevier.
Atal, B.S., 1972. Automatic speaker recognition based on pitch contours. J. Acoustic. Soc. Am. 52 (6), 1687–1697.
Audacity Team, 2006. Audacity®: Free Audio Editor and Recorder [Computer program]. Version 1.2.6 retrieved November 16th 2006 from <http://audacity.sourceforge.net/> (accessed 25.05.16.).

Bajpai, A., Yegnanarayana, B., Exploring features for audio clip classification using LP residual and AANN models, in Proceedings on International Conference on Intelligent Sensing and Information Processing, pp. 305–310, January, 2004.

Balleda, J., Murthy, H.A., Nagarajan, T., Language identification from short segments of speech, in International Conference on Spoken Language Processing, pp. 1033–1036, October, 2000.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Proc. 19 (4), 788–798.

Dehak, N., Carrasquillo, P.A.T., Reynolds, D., Dehak, R., Language recognition via i-vectors and dimensionality reduction, in Proceedings of Interspeech, August, 2011b.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B 39 (1), 1–38.

Dominguez, J.G., Moreno, I.L., Sak, H., Rodriguez, J.G., Moreno, P.J., Automatic language identification using long short-term memory recurrent neural networks, in Proceedings of Interspeech, September, 2014.

Gray, A.H., Markel, J.D., 1974. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Trans. Acoustic Speech Signal Proc. 22 (3), 207–217.

Gupta, C.S., Prasanna, S.R.M., Yegnanarayana, B., Autoassociative neural network models for online speaker verification using source features from vowels, in IEEE International Joint Conference on Neural Networks, May, 2002.

Hayakawa, S., Takeda, K., Itakura, F., 1997. Speaker identification using harmonic structure of LP-residual spectrum. In: Biometric Personal Authentication, vol. 1206. Lecture Notes. pp. 253–260.

Informer Technologies Inc., VentiTV software. <http://ventitv.software.informer.com/> (accessed 31.05.15.).

Jothilakshmi, S., Ramalingam, V., Palanivel, S., 2012. A hierarchical language identification system for Indian languages. Digital Signal Proc. 22 (3), 544–553. Elsevier.

Li, M., Narayanan, S., 2014. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification. Comput. Speech Lang. 28, 940–958. Elsevier.

Maity, S., Vuppala, A.K., Rao, K.S., Nandi, D., IITKGP-MLILSC speech database for language identification, in National Conference on Communications, February, 2012.

Makhoul, J., 1975. Linear prediction: a tutorial review. IEEE Proc. 63 (4), 561–580.

Martinez, D., Plchot, O., Burget, L., Glembek, O., Matejka, P., Language identification in i-vectors space, in Interspeech, Florence, Italy, 2011.

Mary, L., 2006. Multilevel implicit features for language and speaker recognition (Ph.D. dissertation). Indian Institute of Technology Madras, India.

Mary, L., Yegnanarayana, B., Autoassociative neural network models for language identification, in Proceedings in International Conference on Intelligent Sensing and Information Processing, pp. 317–320, 2004.

Matlab and Statistics Toolbox Release, 2012b. The MathWorks, Inc. Natick, Massachusetts, United States.

Moreno, I.L., Dominguez, J.G., Plchot, O., Martinez, D., Rodriguez, J.G., Moreno, P., Automatic language identification using deep neural networks, in IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), May, 2014.

Morgan, D., Riek, L., Mistretta, W., Scofield, C., Grouin, P., Hull, F., Experiments in language identification with neural networks, in International Joint Conference on Neural Networks (IJCNN), vol. 2, pp. 320–325, 1992.

Murthy, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signal. IEEE Trans. Audio Speech Lang. Proc. 16 (8), 1602–1613.

Murthy, K.S.R., Yegnanarayana, B., 2009. Characterization of glottal activity from speech signal. IEEE Signal Proc. Lett. 16 (6), 469–472.

Murty, K.S.R., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Proc. 16 (8), 1602–1613.

Muthusamy, Y.K., Cole, R.A., Oshika, B.T., The OGI Multilanguage Telephone Speech Corpus, in Spoken Language Processing, pp. 895–898, October, 1992.

Naylor, P., Kounoudes, A., Gudnason, J., Brookes, M., 2007. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio Speech Lang. Proc. 15 (1), 34–43.

Pati, D., Prasanna, S.R.M., 2011. Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. Int. J. Speech Technol. 14 (1), 49–63. (Springer).

Pati, D., Prasanna, S.R.M., 2013. A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information. Sadhana 38 (4), 591–620.

Plumpe, M.D., Quatieri, T.F., Reynolds, D.A., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Proc. 7 (5), 569–586.

Qi, Y., Bi, N., 1994. A simplified approximation of the four-parameter LF model of voice source. J. Acoust. Soc. Am. 96 (2), 1182–1185.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Rao, K.S., Koolagudi, S.G., 2013. Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. 16, 181–201. (Springer).

Rao, K.S., Maity, S., Reddy, V.R., 2013. Pitch synchronous and glottal closure based speech analysis for language recognition. Int. J. Speech Technol. 16 (4), 413–430. (Springer).

Reddy, V.R., Maity, S., Rao, K.S., 2013. Identification of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. 16 (4), 489–511. (Springer).

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Proc. 3 (1), 72–83.

Singh, O.P., Haris, B.C., Sinha, R., Language identification using sparse representation: a comparison between GMM supervector and i-vector based approaches, in IEEE India Conference (INDICON), December, 2013.

Sugiyama, M., Automatic language recognition using acoustic features, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 813–816, May, 1991.

Travadi, R., Segbroeck, M.V., Narayanan, S., Modified-prior i-vector estimation for language identification of short duration utterances, in Proceedings of Interspeech, September, 2014.

Vanishree, V.M., 2011. Provision for linguistic diversity and linguistic minorities in India (Master's thesis). Applied Linguistics, St. Mary's University College, Strawberry Hill, London.

Website cell, IT Unit, NSD, All India Radio. <http://www.newsonair.nic.in/Regional_NSD_Search_MP3.aspx> (accessed 17.06.16.).

Wolf, J.J., 1972. Efficient acoustic parameters for speaker recognition. J. Acoustic. Soc. Am. 51 (2), 2044–2055.

Yegnanarayana, B., Avendano, C., Hermansky, H., Murthy, P.S., Processing linear prediction residual for speech enhancement, in European Conference on Speech Communication and Technology, pp. 1399–1402, September, 1997.

Yegnanarayana, B., Prasanna, S., Zachariah, J., Gupta, C., 2005. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. Speech Audio Proc. 13 (4), 575–582.

Yegnenarayana, B., Murthy, K.S.R., 2009. Event based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Proc. 17 (4), 614–624.

Zissman, M.A., Automatic language identification using Gaussian mixture and hidden Markov models, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 399–402, 1993.