Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 68–87

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Implicit processing of LP residual 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, Dimapur 797103, Nagaland, India

Received 27 September 2013; received in revised form 6 June 2016; accepted 10 June 2016

Available online 16 June 2016

Abstract

Present work explores the excitation source information for the language identification (LID) task. In this work, excitation source information is captured by implicit processing of linear prediction (LP) residual signal for discriminating the languages. Raw samples of LP residual signal, its magnitude, and phase components are processed independently at sub-segmental, segmental and suprasegmental levels for extracting the language-specific excitation source information. The LID studies are carried out using 27 Indian languages from Indian Institute of Technology Kharagpur-Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) and 11 international languages from OGI-MLTS corpus. The Gaussian mixture models (GMMs) are used in this work to model the language-specific excitation source information for LID task. From the experimental results, it can be observed that, features extracted from segmental level yields better identification accuracy (50.92%), compared to sub-segmental (47.77%) and suprasegmental levels (43.88%). Further, the evidence from all three levels is combined to obtain the complete excitation source information. Finally, we have investigated the existence of non-overlapping language-specific information present in excitation source and vocal tract features.

Keywords: Hilbert envelope; Residual phase; LP residual; Language identification (LID); Subsegmental; Segmental; Suprasegmental; Excitation source information; IITKGP-MLILSC; OGI-MLTS; Implicit processing of LP residual

1. Introduction

Speech consists of a sequence of sound units. These sound units have respective language-specific constraints and are also influenced by speaking style. Hence, speech signal not only carries the message information, but also the speaker and language information. The objective of language identification (LID) system is to determine the language identity from the uttered speech accurately. Indian languages belong to several language groups and subgroups. Two major language groups are the Indo-Aryan languages spoken by 76.86% of Indian citizens, and the Dravidian languages spoken by 20.82% of Indians (Vanishree, 2011). The Indo-Aryan languages are highly influenced by Sanskrit, whereas the Dravidian languages have a history, independent of Sanskrit (Vanishree, 2011). However, Dravidian languages Telugu and Malayalam are influenced by Sanskrit. Most of the Indian languages have a common set of phonemes and also follow similar grammatical structure. To develop an LID system, it is necessary to derive non-overlapping language-specific information for each language. Therefore, building a robust automatic LID system in

* 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.2016.06.002

0885-2308/

the Indian context is a challenging task. Developing automatic LID system has become a research issue due to its several real-life applications such as speech to speech translation systems, information retrieval from multilingual audio databases and multilingual speech recognition systems.

Speech can be viewed as a convolution of vocal tract system response and the excitation source signal. The quasi-periodic air pulses generated by the vibration of the vocal folds constitute the major source for exciting the vocal tract resonator during the production of voiced sounds. For unvoiced speech production, the expiration of air is constrained either completely (e.g. unvoiced stops) or partially at different places along the vocal tract, acting as a source of excitation to the vocal tract resonator (e.g. fricatives). The characteristics of both vocal tract system and excitation source are embedded in the speech signal and have significant contribution in producing sound units of a particular language. State-of-the-art automatic LID systems mostly exploit the vocal tract information represented by spectral features such as Mel-frequency cepstral coefficients (MFCCs) (Rao et al., 2013), linear prediction cepstral coefficients (LPCCs) (Sugiyama, 1991) and suprasegmental level information such as prosodic features (Mary and Yegnanarayana, 2004; Reddy et al., 2013) to capture the language-specific knowledge from the speech signal. Recent works have also explored the language-specific phonetic information explicitly at phoneme level (Siniscalchi et al., 2013), for LID task. However, no systematic studies are using the characteristics of excitation source in the context of language identification. Therefore, in the present work, we have focused on modeling the language-specific excitation source information. The primary objective of this paper is to examine whether excitation source features contain any language-specific information.

Linear prediction (LP) residual signal, obtained by inverse filtering the speech signal, mostly represents excitation source information (Makhoul, 1975). Linear prediction analysis represents the second order statistical features in terms of the autocorrelation coefficients. So, the LP residual (LPR) signal does not represent any significant second order relations corresponding to the vocal tract resonator, and it contains only the higher order relations (Prasanna et al., 2006). We hypothesize that language-specific information may present in the higher-order relations of the LP residual samples, and it's hard to capture the higher order relations present in the LP residual signal using parametric techniques. In this paper, the implicit relations among the raw LP residual samples are analyzed to model the language-specific excitation source information at different levels. In the present work, raw LP residual samples are analyzed in three different frame sizes: 5 ms (subsegmental level), 20 ms (segmental level) and 250 ms (suprasegmental level). Processing LP residual samples at the subsegmental level will provide the crucial information about consonant and transition regions. Processing of LP residual at segmental level (frame size of 20 ms) may capture the language-specific co-articulation knowledge. Suprasegmental level processing (frame size of 250 ms) of LP residual derives the language-specific contextual or phonotactic information.

However, the magnitude component of LP residual signal may predominate over phase component of LP residual signal during the direct processing of raw samples. The phase component of LP residual signal may also contain information about the excitation source. Therefore, we have processed raw LP residual samples, magnitude and phase components of LP residual signal independently to capture language-specific excitation source information. The magnitude and phase components can be separated by deriving the analytic signal of LP residual (Cohen, 1989). Finally, the confidence scores obtained from sub-segmental (*sub*), segmental (*seg*) and suprasegmental (*supra*) levels are combined to enhance the LID performance. The complementary nature of excitation source and vocal tract features is also investigated by combining the scores of these two features for enhancing the performance accuracy of LID.

The rest of the paper is organized as follows: Section 2 describes prior works on automatic language identification area. The motivation for the present work is represented in Section 3. Section 4 discusses the details of the language databases used in this work. Section 5 describes the proposed excitation source features and their extraction process. A succinct description of vocal tract features is also given in this section. Section 6 explains the experimental methodology. Development of LID systems using spectral and excitation source features is discussed in Section 7. Language identification results using proposed excitation source features on IITKGP-MLILSC are given in Section 8. Section 9 provides the LID results on well-known OGI-MLTS database. Section 10 summarizes and concludes the present work.

2. Previous works

Most of the modern LID systems are developed using vocal tract features. Sugiyama (1991) has explored linear prediction coefficients (LPCs) and cepstral coefficients (LPCCs) for language recognition. Zissman and Singer (1994)

have proposed the Gaussian mixture models (GMMs) (Reynolds and Rose, 1995) for language identification study. Ma et al. (2007) have proposed ensembles of binary classifiers for language recognition. They have performed polynomial expansion of cepstral features for LID. In 2013, Siniscalchi et al. (2013) proposed a novel universal acoustic characterization approach for language recognition (LRE). They have proposed a *universal* set of fundamental units which can be defined across all the languages. This work has exploited some speech attributes like manner and place of articulation of sound units to define the *universal* set of language-specific fundamental units. Roy and Das (2013) have proposed a hybrid approach using the combination of VQ and GMM for identifying a language. They have carried out LID experiment in four Indian languages. This work concludes that hybrid approach gives better results when compared with the baseline GMM system.

In the Indian context, Balleda et al. (2000) have first attempted to identify Indian languages. Mel-frequency cepstral coefficients (MFCCs) have been explored for LID task. In 2004, Mary and Yegnanarayana (2004) have explored the auto-associative neural networks (AANN) for capturing language-specific features. 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 LID task. 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 determines the language group of a given test utterance and then identifies the target language inside 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.

From the prior works, it is observed that most of the works on LID 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 (Pati and Prasanna, 2011; Yegnanarayana et al., 2005), speech enhancement (Yegnanarayana et al., 1997) and emotion recognition (Rao and Koolagudi, 2013) tasks. In the present study, we analyze the LP residual signal at subsegmental, segmental and suprasegmental levels to capture different aspects of excitation source for language discrimination task.

3. Motivation

State-of-the-art LID systems mostly approximate the dynamics of vocal tract shapes and use this 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 explored heretofore for identifying the languages. The motivation of the present work is to study the excitation source information for LID task. The motivation for using excitation source features for LID task can also be observed through the correlation coefficients presented in 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 languages. Correlation determines the degree of similarity between two signals. Suppose we have two real signals $x(n)$ 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, 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 $x(n) = y(n)$, the procedure is known as *autocorrelation* of $x(n)$. From each language database, one male speaker is considered, and the LP residual (LPR) feature has been extracted. Segmental (*seg*) level LP residual feature (see Section 5.2.2 for detail explanation) is used for analyzing the correlation. To normalize the speaker variability between the languages, the mean subtraction is imposed to all the feature vectors across all languages. Then the *seg* level feature vectors are modeled with 128 Gaussian mixtures for each language. The average

Table 1
Correlation coefficients across the languages derived using segmental level LP residual samples.

Languages	Correlation coefficients																										
Arunachali	**3.8**	0.83	0.52	0.8	1.02	0.97	0.7	0.49	0.38	0.87	0.93	0.5	0.66	0.42	0.98	0.32	0.78	1.21	0.48	1.21	1.98	0.63	1.1	0.64	0.5	1.4	1.59
Assamese	0.83	**3.26**	0.59	1.18	1.32	0.69	1.68	0.94	0.69	1.09	1.32	0.58	0.91	1.06	1.7	0.78	1.07	1.46	0.41	1.31	1.3	1.4	1.72	1.23	0.83	1.53	1.88
Bengali	0.52	0.59	**2.9**	0.88	0.61	0.57	0.56	0.82	0.82	0.71	0.57	0.63	0.66	0.51	0.54	0.39	0.36	0.74	0.54	0.67	0.87	0.67	0.57	0.36	0.74	0.8	0.68
Bhojpuri	0.8	1.18	0.88	**3.3**	0.72	0.72	0.67	0.95	0.82	1.2	1.26	0.91	1.07	0.79	1.29	0.59	1.19	1.07	0.58	1.21	1.53	1.16	1.53	0.66	1.17	1.39	0.92
Chhattisgarhi	1.02	1.32	0.61	0.72	**3.42**	1.21	1.2	0.74	0.63	0.88	1.53	0.63	0.91	0.88	1.47	0.73	0.91	2.73	0.72	2.18	2.71	0.93	2.23	0.84	0.92	2.74	2.46
Dogri	0.97	0.69	0.57	0.72	1.21	**2.6**	0.62	0.38	0.54	0.71	1.11	0.67	0.69	0.57	1.2	0.53	0.71	1.57	0.59	1.19	1.66	0.55	1.47	0.72	0.64	1.53	1.91
Gojri	0.7	1.68	0.56	0.67	1.2	0.62	**4**	0.76	1.11	0.87	1.37	0.49	0.76	1.19	1.41	1.38	1.04	2.8	0.65	0.91	0.04	0.77	0.72	0.94	0.5	1.63	1.29
Gujarati	0.49	0.94	0.82	0.95	0.74	0.38	0.76	**4.01**	0.64	0.69	0.8	0.55	0.57	0.72	0.75	0.42	0.55	0.87	0.54	0.78	1.63	0.86	0.94	0.64	0.72	1.07	0.92
Hindi	0.38	0.69	0.82	0.82	0.63	0.54	1.11	0.64	**3.67**	0.69	1.12	0.46	0.91	0.93	0.59	0.54	0.47	0.83	0.57	0.75	0.86	0.8	0.93	0.38	0.95	1	0.79
Indian English	0.87	1.09	0.71	1.2	0.88	0.71	0.87	0.69	0.69	**3.53**	0.98	0.55	1.09	0.9	0.75	0.74	0.73	1.28	0.5	1.21	0.87	0.83	1.15	0.93	0.74	0.92	1.05
Kannada	0.93	1.32	0.57	1.26	1.53	1.11	1.37	0.8	1.12	0.98	**4.12**	0.53	1.12	1.05	1.18	0.91	1.02	1.96	0.56	1.13	1.14	0.7	1.63	0.83	0.59	1.62	2.33
Kashmiri	0.5	0.58	0.63	0.91	0.63	0.67	0.49	0.55	0.46	0.55	0.53	**2.82**	0.69	0.46	0.85	0.46	0.65	1.23	0.47	0.92	1.2	0.73	0.71	0.44	0.53	1.01	1.02
Konkani	0.66	0.91	0.66	1.07	0.91	0.69	0.76	0.57	0.91	1.09	1.12	0.69	**3.1**	1.15	1.08	0.57	0.6	0.48	0.8	0.76	2.2	0.67	1.32	0.56	0.94	1.35	0.93
Malayalam	0.42	1.06	0.51	0.79	0.88	0.57	1.19	0.72	0.93	0.9	1.05	0.46	1.15	**4.22**	0.96	1.02	0.6	0.76	0.56	0.84	1.09	1.05	1.5	0.67	0.97	1.13	1
Manipuri	0.98	1.7	0.54	1.29	1.47	1.2	1.41	0.75	0.59	0.75	1.18	0.85	1.08	0.96	**3.9**	0.56	1.48	1.84	0.42	1.54	1.59	1.02	2.25	1.06	1.23	1.69	2.42
Marathi	0.32	0.78	0.39	0.59	0.73	0.53	1.38	0.42	0.54	0.74	0.91	0.46	0.57	1.02	0.56	**3.34**	0.52	1.59	0.41	0.61	0.15	0.54	0.36	0.55	0.5	1.06	0.9
Mizo	0.78	1.07	0.36	1.19	0.91	0.71	1.04	0.55	0.47	0.73	1.02	0.65	0.6	0.6	1.48	0.52	**3.67**	1.35	0.48	1.21	1.24	0.63	1.1	0.51	0.68	1.24	1.59
Nagamese	1.21	1.46	0.74	1.07	1.73	1.57	1.3	0.87	0.83	1.28	1.96	1.23	0.48	0.76	1.84	1.59	1.35	**3.21**	0.41	1.001	0.14	0.99	1.12	1.22	1.09	1.12	1.62
Nepali	0.48	0.41	0.54	0.58	0.72	0.59	0.65	0.54	0.57	0.5	0.56	0.47	0.8	0.56	0.42	0.41	0.48	0.41	**3.27**	0.48	0.88	0.6	0.74	0.43	0.61	1.23	0.53
Oriya	1.21	1.31	0.67	1.21	2.18	1.19	0.91	0.78	0.75	1.21	1.13	0.92	0.76	0.84	1.54	0.61	1.21	1.5	0.48	**3.98**	1.84	0.98	1.97	0.92	1.22	2.29	2.23
Punjabi	1.98	1.3	0.87	1.53	1.51	1.66	0.04	1.63	0.86	0.87	1.14	1.2	2.2	1.09	1.59	0.15	1.24	0.14	0.88	1.84	**3.82**	1.79	1.9	0.84	1.57	1.75	2.03
Rajasthani	0.63	1.4	0.67	1.16	0.93	0.55	0.77	0.86	0.8	0.83	0.7	0.73	0.67	1.05	1.02	0.54	0.63	0.99	0.6	0.98	1.79	**3.48**	1.44	0.66	0.62	1.23	1.17
Sanskrit	1.1	1.72	0.57	1.53	2.23	1.47	0.72	0.94	0.93	1.15	1.63	0.71	1.32	1.5	2.25	0.36	1.1	1.12	0.74	1.97	2.02	1.44	**3.82**	0.88	1.18	1.5	0.862
Sindhi	0.64	1.23	0.36	0.66	0.84	0.72	0.94	0.64	0.38	0.93	0.83	0.44	0.56	0.67	1.06	0.55	0.51	1.22	0.43	0.92	0.84	0.66	0.88	**3.57**	0.4	1.19	1.44
Tamil	0.5	0.83	0.74	1.17	0.92	0.64	0.5	0.72	0.95	0.74	0.59	0.53	0.94	0.97	1.23	0.5	0.68	1.09	0.61	1.22	1.57	0.62	1.18	0.4	**3.17**	1.28	0.92
Telugu	1.4	1.53	0.8	1.39	2.74	1.53	1.63	1.07	1	0.92	1.62	1.01	1.35	1.13	1.69	1.06	1.24	1.12	1.23	2.29	1.45	1.23	1.07	1.19	1.28	**4.01**	1.01
Urdu	1.59	1.88	0.68	0.92	2.46	1.91	1.29	0.92	0.79	1.05	2.33	1.02	0.93	1	2.42	0.9	1.59	1.62	0.53	2.23	1.09	1.17	1.62	1.44	0.92	1.82	**3.56**

We have used bold format for marking the diagonal elements, which represent auto-correlation coefficients.

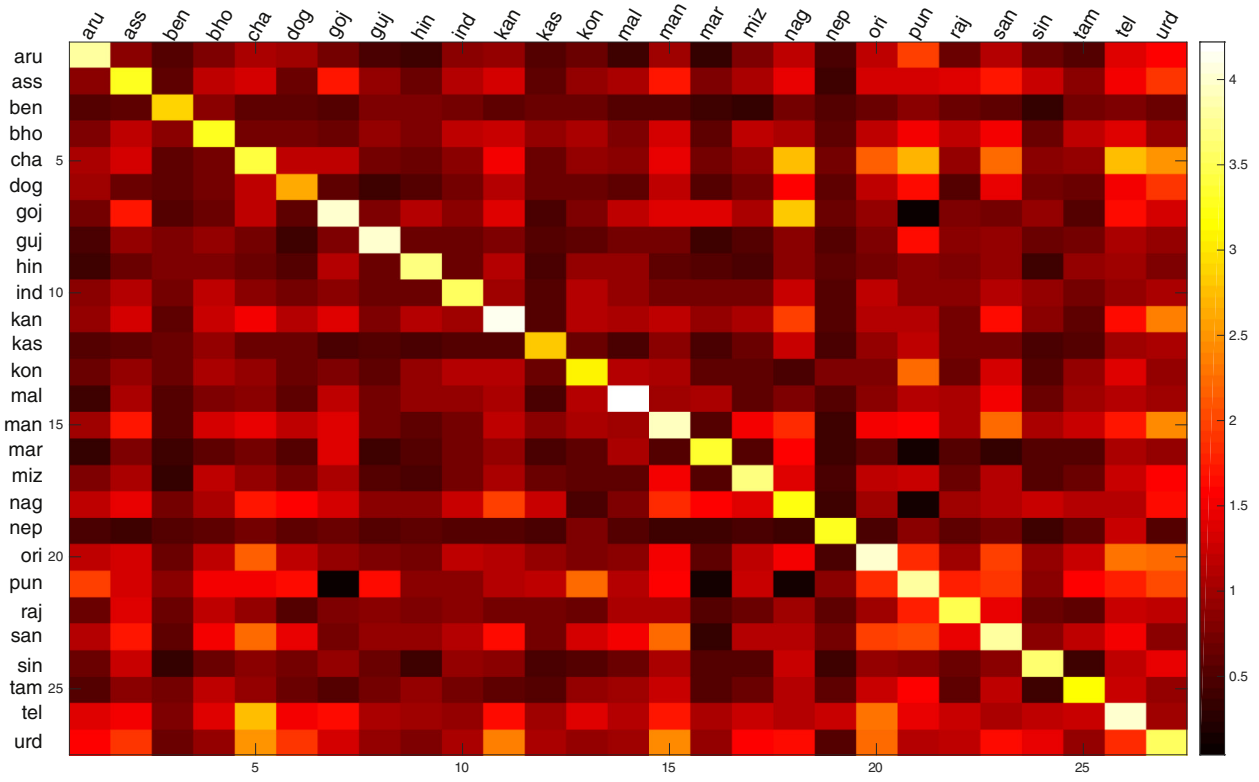

Fig. 1. Heat map representation of correlation coefficients across the languages derived using segmental level LP residual samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of 128 *mean* vectors is considered as the signal for a particular language to compute the correlation coefficients. The significance of *seg* level LP residual feature in language discrimination task has been shown by analyzing these correlation coefficients. The correlation coefficients between two signals are a sequence of length $(2l - 1)$. The average of the coefficients is considered in our work which is shown in Table 1.

The values of the first row of Table 1 indicate the correlation coefficients of the first language with respect to itself and other 26 languages. The first element of the first row shows the auto-correlation coefficient of the first language. The other 26 values of the first row represent the cross-correlation coefficients between the first language and other 26 languages. Lower cross-correlation coefficient value between two languages indicates more dissimilarity between them. We have taken the average of the 26 cross-correlation coefficients from the 3rd column to 27th column of the 1st row which represents average cross-correlation coefficient of the first language (Arunachali) with respect to other 26 languages. This average cross-correlation coefficient value (0.84) is less than the auto-correlation coefficient value (3.8) which resides in the 2nd column of the 1st row. This indicates that the seg level LP residual feature has significant language discriminative capability. If we analyze the other rows of Table 1, similar characteristics can be observed.

The correlation coefficients of 27 languages are shown in heat map format (see Fig. 1). All the diagonal cells represent the correlation coefficients of a particular language with respect to itself. Other cells represent cross-correlation coefficients. Lower cross-correlation coefficient value between two languages indicates more dissimilarity between them. In Fig. 1 the black and white colors represent the lowest and highest correlation coefficient values of Table 1, respectively (see the color index bar right side of Fig. 1). The intensity of color varies smoothly from black through shades of red, orange and yellow to white. Hence, the color of the diagonal blocks should be lighter, compared to the other blocks. It can be observed from Fig. 1 that the color of diagonal blocks is brighter, compared to the other blocks. This indicates that the *seg* level LP residual feature has significant language discriminative capability.

Table 2
Correlation coefficients across the languages derived using LPR *sub*, LPR *supra*, HE *sub*, HE *seg*, HE *supra*, RP *sub*, RP *seg* and RP *supra* level features.

Languages	Correlation coefficients															
	LPR *sub*		LPR *supra*		HE *sub*		HE *seg*		HE *supra*		RP *sub*		RP *seg*		RP *supra*	
	WL	BL	WL	BL	WL	BL	WL	BL	WL	BL	WL	BL	WL	BL	WL	BL
Arunachali	1.86	0.13	0.9	0.02	0.97	0.19	0.46	0.06	0.91	0.02	1.18	0.04	0.7	0.05	0.8	0.09
Assamese	1.14	0.04	0.85	0.01	0.98	0.2	0.8	0.34	0.96	0.07	0.71	0.09	0.78	0.02	0.58	0.1
Bengali	0.6	0.01	0.3	0.06	0.45	0.1	0.56	0.16	1.2	0.4	0.33	0.01	0.31	0.06	0.7	0.08
Bhojpuri	1.09	0.02	0.7	0.2	0.26	0.03	0.9	0.15	0.6	0.19	0.5	0.07	0.4	0.02	0.25	0.07
Chhattisgarhi	0.81	0.21	0.4	0.09	0.62	0.08	0.67	0.08	0.5	0.01	0.8	0.04	0.32	0.09	0.97	0.15
Dogri	0.42	0.11	0.6	0.02	0.74	0.05	0.59	0.02	1.4	0.3	1.24	0.05	0.57	0.06	0.89	0.13
Gojri	0.81	0.42	1.81	0.9	0.42	0.08	0.62	0.08	0.31	0.07	1.1	0.15	0.49	0.05	0.25	0.17
Gujarati	0.5	0.14	0.53	0.1	0.64	0.02	0.8	0.01	0.76	0.02	1.19	0.07	0.8	0.01	0.53	0.15
Hindi	0.48	0.09	0.3	0.12	0.31	0.09	0.91	0.2	0.82	0.04	1.2	0.04	1.32	0.75	1.07	0.3
Indian English	0.8	0.03	0.54	0.21	0.88	0.19	0.79	0.18	0.94	0.07	0.98	0.17	1.5	0.8	0.47	0.2
Kannada	0.69	0.2	0.42	0.08	0.41	0.04	0.46	0.07	0.84	0.28	0.6	0.03	1.26	0.05	1.38	0.2
Kashmiri	0.891	0.08	2.03	0.18	0.9	0.36	0.98	0.2	0.52	0.21	1.4	0.06	1.26	0.07	0.85	0.56
Konkani	0.48	0.01	0.85	0.01	0.71	0.35	0.51	0.11	0.44	0.06	0.6	0.09	0.53	0.02	0.51	0.29
Malayalam	0.46	0.23	0.46	0.19	0.53	0.03	0.6	0.09	1.21	0.2	0.65	0.04	0.58	0.05	0.75	0.12
Manipuri	1.3	0.21	0.56	0.08	0.4	0.09	0.7	0.06	0.43	0.13	1.68	0.4	0.87	0.35	0.51	0.08
Marathi	1.32	0.05	0.31	0.01	0.84	0.33	0.74	0.09	0.39	0.12	0.76	0.08	0.7	0.08	0.4	0.19
Mizo	1.1	0.08	0.71	0.21	0.63	0.09	0.68	0.08	0.82	0.13	0.22	0.04	0.51	0.16	0.67	0.08
Nagamese	0.62	0.15	0.82	0.41	0.72	0.04	0.7	0.33	0.76	0.09	0.91	0.24	0.41	0.03	0.63	0.04
Nepali	0.49	0.15	0.64	0.05	0.4	0.15	0.8	0.07	0.84	0.21	0.4	0.16	1.86	0.07	0.68	0.25
Oriya	1.66	0.19	0.3	0.08	0.68	0.07	0.97	0.27	0.84	0.03	1.13	0.08	1.04	0.1	1.05	0.37
Punjabi	2.1	0.14	2.09	0.4	0.76	0.17	0.83	0.19	0.8	0.04	0.79	0.34	1.5	0.28	0.38	0.08
Rajasthani	0.35	0.04	0.9	0.2	0.65	0.21	0.93	0.26	0.91	0.22	0.6	0.03	0.8	0.04	0.44	0.06
Sanskrit	1.1	0.08	0.46	0.04	0.59	0.22	0.99	0.09	0.75	0.09	0.71	0.16	0.52	0.24	1.32	0.21
Sindhi	0.27	0.03	0.78	0.09	0.62	0.05	0.67	0.12	0.6	0.02	0.42	0.06	0.31	0.17	0.81	0.32
Tamil	0.29	0.11	0.65	0.13	0.93	0.37	0.87	0.2	0.32	0.01	0.81	0.24	0.45	0.07	0.3	0.07
Telugu	0.17	0.01	0.5	0.23	0.7	0.05	0.42	0.07	0.84	0.07	0.85	0.09	1.03	0.19	0.84	0.02
Urdu	0.38	0.1	0.26	0.08	0.5	0.04	0.98	0.27	0.77	0.24	0.74	0.04	0.32	0.07	0.71	0.06

The correlation coefficients are also computed using (i) sub and supra level LPR features and (ii) Hilbert envelope (HE) and Residual phase (RP) features at the sub, seg and supra levels (see Table 2). Correlation coefficient within a language (abbreviated as WL) has been computed from two different speech utterances spoken by a speaker. The first element of the 2nd column of Table 2 indicates the auto-correlation coefficient of Arunachali language computed by using sub level LPR feature. The first element of 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 (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 (0.13) is less than the auto-correlation coefficient value of Arunachali language (1.86). This portrays that the sub level LPR feature has significant language discriminative capability. Similar characteristics have been observed for other languages also. The correlation coefficients computed from HE and RP features also show the significant language discrimination capability (see Table 2). These correlation coefficients are displayed in heat map representation (see Fig. 2) for better visualization. The 1st, 3rd, 5th, 7th, 9th, 11th, 13th and 15th columns represent the auto-correlation coefficients computed by using LPR *sub*, LPR *supra*, HE *sub*, HE *seg*, HE *supra*, RP *sub*, RP *seg* and RP *supra* features, respectively (see Fig. 2). The average of cross-correlation coefficients of a particular language with respect to other 26 languages belong to the even columns (tagged as BL) of Fig. 2.

The "WL" tagged columns of Fig. 2 contain lighter color, as these columns represent auto-correlation coefficients. Whereas, "BL" tagged columns hold comparatively dark colors, as these columns represent average cross-correlation coefficients between the languages. This theoretical discussion elicits the significance of the excitation source features in language identification task, which is the motivation for the present work.

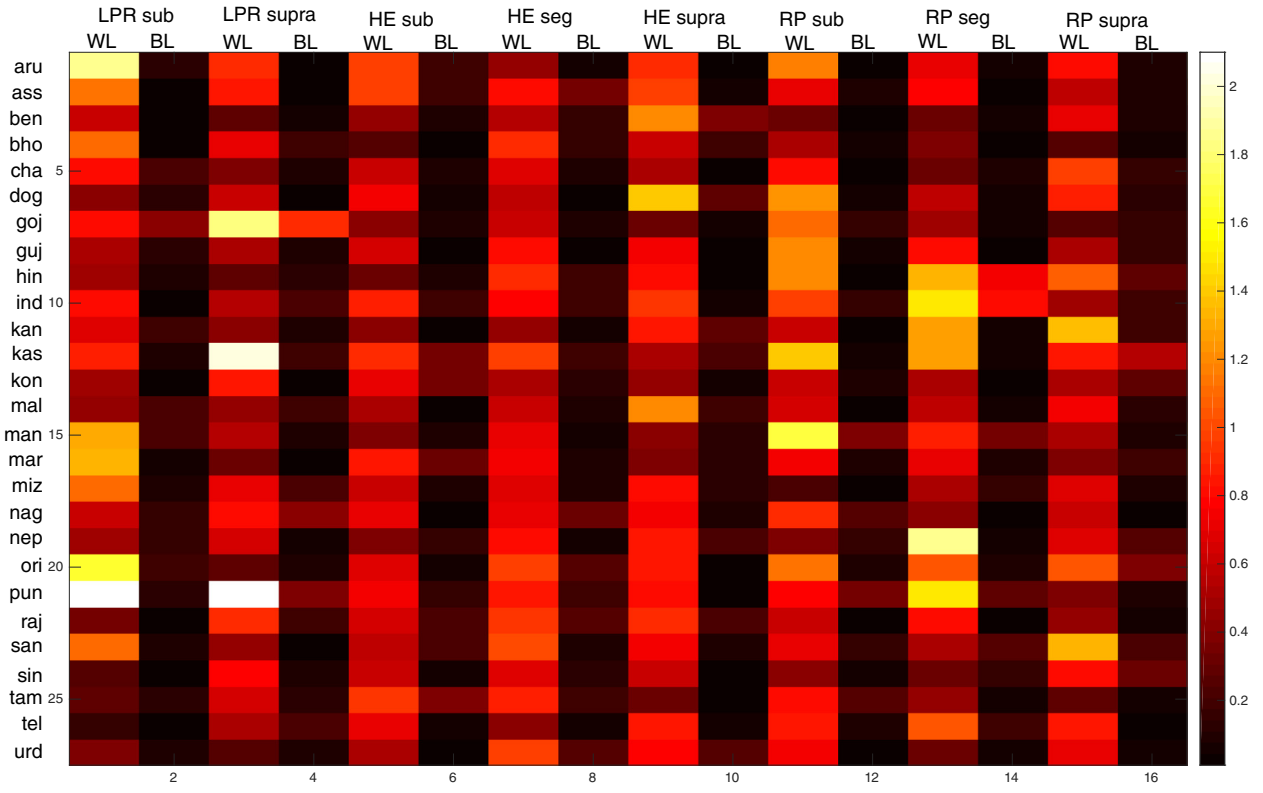

Fig. 2. Heat map representation of correlation coefficients across the languages derived using LPR *sub*, LPR *supra*, HE *sub*, HE *seg*, HE *supra*, RP *sub*, RP *seg* and RP *supra* level features.

4. Speech corpora

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

In the present 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 is recorded from broadcasted TV talk shows, live shows, interviews and news bulletins. Each language database consists of one hour of speech data including both male and female speakers. The broadcasted television channels are accessed using VentiTV software and the Pixelview TV tuner card. 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 Maity et al. (2012).

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

Oregon Graduate Institute (OGI) Multi-Language Telephone-based Speech (MLTS) corpus consists of 11 languages. Muthusamy et al. (1992) have collected the following ten languages: English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese. Later, Hindi language data has been added. In the present paper, we have used speech utterances from the "unrestricted vocabulary speech" category of OGI database. This kind of speech data was obtained by asking the callers to speak on any topic of choice. Each speaker's data belonging to this category consist of two separate utterances with durations of 10 sec and 50 sec. In our work, we have used only the utterances of 50 sec from each speaker for LID study. In each language, we have considered calls from both male and female speakers.

5. Features for developing language identification systems

Current work explores excitation source features for capturing language-specific information. LID systems are also developed using contemporary vocal tract features to make a comparative study between these two features. In the Section 5.1, compendious description of vocal tract features has been given. Section 5.2 provides a detailed description of proposed excitation source features.

5.1. Vocal tract information

During speech production, vocal tract system behaves like a time varying filter, and it characterizes the variations in the vocal tract shape in the form of resonances and antiresonances. Parameterization techniques such as linear prediction cepstral coefficients (LPCCs) and Mel-frequency cepstral coefficients (MFCCs) are available for modeling vocal tract information (Rao et al., 2013). Since Mel-filters are based on human auditory and perception, state-of-the-art LID systems mostly use MFCC features. Since our objective is to explore excitation source information, we use MFCC feature to capture vocal tract information for the comparative study. The MFCC feature vector represents only the shape of the vocal tract at a particular time instant. However, the shape of the vocal tract also varies with respect to time and this time varying nature of the vocal tract can be represented by the delta (velocity) coefficients and delta–delta (acceleration) coefficients. The velocity and acceleration coefficients over a frame span of 2 were appended with the MFCC feature vector to form a resultant 39-dimensional feature vector. All vectors were subjected to cepstral mean subtraction (CMS) followed by cepstral variance normalization.

5.2. Excitation source information

The constriction of the expiration of air acts as excitation source during the production of speech. The quasi-periodic air pulses generated by the vocal folds' vibration act as a source of excitation for voiced speech production. During the production of unvoiced speech, the expiration of air constraints at different places in the vocal tract. This information can be captured by passing the speech signal through the inverse filter (Makhoul, 1975). To capture the excitation source information, linear prediction (LP) residual signal can be analyzed at three different levels: (i) subsegmental level (within a glottal cycle or pitch cycle), (ii) segmental level (within 2–3 successive glottal cycles) and (iii) suprasegmental level (across 50 glottal cycles). In this work, raw LP residual samples, its magnitude, and phase information have been processed separately. Analytic signal representation of LP residual signal is explored to separate magnitude and phase components (Cohen, 1989) of LP residual signal.

5.2.1. Analytic signal representation of linear prediction residual

Speech signal is produced by the convolution of excitation source signal with the vocal tract system response.

$$s(n) = e(n) * h(n) \tag{2}$$

where, $s(n)$ is speech signal, $e(n)$ is the excitation source signal and $h(n)$ is the response of the vocal tract system.

It is necessary to separate the two components for processing them independently. Linear prediction (LP) analysis has been proposed to separate the excitation source and the vocal tract system components of speech (Makhoul, 1975). In the LP analysis, each speech sample is predicted as a linear combination of past p samples, where p is the order of prediction (Makhoul, 1975). Each speech sample $s(n)$ can be predicted as follows:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{3}$$

where $\hat{s}(n)$ is the predicted speech sample and a_k are LP coefficients (LPCs). The error between original and predicted signals is known as prediction error or LP residual (Makhoul, 1975) which is denoted by $r(n)$ and is given by

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{4}$$

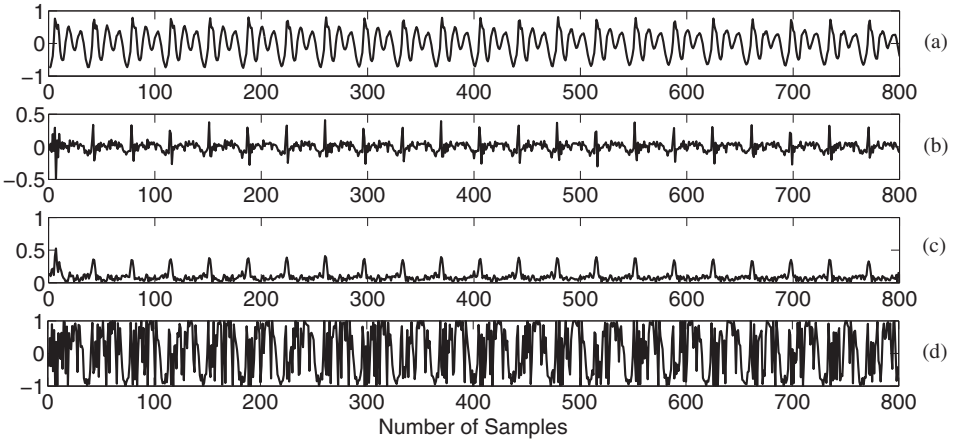

Fig. 3. (a) A segment of voiced speech and its (b) LP residual signal. (c) and (d) are Hilbert envelope and phase of corresponding LP residual signal, respectively.

The LP residual signal $r(n)$ is obtained by passing the speech signal through an inverse filter $A(z)$ (Makhoul, 1975) given by:

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{5}$$

Suppression of vocal tract system response from the original speech signal $s(n)$ gives the LP residual signal $r(n)$. So, LP residual signal mostly represents the excitation source information. We have used LP order of 10 as suggested in Prasanna et al. (2006) followed by inverse filtering the speech signal (sampled at 8 kHz) for estimating the LP residual signal.

The analytic signal of the LP residual signal is denoted as $r_a(n)$, given by:

$$r_a(n) = r(n) + jr_h(n) \tag{6}$$

where, $r(n)$ is the corresponding LP residual signal and $r_h(n)$ is the Hilbert transform of the $r(n)$. The magnitude of the analytic signal is known as Hilbert envelope (HE) of the LP residual signal (Murty and Yegnanarayana, 2006) which is given by:

$$|r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{7}$$

The cosine of the phase is called as residual phase (RP) (Murty and Yegnanarayana, 2006) which is given by:

$$\cos(\theta(n)) = \frac{Re(r_a(n))}{|r_a(n)|} = \frac{r(n)}{|r_a(n)|} \tag{8}$$

In Fig. 3(a), segment of voiced speech, its LP residual signal and corresponding HE and RP have been shown. The HE only reflects the amplitude variation of LP residual signal (see Fig. 3(c)). Whereas, the RP shown in Fig. 3(d) reflects only the phase information of LP residual signal. Proposed implicit processing of LP residual signal, HE and RP are described in Section 5.2.2 and Section 5.2.3.

5.2.2. Proposed implicit processing of linear prediction residual signal

In implicit processing approach, the relations among raw samples of LP residual (LPR) signal are analyzed to model the language-specific phonotactic information. In this work, we focused only on modeling the language-specific excitation source information. Raw LPR samples have been processed at sub-segmental (*sub*), segmental (*seg*) and

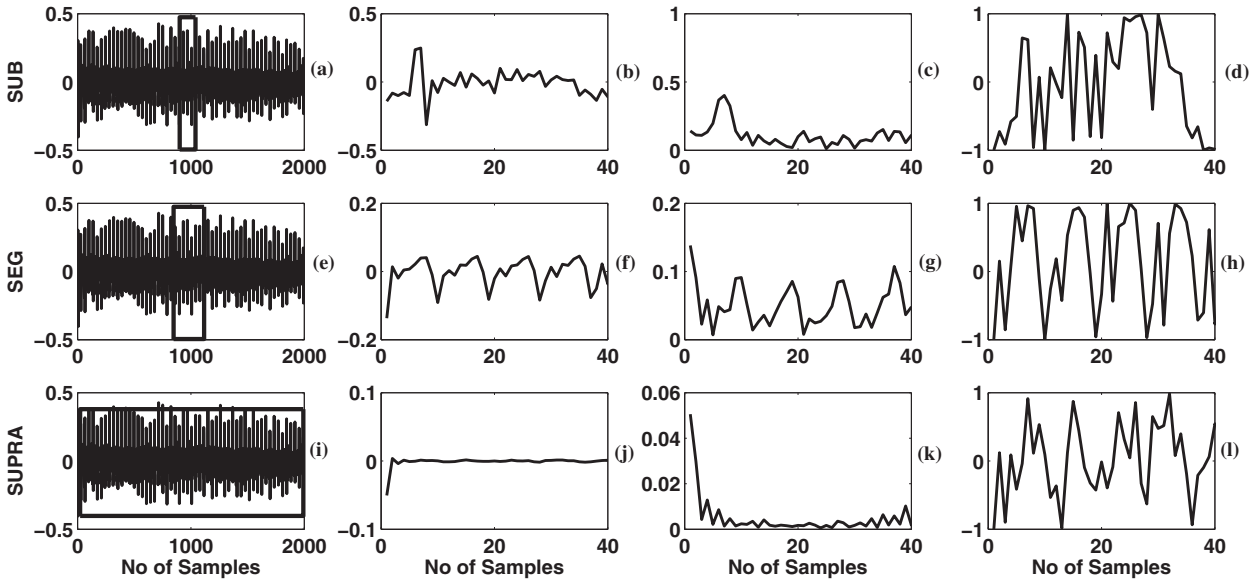

Fig. 4. (a), (e) and (i) are LP residual signals marked by bounding boxes indicating subsegmental, segmental and suprasegmental frames respectively. Sub-segmental level frames of LP residual, Hilbert envelope and residual phase are displayed in figures (b), (c) and (d), respectively. (f), (g) and (h) are LP residual, Hilbert envelope and residual phase at the segmental level, respectively. Suprasegmental level LP residual, Hilbert envelope, and residual phase are shown in (j), (k) and (l), respectively.

suprasegmental (*supra*) levels. At *sub* level, LPR signal is processed at 5 ms block size with a shift of 2.5 ms and corresponding LPR samples are used as *sub* level features to capture the subtle variations present within individual glottal cycles. At *seg* level, first the LPR signal is decimated by a factor 4 to suppress the *sub* level information and then LPR signal is processed with a frame size of 20 ms and frame shift of 2.5 ms. The raw LPR samples obtained by above process are used as *seg* level features. The instantaneous pitch and *epoch* strength of a speech segment are captured by *seg* level processing of LPR signal. At *supra* level processing, LPR signal is first decimated by a factor of 50 to eliminate the *sub* and *seg* level information and then LPR signal is processed in blocks of 250 ms with a shift of 6.25 ms. The raw LPR samples obtained by this process are used as *supra* level features. At *supra* level, the information of slow varying pitch and energy contours is captured. The decimation is performed on LPR signal at *seg* and *supra* levels to suppress the previous level's information and to reduce the dimensionality. In Fig. 4, the LPR signal and its corresponding HE and RP have been shown at *sub*, *seg* and *supra* levels. Fig. 4(a) shows the LPR signal with 5 ms bounding box to indicate *sub* level frame. Fig. 4(b) portrays the 5 ms windowed segment of LPR signal. Fig. 4(e) shows the LPR signal with 20 ms bounding box to indicate *seg* level frame. The 20 ms frame has been decimated by factor 4, which is shown in Fig. 4(f). Fig. 4(i) shows the LPR signal with 250 ms bounding box to indicate *supra* level frame. Fig. 4(j) depicts 250 ms frame after decimated by 50. From Fig. 4(b), (f) and (j), it can be observed that the characteristics of LPR samples are distinct at three different levels. Fig. 4(b) illustrates the minute variation of the LPR samples within one glottal cycle. Periodic nature is observed from Fig. 4(f), which represents the pitch and energy information of 2–3 consecutive glottal cycles. Fig. 4(j) shows the information of pitch and energy contours of several glottal cycles (50 cycles), corresponds to the *supra* level information.

5.2.3. Proposed implicit processing of magnitude and phase components of linear prediction residual

During direct processing of raw LPR samples, magnitude component of LPR signal may predominate over phase component. The amplitude component is not reliable, and it varies with respect to loudness or intensity of speech. Whereas, the phase may be robust to above variations. The phase component of LPR signal contains sequential information of excitation source. Hence, we have separated these two components by the analytic signal representation of LPR signal, which is discussed in Section 5.2.1. The magnitude and phase components of LPR are represented by Hilbert envelope (HE) and residual phase (RP) of LPR signal, respectively. We have proposed the implicit processing approach (see description in Section 5.2.2) for HE and RP independently at three different levels to capture the language-specific excitation source information.

Fig. 4(c), (g) and (k) shows the HE of LPR signal estimated from *sub*, *seg* and *supra* levels, respectively. From this figure, it can be observed that the characteristics of HE of LPR are distinct at three different levels. Similar observation can be made for RP of LPR from Fig. 4(d), (h) and (l). Hence, HE and RP of LPR signal may provide language-discriminative information at *sub*, *seg* and *supra* levels. Therefore, in the present work, we have processed the HE and RP separately to capture different behaviors of excitation source for discriminating the languages.

6. Experimental setup and methodology

Present LID study has been conducted using 27 Indian languages. From each language database one hour of speech data has been taken for building 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 within each language database) from each language database are used to develop language models and other two speakers of each language, who have not participated during training phase, are considered for evaluation. The LID performances reported in this paper are obtained by computing the average of recognition performances across all iterations. 20 test utterances each of 10 sec duration from each language have been considered during evaluation. Detailed procedures to build the language models and to evaluate the developed systems are explained below.

6.1. Training

Proposed implicit excitation source features are modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix (Reynolds and Rose, 1995). The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight for all component densities. The optimum model parameters of a GMM are determined by training a language model using the maximum likelihood (ML) procedure. Maximization is carried out iteratively using an Expectation Maximization (EM) algorithm (Dempster et al., 1977). In each iteration, the posterior probability for the i^{th} mixture is computed and the model parameters are updated as suggested in Reynolds and Rose (1995).

6.2. Testing

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

6.3. Fusion of scores

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

$$C = \frac{1}{k}\sum_{i=1}^{k} w_i c_i \tag{9}$$

where w_i and c_i denote weighting factor and confidence score of i^{th} evidence, and k denotes the number of modalities considered. 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 3 and 2 modalities, respectively. For each set of weighting factors average performance of 27 languages has been calculated. Out of 98 different average performance values for 2 modalities (or 4753 for 3 modalities), best average accuracy and corresponding weighting factors are considered as the optimum one, and has been reported for the combined systems.

7. Development of language identification systems

In this work, LID systems are developed at four phases shown in Fig. 5 and are described below.

7.1. Phase-I

At phase-I, 9 different LID systems are developed:

- Three LID systems are developed by using the LP residual samples directly, extracted from *sub*, *seg* and *supra* levels.
- Three LID systems are developed by using the samples of HE of LP residual signal extracted from *sub*, *seg*, and *supra* levels.
- Three LID systems are developed by using the samples of RP of LP residual signal extracted from *sub*, *seg* and *supra* levels.

The features at *sub*, *seg* and *supra* levels contain partial information about the excitation source. Therefore, to achieve complete excitation source information for discriminating the languages, we have combined the scores from the LID systems developed by partial features. In this study, we have performed 17 different combinations which are shown at phase-II, phase-III and phase-IV.

7.2. Phase-II

In phase-II, we have performed the following five combinations:

- Evidences are combined from the LID systems developed using samples of HE extracted from *sub*, *seg* and *supra* levels.
- Evidences are combined from the LID systems developed using samples of RP extracted from *sub*, *seg* and *supra* levels. Since HE and RP represent the magnitude and phase components of LP residual, the language-specific knowledge from these two components may be different at *sub*, *seg* and *supra* levels. Hence, we have explored the combinations of HE and RP at *sub*, *seg* and *supra* levels separately which is denoted as HE + RP in Fig. 5.
- The evidences of HE and RP features are combined at *sub* level.
- The evidences of HE and RP features are combined at *seg* level.
- The evidences of HE and RP features are combined at *supra* level.

7.3. Phase-III

In phase-III, we have performed the following three combinations:

- Evidences are combined from the LID systems developed using LP residual samples extracted from *sub*, *seg* and *supra* levels.
- The first and fifth combinations shown in phase-II of Fig. 5 refer to LID systems with HE and RP features, respectively. These two features (HE and RP) represent magnitude and phase components of LP residual. Hence, the evidences obtained from these two systems are further combined to achieve the language-specific information completely from the excitation source viewpoint. The LID system-II at phase-III of Fig. 5 is developed by the above-mentioned combination.
- However, the combination of HE and RP at each level (HE + RP) represents partial information about excitation source. Hence, to acquire the complete language-specific excitation source information, we have combined the evidence of HE + RP features from *sub*, *seg* and *supra* levels. The LID system-III at phase-III of Fig. 5 indicates the corresponding combination.
- LID system-IV is developed by using vocal tract information represented by MFCC
- LID system-V is developed by using velocity and acceleration coefficients concatenated with MFCC coefficients (MFCC + Δ + $\Delta\Delta$).

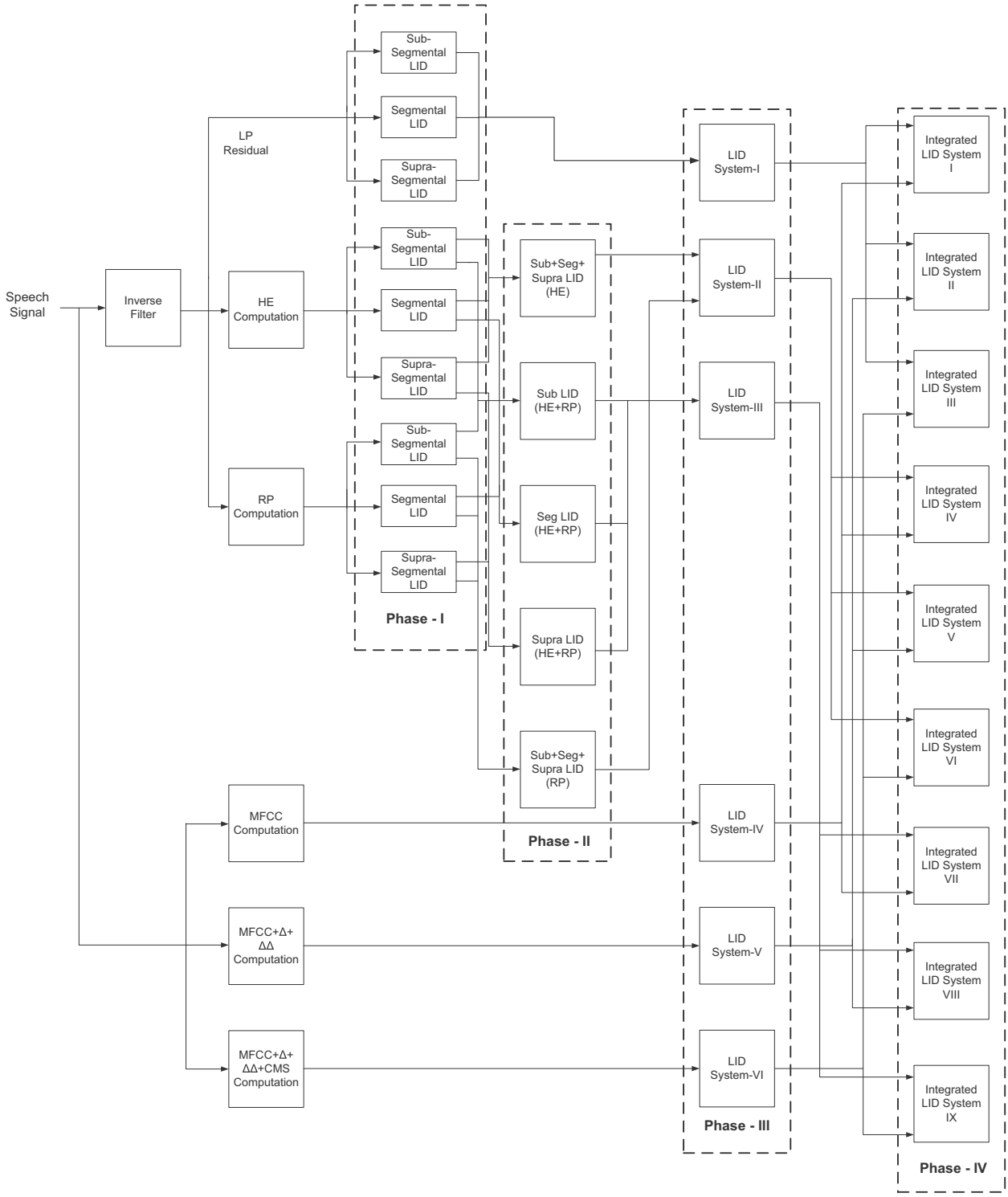

Fig. 5. Development of LID systems using excitation source and vocal tract information.

- LID system-VI is developed by imposing the CMS on MFCC concatenated with dynamic coefficients (MFCC + Δ + ΔΔ + CMS).

To obtain better improvement in LID performance, we have combined the evidence from the LID systems developed by vocal tract information represented by MFCC features and excitation source features, as they represent different aspects of the human speech production system.

7.4. Phase-IV

In phase-IV of Fig. 5, the integrated LID systems developed using vocal tract and excitation source features are illustrated. In phase-IV, we have developed nine different integrated LID systems:

- The LID system-I at phase-III is developed by processing the LP residual samples directly at three levels. The evidence obtained from this system is combined with evidence from the LID system developed by using MFCC feature (LID System-IV) at phase-III to develop the integrated LID system-I at phase-IV.
- The LID system-V at phase-III is developed by concatenating the dynamic coefficients with MFCC. The integrated LID system-II at phase-IV is developed by combining the evidence from LID system-I and LID system-V of phase-III,
- The LID system-VI at phase-III is developed by imposing the cepstral mean subtraction method on MFCC concatenated with dynamic coefficients. The integrated LID system-III at phase-IV is developed by combining the evidence from the LID system-I and LID system-VI of phase-III.

Similarly, the integrated LID system-IV, V, and VI at phase-IV are developed by combining the evidence from LID system-II with LID system-IV, V, VI of phase-III respectively. The integrated LID system-VII, VIII, and IX are developed by combining the evidence from the LID system-III with LID system-IV, V, VI of phase-III respectively.

8. Performance evaluation of LID systems developed using implicit excitation source features

All the language identification systems are evaluated using leave-two-speaker-out approach. The Gaussian mixture models (GMMs) (Reynolds and Rose, 1995) are used to train the language models. Different Gaussian mixtures (32, 64, 128 and 256) have been explored for capturing the language-specific excitation source information.

8.1. Phase-I

At phase-I, nine different LID systems are developed by processing the raw LP residual, HE and RP samples at three different levels. Table 3 portrays the average performances of these LID systems. The first three columns of phase I represent the LID performances obtained by processing the HE samples at *sub*, *seg* and *supra* levels. The average performances of HE feature at *sub*, *seg* and *supra* levels are 22.77%, 32.03% and 25.18%, respectively. Segmental level HE feature gives better accuracy compared to *sub* and *supra* levels. The distinct nature of HE samples at *sub*, *seg* and *supra* levels, which has been portrayed in Fig. 4(c), (g) and (k), is also reflected in the LID performances at three levels. The 4th, 5th and 6th columns of phase I in Table 3 represent the LID performances obtained by processing the RP samples at *sub*, *seg* and *supra* levels, respectively. Average LID performances of RP features at *sub*, *seg* and *supra* levels are 42.03%, 50.00% and 43.14%, respectively. The *seg* level phase information provides better LID accuracy than the phase information present at *sub* and *supra* levels. The processing of RP samples at *sub*, *seg* and *supra* levels also display distinct nature. The phase of LP residual contains more significant language-specific information at each level than the magnitude of LP residual signal, which can be inferred by comparing the LID performances obtained by processing HE and RP features. The 7th, 8th and 9th columns of phase I showed in Table 3 represent the LID performances obtained by processing the raw LP residual samples at three levels. The average LID performances at *sub*, *seg* and *supra* levels are 24.62%, 44.62% and 31.66%, respectively. In this work, we have analyzed the confusion matrices of LID systems developed using different excitation source features. The confusion matrix obtained from seg level LPR feature has been illustrated in Table 4 to analyze the detailed classification among 27 Indian languages. The diagonal elements represent the percentage of correctly identified test utterances (i.e., language wise

Table 3
LID performances using *implicit* excitation source, vocal tract and their combined features evaluated on IITKGP-MLILSC database.

Phase I								
HE			RP			LPR		
sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
22.77	32.03	25.18	42.03	50	43.14	24.62	44.62	31.66

Phase II		
HE (*sub* + *seg* + *supra*)	RP (*sub* + *seg* + *supra*)	HE + RP
		sub / *seg* / *supra*
50.18	63.51	47.77 / 50.92 / 43.88

Phase III					
src1	src2	src3	MFCC	MFCC + Δ + ΔΔ	MFCC + Δ + ΔΔ + CMS
63.51	63.70	58.88	62.40	66.48	70.18

Phase IV								
MFCC + *src1*	MFCC + *src2*	MFCC + *src3*	MFCC + Δ + ΔΔ + *src1*	MFCC + Δ + ΔΔ + *src2*	MFCC + Δ + ΔΔ + *src3*	MFCC + Δ + ΔΔ + CMS + *src1*	MFCC + Δ + ΔΔ + CMS + *src2*	MFCC + Δ + ΔΔ + CMS + *src3*
71.85	72.03	68.89	75.18	75.74	72.59	75.92	76.29	73.70

performance in %). Other elements represent the percentage of misclassification (i.e., percentage of wrongly identified test utterances). The confusion matrix obtained from *seg* level LPR feature is also shown in heat map format (see Fig. 6) for better visualization. The light colors represent high performances, and the dark colors represent poor performances. In this figure, the color of the diagonal blocks is comparatively brighter than other blocks. This indicates the language discrimination capability of *seg* level LPR feature.

8.2. Phase-II

The magnitude component of LPR signal represented by HE and phase component of LPR represented by RP contain partial information about excitation source. Therefore, to acquire the complete excitation source information, we have combined the confidence scores obtained from the LID systems developed by HE and RP features. The score combination can be performed in different ways, which has been shown at phase-II and phase-III in Fig. 5. LID accuracy of 50.18% is obtained by combining the evidences from LID systems developed using HE feature at *sub*, *seg* and *supra* levels (see 1st column of phase II in Table 3). Similarly, the combined evidences obtained from LID systems developed using RP feature at *sub*, *seg* and *supra* levels provides 63.51% average LID performance (see 2nd column of phase II in Table 3). However, these two features contain different information. The evidence obtained from HE and RP features are also combined at each level, which is shown at 3rd, 4th and 5th columns of phase II in Table 3. The combined HE + RP information provides 47.77%, 50.92% and 43.88% identification accuracies at *sub*, *seg* and *supra* levels, respectively. The *seg* level information provides better identification accuracy compared to other levels. The combined HE + RP feature at each level contains distinct information which can be inferred from the average LID accuracies.

8.3. Phase-III

The 3rd column of phase III in Table 3 denoted as *src3* providing 58.88% identification accuracy refers the performance obtained from the LID system-I at phase-III of Fig. 5. The combined HE features at *sub*, *seg* and *supra* levels contain only the magnitude information of LPR signal, which is shown in the first column of phase II of Table 3. The combined RP features at *sub*, *seg* and *supra* levels contain only the phase information of LPR signal, which is shown in the second column of phase II shown in Table 3. Therefore, to obtain the complete information about the excitation source we have again combined the evidences obtained from these two features to develop the LID system-II at phase-III. The first column of phase III in Table 3 represents the average LID performance obtained from LID

Table 4
LID performance (in confusion matrix form) obtained from segmental level LP residual features.

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	35	0	5	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	5	0	0	45	0
Ass	0	35	5	0	0	0	0	5	5	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	45	0	0
Ben	0	5	5	0	0	0	0	0	0	10	0	20	0	0	0	0	0	0	0	25	0	0	0	0	20	15	0
Bho	0	0	10	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	40	0	0
Cha	0	0	0	0	100	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	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	10	0
Goj	0	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
Guj	0	0	5	0	0	0	0	15	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	50	0
Hin	0	40	15	5	0	0	0	0	15	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	20	0	0
Ind	0	0	0	0	5	0	0	0	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kan	0	50	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
Kas	0	0	0	0	0	0	0	0	0	0	5	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kon	0	10	0	0	0	0	0	0	0	0	0	0	85	0	0	0	0	0	0	0	0	0	0	0	5	0	0
Mal	0	0	0	15	0	0	0	0	10	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	0	0	0
Man	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	15	0	0	0
Mar	0	0	20	20	0	0	0	0	0	0	50	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Miz	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	50	0	0	0	0	0	0
Nag	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	70	0	0	0	0	0	0	0	0	0
Nep	0	0	15	0	0	0	0	0	10	0	0	0	0	0	0	5	0	0	0	25	0	0	0	0	0	45	0
Ori	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	10	0	20	0	0	0	0
Pun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Raj	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	35	0	0	0	50	0
San	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	85	0
Sin	0	10	20	0	0	0	0	0	0	0	0	0	0	0	0	50	0	5	0	5	0	0	0	45	0	5	0
Tam	0	15	15	0	0	0	0	10	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	35	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	100	0
Urd	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	0	5	0

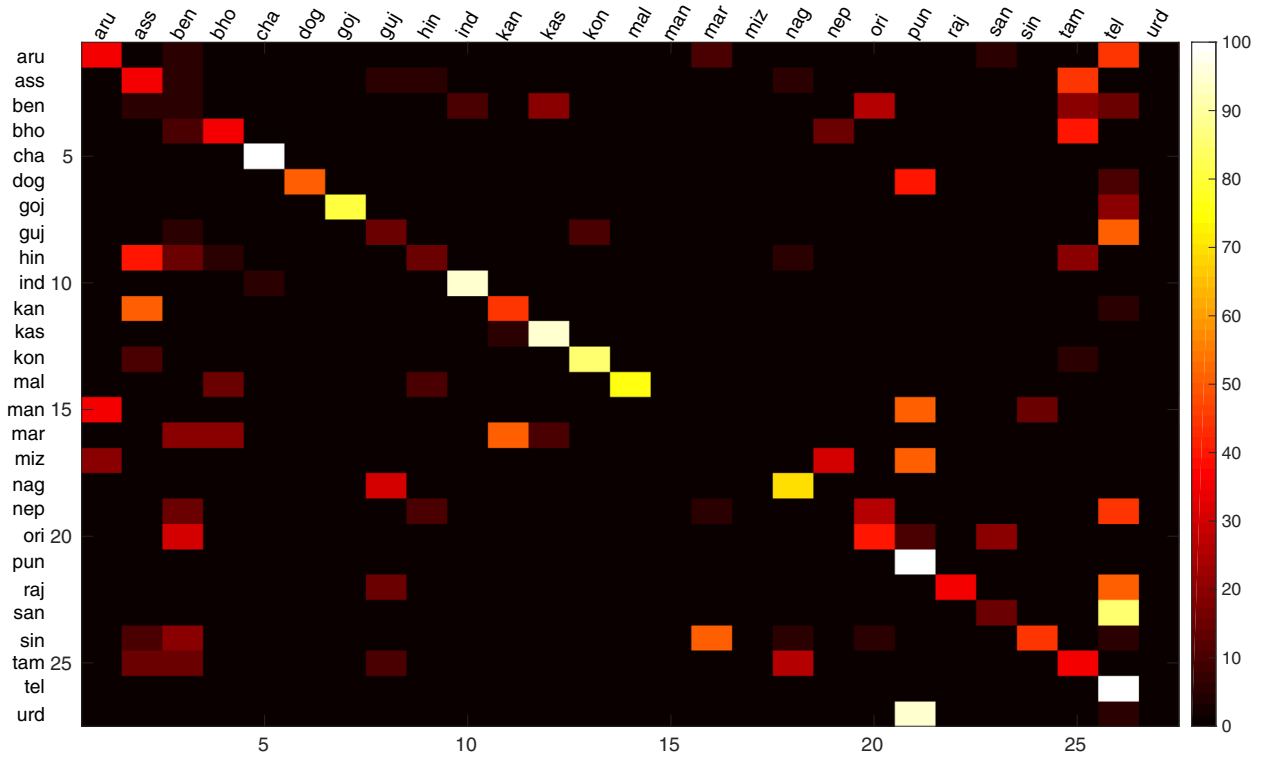

Fig. 6. Heat map representation of LID performances (in confusion matrix form) obtained from segmental level LP residual features.

system-II at phase-III, which is denoted as *src1*. *src1* provides LID performance of 63.51% which is equivalent to the LID performance contributed by only RP feature. This infers that the phase component of LPR signal is more significant for language identification than magnitude component of LPR. The HE and RP features have also been combined at each level which is shown at 3rd, 4th and 5th columns of phase II of Table 3. The combined HE + RP features at *sub*, *seg* and *supra* levels reflect partial information from excitation source point of view. Therefore, we have further combined the evidence obtained from 2nd, 3rd and 4th LID systems of phase-II, to build LID system-III at phase-III in Fig. 5. The average LID performance of 63.70% shown in 2nd column of phase III of Table 3 denoted as *src2* is obtained from the LID system-III at phase-III. The average identification accuracy obtained from *src2* feature is better than the LID performances obtained from *src1* and *src3* features. The average LID performances obtained using MFCC, MFCC + ΔΔ and MFCC + ΔΔ + CMS features are 62.40%, 66.48% and 70.18%, respectively (see 4th, 5th and 6th columns of phase III in Table 3).

8.4. Phase-IV

In speech signal both the vocal tract and excitation source information are present. In our work, we have focused only on capturing the language-specific excitation source information. The vocal tract information represented by MFCC feature contains different information with respect to the excitation source information represented by *src1*, *src2* and *src3*. Therefore, integrated LID systems are developed by combining the scores obtained from the LID system developed by MFCC, MFCC + Δ + ΔΔ and MFCC + Δ + ΔΔ + CMS features with the confidence scores obtained from *src1*, *src2* and *src3* features separately at phase-IV. The combination of evidence from vocal tract and excitation source features provides significant improvement of LID performances compared to identification accuracies obtained from individual features, which indicates their distinct nature. The LID performances of all integrated LID systems are shown at phase IV in Table 3.

In Table 5, the LID accuracies reported in earlier works are compared with our proposed features. Sangwan et al. (2010) achieved 65% accuracy by exploring speech production knowledge on 5 Indian languages (Kannada, Tamil,

Table 5
Comparative study of proposed excitation source features with the contemporary works.

Authors and years	LID accuracy	No of languages	System description
Sangwan et al. (2010)	65%	5 Indian languages	Speech production knowledge
Rao et al. (2013)	58.14%	27 Indian Languages	MFCC extracted from glottal closure regions
Reddy et al. (2013)	35.22%	27 Indian Languages	Word level prosody features
Proposed (2015)	63.70	27 Indian languages	Implicit excitation source features

Table 6
LID performances using *implicit* excitation source, vocal tract and their combined features evaluated on OGI-MLTS database.

Phase – I								
HE			RP			LPR		
sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
20	47	33	40	50	37	31	44	30

Phase – II					
HE (*sub* + *seg* + *supra*)	RP (*sub* + *seg* + *supra*)	HE + RP			
		sub	*seg*	*supra*	
52	63	46	56	42	

Phase – III				
src1	*src2*	*src3*	MFCC	
65	68	51	76	

Phase – IV		
MFCC + *src1*	MFCC + *src2*	MFCC + *src3*
83	84	80

Telugu, Malayalam, and Marathi). Rao et al. (2013) have explored spectral features from glottal closure region for LID task. LID accuracy of 58.14% has been obtained on 27 Indian languages from IITKGP-MLILSC database. Reddy et al. (2013) have exploited word-level prosody features to evaluate 27 Indian languages and achieved 35.22% LID accuracy. In the reported works, LID accuracy decreases with increasing the number of languages. The contemporary features do not carry the desired amount of non-overlapping language-specific information to discriminate a large number of Indian regional languages. Hence, LID systems based on Indian regional languages demand a language-specific feature which has potential capability to capture the non-overlapping language discriminative information. In the present study, we have introduced excitation source features in the context of language identification and achieved 63.70% LID accuracy by analyzing 27 Indian regional languages.

9. Evaluation of implicit excitation source features on OGI-MLTS database

The proposed implicit features of excitation source are also evaluated on OGI-MLTS database. The LID performances obtained from OGI-MLTS database are shown in Table 6. The *seg* level RP feature provides better accuracy, compared to others. It is observed from the results shown in columns 1 to 6 of phase I that magnitude and phase components of LPR signal provide distinct language-specific information.

It can be stated from the LID performances presented in Table 6 that the language-specific knowledge present at *sub*, *seg* and *supra* levels are fundamentally distinct. Hence, the evidence from above three different levels is combined to capture complete excitation source features. The best average LID accuracy of 68% is obtained by processing complete excitation source information (see column 2 of phase III in Table 6). The LID accuracies of excitation source features are also compared with the accuracies of vocal tract features represented by MFCCs. To investigate the complementary nature of these two features, we have combined the scores obtained from these two features. Phase IV of Table 6 represents the accuracies obtained from integrated LID systems. The LID accuracies of integrated LID

systems are enhanced, compared to individual features. This shows the significance of excitation source information for language identification task.

10. Summary and conclusions

In this work, a unified framework has been proposed for extracting language-specific excitation source information from the speech signal. Raw samples of LP residual signal, its magnitude, and phase components are analyzed at three different levels to derive distinct aspects of excitation source information for language discrimination task. Empirical analysis of excitation source features on 27 Indian and 11 international languages shows their importance in language identification task. We report that segmental level feature contains more language discriminative information (50.92%), compared to sub-segmental (47.77%) and suprasegmental (43.88%) levels. Complementary nature of vocal tract and excitation source features is also investigated in this study. Combined excitation source and vocal tract features provide better accuracy (76.29%), compared to the individual features. This indicates that excitation source contributes non-overlapping language-specific information with respect to the vocal tract features. The experimental results obtained from OGI-MLTS database also portray similar trends in LID accuracies comparable to Indian languages. In the future, high-level phonotactic features can be explored in addition to proposed excitation source features to enrich language identification accuracy.

Acknowledgment

This research is funded by the Department of Information Technology (DIT) (11(6)/2011-HCC(TDIL)), the government of India, through the project “Prosodically guided phonetic engine for searching speech databases in Indian languages (PSI)”.

References

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).

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

Cohen, L., 1989. Time frequency distribution: a review. IEEE Proc. 77, 941–979.

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.

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. Digit. Signal Process. 22 (3), 544–553. Elsevier.

Ma, B., Li, H., Tong, R., 2007. Spoken language recognition using ensemble classifiers. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2053–2062.

Maity, S., Vuppala, A.K., Rao, K.S., Nandi, D., IITKGP-MLILSC speech database for language identification. In: National Conference on Communication, 2012.

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

Mary, L., Multilevel Implicit Features for Language and Speaker Recognition (Ph.D. dissertation), Indian Institute of Technology Madras, India, 2006.

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

Murty, K.S.R., Yegnanarayana, B., 2006. Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett. 13 (1), 52–55.

Muthusamy, Y.K., Cole, R.A., Oshika, B.T., The OGI multi-language telephone speech corpus. In: Spoken Language Processing, pp. 895–898, 1992.

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

Prasanna, S.R.M., Gupta, C.S., Yegnanarayana, B., 2006. Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. 48, 1243–1261.

Rao, K.S., Koolagudi, S.G., 2013. Characterization and recognition of emotions from speech using excitation source information. Int. J Speech Tech. 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 Tech. 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 Tech. 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 Process. 3 (1), 72–83.

Roy, P., Das, P., 2013. A hybrid VQ-GMM approach for identifying Indian languages. Int. J Speech Tech.16 (1), 33–39.

Sangwan, A., Mehrabani, M., Hansen, J., Automatic language analysis and identification based on speech production knowledge. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5006–5009, 2010.

Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H., 2013. Universal attribute characterization of spoken languages for automatic spoken language recognition. Comp. Speech Lang. 27 (1), 209–227. Elsevier.

Sugiyama, M., Automatic language recognition using acoustic features. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 813–816, 1991.

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

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

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, 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 Process. 13 (4), 575–582.

Zissman, M.A., Singer, E., Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I/305–I/308, 1994.