



Modelling multi-level prosody and spectral features using deep neural network for an automatic tonal and non-tonal pre-classification-based Indian language identification system

Chuya China Bhanja¹ · Mohammad Azharuddin Laskar¹ · Rabul Hussain Laskar¹

Accepted: 30 December 2020 / Published online: 20 January 2021

© The Author(s), under exclusive licence to Springer Nature B.V. part of Springer Nature 2021

Abstract In this paper an attempt has been made to prepare an automatic tonal and non-tonal pre-classification-based Indian language identification (LID) system using multi-level prosody and spectral features. Languages are first categorized into tonal and non-tonal groups, and then, from among the languages of the respective groups, individual languages are identified. The system uses syllable, word (tri-syllable) and phrase level (multi-word) prosody (collectively called multi-level prosody) along with spectral features, namely Mel-frequency cepstral coefficients (MFCCs), Mean Hilbert envelope coefficients (MHEC), and shifted delta cepstral coefficients of MFCCs and MHECs for the pre-classification task. Multi-level analysis of spectral features has also been proposed and the complementarity of the syllable, word and phrase level (spectral + prosody) has been examined for pre-classification-based LID task. Four different models, particularly, Gaussian Mixture Model (GMM)-Universal Background Model (UBM), Artificial Neural Network (ANN), i-vector based support vector machine (SVM) and Deep Neural Network (DNN) have been developed to identify the languages. Experiments have been carried out on National Institute of Technology Silchar language database (NITS-LD) and OGI Multi-language Telephone Speech corpus (OGI-MLTS). The experiments confirm that both prosody and (spectral + prosody) obtained from syllable-, word- and phrase-level carry complementary information for pre-classification-based LID task. At the pre-classification stage, DNN models based on multi-level (prosody + MFCC) features, coupled with score combination technique results in the lowest EER value of 9.6% for NITS-LD. For OGI-MLTS database, the lowest EER value of 10.2% is observed for multi-level (prosody + MHEC). The pre-classification module helps to improve

✉ Chuya China Bhanja
chuya@rs.ece.student.nits.ac.in

¹ Department of Electronics and Communication Engineering, National Institute of Technology Silchar, Assam 788 010, India

the performance of baseline single-stage LID system by 3.2% and 4.2% for NITS-LD and OGI-MLTS database respectively.

Keywords Tonal and non-tonal languages · Multi-level analysis · Prosody and spectral features · Databases · Classifiers

1 Introduction

The main purpose of LID system is to automatically recognize the spoken language from a given portion of speech. One application of LID systems is to prepare a system for routing an incoming phone call to an appropriate human switchboard operator who is well versed in a particular language. In multilingual countries like, India, a multilingual spoken-dialog system that can serve in multiple languages, finds application in various fields (Mary 2006). In this type of multilingual operation, the machine should be capable of distinguishing among different languages. Several approaches and computationally advanced methods have been proposed in the literature for language distinguishing task with state-of-the-art performance.

The number of target languages has direct bearing on the performance of an LID system. Also, in countries, like India where the languages share common phoneme sets, distinguishing among languages become more challenging. Several researchers attempted to identify the closely related Indian languages. In one such case, Jothilakshmi et al. (2012) presented a hierarchy-based LID system for 9 Indian languages using spectral features, namely MFCC, delta/ double delta and SDCs of MFCC. Here, in the first level, the languages were divided into two language families, namely Indo-Aryan and Dravidian and then individual languages were identified from languages of the corresponding language family. They studied the efficacy of the two-level LID system in discriminating languages having the same origin. The authors also studied the effectiveness of MFCC features and they reported an accuracy of 80.56% (9 target languages) for the GMM-UBM model and (MFCC + delta-double delta features). However, they didn't study the complementarity of prosodic features with MFCC for hierarchical Indian language identification system.

Reddy et al. (2013) proposed another LID system for 27 languages of Indian origin using spectral (MFCC) and prosodic features. Here, prosodic features extracted for different levels, namely syllable, word, and phrase were used and then the final score was obtained by combining the scores obtained for different levels. Complementary nature of prosodic and spectral features at the utterance level was exploited and the evidences from spectral and prosodic features were fused to obtain better language recognition accuracy. In this case, the average accuracy for the identification of 27 Indian languages was 62.13%. To build a more accurate system with larger number of target languages, a module may be added to initially pre-classify the languages into different categories or sub-language families. Also, to accurately identify languages, which are closely related or are having the same origin, properly, a reliable pre-classification module is required.

In one such case, Wang et al. (2007) outlined a tonal/non-tonal pre-classification-based LID system for 16 world's distinct languages using prosodic features only and reported accuracies of 77.9% and 49.2% for 30 s and 10 s test data respectively. However, the system is dependent on phonetically labelled data, which is not always available and requires expertise in linguistics. Countries, like India where language diversity is very high, it is even more difficult to obtain the phonetically labelled data for all the languages. Also, in (Wang et al. 2007) the researchers examined the efficacy of pre-classification module with only the world's distinct languages. No study, however, has been reported on pre-classification of closely related Indian languages.

Additionally, in Wang et al. (2007) only a few parameters of prosodic features, like pitch and duration have been used for both pre-classification and pre-classification-based LID task and the features are extracted considering the whole utterance as a unit. However, literature study confirms the alignment of tonal events with syllables for tonal languages (Atterer and Ladd 2004; Zhang 2014). Also, most of the Indian languages are syllable-centric (Singh 2006), and so language-specific information are manifested at syllable level itself. For tonal languages the pitch changes within a syllable is of a regular pattern (Maddieson et al. 2013). In one of the recent works (China Bhanja et al. 2018), we have observed the usefulness of syllable-level features for pre-classification-based LID system. Some new parameters of prosody have also been proposed to boost the performance of tonal/non-tonal language classification task. However, the system performance was analysed only for seven Northeast (NE) Indian languages. Moreover, the tonal languages included in this database are mostly having monosyllabic words. However, for di-syllabic or poly-syllabic tonal languages, all the syllables may not carry tone information. In case of those languages, features obtained from other levels, like word (three consecutive syllables) or phrases (multi-word) may provide better tone information. In paper (Reddy et al. 2013) researchers observed the complementarity among different levels of prosodic features when identifying individual languages. However, no such study analysed either the complementarity of syllable, word and phrase level prosody or the combining effects of the different levels of prosody for tonal/non-tonal classification or tonal/non-tonal pre-classification-based LID system. Literatures reveals that MFCC carries tone information (Le et al. 2009; Ryant et al. 2014). However, for tonal/non-tonal pre-classification-based LID system, frame level analysis of MFCC has not been explored so far. Since tones in tonal languages lie within a syllable (China Bhanja et al. 2018), MFCC frames corresponding to a syllable are further modelled using Legendre coefficients to obtain syllabic level characterization. However, for di-syllabic or poly-syllabic tonal languages any of the utterance or syllable level MFCC modelling may not be the most suitable in terms of capturing the tone information. To study this aspect, MFCC feature modelling can be analysed at multiple levels so as to explore both the local and global characteristics of the speech signal. Also, its combination with multi-level prosodic features is studied.

MFCCs, though have been the most extensively used features for language identification (Burgos 2012), are sensitive to background noise, acoustic mismatched training and testing environments, room reverberation etc. In another

study, researchers showed that the performance of MFCC features reduces significantly with the increase of noise power (Li and Huang 2011). Several researchers (Li and Narayanan 2014; Sadjadi and Hansen 2015) have worked, in recent times, towards developing a front-end system, robust to noise and mismatched acoustic training and testing environments. In Sadjadi and Hansen (2015), the authors proposed another noise robust LID system which works well on noisy data of DARPA-RATS database, utilizing MHEC features extracted from the frames of an utterance. However, no study discusses if MHEC carries tone information that could be useful for discrimination of tonal languages from non-tonal. Since MHEC carries finer details of the information of human auditory perception which may be useful for identifying different tones, therefore it may be more effective for discriminating tonal/non-tonal languages with higher accuracy. Syllable-level representation of MHEC feature may provide better tonal/non-tonal language discriminating information. Multi-level MHEC may provide complementary information and they may be used as complementary features with prosody to improve the system performance at the pre-classification stage which would further improve the overall performance. Further, this paper studies the system performance for two different datasets that have been collected using two different channels. These experiments may thus help study the effectiveness of acoustic features for two different channel conditions.

In the back-end, a significant advancement can be observed in the context of LID task. Several research efforts have been made in the form of Joint Factor Analysis (JFA), i-vector based approach (Dehak et al. 2011) etc. i-vector based approach uses GMM-UBM to model the acoustic features and various scoring methods, namely probabilistic linear discriminant analysis (PLDA) (Prince and Edler 2007), SVM (Dehak et al. 2011) and cosine distance (CD). In recent study, the effectiveness of DNN (Richardson et al. 2015a, b; Mounika et al. 2016) has also been studied in LID task. In existing literature (Dehak et al. 2011; Prince and Edler 2007; Richardson et al. 2015a, b; Mounika et al. 2016) i-vector based SVM and DNN are used to model the frame-level features. Also, in Martinez et al. (2013), researchers presented a method whereby an utterance is first divided into fixed length segments. Then the segment-level features are used to compute the i-vector of that utterance. However, in tonal languages, features should preferably be extracted syllable by syllable. Nevertheless, the use of multi-level features and subsequent score combination can also be helpful. Syllable-level or multi-level features have not been used with DNN or i-vector SVM frameworks so far.

This paper particularly focusses on the identification of closely related Indian languages. The influence of one language on other is very high in India. Also, there are several under-resourced and/or well-resourced languages in India. Very few databases (ciil-spokencorpus 2009; Maity et al. 2012) have been prepared for Indian languages. Moreover, in India, the existing databases either include less number of target languages or are not commercially available. This makes it important to prepare a database which may cover many more Indian languages.

2 Motivation

It can be observed from the literature that the existing tonal/non-tonal pre-classification-based LID system (Wang et al. 2007) for the world's distinct languages depend on phonetically labelled data. Also, only syllable-level MFCCs and prosody have been explored in a similar system for NE Indian languages (China Bhanja et al. 2018). No study of such a system is available using multi-level MFCCs which carry useful information for di-syllabic or polysyllabic tonal languages and no work studies the effectiveness of MHEC features or their complementarity with prosody in any pre-classification-based LID task. No pre-classification-based LID system has so far explored multi-level analysis of MHECs and prosody. Additionally, most of the LID systems (Sadjadi and Hansen 2015; Dehak et al. 2011; Prince and Edler 2007; Richardson et al. 2015a; Richardson et al. 2015b; Mounika et al. 2016) reported so far, has given emphasis on the modelling of utterance-level features extracted from the frames of an utterance. Modelling of multi-level features using i-vector based SVM or DNN has not been explored. Commercially available databases for Indian languages are very less in number. This paper tries to address the above-mentioned issues. The main contributions of this paper may be summarized as follows:

- A tonal/non-tonal pre-classification-based LID system has been developed for languages of Indian origin using multi-level prosody spectral features. This system does not use phonetically labelled data.
- Comparative study of the frame-level and syllable-level spectral features has been done. Also, performance analysis of MHEC and MFCC features has been carried out. Complementarity of MHEC and MHEC + SDC with prosody has also been studied.
- Comparative performance analysis of systems based on multi-level prosody and multi-level (prosody + spectral) feature with respect to that based on syllable-level feature has been carried out in this work.
- A comparative analysis of the various modelling techniques, namely GMM-UBM, i-vector based SVM, ANN and DNN has been done for a pre-classification-based LID system using multi-level features.
- NITS-LD (Studio-quality) has been prepared in-house and it covers twelve closely related Indian languages, namely Bengali, Assamese, Indian English, Hindi, Nagamese, Odia, Tamil, Mizo, Punjabi, Manipuri, Bodo, and Gojri of different language families. The data has been acquired from news archives of AIR (All India Radio). Moreover, the systems are also evaluated on OGI-MLTS (telephonic) database that consists of world's distinct languages of different families.

The rest of the paper is organized as follows: Sect. 3 provides the description of the databases; the description of the proposed language identification system is given in Sect. 4—system architecture, features and the language modelling

Table 1 Comparison between OGI-MLTS and NITS-LD databases

Characteristics	OGI-MLTS database	NITS-LD
Number of languages	11	12
Channel characteristics	Different	Similar
Channel conditions	Noisy	Non noisy
Types of speech	Spontaneous	Scripted
Recording environment of data	Realistic (telephonic)	Studio
Sampling rate	8KHz	8KHz
Languages and number of speakers present per language	German, Farsi, French, Japanese, English, Korean, Spanish, Mandarin Chinese, Vietnamese, Tamil and Hindi 90 speakers for each language present in the database	As-35, Be-24, En-22, Hi-26, Na-9, Od-16, Ta-17, Ma-13, Mi-13, Bo-11, Go-9 and Pu-17
# h per language	Around 2 h data for each of the language	As-12, Be-14, En-13, Hi-11, Na-11, Od-11, Ta-15, Ma-10, Mi-9, Bo-10, Go-9 and Pu-11

techniques. Experimental results and analysis are given in Sects. 5 and 6 concludes the work by mentioning the future works.

3 Dataset details

In this work, two databases namely, OGI-MLTS database and NITS-LD have been used for validation of the systems.

3.1 OGI-MLTS database

OGI-MLTS speech (Muthusamy et al. 1992) corpus is made up of spontaneously spoken fixed-vocabulary utterances of 11 different languages: Spanish, Farsi, Mandarin Chinese, French, English, German, Vietnamese, Korean, Japanese, Tamil and Hindi. The Japanese language has been not been considered in the experimentation because of the uncertainty of its tonal/ non-tonal nature (Beckman and Pierrehumbert 1986). 90 speakers of each of the languages have been used to prepare the database. It is collected over a telephone line at a sampling frequency of 8 kHz. It covers two tonal languages (Vietnamese and Mandarin) and nine non-tonal languages. The systems have been evaluated for 10 languages (after omitting the Japanese language). Only two Indian languages have been covered in OGI-MLTS database, which is why the NITS-LD database which covers 12 Indian languages has been prepared.

3.2 NITS-LD

NITS-LD includes 12 Indian languages, namely, Bengali (Be), Assamese (As), Hindi (Hi), Indian English (En), Nagamese (Na), Odia (Od), Tamil (Ta), Manipuri (Ma), Mizo (Mi), Bodo (Bo), Gojri (Go) and Panjabi (Pu).

In this database five languages (Manipuri, Mizo, Bodo, Gojri and Panjabi) are tonal and the rest seven are non-tonal. AIR news archives have been used for data preparation. It involves well matured and highly professional speakers. Thus, the speech extracts are all well-articulated and spoken with standard speaking rate and pronunciation. Table 1 compares the OGI-MLTS and the NITS-LD databases. The database prepared using speech samples of AIR news archives have some issues like lesser number of speakers for some of the languages and lesser variability across different sessions in terms of vocabulary. In short, data variability across speakers and words are limited. In order to have sufficiently large training set, a subset of Indic database is also used in addition to NITS-LD database. From the Indic database, around 5 h data of each of the 5 languages namely Hindi, Bodo, Odia, Tamil and Manipuri have been used for this experiment. The details of Indic database are given in (Baby et al. 2016).

4 Language identification system

A pre-classification-based LID system with a tonal/non-tonal pre-classification module has been developed in this work to achieve better performance over the baseline single-stage LID system. This section describes the proposed pre-classification and the pre-classification-based LID systems prepared using multi-level prosody and the spectral features. It also describes the extraction and parameterization of different features and the language modelling techniques used in this experiment.

4.1 Language pre-classification

The proposed tonal/non-tonal language pre-classification system is described in this section. In one of the recent works (China Bhanja et al. 2018), one such system has been prepared using syllable-level prosody and MFCC features. In this paper, a systematic analysis of system performance has been done for all the syllable, word and phrase levels features. It also explores score combination using multi-level features-based models. Two systems have been developed which both use multi-level prosody but use spectral features of different levels. System-I is developed using prosody of syllable, word and phrase levels, and spectral features, particularly MFCC, MFCC + SDC, MHEC, MHEC + SDC extracted from all the speech frames constituting an utterance (utterance-level). The second system, system-II, uses multi-level spectral features along with multi-level prosody.

4.1.1 Description of system-I

In system-I, we propose multi-level analysis of prosodic features and utterance-level analysis of spectral features for tonal/non-tonal pre-classification of different languages. Working of system-I is shown in Fig. 1. The first step of this system is to detect the vowel onset points (VOPs) (Prasanna et al. 2009) locations that correspond to the time instants of starting of vowel regions in a speech signal. The speech samples between two consecutive VOPs are said to constitute a syllable. Here, Pitch is calculated (Talkin 1995) from the spontaneous speech signal to obtain the pitch contour of the whole utterance. The energy contour is formed by the energy values calculated from 10 ms frames of the utterance. A 5th order median filter is used to smoothen the pitch and energy contours. Then the identified VOPs are mapped with the smoothened contours. The contour portions between every consecutive VOPs are considered as syllables. A duration feature is calculated for every syllable, based on the number of frames between its two VOPs. The pitch and energy contours and the duration features pertaining to every syllable are all parameterized. The parameters are row-wise concatenated to form the final feature vector to represent prosody for that syllable. As discussed in (Mary 2006; Reddy et al. 2013), word-level features, obtained from preceding and succeeding syllables along with the present syllable, effectively represent temporal dynamics. Thus, a word-level feature vector is extracted for every “M” syllable as shown in Fig. 1.

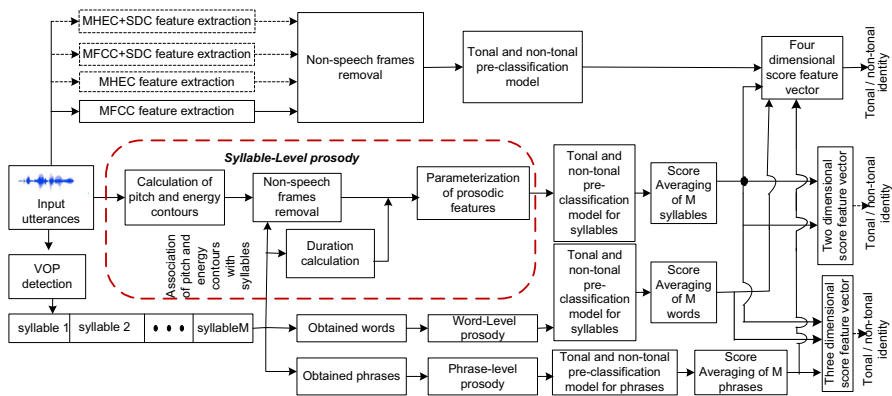


Fig. 1 Working of proposed system-I

System performance has also been analysed for phrase-level prosody which are obtained considering a sequence of 12 syllables of the speech signal. The reason for choosing 12 syllables for representing phrase is that the gross observation of the languages of NITS-LD suggest that a 3 s duration utterance may contain 9–15 syllables. The mean of the phrase length distribution is around 12 syllables for most of the languages.

For 10 s and 30 s duration data number of syllables will always be higher than 12 syllables. Thus, a sequence of 12 syllables may be considered as phrase length for three different duration of data. For phrase-level analysis of the system, the combination of three parameters, namely Δ pitch contour, Δ energy contour and duration contour has been used as in the case of tonal languages, change in pitch within a syllable is an important phonological cue and there is a correlation with change in energy and duration with the pitch change. A shift of one syllable at a time is considered such that the number of phrase-level feature vectors is equal to the number of syllables in the utterance. In case the number of syllables is fewer than 12, as an approximation, the last syllable feature vector is replicated to meet the required count of syllables. In this manner, a phrase-level feature vector is obtained for every “M” syllable. The spectral features (MFCC, MFCC + SDC, MHEC and MHEC + SDC) are extracted from the utterance frames. Only the voiced frames of the utterance are considered for further processing. After normalizing different features, seven separate GMM-UBM models are trained for syllable-level prosody, word-level prosody, phrase-level prosody, utterance-level MFCC, MFCC + SDC, MHEC and MHEC + SDC. At the testing phase, the average prosody scores corresponding to the syllables, words and phrases of an utterance are obtained separately. Thus, three different scores are obtained for an utterance for syllable-, word- and phrase-level prosody and are used in different combinations to form feature vectors. Similarly, for the utterance-level spectral features, the score obtained from all the frames of an utterance are averaged to obtain the final score corresponding to that utterance. At a time, only one of the spectral features, i.e. MFCC, MFCC + SDC, MHEC, or MHEC + SDC is used along with multi-level

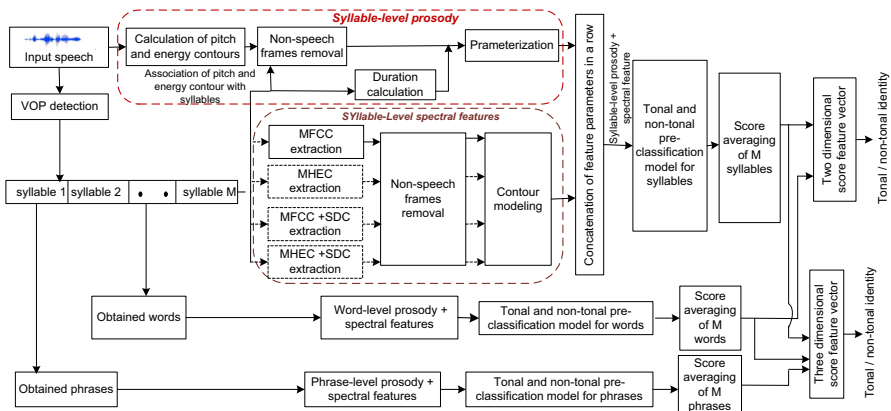


Fig. 2 Working of proposed system-II

prosody. The dotted boxes in Fig. 1 indicate the other features besides MFCCs, which have been explored in this work. Thus, four different scores, one for spectral feature and three others for prosody, are obtained against every test utterance. Depending on the classification scores, individual performances of different features are calculated. To analyse the effect of score combination, experiments have been carried out in three phases. In the first phase, scores obtained for syllable and word levels prosody are stacked together to form a two-dimensional feature vector. A classifier is then trained on these feature vectors to study this score combination. In the second phase, a third score from phrase-level prosody model is concatenated with the two-dimensional score vector of the first phase, yielding three-dimensional feature vector. Another classifier is trained on these three-dimensional features to study the complementarity of the phrase-level score. Lastly, the score feature vector is augmented with one of the utterance-level spectral model scores and a third classifier is built on four-dimensional features. Three different types of classifiers are explored with the scores-based features. The classifiers used are Logistic regression (LR) (Lee et al. 2011), SVM (Campbell et al. 2006) and Random forest (Casale et al. 2008). Here, the scores are used as inputs to any one of the classifiers at a time. The test set data has been split into two parts—one to train the classifier models as described above and other to test the models.

4.1.2 Description of system-II

Unlike the use of only utterance-level spectral features as in system-I, spectral features are analysed at multiple levels in system-II. Figure 2 shows the working of system-II.

Here, similar to system-I, multi-level prosody is used. Spectral features namely, MFCC, MFCC + SDC, MHEC and MHEC + SDC are extracted at syllable, word and phrase levels. A speech activity detection algorithm is used to find the speech frames. For syllable-level features, the MFCCs of all speech frames of a syllable are

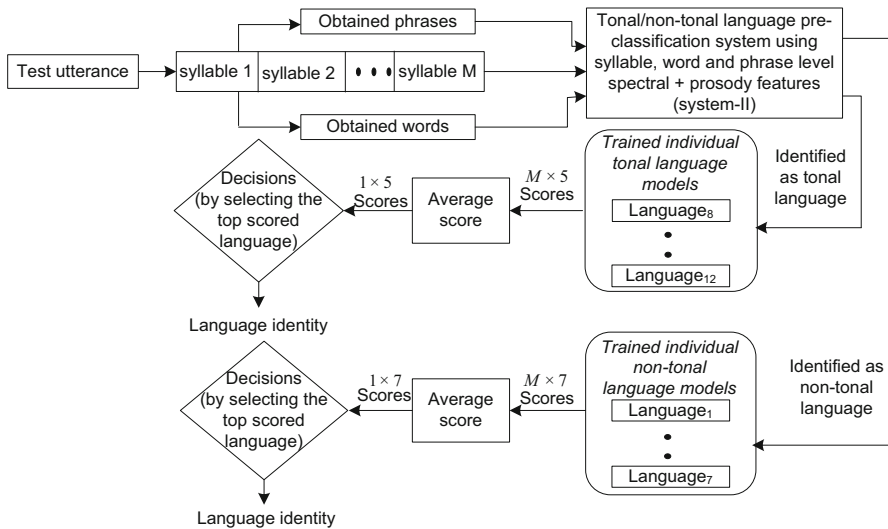


Fig. 3 Working of proposed pre-classification-based LID system (system-III)

further parameterized coefficient-wise using Legendre polynomial. That is, the contours of the different dimensions of MFCC features of a syllable are parameterized. Similarly, the other spectral features are further parameterized using Legendre polynomial. The final feature vector of a syllable or word is obtained by concatenating the parameters of prosody with that of MFCC, MFCC + SDC, MHEC or MHEC + SDC. In the case of phrase-level analysis, Δ MFCC or Δ MHEC is also used along with the phrase-level prosody.

Separate GMM-UBM models are trained for syllable-level MFCC, syllable-level (MFCC + SDC), syllable-level MHEC, syllable-level (MHEC + SDC), syllable-level (prosody + MFCC), syllable-level (prosody + MHEC), word-level (prosody + MFCC), word-level (prosody + MHEC), phrase-level (prosody + MFCC) and phrase-level (prosody + MHEC). Besides this, three other modelling techniques have been explored using the same set of features as is described in Section 4.6. As done in the case of system-I, firstly individual performances are calculated for the various feature levels and subsequently, classifiers are trained on score-based features in two different phases.

4.2 Pre-classification-based LID system (system-III)

In pre-classification-based LID system, languages are first pre-classified into tonal or non-tonal categories and then individual languages are identified from their respective categories. Here, depending on the decision on the pre-classification stage, the languages are processed through either the tonal or the non-tonal language identification modules in the second stage. Figure 3 depicts the pre-classification-based LID. The first stage is constituted by a tonal/non-tonal classification module (i.e. system-II) and it is followed by individual language models in two different

modules corresponding to the tonal and non-tonal categories. Multi-level prosody and spectral features are used in the second stage of the system. The language modelling technique (described in Sect. 4.6) that provides the best performance for the pre-classification task is used for the individual language identification task as well. Score combination of the different models developed using features of various levels is also done in way similar to that in system-II. Then, final scores from individual language models are obtained and the decision is taken based on the top-scored language.

4.3 Feature extraction

In one of the recent works (China Bhanja et al. 2018), some parameters of prosody have been proposed which are found to be effective for the tonal/non-tonal language discrimination of seven NE Indian languages. In this work, their effectiveness in pre-classifying Indian languages has been investigated for a larger set of 12 languages.

4.3.1 Syllable-level prosody

Different parameters of prosodic features used to represent syllable level prosody are: (China Bhanja et al. 2018) F_1 : mean pitch, F_2 : pitch changing level, F_3 : amplitude tilt for pitch contour, F_4 : duration tilt for pitch contour, F_5 : change in pitch, F_6 : distance of pitch contour's peak from VOP, F_7 : distance of the pitch contour point, reading 60% of the peak value, from VOP, F_8 : mean energy, F_9 : change in log energy, F_{10} : normalized energy changing level, F_{11} : distance of energy contour's peak from VOP, F_{12} : amplitude tilt of energy contour, F_{13} : duration tilt of energy contour, F_{14} : distance of the pitch contour point, reading 60% of the peak value, from VOP, F_{15} : syllable duration, F_{16} : ratio of the duration of voiced segment to that of the total segment (Rhythm).

Tonal language has a definite set of tones. For example, Manipuri language has two tones, Mizo language has four tones, Mandarin has four tones and so on. Generally, in level tone system, tones are distinguished by the pitch level, like, high (H) or low (L), relative to each other. In contour tone system, tones are distinguished by their pitch contours, like fall, rise, fall-rise or rise-fall relative to each other. These contours help in characterization of the different languages. Amplitude tilt (F_3) and duration tilt (F_4) explain the contour dynamics (Adami et al. 2003). Some work suggests the presence of a relation between tone height, the jaw's movements and lingual articulation and studied their roles in effecting different degrees of emphasis. Tone height which is represented by the parameter, change in pitch (F_5) (Dediu and Ladd 2007) can be used to differentiate tonal and non-tonal languages. Also, the pitch contour peak and the onset point of the accented syllable are consistently aligned for some languages, namely Greek and English, which are non-tonal in nature (Qu and Goad 2012). And in Mandarin, which is a tonal language, the peak is aligned with the tone-bearing syllable's offset point (Reddy et al. 2013). Therefore, distance of the peak location from VOP may provide

information for discriminating tonal from non-tonal languages. On the other hand, for some Tibeto-Burman family languages, like Dimasa and Mizo (Sarmah and Wiltshire 2010), the influence of the place features of consonants penetrates substantially into the contour of the next tone. Also, the interaction that exist between tones and segments (syllables) may cause shift in the tonal onset point (Gandour 1977). It has been experimentally found that the phenomenon of the tonal onset point propagating into the segment can be roughly parameterize by the distance of the pitch contour point, reading 60% of the peak value, from VOP. This feature parameter denoted by F_7 , may be used for language pre-classification.

Stress may be present up to a certain extent in any language. In certain scenes, some syllables are perceptually more prominent than others. These are termed as stressed syllables. The phonetic correlates of other features like duration and pitch also result in stress. Stress manifestation in the speech signal is often language-dependent and the energy parameter is commonly used to quantify it. Tonal languages which have register tones manifest correlation between tone and stress (Dusan and Deng 1998). Otherwise, most of the tonal languages don't have obvious stress (Qu and Goad 2012). Non-tonal languages, like English, on the other hand, has definite stress. Thus, stress is found to be a language-specific characteristic and can be used as a complementary feature with pitch contour. Stress is computed using the energies of all the speech (voiced) frames of the syllable. In this work, seven parameters (F_8 – F_{14}) have been used for stress quantization and two parameters (F_{15} and F_{16}) have been used to parameterize the duration characteristics. In this experiment, each individual parameters denotes single dimension and thus 16-dimensional prosodic feature vector is used for each syllable. Feature vectors corresponding to each word units are represented by 48-dimensional prosody.

4.3.2 *Phrase-level prosody*

In order to obtain the pitch feature at phrase-level (Δ pitch contour), an average of the successive differences of pitch values within a syllable are calculated. These average values for the sequence of 12 syllables are used to represent the Δ pitch contour of that phrase. Phrase-level energy contour (Δ energy contour) is obtained in a similar way. Phrase-level duration contour (duration contour) is formed using the duration values corresponding to 12 syllables in sequence. Then these parameters are concatenated in a row, resulting in 36-dimensional phrase-level prosody.

4.3.3 *MFCC and MFCC + SDC features*

MFCCs are the most widely used features for any LID task. It models the vocal tract information. Study has revealed that the vocal tract changes observed to be associated with the different tones of languages like, Mandarin and Vietnamese, are found to have a strong correlation with MFCC features (Le et al. 2009). It also reflects the human auditory perception. Besides, MFCCs have been found to carry

complementary information to pitch (Le et al. 2009), which is a robust feature widely used for LID task. MFCC feature extraction is performed according to the standard algorithm described in (Steven and Mermelstein 1980). In this experiment, first 7 MFCC coefficients have been considered as used in (Yin et al. 2006).

SDC features are used to exploit the temporal context information of the speech signal. It is computed using parameters of the standard configuration 7–1–3–7 (N–d–P–k) (Torres-Carrasquillo et al. 2007). The static Cepstral coefficients are augmented with these SDC features resulting in 56-dimensional features.

4.3.4 MHEC and MHEC + SDC features

Another spectral features used in this experiment is MHEC (Sadjadi and Hansen 2015) where instead of using conventional triangular filter banks, Gammatone filter banks can be employed to replicate the human ear cochlea response (Patterson et al. 1987). The amplitude modulation spectrum of a subband is calculated based on the Hilbert envelope of the output of the Gammatone filter bank. Here, 32-channel Gammatone filter banks are used and with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 200 and 3400 Hz. Hilbert envelope obtained in this experiment is divided into frames of 20 ms duration with 50% shift. Each frame is subjected to Hamming window to minimize the abrupt discontinuities that occur at the edges due to truncation of the signal. The windowing also reduce the correlation that may exist between adjacent frames. In order to compress the dynamic range of the envelope, root compression by a factor of 1/15 is done. First 7 coefficients (including c_0) of the MHEC has been used in our experiment. SDC of MHEC is calculated in the similar way as explained in Sect. 4.3.3 and a 56-dimensional feature vector is obtained for SDC features of 7 coefficients of MHEC.

4.4 Contour modelling of spectral features for system-II

After extracting the spectral features for all frames of the syllable, a contour of every cepstral coefficient is obtained from values across all the frames of the syllable. It is modelled as a linear combination of Legendre polynomials as given in Eq. (1).

$$f(t) = \sum_{i=0}^M a_i P_i(t) \quad (1)$$

where $f(t)$ represents the contour, $P_i(t)$ is the i th Legendre Polynomial and a_i encodes a characteristic feature of the contour shape (Martinez et al. 2013); a_0 is the mean, a_1 represents the slope, a_2 defines the curvature, and the more precise details of the contour are encoded by the higher order coefficients. With fourth order Legendre polynomials, this method gives 35-dimensional MFCCs, 280-dimensional MFCC+ SDC, 35-dimensional MHECs and 280-dimensional MHEC+ SDC for a syllable. In order to represent word-level MFCC/MHEC, 105-dimensional feature vectors are used. Δ MFCCs or Δ MHECs at phrase-level are obtained in a similar

way to that in case of phrase-level prosody where each syllable is represented by 12-dimensional Δ MFCCs or Δ MHECs.

4.5 Data normalization

The features require to be normalized for speaker variation, channel variation etc. GMM-UBM and i-vector based SVM classifiers have been trained on z-normalized (Ng et al. 2009) feature, while, for training ANN classifier, features are normalized to values in the range of -1 to $+1$.

4.6 Language modelling

In this work, four different modelling techniques namely, GMM-UBM (Reynolds 2015), i-vector based SVM (Dehak et al. 2011), ANN (Dorofki et al. 2012), and DNN (Mounika et al. 2016) are used. GMMs are used to model the language specific characteristics from the given feature set. UBM model is trained using EM algorithm and a specific GMM for each of the languages are adapted using maximum a posterior adaptation. GMM models are trained for the prosodic and spectral feature vectors obtained from each syllable. ANN has also been used for identifying different languages. A shallow architecture of ANN with one or two hidden layers is one of the approaches used for language identification. Generally, in case of ANN, the numbers of hidden neurons are chosen using trial and error method so that the best possible result may be obtained from the system. In this experiment feed forward neural networks are used which uses gradient descent algorithm. However, they have a poor convergence rate and no definite rules for choosing the optimal parameters of the training stage are available.

4.6.1 i-vector based SVM

Utilizing the same UBM (GMM), as explained in Sect. 4.6, i-vector extractor is trained. i-vector is a compact form representation of acoustic features. It is based on a factor analysis model given by (Dehak et al. 2011).

$$s = m + Tw \quad (2)$$

where s represents the GMM supervector of the speech segment with respect to the UBM. m is the supervectors' mean, T is the total variability matrix and w represents the i-vector.

An utterance can be considered as a sequence of a number of syllables. In order to improve the system performance, in one study (Dey et al. 2017), the i-vector with reference to a particular frame of an utterance has been computed with a left and a right context of L frames. Motivated by their study, in this work, i-vectors are computed for each of the M syllables in the utterance with a left and a right context of L syllables each. That is, the Baum-Welch statistics for the N^{th} syllable are calculated using syllables from $N - L$ to $N + L$. The i-vector sequence obtained as a result may be denoted by $w = [w_1, w_2, \dots, w_M]$.

The total variability matrix is trained on similarly short segments of speech obtained by dividing the training utterances. For word-level and phrase-level i-vectors a similar procedure is followed. i-vectors entirely contain the language, channel and speaker variability's. Variability compensation techniques, like linear discriminative analysis (LDA), WCCN or nuisance attribute projection (NAP) are required to reduce the variability in i-vector space for the following SVM classifier. In this case, within-class covariance normalization (WCCN) (Hatch et al. 2006) of the data has been done to generalize the linear kernel of SVM classifier.

4.6.2 DNN

For ANN, It would be problematic to train neural networks with more than 2 hidden layers. In a recent study, several advanced optimization software or fast computing hardware make it possible to train much deeper networks. DNN can have 5 or more hidden layers and training of DNN uses stochastic gradient descent (SGD) algorithm with a mini-batch for updating the DNN. It uses backpropagation algorithm to estimate the gradients of the DNN parameters for each mini batch. In this case, input to the DNN is a stacked set of features obtained from the syllables. Here, $\pm L$ syllables/ words/phrases of context around the current syllable or word have been stacked together to obtain input feature vectors of different dimensions for different features.

5 Experiments, results and discussions

Several experiments have been carried out to evaluate the performance of the proposed system implemented using multi-level prosody and spectral features. Equal Error Rate (EER) based performance analysis has been presented in this paper. Besides, Detection Error Trade-off (DET) curve is also presented. Systems have been tested for 30 s, 10 s and 3 s data.

5.1 Experimental setup

In this experiment, the whole dataset (described in Table 1) of NITS-LD has been split into three parts: NITS-train, NITS-development and NITS-test data set. Around 7–10 h data of each of the 12 languages totalling to around 100 h data forms the NITS-train set for this experiment. Out of the 100 h NITS-train set, 39 h of data has been obtained from five tonal languages and the remaining 61 h of data has been obtained from seven non-tonal languages. Also, 1 h of data of each of the 12 languages, totalling to around 12 h makes up the NITS-development set and 1 h data from each of the 12-languages totalling to around 12 h makes up the NITS-test set. NITS-train, NITS-development and NITS-test set are mutually exclusive. Performance of the system is dependent on the duration of the test utterances. Experiments, therefore, have been conducted with three different durations of test data, namely 30 s, 10 s and 3 s data.

In case of GMM-UBM, UBM has been built using NITS-development set and a GMM has been adapted using NITS-train set for each of the language categories (39 h of data of tonal language and 61 h of data of non-tonal language). The models have been tested for three different durations of test utterances. i-vector extractor is trained using the same NITS-development set. For i-vectors based SVM, a context size of $L = 3$, leading to a sliding window of 7 syllables/words/phrases is used with a shift step size of 1. In the case of ANN, NITS-train set has been used to train the models and NITS-development set to validate the models after each epoch. fivefold cross-validation approach is used in this experiment.

Feed-forward DNNs with fully connected layers have been used in these experiments. There are three hidden layers in the network and each hidden layer contains 512 hidden units with ReLU activation functions for the prosody model and 1024 hidden units with the same activation functions for rest of the DNN models. The learning rate and the L_2 regularization parameters have been set at 0.001 and 10^{-7} respectively. The output has been configured as a single output neuron with sigmoid activation to produce the class probabilities of the pre-classification stage and as a softmax layer with a dimension equal to the number of target languages for the second stage. Also, the model minimizes cross entropy loss using Adadelta optimizer. The models have been trained for 100 epochs with a mini-batch size of 256. NITS-train set has been used to train the models and NITS-development set to validate the models after each epoch. fivefold cross-validation approach has been used in this experiment and the model that gives the best accuracy at validation stage has been used as the final model. Finally, the NITS-test has been used for evaluating and reporting results on developed models.

To analyse the performance of OGI-MLTS database, the same procedure has been followed for all the modelling techniques. 15 h of training data (9 h data from non-tonal category and 6 h data from tonal category), 5 h of development data and 5 h of test data (30 min data from each language) of this database are used to perform the experiments.

5.2 Results of the tonal and non-tonal language pre-classification system

5.2.1 Results of system-1

In the case of GMM-UBM, experiment has been conducted using different number of Gaussian components particularly 2, 4, 8, 16, 32, 64, 128, 256, 512, etc. and it is observed that for syllable-level prosody, 16; for word-level prosody, 32; for phrase-level prosody, 16; for utterance-level MFCC or MHEC, 256; and for utterance-level (MFCC + SDC) or (MHEC + SDC), 512 result in the lowest individual EERs. Here, for syllable /word /phrase level analysis, the likelihood of all the syllables / words /phrases of a test utterance are averaged to obtain the score for the utterance. The top scored language is considered to be the identified language. EERs are calculated for individual languages. The average EER of the pre-classification stage is also calculated.

From Table 2, it can be observed that among syllable, word and phrase level prosody, syllable-level prosody provides the lowest EERs, followed by word-level

Table 2 Feature-wise performance of pre-classification stage (system-I) on NITS-LD database

Features	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu	Avg
Syllable-level prosody	45.2	53.7	54	27.7	28.2	37.9	39.6	25.7	25.2	38	25.8	28.7	35.8
	45.3	58.8	57.8	28.6	33.1	38.9	40.8	28	26.4	40.9	26.2	32.7	38.1
	47.8	59.2	58.4	30.7	35.8	39	42	28.8	28.9	41.5	26.7	33.2	39.3
Word-level prosody	42.4	55.1	53.6	26.4	25.5	36.7	46	29.4	29.5	42.7	32.2	27	37.2
	45.1	61.7	57.6	28.1	32.9	39.8	48.6	35.4	33.6	47.2	33.3	33.1	41.3
	47.4	62.1	58.3	30.5	35.4	43.3	48.2	36.2	36.7	47.6	34.9	32	42.7
Phrase-level prosody	49.4	55.4	57.4	29.6	31.4	41.2	41.3	27.1	26.7	39.7	27.6	30.5	38.1
	45.6	61.8	58.4	29.3	33.7	41.1	48.9	35.6	34.7	47.5	34.9	32.3	41.9
	49.6	63	59.4	31.3	35.7	43.7	48.9	36.4	37.2	48	35.5	33.2	43.4
Utterance-level MFCC	26.5	32.6	34.8	17.9	36.6	19.5	22.4	19.9	18.8	28.3	20.6	23.7	25.1
	27.2	38.9	37.4	20.5	37.6	22.2	27.8	23.4	24.7	33.5	23.6	26.9	28.6
	29.3	44.9	37.7	22.4	38.9	25.6	28.9	25.7	26.6	34.2	26.7	27.6	30.7
Utterance-level (MFCC + SDC)	24.8	30.5	32.6	15.6	34.7	17.4	20.3	17.7	16.3	26.4	18.7	21.6	23.0
	25.6	36.4	36.1	18.8	35.7	20.5	25.6	21.9	23	31.6	22	24.3	26.7
	28.4	43	36.5	21	37.1	24	27.4	24.1	25.1	33.5	25.3	26.1	29.2
Utterance-level MHEC	30.6	39.6	41.8	22.9	34.6	29.3	30.9	19.2	21.4	25.3	19.9	20.4	27.9
	36.6	42.6	42.3	23.4	35.1	30.7	32.8	21.9	22.3	29.8	21.7	21.6	30.0
	38.7	44.6	43.1	26.3	37.8	32.9	36.5	25	22.7	32.6	22.8	23.5	32.2
Utterance-level (MHEC + SDC)	29.1	37.7	39.9	20.5	32.7	27.6	29.1	17.5	19.6	24.2	18	18.8	26.2
	35.3	41.3	41	22.1	33.8	29.4	31.5	20.6	21	28.5	20.4	20.3	28.7
	37.7	43.6	42.3	25.1	36.6	31.8	35.1	23.9	21.5	31.6	21.1	22.2	31.0

Table 3 Performance of pre-classification stage (system-I) on NITSLD database for different score feature combination

Features	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu	Avg
EERs of GMM-UBM for the score feature combinations for 30 s, 10 s and 3 s test data respectively (%)													
Score combination of syllable and word levels prosody	40.7	46.6	50.1	26.3	35.8	34.9	36.5	25	22.7	32.6	22.8	23.5	33.1
	42.6	54.6	53.3	27.4	38.1	36.7	38.8	27.9	25.3	36.8	25.7	28.6	36.3
	43.6	55.6	55.8	29.9	41.6	37.3	40.9	28.2	28.4	35.3	24.9	30.4	37.6
Score combination of syllable, word and phrase levels prosody (Multi-level Prosody)	40.4	48.9	49.2	22.9	23.4	33.1	34.8	21.9	20.4	32.2	22	23.1	31
	41.5	53.7	52.7	23.5	30	33.8	35.7	24.9	21.3	35.8	21.1	27.6	33.5
	43.5	54.9	54.1	26.4	31.5	34.7	37.7	24.5	24.6	37.2	22.4	28.9	35
Multi-level Prosody, Utterance-level (MFCC + SDC)													
SVM	21.5	30.3	33.6	13.3	27.2	20.3	22.6	11.7	11.8	16.7	10.3	12.5	19.3
	23.8	34.1	32.6	15.9	27.6	22.3	23.8	13.1	13.8	21.6	10.8	12.6	21.0
	26.5	35.7	34.1	18.8	29.4	24.3	27.3	14.9	14.3	23.4	13.6	16.3	23.2
LR	22.2	31	34.3	14	27.9	21	23.3	12.4	12.5	17.4	11	13.2	20.0
	24.7	35.0	33.5	16.8	28.5	23.2	24.7	14.0	14.7	22.5	11.7	13.5	21.9
	27.3	36.5	34.9	19.6	30.2	25.1	28.1	15.7	15.1	24.2	14.4	17.1	24.0
Random forest	21.9	30.7	34	13.7	27.6	20.7	23	12.1	12.2	17.1	10.7	12.9	19.7
	24.4	34.7	33.2	16.5	28.2	22.9	24.4	13.7	14.4	22.2	11.4	13.2	21.6
	27	36.2	34.6	19.3	29.9	24.8	27.8	15.4	14.8	23.9	14.1	16.8	23.7
Multi-level Prosody, Utterance-level (MHEC + SDC)													
SVM	23.6	32.4	35.7	15.4	29.3	22.4	24.7	13.8	13.9	18.8	12.4	14.6	21.4
	25.9	36.2	34.7	18	29.7	24.4	25.9	15.2	15.9	23.7	12.9	14.7	23.1
	28.3	37.5	35.9	20.6	31.2	26.1	29.1	16.7	16.1	25.2	15.4	18.1	25.0

Table 3 continued

Features	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu	Avg
LR	24.5	33.3	36.6	16.3	30.2	23.3	25.6	14.7	14.8	19.7	13.3	15.5	22.3
	26.7	37	35.5	18.8	30.5	25.2	26.7	16	16.7	24.5	13.7	15.5	23.9
	29	38.2	36.6	21.3	31.9	26.8	29.8	17.4	16.8	25.9	16.1	18.8	25.7
Random forest	24.1	32.9	36.2	15.9	29.8	22.9	25.2	14.3	14.4	19.3	12.9	15.1	21.9
	26.3	36.6	35.1	18.4	30.1	24.8	26.3	15.6	16.3	24.1	13.3	15.1	23.5
	28.7	37.9	36.3	21	31.6	26.5	29.5	17.1	16.5	25.6	15.8	18.5	25.4

and then phrase-level prosody. Therefore, it can be inferred that tones in tonal languages are coded more distinctly in syllables than in words or phrases. Moreover, in this experiment, some of the tonal languages are monosyllabic and possibly because of that the word-level prosody provides the second lowest EER readings. Here, the average EERs for tonal languages are lesser for MHECs than MFCCs whereas, average EERs for non-tonal languages are lesser for MFCCs than MHECs. It is possibly because MHECs are based on the Hilbert envelop of the Gammtone filterbanks output which is the time–frequency representation of the cochleagram. Human cochlea which is a part of inner ear perceives the pitch variation more accurately than outer ear. Moreover, at low frequency range, the ERB scale, involved in the extraction of MHECs has finer resolution than Mel scale (Zhao and Wang 2013) and hence the better performance for tonal languages. It can also be observed that MFCC provides the lowest EERs among all individual features and the EERs further reduce when their SDC coefficients are considered. Here, SDC coefficients of both MFCCs and MHECs are found to carry useful information relevant to the pre-classification task. Also, for NITS-LD, SDCs of MFCC features are more effective than the SDCs of MHEC features.

Table 3 shows the performance of system-I for NITS-LD when combining the scores obtained using GMM-UBM for different levels of prosody and also for the combination of the scores of spectral and multi-level prosodic features. From Tables 2 and 3 it can be observed that, combining the scores of GMM-UBM obtained for syllable and word level prosody, improves the system performance by 2.7%, 1.8% and 1.7% over the syllable-level prosody and it shows further improvements of 4.8%, 4.6% and 4.3% (30 s, 10 s and 3 s test data respectively) after combining the scores from syllable, word and phrase level prosody. This may be due to the presence of some non-overlapping tonal/ non-tonal language discriminating information in the individual features. Moreover, some of the languages like Manipuri, Mizo, Bodo etc. are monosyllabic or are having most of the words monosyllabic in nature. On the other hand languages like Assamese, Indian English etc. are polysyllabic. Therefore, individual EERs of monosyllabic languages are lesser for syllable-level prosody than word-level prosody and for some of the polysyllabic languages, word-level prosody provides lesser EERs than syllable-level prosody. Hence the better performance obtained on combining the scores of different levels of prosody.

It can also be observed that both utterance-level (MFCC + SDC) and (MHEC + SDC) perform better than multi-level prosody and EER values reduce on combing the scores of GMM-UBM model developed using either utterance-level (MFCC + SDC) and multi-level prosody or (MHEC + SDC) and multi-level prosody. Additionally, for system-I, lowest EERs have been obtained for the combination of multi-level prosody and utterance-level (MFCC + SDC) and for this combination EERs reduce by 11.7%, 12.5% and 11.8% with respect to that for multi-level prosody for 30 s, 10 s and 3 s test data respectively. Of the classifiers trained on scores, SVM with polynomial kernel outperforms the rest with slightly better EER readings.

Similar experiments have been conducted on OGI-MLTS database and the experimental results are shown in Fig. 4. In this case, the number of Gaussians of

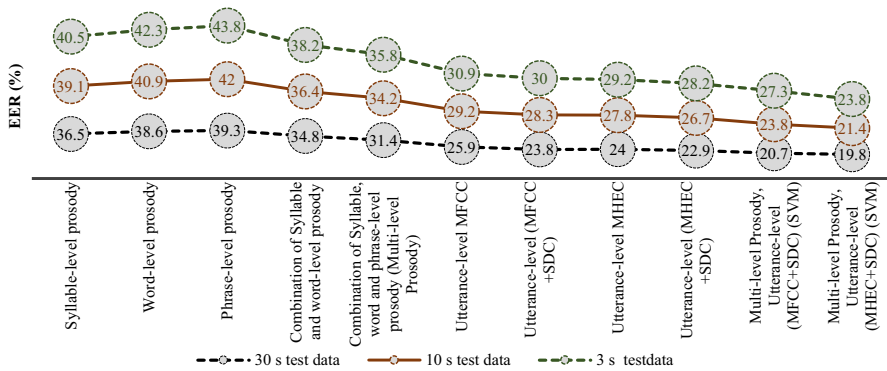


Fig. 4 Performance of system-I for OGI-MLTS database

GMM-UBM modelling techniques used for different features is same as used in the case of NITS-LD. From Fig. 4 it can be observed that syllable-level prosody carries more tonal/non-tonal language discriminating information than word or phrase level prosody. In this case, relative to syllable-level prosody, EER values reduce by 1.7%, 2.7% and 2.3% (30 s, 10 s, and 3 s test data) when the scores of syllable and word level prosody are used in combination, and 5.1%, 4.9% and 4.7% when the scores of syllable, word and phrase level prosody are combined. Like NITS-LD, both utterance-level MHECs and MFCCs carry complementary information with respect to multi-level prosody for OGI-MLTS database. The system provides the lowest EERs of 19.8%, 21.4% and 23.8% for the three respective test data for the combination of scores of multi-level prosody and utterance-level (MHEC + SDC) features which are 11.6%, 12.8% and 12% lesser than that obtained for multi-level prosodic features. In this case, score combination has been done using SVM with polynomial kernel. The following observations can be made from Tables 2, 3 and Fig. 4:

- Both utterance-level MHECs and MFCCs are useful for tonal and non-tonal discrimination. Also, both utterance-level MFCC and MHEC carry complementary information with respect to multi-level prosody.
- Prosody obtained from syllable, word and phrase levels carry complementary information.
- MFCC performs better than MHEC for NITS-LD, however, MHEC performs better than MFCC for OGI-MLTS database. It may be due to the fact that MHEC extraction uses nonlinear rectification step prior to the DCT, whereas MFCC uses log scale. This log operation on the Mel-power-spectrum makes the MFCCs scale-invariant whereas the MHECs are not scale-invariant because of the power-law compression with the factor of 1/15. This could make MHECs more noise robust and it provides better performance for the noisy OGI-MLTS database.
- System-I provides the EER of 19.8%, 21.4% and 23.8% for OGI-MLTS database using 30 s, 10 s and 3 s test data respectively which are slightly higher

Table 4 Performance of system-II for NITS-LD using GMM-UBM

Test data (s)	EER (%) for multi-level features						EERs(%) for combination system								
	Syllable-level			Word-level			Phrase-level			Syllable and word levels			Multi-level		
	MF CC	MF CC + SDC	MH EC + SDC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MH EC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC	Pro- sody + MF CC
30	22.0	22.8	24	25.2	20.2	22.4	22.6	24	23.9	25.3	18.2	20	15.8	17.6	
10	24.0	25.1	26.6	27	23.7	25.1	25	27.4	26.3	28.6	21	23.2	19	20.4	
3	26.7	27.4	28.7	29.1	25.6	28	27.1	30.8	28	30.2	24.1	25.4	22.1	23	

than the lowest EERs obtained for NITS-LD (19.3%, 21% and 23.2%). This is possibly because the collected speech samples of OGI-MLTS database are noisy whereas in case of NITS-LD the speech samples are noise free.

5.2.2 Results of system-II

This section shows the performances of different models namely, ANN, GMM-UBM, i-vector based SVM and DNN for multi-level spectral + prosodic features. It also analyses the system performance for the score combination of different models developed using multi-level features.

In case of GMM-UBM, for syllable-level MFCCs or MHECs, 256; for syllable-level (MFCC + SDC) or (MHEC + SDC), 512; for syllable-level (prosody + MFCC) or (prosody + MHEC), 512; for word-level (prosody + MFCC) or (prosody + MHEC), 512; and for phrase-level (prosody + MFCC) or (prosody + MHEC), 256 result in the lowest individual EERs. Score combination is performed in a similar way to that of system-I. From Tables 2, 3 and 4 it can be observed that syllable-level MFCC or MHEC features provide better performance than the MFCCs or MHECs of utterance-level. The improvements are 3.1%, 4.6% and 4% for MFCCs and 3.9%, 3.4% and 4.8% for MHECs using 30 s, 10 s and 3 s test data respectively. Here SDCs of both MFCC and MHEC extracted of syllables are identified to be insignificant for syllable-level analysis of pre-classification task. In this case, the system provides the lowest EER readings on combining scores corresponding to multi-level prosody + MFCC features which are 3.5%, 2% and 1.2% (for the three respective test data) lesser than the respective lowest EER readings of system-I. As explained in Sect. 5.2.1, the score for a particular utterance is calculated by averaging the likelihood scores of all the syllables/words/phrases of that utterance. The score combination among different levels of features has been performed in a similar way to system-I.

System performance has been analysed using ANN classifier. Several experiments have been carried out with different network structures and 16L–29N–8N–2L, for syllable-level prosody; 35L–50N–12N–2L, for syllable-level MFCC or MHEC; 280L–150N–40N–2L for syllable-level (MFCC + SDC) or (MHEC + SDC); 51L–82N–35N–2L, for syllable-level (prosody + MFCC) or (prosody + MHEC); 153L–182N–63N–2L, for word-level (prosody + MFCC) or (prosody + MHEC) and 48L–70N–16N–2L for phrase-level (prosody + MFCC) or (prosody + MHEC) prove to be the most effective network structures in case of NITS-LD. Here, L represents the linear units, N represents the non-linear units and selection of the most effective architecture of ANN is based on the trial and error procedure. For instance, in case of network structure 16L–29N–8N–2L, the number of nodes in the input layer is 16; the number of nodes at the output layer is 2 and the number of hidden units in the two hidden layers are 29 and 8 respectively. In case of all the network structure, same convention has been followed. In this experiment, the epochs limit is set to 500. Tan-sigmoid is used as the activation function. In this case, like GMM-UBM, the output scores corresponding to all the syllables/words/

phrases of a test utterance are averaged to calculate the final score for that utterance. Here, syllable-level prosody provides EERs of 29.8%, 32.3% and 34.6% for 30 s, 10 s and 3 s test data respectively. Performances of ANN classifier for system-II using different features are given in Table 5.

In order to get i-vectors corresponding to different features, same number of Gaussian mixtures have been used as used in case of GMM-UBM in system-II. Linear kernel of SVM and TV matrix dimensions of 100, 200, 400, 200, 250 and 200 for syllable-level prosody, syllable-level MFCC or MHEC, syllable-level (MFCC + SDC) or (MHEC + SDC), syllable-level (prosody + MFCC) or (prosody + MHEC), word-level (prosody + MFCC) or (prosody + MHEC) and phrase-level (prosody + MFCC) or (prosody + MHEC) lead to the lowest individual EERs. The scores of the SVM are converted into posterior probabilities based on optimal sigmoid transformation and then in a way similar to GMM-UBM, scores of all the syllables/words/phrases are averaged to calculate the score of the utterance. In this case, syllable-level prosody provides EERs of 32.3%, 34.6% and 37.4% for 30 s, 10 s and 3 s test data respectively. Score combination has been performed in a way similar to the ANN models. Table 6 shows the performances of system-II using i-vector based SVM and different features.

In the case of syllable-level analysis, input to the DNN is a stacked set of features obtained from the syllables. Here, ± 3 syllables around the current syllable have been stacked together to obtain input feature vectors of dimensions ($16*7 = 112$) for prosody, ($35*7 = 245$) for MHEC or MFCC, ($280*7 = 1960$) for (MHEC + SDC) or (MFCC + SDC) and ($51*7 = 357$) for (MHEC + prosody) or (MFCC + prosody). Similarly, for words and phrases, a context size of ± 3 lead to input feature vectors with the dimension of ($153*7 = 1071$) and ($48*7 = 336$) respectively, both for (prosody + MHEC) and (prosody + MFCC). Again, the scores corresponding to all of the syllables/words/phrases of a test utterance are averaged to calculate the score for that utterance. EERs of DNN model for different features are given in Table 7. Here, syllable-level prosody provides EERs of 27.2%, 29.6% and 31.2% for 30 s, 10 s and 3 s test data respectively.

From Tables 4, 5, 6 and 7 following observations can be made:

- Both syllable-level MFCC and MHEC perform better than the utterance-level.
- At syllable level also, both MFCC and MHEC carry commentary information with prosody for tonal/non-tonal discrimination task.
- Prosody + spectral features obtained from syllable, word and phrase levels carry non-overlapping tonal and non-tonal language discriminating information and score combination of different levels of features provides the best performance.
- SDC coefficients of both MFCC and MHEC do not carry any tonal and non-tonal language discrimination information at syllable level. This is possibly because the performance of SDC features are affected as the time span of the syllable units which often prove too short to capture any significant spectral transition cues between frames.
- DNN outperforms the other classifier, followed by ANN, GMM-UBM and then i-vector based SVM. Modelling with DNN helps to reduce the EERs of system-

Table 5 Performance of system-II for NITS-LD using ANN

Test data (s)	EER (%) for different features extracted of different analysis units										EERs(%) for combination system			
	Syllable-level			Word-level			Phrase-level			Syllable and word levels			Multi-level	
	MF	MFC + SDC	MH EC	MH EC + SDC	Pro- sody + MFCC	Prosody + MHEC	Pro- sody + MFCC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC
30	23.0	23.8	25.4	26	20	22	21.9	25.1	23.8	27.8	17.8	20.6	15.1	17
10	25	26.1	27.8	28.1	22.4	24.8	24	26.1	26.2	29.4	20	22	18.2	19.1
3	28.6	29	31.2	32	25.1	27.7	27.3	29	29.5	31.3	22.8	25.4	21.4	22

Table 6 Performance of system-II for NITS-LD using i-vector based SVM

Test data (s)	EER (%) for different features extracted of different analysis units										EERs(%) for combination system								
	Syllable-level					Word-level					Phrase-level			Syllable and word levels			Multi-level		
	MF	MF	MF	MF	MF	Pro-sody + MFCC	Pro-sody + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC	Pro-sody + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC	Pro-sody + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC	Pro-sody + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC	Pro-sody + MHEC
30	22.4	23.2	24.8	25.4	21.6	23.1	22.1	24.9	23.9	25.3	18.4	20.2	16.4	18					
10	26.2	27.1	28	29.1	25.1	27.4	27.8	30.4	28.4	32	22.9	24.7	21.6	22.1					
3	29	30.2	30.6	31.6	28	29	29.6	32.3	30.3	33.6	25	26.8	23.8	24.2					

Table 7 Performance of system-II for NITS-LD using DNN

Test data (s)	EER (%) for different features extracted of different analysis units													
	Syllable-level			Word-level			Phrase-level			EERs(%) for combination system				
	MF + SDC	MH EC	MHEC + SDC	Pro-sody + MFCC	Pro-sody + MHEC	Pro-sody + MFCC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC + MHEC	Pro-sody + MFCC + MHEC		
30	16.2	17.6	19.8	20.6	13.4	16.9	15	18.8	16.8	20	12.1	13.9	9.6	11
10	22.6	24	24.3	25.1	20.8	22.6	21.8	23.4	22.6	24.5	18.6	20	15.1	17.4
3	26.2	27.1	27.2	28	24	27.2	25.6	28.5	27	29.7	22.4	24.1	19.2	21.2

II over system-I with the values of 9.7%, 5.9% and 4% for 30 s, 10 s and 3 s test data respectively.

Same experiments have been conducted for OGI-MLTS database. Performances of OGI-MLTS database with GMM-UBM and DNN models are given in Tables 8 and 9 respectively. Modelling parameters remain the same as that used in case of NITS-LD. Like NIT-LD, for OGI-MLTS database, both MFCC and MHEC perform better when they are extracted at syllable-level than at utterance-level. Also, SDCs of both MFCC and MHEC do not carry any significant information at syllable-level. Here, the combined scores (using the polynomial kernel of SVM) of GMM-UBM developed using multi-level (prosody + MHEC) feature provides the lowest EER values, which are 3.5%, 2.2% and 1% (for the three respective test data) lesser than the lowest EERs of system-I. EERs further reduce when the features are modelled using DNN and the reductions are 9.6%, 5.6% and 3.8% for the three respective test data.

5.3 Results of the tonal and non-tonal pre-classification-based LID system

Two different case studies have been considered in this paper to study the effectiveness of pre-classification module in language identification.

5.3.1 Case study I (system-III)

In this case, we implement a pre-classification-based LID system for OGI-MLTS and NITS-LD databases (system-III). Tonal/non-tonal language pre-classification module as described in system-II (DNN-based) is used as a front-end of the individual language identification system. A baseline system has been prepared for identifying individual languages without any pre-classification module and the combining scores of multi-level (prosody + MFCC) features have been used. The performance of the proposed pre-classification-based LID system is compared with respect to the baseline system. For the baseline system, a single discriminative model corresponding to each of the individual feature set is trained using DNN to classify the participating languages. Thus, five separate DNN models for the features, namely syllable-level prosody, syllable-level MFCC, syllable-level (prosody + MFCC), word-level (prosody + MFCC) and phrase-level (prosody + MFCC) have been prepared. At the testing phase, identification is done based on output of the DNN model. Here, for NITS-LD, prosody, MFCC and their combinations are used as features. And the features are extracted for different levels (syllable, words and phrases) in a way similar to that in the pre-classification stage (system-II). Score combination and decision making are also done as explained earlier. In case of OGI-MLTS database prosody, MHEC and their combination are used to analyse the system performance. The features have been chosen based on their performance measures obtained for the respective databases. The performances of the baseline system and the pre-classification-based language identification system are given in Table 10.

Table 8 Performance of system-II for OGI-MLTS database using GMM-UIBM

Test data (s)	EER (%) for different features extracted of different analysis units																			
	Syllable-level					Word-level					Phrase-level					EERs (%) for combination system				
	Pro- sody	MF CC	MH EC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	
30	31.4	22.6	20.1	20.8	18.3	21.6	21.6	19.4	22.4	20.1	22.4	20.1	19.6	17.4	18.4	16.1	19.2	22.8		
10	34.2	24.6	22.8	22	21.6	23.4	22.8	22.8	25	23.6	25	23.6	21.2	20.8	20	19.2	23.2	22.8		
3	35.8	27	25.2	25.4	24	28	25.3	25.3	29.4	26.2	29.4	26.2	24.3	23.6	23.2	22.8	23.2	22.8		

Table 9 Performance of system-II for OGI-MLTS database using DNN

Test data (s)	EER (%) for different features extracted of different analysis units						EERs (%) for combination system								
	Syllable-level			Word-level			Phrase-level			Syllable and word levels			Multi-level		
	Pro- sody	MF CC	MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	Pro- sody + MFCC	Pro- sody + MHEC	
30	28.1	18.6	17.8	16.4	14.2	18.2	16	19.3	17.4	14.1	13.2	12.4	10.2		
10	30	22.8	20.6	20.8	17.1	21.4	19.2	22.8	21	18.7	16.4	16.7	15.8		
3	32.4	26.4	25.3	23	22.6	25.7	24.3	26.4	25.2	22.1	21.3	21.8	20		

Table 10 Performance of baseline system and system-III

Baseline system (without pre-classification language identification)	EERs of individual for 30 s, 10 s and 3 s test data respectively (%)												
	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu	Avg
Syllable-level Prosody	32.5	18.9	23.8	32.6	43.3	44.0	27.8	40.3	38.9	19.2	24.3	15.2	30
	35.4	20.5	26.2	34.1	46.4	46.1	29.3	42.5	40	22.4	26.8	18.8	32.3
	37.9	23.3	28.9	37.6	49	49.4	32.6	44.2	43.7	25.1	30.5	22.5	35.4
Syllable-level MFCC	34.6	18.3	24	27.6	6.4	20.8	23.1	28.1	6.6	12.7	25.8	21.9	20.8
	36.3	22.3	28.2	31.1	12.3	22.6	26.9	35.4	12.8	14.6	29.3	23.5	24.6
	38	26.2	30.1	33.9	14.6	27.4	30.8	37.4	15.6	17.3	32.6	24.8	27.3
Syllable-level (Prosody + MFCC)	33.7	15.4	22.9	26.4	5.9	18.3	21.8	26.5	5.5	11.9	22.3	14	18.7
	35.9	18.7	25.4	29.7	12	21.9	25.7	33.7	12.1	13.3	25.3	16.9	22.6
	36	21.6	26.4	32.8	13.4	26.8	29.1	36.2	13.3	15.8	27.3	21.8	25
Word-level (Prosody + MFCC)	31	14.8	19.4	25.5	6.2	14.3	18.1	25.8	6.8	9.2	18.6	14.3	17
	33	16.1	23	28.1	8.8	17.9	21.4	28.9	10.1	11.5	21	13.4	19.4
	35.1	19.2	24	30.4	13.2	22.9	26.7	34.8	14	11.6	22.8	19.7	22.9
Phrase-level (Prosody + MFCC)	34.2	16.3	23.4	28.1	10.4	19.3	22.9	24.3	10.6	10.2	19.5	13.8	19.4
	37.1	19.7	27.6	29.3	14	23.9	27.1	33.3	14.5	14	26.1	17.9	23.7
	40	22.8	32	34.1	17.9	27.4	30.2	36.1	19.2	20.3	30.5	21.1	27.6
Score combination of syllable and word levels (Prosody + MFCC)	30.3	14	19.3	25.1	5.1	13.5	17.9	24.2	4.9	8.6	18	13.6	16.2
	32.1	15.8	22.6	27.2	8	17.2	20.8	27.8	9.3	10.8	20.4	12.9	18.7
	34.3	18.5	23.9	28.4	10.3	18.6	24.9	31.4	12.1	10.6	24.4	20.9	21.5
Score combination of multi-level (Prosody + MFCC)	27.9	12.3	15.8	21.6	4.6	10.6	15.7	22.6	4.8	8.1	16.4	12.3	14.4
	30.1	13.1	18.2	22.9	6	11.9	18.1	23.6	6.2	9.2	18.1	12.4	15.8
	31.6	15.8	21.6	23.9	9.6	15.3	21.5	30.9	10.9	10.1	20.9	16.3	19
With pre-classification language identification (system-III)													

Table 10 continued

Baseline system (without pre-classification language identification)	EERs of individual for 30 s, 10 s and 3 s test data respectively (%)												
	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu	Avg
Score combination of multi-level (Prosody + MFCC)	11.8	12.8	16.8	9.6	12.8	15.4	13.6	10.1	13	4.9	8	5.8	11.2
	13.5	14.2	18.3	12.4	14.2	16.7	16.2	12.6	14.8	6.3	9.4	8.3	13.0
	17	17.7	22.5	14.8	17.7	20.2	19.7	16	17.5	10.7	13.5	12.1	16.6

Bold values indicate the lowest EER values obtained for baseline system and System-III when NITS-LD has been used

From Table 10, it can be observed that the languages like Hindi, Nagamese, Odia, Manipuri, Mizo etc. are identified better using MFCC feature whereas the languages like Assamese, Gojri, Punjabi etc. are well distinguished by prosodic features. Thus, it confirms that prosody and MFCC are complementary with each other and their combination helps to improve the language identification performance. In this case, the system performs better for the word-level (prosody + MFCC) features than syllable-level (prosody + MFCC) features and the overall EERs, in this case, reduce by 1.7%, 3.2% and 5.6% for 30 s, 10 s and 3 s test data respectively. This is possibly due to the presence of dynamic information of different features at word-level where 3 consecutive syllables are concatenated. System performances are further improved for the score combination of the DNN models prepared for syllable and word level (prosody + MFCC) features or for syllable, word and phrase level (prosody + MFCC) features. The baseline system provides the lowest EERs of 14.4%, 15.8% and 19% for 30 s, 10 s and 3 s test data after combining the scores from syllable-, word- and phrase-level (prosody + MFCC).

From Table 10 it can also be observed that on the addition of a pre-classification module, system performance improves. The pre-classification-based LID system provides the EERs of 11.2%, 13% and 16.6% for the three respective test data. Thus, the proposed system shows absolute improvement of 3.2%, 2.8% and 2.4% in EER over the baseline system. It is also evident that the pre-classification module may not necessarily improve the performance of all languages. The languages like English, Nagamese, Odia, Mizo etc. show higher EERs when pre-classification is done before individual language identification. The error in decision that occurs in the pre-classification module is carried over to the second stage, thereby increasing the EERs of certain languages. Nevertheless, those languages which are correctly pre-classified are identified with significantly reduced EERs, and as a result the overall EERs of the system reduce by a good margin.

The system performance has been analysed further by considering a hundred percent accurate pre-classification module at the first stage. The system is found to have lowest EER values of 9%, 10.8% and 14.4% for three different duration of test data after combining the scores from syllable-, word- and phrase-level (prosody + MFCC). Thus, the system provides the absolute improvements of 5.4%, 5% and 4.6% over the baseline system.

Experimental results given in Table 11 show that pre-classification also help boost the identification performance of the world's distinct languages of OGI-MLTS database. In this case the improvements are 4.2%, 4.1% and 3.2% for the three respective test data. The system performance further improves when a hundred percent accurate pre-classification module at the first stage has been considered. In this case, after combining the scores of the syllable-word- and phrase- level (prosody + MHEC), the system provides the EERs of 10.1%, 13.7% and 15.4% for 30 s, 10 s and 3 s test data. Here, the system performance improves by 6.1%, 5.6% and 5.2% over the baseline system. From Table 10 and Table 11 it is observed that the pre-classification module results in more significant performance improvement for OGI-MLTS database than NITS-LD. This observation is also made in the case when a hundred percent accurate pre-classification module is considered at the first

stage. This is possibly because OGI-MLTS database involves World's distinct languages, while NITS-LD languages are of the same origin and are closely related.

Table 12 shows the confusion matrix and Table 13 shows the FPR and FNR values of the system when the languages are identified without pre-classifying them. Experimental results are given for DNN model (fivefold cross validation) based on multi-level (prosody + MFCC) feature (after score combination). Table 12 and Table 13 show that most of the languages are confused with Hindi and hence the system reports the highest FPR value for it. The system shows the lowest FNR for Nagamese language in this case and also, among all the languages Nagamese language reads the lowest EER. The system reports the lowest FPR for Tamil language indicating that languages are least confused with it.

Table 14 shows the confusion matrix and Table 15 depicts the FPR and FNR values of the system for pre-classification-based LID task of NITS-LD. Following observations can be made from Tables 14 and 15:

- It may be said that confusion of other languages with Manipuri is very less as the system reports the lowest FPR for this language. Additionally, only Mizo language is confused with it. It may be because of the fact that both Manipuri and Mizo belong to the Sino-Tibetan family and hence the confusion.
- The system provides the lowest FNR for Bodo language. It is only confused with Gojri and Hindi language and other languages are less confused with it. Therefore it provides the lowest EER for this system.
- The system provides the highest FPR for Hindi language. It can be observed that use of pre-classification module helps to reduce the FPR value to certain extent for this language. Since it is considered as the parent language of most of the languages, FPR value turns out to be the highest for it.

From Tables 12, 13, 14 and 15 it can be clearly observed that an accurate pre-classification stage help boost the system performance. FPR and FNR values are reasonably high for the system where no such pre-classification module is present. FPR values of some of the languages like Bengali, English, Hindi, Tamil, Manipuri, Mizo etc. reduce for the pre-classification module and for some of the languages like Hindi, Bengali, English, Tamil, Manipuri, Bodo, Gojri etc. both FPR and FNR value reduce. The misclassification of some of the languages at pre-classification stage leads to an increase in overall FNR for some languages. Therefore, it can be inferred that if a perfectly accurate pre-classification module is present, languages of tonal category would not confuse with the languages of non-tonal category and cause the FPR and FNR values for all the languages to reduce.

5.3.2 Case study II

A comparative study of the pre-classification-based LID system with the system proposed in (Reddy et al. 2013) is presented in this section. In the first experiment, the performance of 12 individual languages of NITS-LD has been obtained using the system described in Reddy et al. (2013). In the second experiment, an LID

Table 11 Performance of OGI-MLTS database

	EERs (%)		
	30 s test data	10 s test data	3 s test data
Baseline (Without pre-classification language identification)			
Syllable-level Prosody	32.3	34.2	37.1
Syllable-level MHEC	21.4	25	28.6
Syllable-level (Prosody + MHEC)	19.8	23.6	25.9
Word-level (Prosody + MHEC)	18.6	21.7	23.3
Phrase-level (Prosody + MHEC)	22.3	25.4	29.1
Score combination of syllable and word-level (Prosody + MHEC)	17.8	20.4	21.9
Score combination of multi-level (Prosody + MHEC)	16.2	19.3	20.6
With pre-classification language identification (system-III)			
Score combination of multi-level (Prosody + MHEC)	12	15.2	17.4

Table 12 Confusion matrix obtained for NITS-LD when no pre-classification module is present. (Rows list the actual class and columns represent the assigned class)

	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu
As	360	71	0	0	69	0	0	0	0	0	0	0
Be	30	418	5	5	0	32	10	0	0	0	0	0
En	0	0	415	29	0	0	0	0	30	26	0	0
Hi	0	38	0	390	0	15	10	0	0	0	0	47
Na	20	0	0	12	468	0	0	0	0	0	0	0
Od	10	34	0	20	0	436	0	0	0	0	0	0
Ta	0	0	0	24	0	0	418	0	0	0	38	20
Ma	0	0	17	0	0	0	0	384	74	25	0	0
Mi	0	0	0	0	7	0	0	21	467	5	0	0
Bo	0	0	0	30	0	0	0	0	21	449	0	0
Go	0	0	0	68	0	0	0	0	0	0	412	20
Pu	0	0	0	31	0	0	0	0	0	0	35	434

Table 13 FNR and FPR analysis (in %) for without pre-classification-based LID system using combination of prosody and MFCC features and DNN

Languages	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu
FNR	28	16.4	17	22	6.4	12.8	16.4	23.2	6.6	10.2	17.6	13.2
FPR	1.09	2.6	0.40	4.0	1.3	0.85	0.36	0.38	2.27	1.01	1.3	1.5

system using the pre-classification module, system-I, which is based on the same set of features as in (Reddy et al. 2013), is evaluated.

Average EERs obtained, in this case, are given in Table 16. EERs for prosodic features are obtained after score combination of different levels of syllable-, word- and phrase-level prosody. It can be observed that in this case also, pre-classification help boost the system performance and the improvements are 2.8%, 3.2% and 3.4% for the three respective test data. The system performance further improves when a hundred percent accurate pre-classification module is considered and the system shows the improvements of 5.6%, 5.2% and 4.8% over the system where no such pre-classification module is present.

6 Conclusions

This paper presents an automatic pre-classification-based LID system for the same originated Indian languages based on multi-level prosody and spectral features. Here, features are extracted from syllable, word, and phrase levels of the speech signal and it is observed that syllable-level features are the most appropriate for

Table 14 Confusion matrix for pre-classification-based LID system for NITS-LD using DNN in both stages (five-fold cross validation), (prosodic + MFCC) feature and 30 s test data.(Rows list the actual class and columns represent the assigned class)

	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu
As	454	22	0	29	10	0	0	0	0	0	0	0
Be	25	434	0	0	0	39	0	0	0	0	0	0
En	44	0	441	10	0	0	0	0	0	0	0	20
Hi	0	27	0	449	0	25	8	0	0	0	0	0
Na	39	0	0	19	428	0	0	0	5	0	0	0
Od	0	42	0	35	0	400	0	0	0	0	0	0
Ta	25	0	0	0	0	0	458	0	0	0	6	47
Ma	6	0	10	0	0	0	0	449	34	0	0	0
Mi	0	0	0	38	4	0	0	4	394	13	0	0
Bo	0	0	0	19	0	0	0	0	0	485	17	0
Go	0	0	0	41	0	0	0	0	0	0	453	0
Pu	0	0	0	19	0	0	0	0	0	0	20	453

tonal/non-tonal pre-classification-based LID task. Experimental results also suggest that multi-level features carry complementary information with each other and the

Table 15 FNR and FPR analysis (in %) for pre-classification-based LID system using combination of prosody and MFCC features and DNN

Languages	As	Be	En	Hi	Na	Od	Ta	Ma	Mi	Bo	Go	Pu
FNR	11.8	12.8	14.3	11.7	12.8	16.4	14.5	10.02	13	6.9	8.2	7.9
FPR	2.5	1.6	0.18	3.8	0.25	1.15	0.14	0.07	0.703	0.23	0.78	1.21

system performance improves when the scores obtained from different models using multi-level features are combined. It is observed that at pre-classification stage, fusion of scores from the prosodic features derived from multiple levels of speech signal help to reduce the EERs from that of syllable-level prosody system by 4.8%, 4.6% and 4.3% (30 s, 10 s and 3 s) for NITS-LD and 5.1%, 4.9% and 4.7% for OGI-MLTS database. Score combination of multi-level (spectral + prosody) helps to reduce the EERs by 3.8%, 5.7% and 4.8% for NITS-LD and 4%, 1.3% and 2.6% for OGI-MLTS compared to the system developed using syllable-level (spectral + prosody). The effectiveness of MHEC/MHEC + SDC features and their complementarity with prosody for pre-classification-based LID task has been analysed. It can also be noticed that MHEC performs better than MFCC for noisy data condition. SDCs of both MHEC and MFCC are not effective for syllable level analysis. However, they carry complementary information for utterance level analysis of the system. Experiments have been conducted for twelve languages of

Table 16 Performance of Case study II

EER (%)	Multi-level Prosody		Utterance-level MFCC		Score combination of multi-level Prosody and utterance-level MFCC				
	30 s	10 s	3 s	30 s	10 s	3 s			
Without pre-classification module	35.4	38.2	40.8	28.6	31.3	34	23.4	27.6	30.1
With pre-classification module	31.3	33.7	37.6	24.9	28.1	30.6	20.6	23.4	26.7
With a 100% accurate pre-classification module	29.8	31.6	35	22.5	26.3	28.4	17.8	22.4	25.3

NITS-LD and ten languages of OGI-MLTS database. Four different models, namely ANN, GMM-UBM, i-Vector based SVM, and DNN have been explored along with different types of features. 30 s, 10 s and 3 s test data are used for validation of the proposed systems. At the pre-classification stage, 30 s duration test data and the choice of DNN models provide the lowest EERs of 9.6% and 10.2% for NITS-LD and OGI-MLTS respectively. We have done two different case studies and observed the effectiveness of pre-classification module in both the case studies. Our proposed pre-classification-based LID system provides EERs of 11.2%, 13.0%, and 16.6% for NITS-LD and 12%, 15.2% and 17.4% for OGI-MLTS database. Also, experimental results indicate that the efficacy of pre-classification module for the world's distinct languages of OGI-MLTS database is more prominent than closely related Indian languages of NITS-LD.

In future, some extra features may be added to improve the language classification performance for short duration test data. Syllable boundary determination being an important part of the techniques presented in this paper, any error therein can lead to performance degradation. Therefore, more accurate syllable determination is an important area of research. Also, to improve system performance of closely related Indian languages, instead of preparing a two-level LID system, a hierarchical LID system can be prepared. In future, session, channel and speaker variability may be dealt with more extensively to develop more effective language classification system.

References

- Adami, A. G., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003, April). Modeling prosodic dynamics for speaker recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03)*. (Vol. 4, pp. IV-788). IEEE.
- Atterer, M., & Ladd, D. R. (2004). On the phonetics and phonology of “segmental anchoring” of F0: Evidence from German. *Journal of Phonetics*, *32*(2), 177–197.
- Baby, A., Thomas, A. L., & Nishanthi, N. L. (2016). T. Consortium, “Resources for Indian languages,” CBBLR-Community-Based Building of Language Resources. Brno, Czech Republic: Tribun EU, 37–43.
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255–309.
- Burgos, W. (2014). Gammatone and MFCC Features in Speaker Recognition (Doctoral dissertation).
- Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*, 210–229.
- Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008, August). Speech emotion classification using machine learning algorithms. In *The IEEE international conference on semantic computing* (pp. 158–165). IEEE.
- China Bhanja, C., Laskar, M. A., & Laskar, R. H. (2018 October). A pre-classification-based language identification for Northeast Indian Languages using prosody and spectral features. *Circuits System and Signal Processing*. <https://doi.org/10.1007/s00034-018-0962-x>.
- Dediu, D., & Ladd, D. R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences*, *104*(26), 10944–10949.
- Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*.

- Dey, S., Motlicek, P., Madikeri, S., & Ferras, M. (2017). Template-matching for text-dependent speaker verification. *Speech Communication*, 88, 96–105.
- Dorofki, M., Elshafie, A. H., Jaafar, O., Karim, O. A., & Mastura, S. (2012). Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data. *International Proceedings of Chemical, Biological and Environmental Engineering*, 33, 39–44.
- Dusan S, & Deng L. (1998). Recovering vocal tract shapes from MFCC parameters. In *Fifth International Conference on Spoken Language Processing*.
- Gandour, J. (1977). Counterfeit tones in the speech of Southern Thai bidialectals. *Lingua*, 41, 125–143.
- Hatch, A. O., Kajarekar, S., & Stolcke, A. (2006). Within-class covariance normalization for SVM-based speaker recognition. In *Ninth international conference on spoken language processing*. <https://www.iitm.ac.in/donlab/tt/index.php>
- Jothilakshmi, S., Ramalingam, V., & Palanivel, S. (2012). A hierarchical language identification system for Indian languages. *Digital Signal Processing*, 22(3), 544–553.
- Le, P. N., Ambikairajah, E., & Choi, E. H. (2009, July). Improvement of Vietnamese tone classification using FM and MFCC features. In *Computing and communication technologies*, 2009. RIVF'09. International Conference on (pp. 1–4). IEEE.
- Lee, C. C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9–10), 1162–1171.
- Li, M., & Narayanan, S. (2014). Simplified supervised i-vector modelling with application to robust and efficient language identification and speaker verification. *Computer Speech and Language*, 28, 940–958.
- Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6), 1791–1801.
- Maddieson, I., Dryer, M. S., & Haspelmath, M. (2013). *The world atlas of language structures online*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology.
- Maity, S., Vuppala, A. K., Rao, K. S., & Nandi, D. (2012, February). IITKGP-MLILSC speech database for language identification. In *National conference on communication*.
- Martinez, D., Lleida, E., Ortega, A., & Miguel, A. (2013, May). Prosodic features and formant modelling for an i-vector-based language recognition system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6847–6851). IEEE.
- Mary, L. (2006). Multilevel implicit features for language and speaker recognition. Ph.D. dissertation. IIT Madras, India.
- Mounika, K. V., Achanta, S., Lakshmi, H. R., Gangashetty, S. V., & Vuppala, A. K. (2016, June). An investigation of deep neural network architectures for language recognition in Indian languages. In *INTERSPEECH* (pp. 2930–2933).
- Muthusamy, Y. K., Cole, R. A., & Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. In *Second International Conference on Spoken Language Processing*.
- Ng, R. W. M., Lee, T., Leung, C. C., Ma, B., Li, H. (2009). Analysis and selection of prosodic features for language identification. In *Proc. Asian Language Processing*, pp 123–128.
- Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987, December). An efficient auditory filterbank based on the gammatone function. In *A meeting of the IOC Speech Group on Auditory Modelling at RSRE* (Vol. 2, No. 7).
- Prasanna, S. R. M., Reddy, B. S., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4), 556–565.
- Prince, S. J., & Elder, J. H. (2007, October). Probabilistic linear discriminant analysis for inferences about identity. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*. (pp. 1–8). IEEE.
- Qu, C., & Goad, H. (2012). *The interaction of stress and tone in standard Chinese: Experimental findings and theoretical consequences*. Tone: Theory and Practice, Max Planck Institute for Evolutionary Anthropology.
- Reddy, V. R., Maity, S., & Rao, K. S. (2013). Identification of Indian languages using multi-level spectral and prosodic features. *International Journal of Speech Technology*, 16(4), 489–511.
- Reynolds, D. (2015). Gaussian mixture models. *Encyclopedia of biometrics*, 827–832.
- Richardson, F., Reynolds, D., & Dehak, N. (2015a). A unified deep neural network for speaker and language recognition. In: *proc of International Speech Communication Association*.

- Richardson, F., Reynolds, D., & Dehak, N. (2015b). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10), 1671–1675.
- Ryant, N., Yuan, J., & Liberman, M. (2014, May). Mandarin tone classification without pitch tracking. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 4868–4872). IEEE.
- Sadjadi, S. O., & Hansen, J. H. L. (2015). Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Communication*, 72, 138–148.
- Sarmah, P., & Wiltshire, C. R. (2010). A preliminary acoustic study of Mizo vowels and tones. *Journal of Acoustic Society of India*, 37(3), 121–129.
- Singh, A. K. (2006, October). A computational phonetic model for Indian language scripts. In *Constraints on spelling changes: Fifth international workshop on writing systems*.
- Steven, D., & Mermelstein, P. (August 1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28(4), 357–366.
- Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Klein & K. K. Paliwal (Eds.), *Speech coding and synthesis*. New York: Elsevier.
- Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & Deller Jr, J. R. (2007). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *Seventh international conference on spoken language processing*.
- Wang, L., Ambikairajah, E., & Choi, E. H. (2007, September). Automatic language recognition with tonal and non-tonal language pre-classification. In *Signal Processing Conference, 2007 15th European* (pp. 2375–2379). IEEE.
- www.ciiil-spokencorpus.net [Online, Retrieved January 20, 2009].
- Yin, B., Ambikairajah, E., & Chen, F. (2006). Combining cepstral and prosodic features in language identification. In *18th international conference on pattern recognition (ICPR'06)* (Vol. 4, pp. 254–257). IEEE.
- Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. In C.-T. James Huang, Y.-H. Audrey Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 443–464). Oxford: Wiley Blackwell.
- Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *2013 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, (pp. 7204–7208). IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.