Circuits, Systems, and Signal Processing (2019) 38:2266–2296
https://doi.org/10.1007/s00034-018-0962-x

A Pre-classification-Based Language Identification for Northeast Indian Languages Using Prosody and Spectral Features

Chuya China Bhanja[1] · Mohammad Azharuddin Laskar[1] · Rabul Hussain Laskar[1]

Received: 20 February 2018 / Revised: 4 October 2018 / Accepted: 6 October 2018 /
Published online: 12 October 2018

Abstract

This paper is aimed at developing a two-stage language identification (LID) system for Northeast Indian languages. In the first stage, languages are pre-classified into tonal and non-tonal categories, and in the second stage, individual languages are identified from languages of the corresponding category. In this work, new parameters to model the prosodic characteristics of the speech signal have been proposed for pre-classification as well as individual language identification. Also, the effectiveness of spectral features, namely Mel-frequency cepstral coefficient (MFCC) and their combination with prosodic features, has been studied for pre-classification task. The usefulness of MFCC with their delta and acceleration coefficients in combination with prosodic features has been investigated for individual language identification. The performance of the system is analyzed for the features extracted of different analysis units, such as syllable, disyllable, word, and utterance. Comparative performance analysis of three different classifiers, namely artificial neural network (ANN), Gaussian mixture model–Universal background model (GMM–UBM), and *i*-vector based support vector machine (*i*-vector based SVM), has been made for pre-classification as well as individual language identification. A new database, NIT Silchar language database (NITS-LD), has been developed for seven NE Indian languages using All India Radio broadcast news. The experimental analysis suggests that the parameters proposed to represent the prosodic characteristics help to improve the performance of both the stages and show improvements over existing parameters by as much as 7.4%, 11.9%, and 9.1% for 30 s, 10 s, and 3 s test data, respectively, in the pre-classification stage. Of the baseline single-stage systems, GMM–UBM provides the highest accuracies of 80%, 76.8%, and 72% for 30 s, 10 s, and 3 s test data, respectively. In the proposed system, the combination of the ANN model in pre-classification stage and the

✉ Chuya China Bhanja
chuya.bhanja@gmail.com

Extended author information available on the last page of the article

GMM–UBM model in individual language identification stage provides the highest accuracies, and it shows the improvements over the baseline system by 7.2%, 7%, and 4.9% for 30 s, 10 s, and 3 s test data. For OGI-Multilingual (OGI-MLTS) database, improvements of 8.1%, 7.4%, and 5.7% for 30 s, 10 s, and 3 s test data, respectively, are observed over the baseline LID system.

Keywords Language identification · Pre-classification of tonal and non-tonal languages · Syllables · Features · Classifiers · Database

1 Introduction

The main objective of automatic LID systems is to identify the language correctly from a given speech sample [4]. An ideal LID system should accurately utilize different aspects of speech information which are useful for distinguishing languages from a huge number of target languages. In the practical scenario, performance of an LID system largely depends on the number of target languages. In order to get higher accuracy for system involving large number of target languages, pre-classification of languages into different sub-language families or into different categories can be done. Also, to identify closely related languages or the languages of same origin, a highly accurate pre-classification module is required.

In order to address this aspect, Wang et al. [44] outlined a novel system for pre-classifying languages into tonal and non-tonal categories at utterance level, using different parameters of pitch contour and durations features and ANN as classifier. They have extended their work further to show the impact of the pre-classification task on performance of the system [45]. Here they showed that the performance of the system improves by 4–5% when pre-classification of languages into tonal and non-tonal category is done before doing individual language classification. Additionally, they reported that computation time of CPU reduces for the prosody-based two-level language identification systems. However, this system has several disadvantages. The main drawback of this system is that the use of phonetically labeled data makes the system unusable where either any linguistic expert or phonetically labeled data are not available. Also, extending such a system to include a new language would be a nontrivial task. In [44, 45], researchers studied the effectiveness of pre-classification module in distinguishing world's distinct languages. However, no work has so far studied the usefulness of such a system in distinguishing closely related Indian languages. Also, in [44, 45] feature parameters are first extracted from each of the voiced segments constituting an utterance, and then feature representation of that utterance is estimated. However, the literature confirms that for tonal languages, the tonal events are aligned with segmental events [5]. The peak and valley of pitch contour are aligned to the onset and offset of a segment [46], and therefore, pitch can be utilized to segment the continuous speech into smaller analysis units, which closely correspond to syllable-like units [24]. Accordingly, either open or sonorant closed syllables can be considered as tone bearing units of tonal languages [51]. Thus, for tonal/non-tonal classification of languages, syllable-level analysis may lead to more discriminative feature representation. Besides, the NE Indian languages are known to be syllable centric [39], that is, the language-specific cues are more evident at syllable level itself.

This paper therefore proposes a syllable-level tonal/non-tonal pre-classification based LID system for NE Indian languages that may not depend on the use of phonetic engine.

Attributes like pitch, duration, and energy render the naturalness of speech collectively called prosody, are less affected by noise. Prosodic features cannot be derived from the phoneme structure of human utterance and is also very difficult to replicate. Even in speech recognition, human being makes use of prosodic information also to discern the distinctness in the perceived sounds [33]. Also, in several LID task [24, 32], prosodic features have been used as a complementary information with vocal tract information. Literature reveals that a vast population (almost half) of world languages is tonal [6, 13, 22]. For tonal languages, pitch is an important phonological cue, and it changes in a regular manner within a tone bearing unit. Moreover, tone has an effective correlation with other prosodic features like energy profile and duration [31]. However, the parameters of prosodic features proposed in [24, 32] are not sufficient for tonal and non-tonal language discrimination task. Effective parameterization of prosody can prove to be a viable way to improve the performance of the pre-classification system even though prosody-only based LID system is still far from the state-of the art cepstral feature-based LID system.

On the other hand, spectral features, namely MFCC, persist as a de facto feature for any language identification system. It has also been identified to be quite useful for carrying tone information [21, 37]. Also in two-stage language identification system, MFCC persisted as the most useful features probably due to their admissible performance. It has been proven to be quite useful for identification of Indian languages [42]. In [17], also, Jothilakshmi et al. reported a hierarchical LID system for nine Indian languages using MFCC, MFCC along with delta and double delta coefficients (Δ and $\Delta - \Delta$), and shifted delta coefficient (SDC) features, and noticed that GMM–UBM model with MFCC along with Δ and $\Delta - \Delta$ features provides the highest accuracy among the other features. They also noticed the usefulness of two-level LID system for identifying the languages from same origin. Also, in [2], authors used MFCC and SDC features to identify four under-resourced and closely related South-Asian languages. They reported a good accuracy in identifying the languages from 3 s test utterance. In another approach, Yin et al. [50] proposed a hierarchical LID system (HLID) where a tree structure is followed to identify languages with higher accuracy. Here, instead of using a two-level identification system, a test utterance is classified level by level, depending on the most distinguishing information at each level. Also, they showed that because of hierarchy, system performance improves upon not only baseline system, but also likelihood score fusion-based system. However, the authors did not study the impact of hierarchy-based approach for identification of closely related NE Indian languages. Moreover, in [2, 17, 50], researchers extract MFCC features from the frames constituting an utterance. This type of representation may not be the most suitable way to represent the tonal characteristics that generally lie at the syllable level. In order to overcome these difficulties, this work proposes a syllabic-level representation of MFCC features by fitting individual coefficient of the MFCC vectors across all the frames of a syllable using Legendre polynomial. The existing literatures [2, 17, 21, 37, 42, 50] did not study the effectiveness of MFCC features for tonal/non-tonal language classification and also, pre-classification-based LID task. This paper also

studies the complementarity of MFCC features with prosody at syllable level for pre-classification and proposes the use of MFCC with their Δ and $\Delta - \Delta$ coefficients in combination with prosodic features for pre-classification based LID task.

The systems [25, 32] process the language-specific information lying at different levels (sub-segmental, segmental, and supra-segmental) using individual models and then combine the scores to generate the final decision. In contrast, the proposed syllabic-level MFCC representation enables us to explore feature-level combination of spectral and prosodic features.

As per literature study, both generative and discriminative modeling approaches have been used for LID task [23, 24, 32]. In this study, ANN, a discriminative classifier; GMM–UBM, a generative one, and *i*-vector based SVM, which exploits the goodness of both the approaches, have been explored. In [23], researchers reported a system where they divided the whole utterance into fixed length segments, and then *i*-vector corresponding to that utterance is obtained from spectral and prosodic features extracted from the segments constituting that utterance. This approach to segmentation does not consider the actual syllable boundaries and may lead to inaccurate representation of acoustic events within a segment. As tonal events are prominently characterized at syllable level [28], features should preferably be extracted from syllable or syllable-like units. This work thus explores using syllable-level framework with all the three classifiers, namely ANN, GMM–UBM, and *i*-vector-based SVM.

In this paper, focus has been laid on closely related languages of NE India. The ethnic mix of this region affects the languages that they share to communicate among each other. The language diversity is one of the interesting phenomena in NE states of India. The influence of one language on other as well the languages of bordering countries is very high in NE India, and therefore, distinguishing among these languages with a higher accuracy is difficult as compared to other distinct languages. In India, available language resource hardly includes the NE Indian languages. This necessitates the preparation of a database including NE Indian languages to be used for building a good LID system, and this is quite a challenging task.

The contributions of this paper are as follows:

- An automatic tonal/non-tonal language pre-classification based LID system has been proposed for closely related NE Indian languages without using any phonetic information.
- A more effective way of parameterization of prosodic features has been proposed so that it helps boost the performance at pre-classification stage as well as individual language identification stage. Syllable-level representation of MFCC features using Legendre coefficients has been proposed. Complementarity of MFCC and prosodic features extracted at syllable level, and also their combination has been explored for pre-classification based LID task.
- The syllables are known to be the most appropriate tone bearers for tonal languages. This work therefore explores using syllable-level feature representation for tonal/non-tonal pre-classification based LID system.
- NIT Silchar language database (NITS-LD) has been prepared, covering seven NE Indian languages to carry out our experiment. The seven languages are Assamese, Bengali, Indian English, Hindi, Manipuri, Mizo, and Nagamese. The data have

been collected from All India Radio news, and a total of 4 h of data for each of the languages is considered. These languages are closely related, and speakers from the regions are usually multilingual.
- A comparative performance analysis has been done for pre-classification and individual language identification among three different classifiers, namely GMM–UBM, ANN, and *i*-vector based SVM using syllable-level features. Experiments have been carried out for the combination of ANN model in pre-classification stage and each one of the three models in the individual language identification stage to obtain the best possible performance of the system.

The rest of the paper is organized as follows: Section 2 describes the proposed system for language identification. Section 3 discusses about the development of the language identification system to perform the experiments. Experimental results and analysis of the proposed system are given in Sect. 4, and Sect. 5 concludes the work by mentioning the future works.

2 Proposed System for Language Identification

This section describes the workings of the proposed pre-classification-based LID system. It consists of a tonal/non-tonal language pre-classification stage, followed by two parallel modules in the second stage, one for identification of tonal languages and the other for non-tonal languages. To make performance analysis, experiments have been carried out considering three different cases, namely Case I, Case II, and Case III. Case I represents the baseline system, where language identification is done in a way similar to a conventional LID system. Here, in training stage, either a separate model (L_1, L_2, …, L_M) is built for each of the M number of languages, or a single discriminative model, like a neural network, is trained to distinguish among different languages. At testing, identification is done by comparing likelihood scores of the trial utterance with respect to the all different models, or simply based on the decision of the discriminative model. The front-end features considered for this system are prosody + MFCC and its Δ and $\Delta - \Delta$ coefficients.

In Case II and Case III, firstly, languages are pre-classified into tonal and non-tonal categories, and then, individual languages are identified at the next stage. Further, for Case II, irrespective of whether the test trial gets correctly categorized or not at the pre-classification stage, it is processed to the next stage of classification. However, in Case III, only the correctly categorized test trials (separated manually) are processed by the individual language identification stage.

In Case II and Case III, combination of prosody and MFCC is used as front-end feature at pre-classification stage, and the combination of prosody, MFCC, and its Δ and $\Delta - \Delta$ coefficients is used at individual language identification stage. The block diagram representations of Case II and Case III are shown in Fig. 1.

Figure 2 illustrates the two-stage LID system, highlighting the distinct features used at the different stages. Here, L_1 (Assamese), L_2 (Bengali), L_3 (Indian English), L_4 (Nagamese), L_5 (Hindi), L_6 (Mizo), and L_7 (Manipuri) are the languages involved in the experiment. The details of the pre-classification module are in Sect. 2.1.

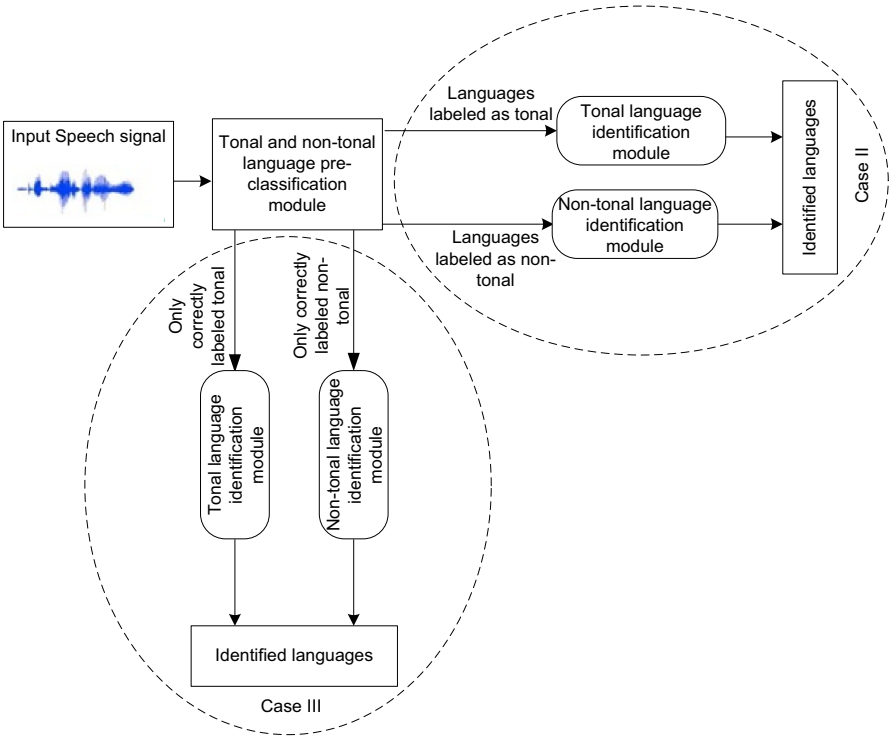

Fig. 1 Block diagram representation of the proposed system

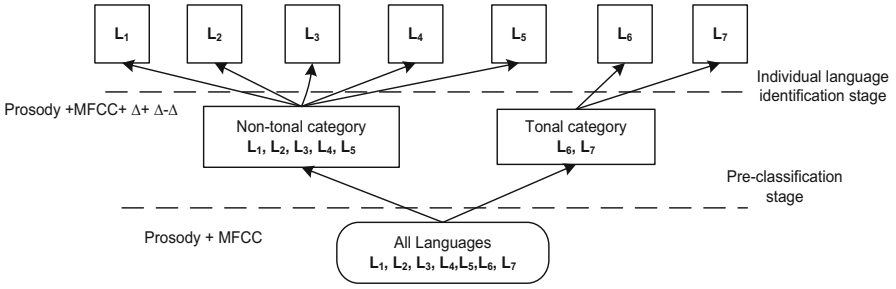

Fig. 2 Use of different features at different stages of the proposed system

2.1 Tonal and Non-tonal Language Pre-classification System

Block diagram representation of the language pre-classification system using prosodic and spectral features extracted from syllables of the speech signal is shown in Fig. 3. Syllables can be treated as context dependent unit, and also it has the ability to capture some co-articulation which is useful for language discrimination [20]. Syllables, in general, follow a common structure like vowel (*V*), vowel consonant (VC), consonant vowel consonant consonant (CVCC), and vowel consonant consonant (VCC). In case of Indian languages, most of the syllables are CV types [18]. Sometimes, the tonal properties can be associated with the onset or/and offset of the syllables [5, 47]. In

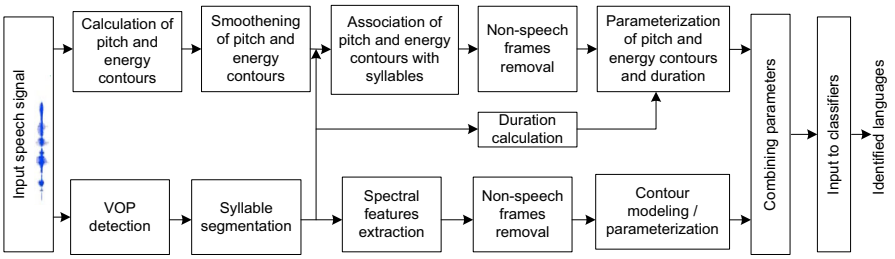

Fig. 3 Block diagram representation of the tonal and non-tonal language pre-classification module

order to get prosodic features corresponding to each syllable-like units, firstly, the pitch and energy contours of whole utterances are obtained. In this study, pitch is calculated through autocorrelation method using robust algorithm for pitch tracking (RAPT) algorithm [43]. It detects the unvoiced frames of an utterance of the speech signal. Energy values calculated from each 10 ms frame of an utterance constitute the energy contour of that utterance. The contours are then smoothened using fifth-order median filter, after which the identified vowel onset points (VOPs) [30] are associated with the smoothened pitch and energy contours. The pitch and energy contours between every consecutive VOPs are obtained and then parameterized to obtain the feature vectors. However, the contours whose lengths are less than 50 ms are not considered.

Here, duration of each syllable has been calculated by considering the number frames between two consecutive VOPs. Duration is then parameterized by rhythm parameter and is taken as another feature. Spectral features are then extracted from the overlapping frames of each syllable. The feature vectors for all the frames corresponding to a syllable are stacked together. In the next step, voiced/unvoiced algorithm is used to identify the frames where speech is present and features corresponding to only the voiced frames are retained. In this experiment, contours corresponding to each dimension of the spectral features for a syllable are parameterized. Parameters of prosodic and spectral information, thereby obtained, are then concatenated to form the final feature vector of a syllable. These combined feature vectors are then fed into the classifiers.

3 Development of Language identification System

In this work, a pre-classification based language identification system has been proposed for Northeast Indian languages. Generally, a language identification system consists of two important components: feature set and classifiers. This section describes the features and classifiers considered in this work.

3.1 Extraction and parameterization of different features for language identification

Here, pitch contour, energy contour, duration of the speech segments, and MFCC are used as front-end features for the two-stage LID task. Different parameters of these features are discussed in this section.

3.1.1 Parameterization of Prosody for Language Identification

Existing parameters for tonal and non-tonal language classification:

The existing parameters of prosody, such as, A_1: mean pitch [44], A_2: pitch changing speed [44], A_3: pitch changing level [44] are calculated from each syllables of the speech signals. For utterance-level analysis, these parameters are obtained using the same process as discussed in paper [44].

Proposed prosody parameters for tonal/non-tonal pre-classification based LID system

- Parameterization of pitch contour

In this work, following parameters are used to parameterize the pitch contour.

A_4: Amplitude tilt for pitch contour (FA_t).
A_5: Duration tilt for pitch contour (FD_t).

Level tones, namely the high (H) or low (L) tones, and contour tones, such as rise, fall, fall–rise or rise–fall tones, dictate the lexical meaning in case of tonal languages. However, in case of non-tonal languages, the lexical meaning does not change with change of pitch contours. Besides, the different tonal languages have their own fixed set of tones. For example, Mizo language is known to have four tones, Manipuri has two tones, Mandarin has four tones, and Vietnamese has six tones. These contours can therefore help characterize the different languages. To represent the dynamics of these contours, generally, amplitude tilt (A_5) and duration tilt ($A6$) parameters are used [1]. These quantities are defined as:

$$FA_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|} \tag{1}$$

$$FD_t = \frac{|D_r| - |D_f|}{|D_r| + |D_f|} \tag{2}$$

where A_r and A_f are the rise and fall and fall of the pitch contour, respectively, with respect to the peak of the contour. Similarly, D_r and D_f are the duration corresponding to rise and fall, respectively.

A_6: Change in pitch (ΔF_0)

Several researchers have investigated the relation of tone height (height of the peak of the pitch contour) with lingual articulation and the jaw's movements and their role in expressing different degrees of emphasis. In case of non-tonal languages, pitch can be freely varied, while, in a tonal language, pitch is phonemically contrastive. Therefore, tone height which will be different for tonal and non-tonal languages and hence can be used as a feature for this system. It is estimated from the difference between the pitch values of peak (F_{0p}) and valley point (F_{0v})

$$\Delta F_0 = F_{0p} - F_{0v} \tag{3}$$

A_7: Distance of peak of pitch contour with respect to VOP (D_r)

Literature shows that [32] the alignment of the peak of the pitch contour has bearing on the perceptual prominence [31]. For non-tonal languages, like English, Greek, the peak is consistently aligned with the onset of the accented syllable, while, for tonal languages, like Mandarin, it is aligned to the offset of the tone bearing syllables. Therefore, peak locations of the pitch contour with respect VOPs (D_r) may help characterize different languages.

A_8: Distance of 60% of the peak value of the pitch contour with respect to VOP

Researchers observed significant effect of place features of consonants and the manner of articulation of consonants on the tonal onset for languages, like Dimasa and Mizo, of the Tibeto-Burman family [38]. And the effect permeates to a great extent into the contour of the following tone. In another study, it has been noticed that the tonal onset can shift due to the interaction between tones and segments (syllables) [28]. This characteristic behavior of languages can help in distinguishing one language from another. It has been experimentally found that the extent to which these effects on tonal onset, propagate into the segment can be roughly approximated by the location of 60% of the peak value of the pitch contour. This work therefore proposes to use the distance of the location with 60% of the peak value, with respect to VOP, as a feature for language classification.

- Parameterization of energy contour

Stress has been assumed to be present up to a certain level in all languages. Some syllables are considered as stressed syllables since they are in some scene perceptually more prominent than others. Stress is parasitic; it can be produced by the phonetic correlates of other phenomena, like pitch and duration. In most of the cases, syllables with higher pitch variation and longer duration are considered as the stressed syllables. The way stress arises in the speech signal is vastly language dependent, and it is quantified by the energy parameter. In case of tonal languages especially, where register tones occur, a direct correlation between tone and stress exists [31]. However, for most of the tonal languages, stress is much less obvious [11] On the other hand, for non-tonal languages, like English, stress is obvious. So, stress is yet another language-dependent trait and can be used to complement the pitch contour cues. Stress is calculated from the energy values of all the voiced frames present within a syllable. Six parameters have been used to quantify the stress characteristics in this work.

A_9: Mean energy

Mean energy is calculated by averaging the energy values of the energy contour corresponding to the syllable.

A_{10}: Change in log energy

Using the quantitative measure quantifying stress characteristics, described in [24], log energy has also been considered in this work. Log energy is more akin to human perception of stress variation.

A_{11}: Energy changing speed

As similar to pitch contour, the energy contour of a syllable is also found to characterize languages [35]. Literature study reveals that that there is an interaction between

tone and stress for register tone languages or the languages which contain level tone and contour tone [31] like Standard Chinese language. Also, in case of tonal languages, it is observed that stress does not necessarily coincide with high tone. Hence, like pitch changing speed, energy changing speed may so be used to characterize languages. It is estimated according to the equation:

$$\mathrm{EV}_j = \sum_{i=1}^{N-1} |E_{i+1} - E_i| \tag{4}$$

Here j represents the index of each segment; N represents the number of frames present in the segment, and $E_1, E_2, \ldots, E_N$ represent the energy values of each frame within a segment. The normalized energy changing speed is given by:

$$\mathrm{EV}_j = \frac{\mathrm{EV}_j}{\text{mean energy} \times \text{number of voiced frames}} \tag{5}$$

A_{12}: Energy changing level

The energy changing speed is a local parameter of energy contour and does not provide the gross level change across a syllable. Therefore, to model the global nature of speed change, another parameter called energy changing level is introduced and is given by

$$\left(\breve{\sigma}_{\mathrm{e}}\right)_j = \frac{\sigma_{\mathrm{e}_j}}{\text{mean energy} \times \text{number of vowels}} \tag{6}$$

where σ_{e_j} is the standard deviation of jth segment (syllables), $\left(\breve{\sigma}_{\mathrm{e}}\right)_j$ is the normalized energy changing level. Energy changing level is a global parameter, and it can be used to discriminate tonal languages from non-tonal.

A_{13}: Amplitude tilt for energy contour (EA_{t})

The dynamics of the energy contour are usually defined using tilt parameters. Amplitude tilt for energy contour is calculated as follows:

$$EA_{\mathrm{t}} = \frac{|A_{\mathrm{er}}| - |A_{\mathrm{ef}}|}{|A_{\mathrm{er}}| + |Ae_{\mathrm{f}}|} \tag{7}$$

Here A_{er}, A_{ef} are the rise and fall point of the energy contour, respectively, with respect to the peak value of the contour .

A_{14}: Duration tilt for energy contour (ED_{t})

Duration tilt can also be used for quantitative representation of the energy contour dynamics. It can be expressed as per the equation:

$$ED_{\mathrm{t}} = \frac{|De_{\mathrm{r}}| - |D_{\mathrm{ef}}|}{|De_{\mathrm{r}}| + |De_{\mathrm{f}}|} \tag{8}$$

Here D_{er} and D_{ef} are the duration for rise and fall, respectively.

A_{15}: Distance of peak of energy contour with respect to VOP

As similar to the case of pitch, the parameter defined by the distance of the peak of the energy contour with respect to VOP is also used, in this work, as a possible cue for language classification. It may be reasoned that such a parameter may be useful, given that the stress and pitch are correlated in case of certain languages and uncorrelated in other cases.

A_{16}: Distance of 60% of the peak value of the energy contour with respect to VOP

Going by the same reasoning of language-dependent correlation between stress and pitch, yet another parameter is introduced, that is defined in a way, similar to A_9 parameter of pitch contour. Here, the distance of the location with 60% of the peak value of the energy contour with respect to the VOP location is calculated.

- Parameterization of duration

In this work, two parameters have been used to parameterize the duration characteristics

A_{17}: Syllable duration

Tone is the phonologically contrastive use of pitch within a segment or a syllable. Tonal contrasts are realized not only by differences in pitch contour, but also by systematic differences in duration [19]. From studies, it can be observed that dynamic tones tend to be confined to phonetically long sonorous segments [3]. Also, the vowels on low tones are longer than those on high tones, and on the contrary, vowels on rising tones are longer than those on falling tones [14]. Thus syllable duration has characteristic information about tonal languages and can therefore be used as a discriminating cue for tonal and non-tonal language classification task. In this experiment, syllable duration is calculated by counting the total number of frames present in a syllable.

A_{18}: Ratio of voiced region duration to total segment duration (Rhythm)

Here, rhythm of each syllable is represented as the ratio of voiced region duration within a syllable to the total syllable duration. This is approximated by the ratio of duration of vowels to the duration between two consecutive vowels are used as rhythm. Mean number of vowel qualities [22] are different for tonal and non-tonal languages. Vowels can be classified as high/low or close/open and duration of each type of vowels will be different for tonal and non-tonal languages. Hence, rhythm can be used as a distinguishing parameter for this system.

A_{19}: Vowel counts

Number of vowel inventories of tonal languages is significantly different from non-tonal, and hence, this can be used as an important parameter in this classification task [22]. Vowel counts are obtained by counting the number of VOPs present in the analysis units (utterance). For a syllable, vowel count would be always equal to 1. Therefore, use of this parameter is insignificant for a syllable. VOPs can be obtained for a spontaneous speech signal by using VOP detection algorithm of [30].

Though some the above-mentioned feature parameters (A_4, A_5, A_6, A_7, and A_{10}) have previously been used in tasks of language identification [24], effect of these parameters for discriminating tonal/non-tonal languages has not been studied so far. This work analyzes the effect of these parameters for tonal and non-tonal language pre-classification-based LID task.

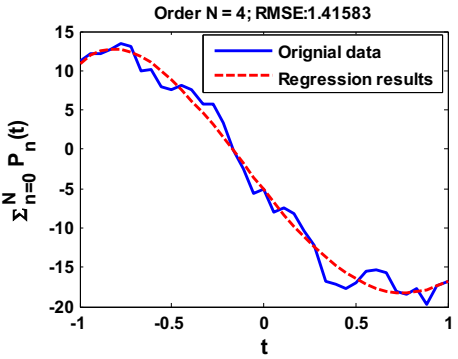

Fig. 4 Fitting of Legendre polynomial to the first coefficient of MFCC of a syllable

3.1.2 Contour Modeling/Parameterization of Spectral Features

MFCC features represent human auditory perceptions for languages and are a predominantly used feature for LID task. MFCC features are known to represent the vocal tract information. Researchers have observed that the vocal tract changes associated with different tones of languages, like Mandarin and Vietnamese, have strong correlation with MFCCs [12]. Literature study [48] also suggests that proper recognition of tones depends not only upon the tone production process but also on the human perception ability. And as MFCCs model the human auditory perception, it serves as a suitable feature for the system. Besides, MFCCs have complementary information with respect to pitch [21], which is known to be a robust feature for language identification.

Extraction of MFCC features is done using standard algorithm which is explained in [40]. In addition to MFCC features, Δ and $\Delta - \Delta$ coefficients are also explored. In this experiment, the feature vectors for all the frames of a syllable are stacked together, and the contour corresponding to each cepstral coefficient is modeled as a linear combination of Legendre polynomials according to Eq. (9)

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \tag{9}$$

where $f(t)$ is the contour being modeled, $P_i(t)$ is the ith Legendre polynomial and coefficient a_i represents a characteristic of the contour shape [23]; a_0 corresponds to the mean, a_1 to the slope, a_2 to the curvature, and higher-order represents more precise detail of the contour. Here, Legendre polynomials of order four lead to 35-dimensional MFCC feature and 105-dimensional MFCC $+ \Delta + \Delta - \Delta$ features for a syllable. In this study, 35-dimensional MFCC features are used at pre-classification stage and 105-dimensional MFCC $+ \Delta + \Delta - \Delta$ are used at individual language identification stage. Fitting of Legendre polynomial to the 1st coefficient of MFCC is shown in Fig. 4. In the present study, the parameters representing prosody and MFCC are concatenated in a row to obtain the combined feature vector of a syllable and also other analysis units.

Table 1 Dimensions of different feature vectors extracted of different analysis units

Analysis units	Prosody	MFCC	MFCC + Δ + $\Delta - \Delta$	Prosody + MFCC	Prosody + MFCC + Δ + $\Delta - \Delta$
Utterance	5 ($A_1, A_2, A_3, A_{17}, A_{19}$)	35	–	40	–
Syllable	18 (A_1–A_{18})	35	105	53	123
Disyllable	36 (concatenating the prosodic parameters of two consecutive syllables)	70	–	106	–
Word	54 (concatenating the prosodic parameters of succeeding and preceding syllables along with the present syllable)	105	–	159	–

Here, each of the parameters represents a dimension of the feature vector. Dimensions of different feature vectors are shown in Table 1.

3.2 Database Used in Language Identification

OGI-MLTS (OGI-MLTS) speech [26] corpus contains spontaneous and fixed-vocabulary utterances of 11 languages: Hindi, Farsi, French, English, German, Korean, Japanese, Spanish, Mandarin Chinese, Tamil, and Vietnamese. Japanese language has not been used in this experiment. The utterances were spoken by individual speakers of each language over telephone line, and the speech was sampled at 8 kHz. This set includes two tonal languages (Mandarin and Vietnamese) and nine non-tonal languages. In this experiment, 10 languages (except Japanese language) have been used to evaluate the system performance. Since the OGI-MLTS database includes only two Indian languages (Hindi and Tamil), NITS-LD has been prepared to study identification of Indian languages. Table 2 shows the details of the NITS database. In this database, two languages (Manipuri and Mizo) are tonal and the rest five are non-tonal. The data were collected from AIR news archives. The speakers of AIR news channels are highly professional and matured. Hence the speech samples collected from this news archives are well articulated and are standard in terms of pronunciation and speaking rate.

The database collected from AIR news archives has some inherent problems, like (i) the number of speakers for individual languages is less, and especially for some languages like Nagamese, the number is noticeably small; (ii) there could be instances of overlapping speech samples from different speakers; and (iii) news headlines may have

Table 2 Different features of OGI-MLTS database and NITS-LD

Characteristics	OGI-MLTS database	NITS-LD
Number of languages	11	7
Channel characteristics	Different	Similar
Channel conditions	Noisy	Non noisy
Types of speech	spontaneous	Scripted
Recording environment of data	realistic	Studio
Sampling rate	8 kHz	8 kHz
Speakers per language	90 speakers for each language present in the database	Assamese (As)-50, Bengali (Be)-40, Indian English (En)-21, Hindi (Hi)-20, Manipuri (Ma)-13, Mizo (Mi)-8 and Nagamese (Na)-6

back ground music. Hence, proper care has been taken while preparing the database. Comparison between OGI-MLTS database and NITS-LD is shown in Table 2.

3.3 Feature Modeling for Language Identification

Feature modeling has been done using different approaches: generative, discriminative, and their combination. Particularly, GMM–UBM [34, 36, 41], ANN [10, 49], and *i*-vector based SVM [7, 8] have been used in this study. Here, *i*-vectors of our training data are normalized by within-class covariance normalization (WCCN) [16] technique to generalize the linear kernel of SVM classifier.

3.4 Data Normalization

Features extracted from different utterances of different speakers need to be normalized to avoid speaker variation, channel variation, etc. In this study, the data are normalized through z-normalization [27] for GMM–UBM and *i*-vector based SVM classifiers. In case of ANN classifier, each parameter of the feature vectors of the training and testing data is normalized to the range of -1 to $+1$.

4 Experimental Results

4.1 Experimental Setup

All the experiments described in this paper have been performed on NITS-LD and OGI-MLTS databases. Training data in case of NITS-LD consists of speech from seven languages, each having 2–3 h of data. In all, there are 14 h of data—8 h data from non-tonal languages and 6-h data from tonal languages. OGI-MLTS database's training set is constituted by 10 h of data—6 h data from non-tonal languages and 4 h

data from tonal category. A syllabic-level approach has been adopted in this work. Fourteen hours of NITS-LD training data amounts to 134,400 syllables, while 10 h of OGI database results in 95,760 syllables. Since the performance of the language identification system varies greatly with change in the duration of the test utterances, the performance has been analyzed for utterances of three different durations, namely 30 s, 10 s, and 3 s. Here, 200 test utterances (100 utterances for tonal language and 100 utterances for non-tonal language) from both the databases are used to analyze the performance of the system. For all of the experiments presented in this paper, training and testing data have been kept mutually exclusive.

The UBM for GMM–UBM and *i*-vector based SVM systems is built using data, in part from NITS-LD and OGI-MLTS databases. Data from 17 languages, each of 1-h duration, are used for this purpose. These data are non-overlapping with either the training or the testing data. An utterance constituted of different number of syllables. Given a speech utterance with N syllables, *i*-vectors are computed with a context size of L syllables. The Baum–Welch statistics have been computed on the sequence of syllables, starting from $Q-L$ to $Q+L$ for obtaining *i*-vector for the Qth syllable. The corresponding sequence of *i*-vectors may be denoted by$w = [w_1, w_2, \ldots, w_N]$. The *i*-vector extractor follows the total variability space model, which is given by [9]

$$s = m + Tw \tag{10}$$

where s is a supervector obtained for the speech segment with respect to UBM. m is the mean value of supervectors, T stands for the total variability subspace, and w is the compact form *i*-vector representation. A context size of $L = 3$ syllables, with a sliding window of 7 syllables has been used with a shift step of 1 syllable. As *i*-vectors are extracted for short sequence of feature vectors, the total variability subspace too is trained on similar short segments.

Figure 5 illustrates the pre-classification based LID system framework. It has a pre-classification stage for tonal/non-tonal classification based on score comparison. It is followed by the second stage, where individual language identification is done by finding the top-scored language.

Four key aspects are addressed in this paper by conducting systematic experiments. Performance analysis of the system for different features has been carried out to study their discriminative power. Also the effectiveness of the proposed parameters of prosody for this system has been studied. Here, existing parameters of prosody are denoted by F_1, existing + proposed parameters of prosody are denoted by F_2, MFCCs are denoted by F_3, existing + proposed parameters of prosody + MFCCs by F_4, and existing + proposed parameters of prosody + MFCC + Δ + $\Delta - \Delta$ by F_5. Performance analysis of different models has been done to find the most suitable model for pre-classification stage and also to the most suitable combination of two different models for this pre-classification-based language identification system. Experimental analysis of different features extracted of different analysis units of the speech sample have been performed to analyze the impact of considering syllables as basic units. Several experiments have been carried out to show the importance of pre-classification stage in LID system.

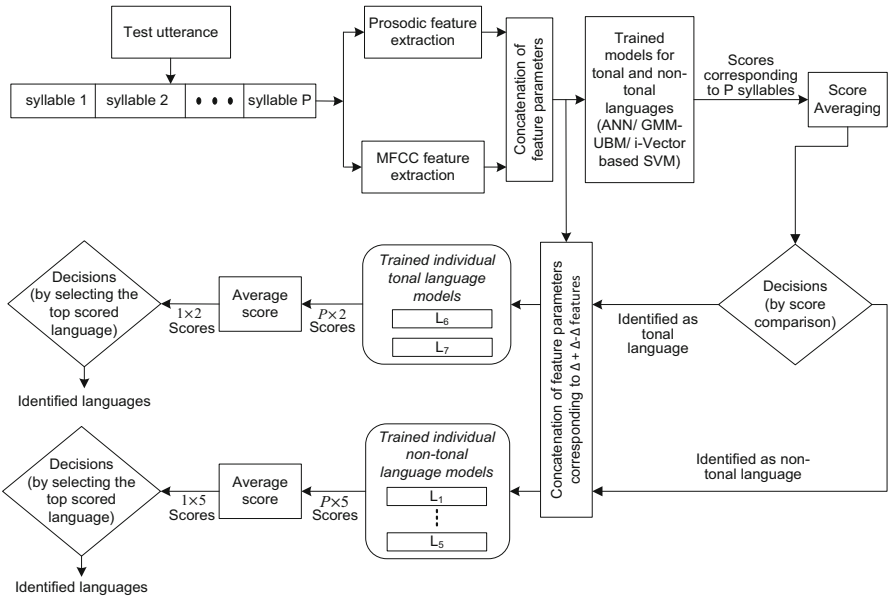

Fig. 5 Pre-classification-based LID system framework showing all intermediate stages of processing

4.2 Experimental Results of Pre-classification Stage

4.2.1 Syllable-Level Performance

In this section, performances of all the three classifiers are analyzed using F_1, F_2, F_3, and F_4 features. The features are extracted from syllabic units of the speech signal. In case of GMM–UBM modeling technique, performance of a system is greatly affected by the change in the number of Gaussian mixtures. Therefore, experiment has been performed for different Gaussian mixtures (2, 4, 8, 16, 32, 64, and 128), and it is observed that for F_1, 2; for F_2, 32; for F_3, 128; and for F_4, 256 result in the highest individual accuracies. Here, the likelihood scores obtained for all of the syllables of a test utterance are averaged to compute the score for that utterance. The decision is taken in favor of the top-scored language. The final accuracy of a language is calculated in terms of the percentage of correctly identified trials.

Similarly, performance of any system using ANN model varies with the different network structures of ANN. Therefore, several experiments have been carried out with different network structures, and it is observed that 5L-8N-1L for F_1, 18L-29N-8N-1L for F_2, 35L-50N-12N-1L for F_3 and 53L-82N-35N-1L for F_4 provide the highest individual accuracies for NITS-LD. Here maximum number of epochs has been set at 500. Tan-sigmoid transfer function is used in the present study. In case of ANN, again, the output scores obtained from neural network for all the syllables of a test utterance are averaged to obtain the score for that utterance.

When modeling of the system is done using *i*-vector based SVM, performance of that system depends upon two major parameters: number of Gaussian mixtures

Table 3 Accuracies of different languages of NITS-LD at pre-classification stage for GMM–UBM classifier

Features	Test data (s)	Accuracy of different languages at pre-classification stage (%)							
		As	Be	En	Na	Hi	Mi	Ma	Average accuracy $= \frac{\sum_{j=1}^{N} Lj}{N}$
F_1	30	53.6	70.2	49	64.1	93.2	57	55.8	64
	10	47	63.8	41.2	59	86	50.2	48	57.2
	3	40.2	57	35.8	51	80.7	46	42.1	51
F_2	30	56.5	76	58.1	66.67	94	60	60	67
	10	48.6	70.7	49.8	62.8	84.6	57.6	53.1	61
	3	47	68	46.4	61.7	73	55.8	51.2	57.5
F_3	30	77	88.8	67.1	76.6	90.8	70.8	65.4	77
	10	72.3	85.8	65.1	72.6	86.8	69.8	64.4	73.9
	3	66.8	82.5	62.7	64.8	81	66.7	61.6	69.8
F_4	30	77.8	89.4	68.2	78.1	96.3	71	66.2	78.9
	10	74.6	88.8	67.8	76.9	89.9	69.9	65.8	76.3
	3	68	83.4	63.8	71	83	67.7	60.4	71

and total variability (TV) matrix dimension. Therefore, a comparative performance analysis has been made using different values for these two parameters. Experimental results show for F_1, 16 Gaussian mixtures, 100-dimensional TV matrix, and linear kernel of SVM; for F_2, 32 Gaussian mixtures, 100-dimensional TV matrix, and linear kernel of SVM; for F_3, 128 Gaussian mixtures, 100-dimensional TV matrix, and linear kernel of SVM; for F_4, 256 Gaussian mixtures, 200-dimensional TV matrix, and linear kernel of SVM, provide the highest individual accuracies.

Here, scores of SVM model is transformed into posterior probabilities using optimal sigmoid transformation, and then scores of all the syllables constituting a test utterance are averaged to obtain the score of that utterance. Accuracies of GMM–UBM, ANN, and *i*-vector based SVM model for the pre-classification task are given in Tables 3, 4, and 5, respectively. Average accuracies corresponding to each feature are calculated by using the formula which is given in Table 3. Here N is the total number of languages present in the individual database (NITS-LD, $N = 7$; OGI-MLTS, $N = 10$) and L_j, accuracies of individual language. Some notable observations made from Tables 3, 4, and 5 are given below:

- Proposed parameters of prosody show the improvements over the existing parameters for all the three classifiers. The improvements are: 3%, 3.8%, and 6.2% for GMM–UBM, 3.4%, 3.7%, and 10% for ANN, and 7.4%, 11.9%, and 9.1% for *i*-vector based SVM, for 30 s, 10 s, and 3 s test data, respectively. Therefore, it may be concluded that better parametric representation of the prosodic characteristics may have helped to improve the performance of the pre-classification module.
- F_3 performs better than F_2.

Table 4 Accuracies of different languages of NITS-LD at pre-classification stage for ANN classifier

Features	Test data (s)	Accuracy of different languages at pre-classification stage (%)							
		As	Be	En	Na	Hi	Mi	Ma	Average accuracy $= \frac{\sum_{j=1}^{N} Lj}{N}$
F_1	30	55.8	69.3	90.1	77	86.7	39.3	37.8	65.8
	10	50.7	64.8	82	71.8	81	36	34	60
	3	42	56.7	73.9	60.7	71.2	30.8	28.6	52
F_2	30	57.2	75.9	98.9	82.1	90	41	40	69.2
	10	49.4	69.2	92.1	76.5	82.2	37.2	36	63.7
	3	47.5	68.8	88.7	75.6	80.3	35.7	34.6	62
F_3	30	74.2	82.9	95.9	85.5	88.9	56.3	55.2	76.9
	10	72.8	79.8	91.6	82	87.8	55.6	54	75
	3	67.7	75.5	85.8	79.8	82	51.2	51	70.4
F_4	30	77.8	85	96.2	89.6	90.3	60.8	60.3	80
	10	74.4	81.8	93.6	82.8	88.6	57.7	56	76.6
	3	68	77.8	89	80.3	83.6	53.5	52.8	71.1

Table 5 Accuracies of different languages of NITS-LD at pre-classification stage for *i*-vector based SVM classifier

Features	Test data (s)	Accuracy of different languages at pre-classification stage (%)							
		As	Be	En	Na	Hi	Mi	Ma	Average accuracy $= \frac{\sum_{j=1}^{N} Lj}{N}$
F_1	30	48.2	67.1	44.3	64.8	48.1	52.6	73	56.8
	10	42	65.4	40.1	59	44	46	68.2	52.1
	3	37	59.3	59	54.8	39	41.6	62.4	50
F_2	30	56	79.8	54.1	73.6	58.3	60.1	82.5	66.3
	10	54.1	77	50.1	72.6	56.3	58.1	81.5	64.2
	3	50	75	45	66.67	52	45.3	80	59.1
F_3	30	66.7	81.6	59.8	76.4	65.8	66.8	84.6	71.6
	10	55.8	78	50.8	71.5	55.2	59.1	80.5	64.2
	3	54	78.8	49.6	72	54	58.9	81	64
F_4	30	68.6	83.6	61.7	78.6	66.67	68.6	85.8	73.4
	10	56.8	79.2	52.1	73.8	58.8	64.8	82	66.7
	3	55.1	79	50	72.7	54.6	60	82.1	64.8

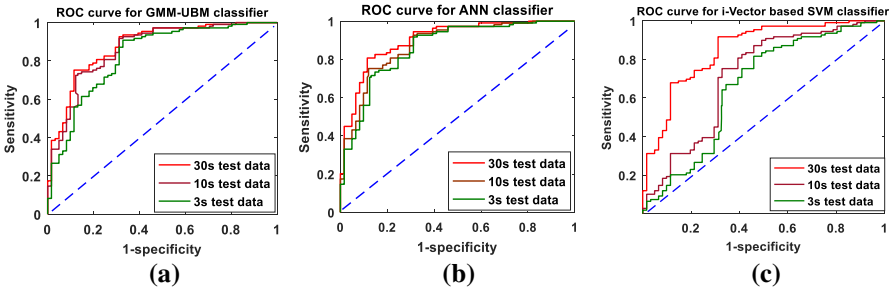

Fig. 6 ROC curves for **a** GMM–UBM, **b** ANN, **c** *i*-vector based SVM classifiers at pre-classification stage

- F_4 provides the highest accuracy.
- In case of GMM–UBM and ANN, performances for the languages of non-tonal category are better than that of the tonal category. However, performance of the languages of tonal category is better than that of non-tonal language category for *i*-vector based SVM classifier.
- 30 s duration test utterance provides the highest accuracy, followed by 10 s and then 3 s. This can be explained as, in a syllable-level implementation, every syllable of an utterance may not give appropriate score for making the right classification decision, due to anomaly that could creep in at any stage of the identification process due to various reasons, like spurious VOP detection, missed VOPs and non-perfect removal of silence frames. When the utterances are short and the syllables are few, the proportional presence of such syllables becomes more influential leading to substantial degradation in the performance. And their influence is much reduced as the number of syllables increase with the increase in the duration of the test utterance.
- ANN outperforms all other classifiers considered in this study. GMM–UBM gives the next best accuracy, followed by *i*-vector based SVM.

Receiver operating characteristics (ROC) curves for three different classifiers using F_4 features are shown in Fig. 6a–c. It represents the (1-specificity) versus sensitivity relation across the range of test trial scores. It can be observed that 30 s utterance score is high in terms of sensitivity and specificity across all the three classifiers. The 10 s utterance results in the next best readings, followed by 3 s utterance. It may also be observed that, in case of *i*-vector model, the effect of test data duration is more prominent, as the plots are farther apart from one another. This may be explained by the fact that, *i*-vectors extracted of shorter speech segments tend to be noisy [29] and hence affect the performance adversely.

Same experiments have been performed on OGI-MLTS database. Accuracies obtained for GMM–UBM, ANN, and *i*-vector based SVM classifiers for OGI-MLTS database using prosody, MFCC, and their combination are shown in Fig. 7a–c. For ANN, it is observed that 18L-29N-10N-1L for F_2, 35L-50N-12N-1L for F_3, and 53L-82N-35N-1L for F_4 features leads to their respective best performances. It can be observed that for proposed parameters of prosody, GMM–UBM shows 4%, 6.5%, 7.1%; ANN shows 3.8%, 2.8%, and 2.8%; *i*-vector based SVM shows 6.4%, 6.6%, and

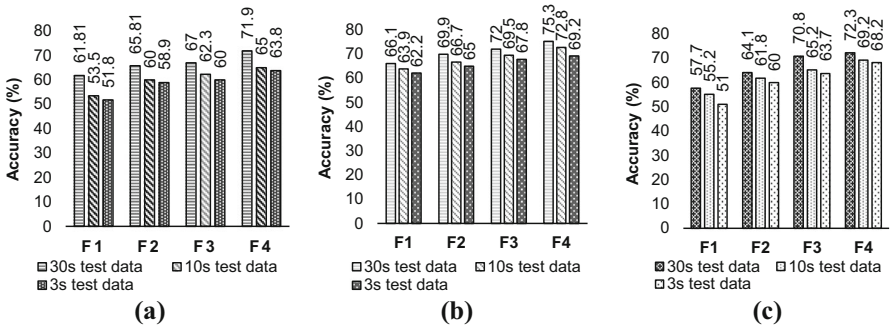

Fig. 7 Performance of **a** GMM–UBM classifier, **b** ANN classifier, **c** *i*-vector based SVM classifier at pre-classification stage for different features obtained from the speech sample of OGI-MLTS database

9% improvements for 30 s, 10 s, and 3 s test data, respectively. Here, ANN performs the best, followed by *i*-vector based SVM classifier and then GMM–UBM classifier. F_4 features and 30 s test data results in the highest accuracies.

The specific observations made from Tables 3, 4, 5, Fig. 7a–c are given:

- Proposed parameters of prosody show significant improvement over the existing parameters for all the classifiers. Improvement is observed for both NITS-LD and OGI-MLTS databases.
- F_4 features provides the highest accuracy for pre-classifying languages of both NITS-LD and OGI-MLTS database.
- ANN model outperforms the other models for both NITS-LD and OGI-MLTS databases. Thus it can be inferred that ANN is able to model more language-specific information than other two models.
- All the classifiers for NITS-LD either perform better for OGI-MLTS database or give comparable performances. This may be attributed to the fact that, the number of target languages are less in NITS-LD as compared to OGI-MLTS database, and the collected speech samples are noise free (whereas in OGI-MLTS database, collected data are noisy).

Another experiment has been performed using F_5 features for pre-classification task. It is observed that the performance improvement obtained in the pre-classification module is not so significant after inclusion of Δ and $\Delta - \Delta$ features, considering the cost of increased dimension of the feature vectors. Therefore, Δ and $\Delta - \Delta$ features have not been considered in the pre-classification module.

4.2.2 Comparison Among the Performances for the Features Extracted of Different Analysis Units of the Speech Signal

An attempt has been made to analyze the performance of the system, built using F_1, F_2, F_3 and F_4, extracted of different analysis units, viz., disyllables, words, or the whole utterance. To obtain the utterance-level performance of the system, features extracted of the whole utterance of the sample are used. Dimensions of the feature vectors obtained from utterance are given in Table 1. Since for 30 s or 10 s duration

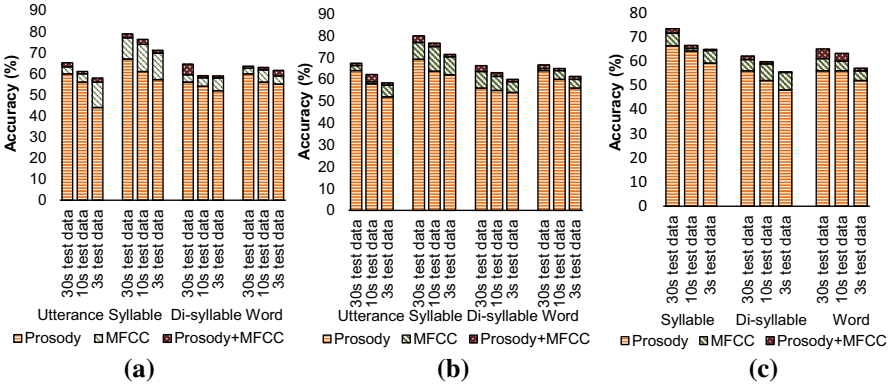

Fig. 8 Comparative analysis of the performances for the features extracted of different analysis units of NITS-LD using **a** GMM–UBM classifier, **b** ANN classifier, **c** *i*-vector based SVM classifier

utterance, calculation of tilt parameter may give erroneous results, and therefore, only the existing parameters (F_1) are explored. Performance of the system is also analyzed for the features extracted of disyllable and word units of the speech samples. The extracted features are then fed into three above-mentioned classifiers. Dimensions of feature vectors corresponding to disyllable and word units are given in Table 1. By utilizing F_1, F_2, F_3, and F_4 of different analysis units, performances of all the three classifiers for NITS-LD are shown in Fig. 8a–c. Table 6 present the classifiers' configurations that resulted in highest accuracies.

Table 7 shows the performances of all the classifiers in pre-classifying the languages of OGI-MLTS database when features are extracted of utterance, syllable, disyllable, and word units. It can be observed that features extracted of syllables are the most useful cues for discriminating tonal and non-tonal languages of both the database. The tones in languages are coded at syllabic level, rather than utterance. Therefore, tone parameters pertaining to individual syllables perform the best. Figure 8a–c and Table 7 also show that the performances of all the classifiers are poorer for disyllables or words as compared to syllables for both the databases.

It may be reasoned that since most of the words of the tonal languages of OGI-MLTS database (Mandarin Chinese and Vietnamese) [15] and of NITS-LD (Manipuri and Mizo) are monosyllabic in nature, the syllable-level features give the best results for all the classifiers. Therefore, syllables are opted as basic units for subsequent experiments. In pre-classification stage, ANN outperforms the other model. Therefore, to perform the two-stage language identification, ANN is adopted for pre-classification stage.

4.3 Experimental Results for Individual Language Identification

In order to show the effectiveness of pre-classification module on language identification, three different approaches have been adopted. In Case I, individual languages are identified without pre-classifying the languages into different categories. In Case II and Case III, pre-classification of languages are done before identifying individual

Table 6 Different parameters of three different classifiers that provides the maximum accuracies for utterance, disyllable, and word-level analysis at the pre-classification stage

Databases used	Classifiers used	Analysis level	Parameters of the three different classifiers (Gaussian mixture numbers/network structure/Gaussian mixture numbers, TV dimensions)		
			F_2	F_3	F_4
NITS-LD and OGI-MLTS database	GMM–UBM	Utterance	2 (for F_1 features only)	8	16 (for $F_1 + F_3$ features)
		Disyllable	16	32	128
		Word	32	64	128
	ANN	Utterance	5L-8N-1L	35L-50N-1L	40L-60N-1L
		Disyllable	36L-50N-1L	70L-105N-30N-1L	106L-150N-50N-1L
		Word	54L-85N-1L	105L-160N-55N-1L	159L-185N-56N-1L
	i-vector-based SVM	Utterance	–	–	–
		Disyllable	32, 200, linear kernel	64, 200, linear kernel	128, 200, linear kernel
		Word	32, 250, linear kernel	128, 200, linear kernel	256, 250, linear kernel

Table 7 Performances of the three classifiers for OGI-MLTS database when features are extracted of different segments

Classifiers	Accuracy of different classifiers for the features extracted of different segments (%)											
	Utterance			Syllable			Disyllable			Word		
	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s
GMM–UBM	62	55.1	54	71.9	65	65.8	69.1	65.8	65	64.8	66.6	63.6
ANN	63.8	57.9	53.2	75.3	72.8	69.2	72.2	61.2	60.8	73.6	68.4	65.2
i-vector-based SVM	–	–	–	72.3	69.2	68.2	69.2	62.1	59	67	61.6	59.8

languages. Scoring and decision making are done in a similar way as in the pre-classification stage.

4.3.1 Individual Language Identification Without Pre-classification

In this case, pre-classification module is absent, and hence individual languages are identified like as in a conventional LID system. Here, in training stage, separate models of each language are trained using the front-end feature vectors. Experimental results obtained for this case are given in Table 8. After analyzing the perfor-

Table 8 Performance of Case I for NITS-LD

Without pre-classification	Accuracy (%)														
	F_1			F_2			F_3			F_4			F_5		
Language identification	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s
GMM–UBM	56.6	50	48.6	65	60	55.7	76	72.7	70.4	76.9	72.6	70.1	80	76.8	72
ANN	61	55.7	52.3	69.8	63.6	62.1	74.8	69.8	65.8	75.9	71.3	69.3	78.1	75	70.2
i-vector-based SVM	52.8	47.1	42.6	62.2	55.1	50	69.6	63.6	60.3	71.2	65.6	62.8	73.6	68.1	64.9

mances of GMM–UBM, ANN, and *i*-vector-based SVM classifiers, it is observed that GMM–UBM outperforms the other models when there is no pre-classification stage. Table 8 also shows the effectiveness of the proposed parameters of prosody on individual language identification. In this case, after using proposed parameters of prosody, GMM–UBM shows 8.4%, 10%, and 7.1%; ANN shows 8.8%, 7.9%, and 9.8%; *i*-vector based SVM shows 9.4%, 8%, and 7.4% improvements for 30 s, 10 s, and 3 s data, respectively, over the existing parameters of prosody. Table 8 also shows the effectiveness of Δ and $\Delta - \Delta$ coefficients for individual language identification task.

4.3.2 Individual Language Identification with Pre-classification

Once the speech samples are classified either as tonal or non-tonal languages, it is fed to the next stage for final language identification. The second-stage identification has been performed in two different ways. In the first way, denoted as Case II, truly detected as well as wrongly detected tonal and non-tonal samples are passed into the next stage for final identification. The second way, denoted as Case III, only truly detected tonal and non-tonal samples are fed into the next stage. The second stage consists of two modules: One classifies individual tonal languages and the other individual non-tonal languages. Several experiments have been performed where ANN is used at pre-classification stage (Case II and Case III) and one of the three classifiers in the second stage. Table 8 shows the accuracy values of Case II and Case III for different features and classifiers. The observations made on Tables 8 and 9 are given:

- The fact that Case III has the best accuracy clearly tells that an accurate pre-classification stage can boost the performance of the system manifold. With the present pre-classification module, as presented in this paper (Case II), the performance of the system is still better than Case I, where no such module is used.
- Here, with respect to Case I, each of the individual features and their combination shows significant performance improvements in Case II and Case III. The combination of ANN model in pre-classification and GMM–UBM model in second stage (Case II and Case III) with F_5 features provides the highest accuracy in pre-classification-based language identification task. This combination shows 3.5%,

Table 9 Accuracy comparison between Case II and Case III for NITS-LD

	Accuracy (%)											
	F_2			F_3			F_4			F_5		
	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s	30 s	10 s	3 s
With pre-classification language identification (Case II)												
GMM–UBM	67.67	64	58.8	77.9	71.9	69.6	80	75.8	71.8	83.5	79	74.1
ANN	69	64.6	62	76.7	73.6	71.4	78	73.4	71.1	80	76.7	73
i-vector-based SVM	63.8	58.1	56	71.8	65.3	63.6	75.8	67.7	64.6	78.3	70	66.7
With pre-classification language identification (Case III)												
GMM–UBM	69	66.8	60.2	78.9	75.9	71.6	84.6	78.9	75	87.2	83.8	76.9
ANN	70.2	66.7	63.6	78.6	76	73.4	83	77.6	73.1	85.1	80.2	74.6
i-vector-based SVM	67.7	66	61.1	76.6	67.7	66.8	80.6	74	72.1	82	74.8	74

Table 10 Confusion matrix for Case I using GMM–UBM

Languages	As	Be	En	Na	Hi	Mi	Ma
As	*16*	1	0	1	0	2	0
Be	0	*18*	0	1	0	0	1
En	0	1	*15*	1	2	1	0
Na	2	0	0	*16*	1	1	0
Hi	1	1	0	1	*17*	0	0
Mi	1	0	0	1	0	*41*	7
Ma	1	4	2	0	0	5	*38*

Italic values indicate true positive trials
Rows correspond to actual class and columns to the assigned class of test data

2.2%, and 2.1% improvements in Case II over Case I, for 30 s, 10 s, and 3 s test data, respectively. Case III reflects the performance of LID system when a 100-percent accurate pre-classification module is available. In this case, the same combination provides 7.2%, 6%, and 4.9% improvements with respect to Case I for 30 s, 10 s, and 3 s test data. Therefore, it can be inferred that this combination possibly captures the most language discriminating cues for this system.

Tables 10 and 11 show the confusion matrices for Case I and Case II. Tables 12 and 13 show the confusion matrices for Case III. Experimental results are given for 30-s duration test data and F_4 feature. From Table 10, it can be observed that even though the performance of Bengali language is good, most of the languages are confused with this and accuracy of Manipuri language is lesser than other languages.

Observations from Tables 12 and 13 are given below:

- Even though the highest accuracy is achieved for Bengali language, all other languages of non-tonal category are confused with it.
- Bengali is least confused with other languages. Except for Hindi (one instance), no other language has been confused with Indian English.

Table 11 Confusion matrix for Case II when ANN is in pre-classification stage and GMM–UBM in second stage

Languages	As	Be	En	Na	Hi	Mi	Ma
As	*18*	3	0	1	0	1	0
Be	1	*13*	0	0	0	0	0
En	1	0	*18*	0	0	0	1
Na	1	0	0	*17*	1	1	0
Hi	1	1	0	2	*16*	1	0
Mi	3	3	0	0	0	*52*	2
Ma	0	2	3	0	0	4	*33*

Italic values indicate true positive trials
Rows correspond to actual class and columns to the assigned class of test data

Table 12 Confusion matrix for non-tonal languages when ANN classifier used in pre-classification stage and GMM–UBM in second stage (Case III)

Languages	As	Be	En	Na	Hi
As	*15*	1	0	1	1
Be	1	*14*	0	0	0
En	1	1	*17*	1	0
Na	1	0	0	*16*	1
Hi	0	1	1	0	*17*

Italic values indicate true positive trials

Table 13 Confusion matrix for tonal languages ANN classifier used in pre-classification stage and GMM–UBM in second stage (Case III)

Languages	Mi	Ma
Mi	*27*	4
Ma	4	*25*

Italic values indicate true positive trials

- Performance of the system in identifying Mizo language is better than that of Manipuri language. This could be possibly because Mizo language has eight tones, whereas Manipuri language has just two tones. The features, as a result, could capture more information with respect to Mizo language than Manipuri.

From Table 10, it can be observed that, in Case I, confusion of each language with others are reasonably high, and Table 14 confirms it, as is implied by the low sensitivity and specificity values. It can be observed from Tables 10, 11, 14, and 15 that pre-classification module helps to reduce the confusion among the languages. However, it may not help boost the performance of all languages necessarily. The error which occurs at the pre-classification stage itself is carried over to the next stage, thereby decreasing the accuracies of certain languages. However, the correctly pre-classified languages are identified with significantly improved accuracy, such that the

Table 14 Sensitivity and specificity (in %) for Case I using GMM–UBM, 30-s duration test data and F_4 feature

Languages	Sensitivity	Specificity
As	80	97.23
Be	90	96.2
En	75	98.9
Na	80	97.3
Hi	85	98.4
Mi	82	95
Ma	76	95.6

Table 15 Sensitivity and specificity (in %) for Case II when ANN is in pre-classification stage and GMM–UBM in second stage, 30-s duration test data and F_4 feature

Languages	Sensitivity	Specificity
As	78.2	96.1
Be	92.8	95.2
En	90	98.4
Na	85	98.4
Hi	76.1	99.95
Mi	86.7	95
Ma	78.5	98.2

Table 16 Sensitivity and specificity (in %) for non-tonal languages when ANN classifier used in pre-classification stage and GMM–UBM in second stage, 30-s duration test data and F_4 feature (Case III)

Languages	Sensitivity	Specificity
As	83.3	97.73
Be	93.3	97.78
En	85	99.93
Na	88.9	98.5
Hi	89.4	98.5

Table 17 Sensitivity and specificity (in %) for tonal languages when ANN classifier used in pre-classification stage and GMM–UBM in second stage, 30-s duration test data and F_4 feature (Case III)

Languages	Sensitivity	Specificity
Mi	87	96.7
Ma	86.2	96.7

overall performance of the system improves. In Case III, there is no possibility to confuse the languages of non-tonal category with the languages of tonal category and therefore, sensitivity, and specificity values for all the languages improve (Tables 16 and 17) significantly.

As a result, Case III reports the highest overall accuracy, followed by Case II and then Case I. So, it may be inferred that improving the accuracy of the pre-classification

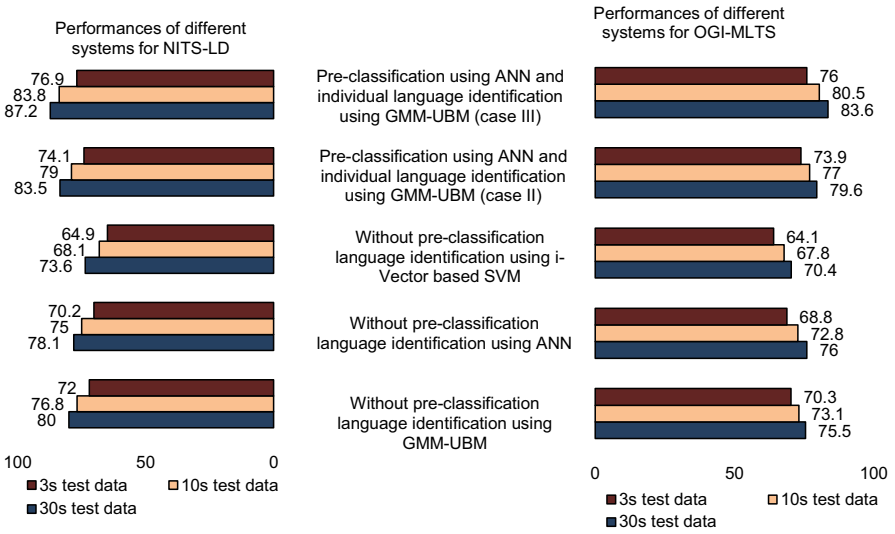

Fig. 9 Comparative analysis between the performances of NITS-LD and OGI-MLTS database

system may help to reduce the confusion and thereby enhance the performance of the individual language identification module.

From Fig. 9, it can be observed that the combination of ANN in first stage and GMM–UBM in second, of Case III, fares the best among all combinations. However, for this combination, accuracy for OGI-MLTS database is lesser than that for NITS-LD. This result reveals two factors: (i) System performance depends largely on the number of target languages. Since the number of target languages in NITS-LD are lesser than that in OGI-MLTS database, this system provides better performance for NITS-LD. (ii) NITS-LD includes well-articulated, noise-free data, and hence, the better performance.

However, from Fig. 9, it can also be observed that because of pre-classification module improvements in the system performance for OGI-MLTS database (8.1%, 7.4%, and 5.7%) are more significant than NITS-LD (7.2%, 7%, and 4.9%). This could be because OGI-MLTS database has been prepared using World's distinct language and NITS-LD includes closely related languages of same origin.

5 Conclusions and Future Scopes

This work proposes a system that provides improved performance over existing language identification systems. The proposed system has a pre-classification stage to distinguish the tonal and non-tonal languages. The performance of the system is analyzed for three different cases, namely Case I: where individual languages are identified without using any pre-classification module; Case II: where individual languages are identified from all detected tonal and non-tonal languages of the pre-classification stage; and Case III: where individual languages are identified from only correctly

detected tonal and non-tonal languages of the pre-classification stage. Also this method eliminates the need of automatic speech recognizer or any phonetic information of the languages. Comparison among the performances for prosody, MFCC with Δ and $\Delta - \Delta$, and their combination at syllabic level has been done. Effectiveness of the proposed parameters of prosody on the pre-classification as well as individual language identification has been examined in this paper. It is observed that for the proposed parameters of prosody, system performance at the pre-classification, and individual languages identification stage improve significantly. This paper also demonstrates that at pre-classification stage MFCC performs better than prosody and their combination leads to further improvement. And MFCC with Δ and $\Delta - \Delta$ proves to be the most effective among the other features used in this experiment for individual language identification task. Seven languages of NITS-LD and 10 languages of OGI-MLTS database have been used for the experiment using features based on different analysis units of the speech signal. It can be inferred from the experiments that syllables are the most appropriate analysis segments to be used for this pre-classification based language identification system. The performance of the proposed system has been analyzed for three different sizes of test data using three different classifiers (GMM–UBM, ANN and *i*-vector based SVM classifiers). Experiment shows that for OGI-MLTS, all the three classifiers perform the identification tasks with slightly lower accuracy compared to NITS-LD. Also, in pre-classification stage, ANN classifier outperforms the other two classifiers for NITS-LD as well as OGI-MLTS. The combination of ANN classifier in pre-classification stage and GMM–UBM classifier in the second stage provides the highest accuracies for both databases.

However, this accuracy might still be not satisfactory for a practical system. Syllables are used here as basic units because of being the most effective tone bearers. However, inaccurate syllable boundaries may cause error in the identification system. As syllables are identified from the locations of VOPs, the accuracy of VOP detection algorithm might affect the system performance. This aspect can be explored in future course of work. Also, the tonal languages used in this experiment are all monosyllabic in nature. In case of monosyllabic languages, most of the syllables bear tone, but for disyllabic or polysyllabic tonal languages, tones are not carried by all the syllables which is why it would demand further processing. In future, an extra module can be added to detect tone bearing syllables to improve the performance of the overall system.

References

1. A.G. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey, Modeling prosodic dynamics for speaker recognition, in *Proceedings*, *IEEE International Conference on Acoustic*, *Speech Signal Process*, vol. 4 (Hong Kong, 2003), pp. 788–791
2. F. Adeeba, S. Hussain, Acoustic feature analysis and discriminative modeling for language identification of closely related South-Asian languages. Circuits System Signal Process. (2017). https://doi.org/10.1007/s00034-017-0724-1
3. C.L. Alan, Tonal effects on perceived vowel duration. Lab. Phonol. **10**(4), 151–168 (2010)
4. E. Ambikairajah, H. Li, L. Wang, B. Yin, V. Sethu, Language identification: a tutorial. IEEE Circuits Syst. Mag. **11**(2), 82–108 (2011)

5. M. Atterer, D.R. Ladd, On the phonetics and phonology of "segmental anchoring" of F_0. J. Phonetics **32**, 177–197 (2004)
6. D. Dan, D. Robert Ladd, Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and microcephalin. PANS (2007). https://doi.org/10.1073/pnas.0610848104
7. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **99**(4), 788–798 (2010)
8. N. Dehak, P. Torres-Carrasquillo, D. Reynolds, R. Dehak, Language recognition via *i*-vectors and dimensionality reduction, in *Interspeech Conference* (Florence, 2011), pp. 857–860
9. S. Dey, P. Motlicek, S. Madikeri, M. Ferras, Template-matching for text-dependent speaker verification. Speech Commun. **88**, 96–105 (2017)
10. M. Dorofki, A.H. Elshafie, O. Gaafar, O.A. Karim, S. Mastura, Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data, in *International Conference on Environment, Energy and Biotechnology* (Singapore, 2012)
11. S. Duanmu, Tone and non-tone languages: an alternative to language typology and parameters. Lang. Linguist. **5**(4), 891–923 (2004)
12. S. Dusan, L. Deng, Recovering vocal tract shapes from MFCC parameters, in *5th International Conference on Spoken Language Processing* (1998)
13. C. Everett, D. Basì, S.G. Roberts, Climate, vocal folds, and tonal languages: connecting the physiological and geographical dots. PNAS **112**(5), 1322–1327 (2016)
14. J. Gandour, Counterfeit tones in the speech of Southern Thai bidialectals. Lingua **41**(2), 125–143 (1977)
15. A. Gelbukh, *Computational Linguistics and Intelligent Text Processing, Part-1* (Springer, Berlin, 2011)
16. A.O. Hatch, S. Kajarekar, A. Stolcke, Within-class covariance normalization for SVM-based speaker recognition, in *Proceeding of the ICSLP* (2006), pp. 1471–1474
17. S. Jothilakshmi, V. Ramalingam, S. Palanivel, A hierarchical language identification system for Indian languages. Digit. Signal Proc. **22**(3), 544–553 (2012)
18. A.N. Khan, S.V. Gangashetty, B. Yegnanarayana, Syllabic properties of three Indian languages: implications for speech recognition and language identification, in *International Conference on Natural Language Processing* (Mysore, 2003), pp. 125–134
19. E. Kidder, Tone, intonation, stress and duration in Navajo. in *En Linguistic Theory at the University of Arizona*, eds. by Mans Hulden y Shannon T. Bischoff (Arizona: University of Arizona Linguistics Circle, 2008), Vol. 16, pp 55–66
20. R.A. Krakow, Physiological organization of syllables: a review. J. Phonetics **27**, 23–54 (1999)
21. P.N. Le, E. Ambikairajah, E.H. Choi, Improvement of vietnamese tone classification using FM and MFCC features, in *International Conference on Computing and Communication Technologies*, (*RIVF'09*) (2009), pp. 1–4
22. I. Maddieson, Tone, in *The World Atlas of Language Structures Online*, ed. by Matthew S. Dryer, Martin Haspelmath (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013)
23. D. Martinez, E.A. Lleida: Ortega and A. Miguel, prosodic features and formant modeling for an *i*-vector based Language recognition system, in *ICASSP* (2013)
24. L. Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition. Speech Commun. **50**(10), 782–796 (2008)
25. L. Mary, Multilevel implicit features for language and speaker recognition. Ph.D. Dissertation (IIT Madras, 2006)
26. Y. Muthusamy, R. Cole, B. Oshika, The OGI multi-language telephone speech corpuses, in *Proceedings of International Conference Spoken Language Processing* (*ICSLP*) (1992), pp. 895–898
27. R.W.M. Ng, T. Lee, C.C. Leung, B. Ma, H. Li, Analysis and selection of prosodic features for language identification, in *Proc. IALP*. (2009), pp. 123–128
28. P. Pittayaporn, Directionality of tone change, in *Proceedings of the 16th International Congress of Phonetic Sciences* (Saarland University, Saarbrücken, 2007), pp. 1421–1424
29. A. Poddar, M. Sahidullah, G. Saha, Improved *i*-vector extraction technique for speaker verification with short utterances. Int. J. Speech Technol. **3**, 1–16 (2017)
30. S.R.M. Prasanna, B.V.S. Reddy, P. Krishnamurthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio Speech Lang. Process. **17**, 556–565 (2009)

31. C. Qu, H. Goad, *The interaction of stress and tone in standard Chinese: experimental findings and theoretical consequences* (Theory and Practice, Max Planck Institute for Evolutionary Anthropology, Tone, 2012)
32. V. Ramu Reddy, S. Maity, K.S. Rao, Identification of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. **16**(4), 489–511 (2013)
33. K.S. Rao, Application of prosody models for developing speech systems in Indian languages. Int. J. Speech Technol. **14**(1), 19–33 (2011)
34. R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**(2), 195–239 (1984)
35. B. Remijsen, The study of tone in languages with a quantity contrast. Language Documentation and Conservation. **8**, 634–651 (2014)
36. D. Reynolds, *Gaussian Mixture Models. Encyclopedia of Biometric Recognition* (Springer, Berlin, 2008)
37. N. Ryant, J. Hong Yuan, M. Liberman, Mandarin tone classification without pitch tracking, in *ICASSP* (2014)
38. P. Sarmah, C.R. Wiltshire, A preliminary acoustic study of Mizo vowels and tones. J. Acoust. Soc. India **37**(3), 121–129 (2010)
39. A.K. Singh, A computational phonetic model for Indian language scripts, in *Constraints on Spelling Changes. Fifth International Workshop on Writing Systems* (Nijmegen, 2006)
40. D. Steven, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**, 357–366 (1980)
41. M.N. Stuttle, A Gaussian mixture model spectral representation for speech recognition. Ph.D. Dissertation (University of Cambridge, 2003)
42. M.J.S. Suresh, S.A. Thorat, Language identification system using MFCC and SDC feature, *Language* (2018)
43. D. Talkin, A robust algorithm for pitch tracking (RAPT), in *Speech Coding and Synthesis*, ed. by W.B. Klein, K.K. Paliwal (Elsevier, New York, 1995)
44. L. Wang, E.E. Ambikairajah, H.C. Choi, Automatic tonal and non-tonal language classification and language identification using prosodic information, in *International Symposium on Chinese Spoken language Processing*. (*ISCSLP*) (2006), pp. 485–496
45. L. Wang, E. Ambikairajah, H.C. Choi Eric, Automatic language recognition with tonal and non-tonal language pre-classification, in *15th European Signal Processing Conference* (2007)
46. Y. Xu, 'Effects of tone and focus on the formation and alignment of F_0 contours. J. Phonetics **27**, 55–105 (1999)
47. Y. Xu, Consistency of tone-syllable alignment across different syllable structures and speaking rates. Phonetica **55**, 179–203 (1998)
48. Y. Xu, Understanding tone from the perspective of production and perception. Lang. Linguist. **5**(4), 757–797 (2004)
49. B. Yegnanarayana, *Artificial Neural Networks* (Prentice-Hall of india Private Limited, New Delhi, 2005)
50. B. Yin, Language identification with language and feature dependency. Ph.D. Dissertation (The University of New South Wales, 2009)
51. J. Zhang, Tones, tonal phonology, and tone sandhi, in *Chinese Linguistics*, ed. by C.-T. James Huang, Y.-H. Audrey Li, A. Simpson (Wiley, Oxford, 2014), pp. 443–464

Affiliations

Chuya China Bhanja[1] · Mohammad Azharuddin Laskar[1] · Rabul Hussain Laskar[1]

Mohammad Azharuddin Laskar
azharlaskar@gmail.com

Rabul Hussain Laskar
rhlaskar@ece.nits.ac.in

[1] Department of Electronics and Communication Engineering, National Institute of Technology Silchar, Silchar, Assam 788010, India