A Comparative evaluation of three automatic language identification approaches for Indian Languages

Sushanta Kabir Dutta [#1], L Joyprakash Singh [#2]

[#]Department of Electronics and Communication Engineering, North Eastern Hill University,
Shillong-793022, Meghalaya, India

[1] sushantatzp@gmail.com
[2] jplairen@gmail.com

***Abstract*— In the present study, three language identification (LID) approaches are compared for a set of three Indian languages. However, the number of languages can be increased. One of the approaches is the Hidden Markov Model (HMM) based Phonetic Engine (PE), another is the Gaussian Mixture Model based Universal Background Model (GMM-UBM) classifier and the remaining one is a prosodic feature based classification system. The PE belongs to the category of Explicit LID techniques while the GMM-UBM system and the prosodic feature based classifier fall into the group of Implicit LID techniques. Basically, Explicit LID techniques require a segmented and phonetically labelled speech corpus, while the Implicit LID techniques do not require any phonetic description of the data. All the systems are used here in identifying a set of test data from three Indian languages, Manipuri, Assamese and Bengali. The selection of these languages is made on the basis of the wide range of usages of them in North Eastern India and at the same time no identification tasks were carried out so far for a database inclusive of these three languages together. The purpose of this comparison is to check the LID efficiency among the three approaches. In the experiments, it is found that the prosodic feature based approach can only be used for a broader classification of languages into two categories such as tonal and non tonal. Such a system can't be used alone to completely identify a language. However, the other two approaches can suitably be used for automatic identification of the languages. The identification rate (IDR) of the PE is superior to the GMM-UBM classifier. The average IDR reported with PE is 99% while for the GMM-UBM classifier it is found to be 96.94% with the same speech corpus being in use. However, the data preparation task is a little more expensive in PE than that of GMM-UBM classifier. Thus, the compensation for accuracy may be paid with the cost incurred in data preparation.**

Keyword- Hidden Markov Model, Phonetic Engine, language identification, Gaussian Mixture Model based Universal Background Model

I. Introduction

Language identification (LID) is becoming a very exciting area of research in today's world [1][2]. This may be due to the globalization of sectors like tourism, agriculture, business etc among the people from different regions using different languages for communication. In some applications like call-centres to pre-short the callers, an automatic LID system is very useful, since it can divert a call from a customer to the call-centre employee who is fluent in his/her native language [3]. Similarly, such a system would be very useful in applications such as automatic dialog systems, travel information retrieval etc. [4][5].

Broadly, there are two categories of LID systems: Explicit LID and Implicit LID [6][7]. The Explicit LID uses a segmented and phonetically labelled speech corpus and therefore provides a good accuracy in language identification applications. However, it is difficult to get a phonetically labelled speech corpus, since labelling is to be done by the experts of the languages. Besides, the accuracy of the system very much relies on the accuracy in phonetic labelling. On the other hand, the complexity of the Implicit LID techniques is less, since they don't require any labelled corpus. Such a system requires only the raw speech data with true identity of the language. But the accuracy of the identification may be low here, as compared to the Explicit techniques.

A Phonetic Engine (PE) is a system that transforms a speech signal to symbolic form [8][9]. These symbols are such that each of them represents a phonetic unit. We have used International Phonetic Alphabet (IPA) (revised in 2005) symbols in the transcription process [10]. Each phonetic unit is later on modelled using a 5-state left to right Hidden Markov Model. The global mean and variance of each of these models have been estimated to get the baseline acoustic model of the PE [11][12]. Now for each target language, the corresponding PE is designed. Next, a set of PEs are put together to build the LID system. Such a system looks similar to a parallel Phone Recognizer (PR) in the first observation, though the use of IPA is not available with the PRs. The use of IPA allows discriminative markings among the phone sets being used by different languages [9][13].

In Gaussian Mixture Model based Universal Background Model (GMM-UBM) approach, a single large GMM is built for the data belonging to all the three languages used. Before this, it is required to estimate the GMM parameters (mean, variance and weight vectors) for the speech data from each individual language [14]. 2000). These are used together in the single large GMM which is the UBM. The Mel-frequency Cepstral coefficients (MFCC) [15][16] are the extracted features from speech samples and the GMM parameters are estimated from the set of these feature vectors.

On the other hand, the prosodic features based approach relies on the use of prosody in languages. To distinguish among languages, the method exploits the variation of pitch information in various samples of the target languages [17]. However, the method can’t alone be used for complete identification of languages.

The rest of the paper is organized as below: section II gives a brief review of language identification cues, section III discusses the speech corpus, section IV presents the overview of PE based LID system, section V presents the overview of GMM-UBM based LID model, and section VI presents the prosodic feature based tonal non-tonal classification system and section VII discusses the results of experiments while section VIII presents the conclusion.

II. A BRIEF REVIEW OF SOME LANGUAGE IDENTIFICATION CUES

A lot of research has been done in the area of language identification research since the year of 1970 to till date. In the initial years the researchers used varied data bases due to which it was difficult to summarize the research findings. However, after the introduction of standards like TIMIT, OGI data-bases etc the research findings became possible to summarize. These details are available in a number of research articles [1][18][19]. In language identification research, the following identification cues [16][20] are widely used in order to distinguish one langue from another.

A. Acoustic cues

It deals with the physical characteristics of the speech signal described by frequency, time and intensity informations [18][19]. These are represented as a sequence of feature vectors where each individual feature vector corresponds to acoustic information for a particular time frame. Acoustic information is one of the most primitive forms of information obtained by speech parameterization. The widely used speech parameterization techniques are linear prediction coefficients (LPC), Mel-frequency Cepstral coefficients (MFCC), and Perceptual Linear Prediction (PLP) and Linear prediction Cepstral coefficients (LPCC).

B. Phonotactic cues

There are various phonological factors governing the distinctiveness of a particular language. Some of these are the phone sets and phonotactic constraints of a language. The word ‘phonotactic’ refers to the rule that govern the different combination of phones or phonemes in a language. Different languages may have different allowable combination of sequence of phoneme combinations [6][13]. For example, Japanese has strict phonotactic constraint prohibiting a consonant following another consonant, while English does not have such rules. Thus phonotactic information may be useful in capturing some dynamic nature of speech signal usually lost in feature extraction [17].

C. Vocabulary cues

Conceptually the most important difference among the languages is the different set of word they use. Therefore, the vocabulary differs. This is an important cue [18][20].

D. Prosodic cues

The stress, intonation and rhythm are all important elements used within the prosodic structure of a spoken utterance. The manner in which these are incorporated varies from language to language [19][21].

There exposed a significant importance of acoustic-phonetic approaches in language identification [7][15]. To get accurate estimation of the information sources some form of detailed modeling is necessary. In PE based language identification systems, continuous density HMMs [12] are used to build language dependent phone recognizers (PRs). An acoustic model is then created by using audio recordings of speech with their corresponding transcriptions which are later complied to get statistical representations of the phone units in the PRs [22][23]. Again, the GMM-UBM method uses a single background model for all languages. Initially, one large GMM is trained with data from all the required languages are considered to represent the characteristics of all of these languages. From the UBM, a separate GMM for each language is derived by Maximum A-Posterior (MAP) adapting the trained GMM-UBM to the acoustic training data of that language. This leads to more robust estimates of language models for those languages with less training data. This model is used to identify the unknown test samples. Besides, the use of prosodic features varies in case of some languages. Based on this, languages can be categories into two different groups such as tonal and non-tonal languages [2]. Thus a classification system can also be built using the prosodic features.

III. THE SPEECH CORPUS

The three languages selected for the close set identification task are Manipuri, Assamese and Bengali. These languages are widely used in the North Eastern part of India. Manipuri is the officially used language in the state of Manipur, while Assamese is the language used in the state of Assam. The Bengali language is officially used in some parts of Assam (Barak valley) and Manipur, and also in the state of Tripura. However, it is the officially used language in the state of West Bengal. The scripts of Assamese and Bengali are closely related, while the Bengali script was also in use with the Manipuri language, until the eighteenth century before adoption of Meitei script. Therefore, the present LID system is built in order to identify these three closely related languages. The spoken data are collected for the three languages for the purpose.

The Zoom H4N recorder had been chosen for recording speech data used in the present work. The recorder has 4-channel simultaneous recording facility with digitally controlled high quality mike fitted with pre-amplifiers for high quality recording. It has a large 1.9 inch LCD screen with user interface for the ease of operation. The sampling frequency of the collected data is set to 16 kHz with 16-bit per sample during different recording sessions. The ‘Read mode’ speech data are recorded from professional news reader in studio. The professional newsreaders of all the three languages (Manipuri, Assamese and Bengali) delivered at least 5 minutes of speech in recording studio (e.g. 5 minutes of the morning 7:30 news). The ‘Lecture mode’, data are collected from experienced teachers from School/College or University. Each speaker delivered speech data for a minimum of 15 minutes. The ‘Conversation mode’ data are recorded from a group of speakers. Here around 3 to 4 the speakers were asked to discuss on a particular topic which was recorded.

Next, the transcription of the entire collected data is done using IPA symbols (Revised in 2005) [6][7][13]. This sequence of symbols is assigned to the sequence of feature vectors which represents the spectral characteristic of a speech segment. This stage is required in development of the PEs for all the languages used [24][25]. Two third of the data are used in the training phase, while the rest of the data are used in the testing process.

IV. PHONETIC ENGINE BASED LID SYSTEM

A Phonetic Engine (PE) performs identification tasks in speech and speaker recognition in the similar manner to that of a phone recognizer [9]. However, the PE uses the IPA symbols [10] containing letters and diacritic marks in describing the phones of the spoken data additionally. The PE can suitably be used in language identification tasks also [25].

The development of PE is described in details in literature [24]. However, when building an LID system with PEs, the acoustic analysis is used as the first step [25]. Acoustic analysis deals with slicing the speech signal into successive frames of 20-25 msec duration each with 10-15 msec overlap, where each frame is later on represented by a set of feature vectors extracted from there [16]. Here the frames with 25 msec duration and 10 msec overlap have been used. The Mel-Frequency Cepstral Coefficient (MFCC) feature vectors [14] are extracted for the purpose. Each frame is then represented by a 39 dimensional feature set containing one energy coefficient, 12 MFCCs, 13 delta and 13 acceleration co-efficients. The delta and acceleration coefficients provide the dynamic information of the signal. Then a 5-state prototype left to right HMM with 32 GMM per state has been used for modeling of each phonetic unit prepared by using International Phonetic Alphabet (IPA). IPA is designed to represent each distinctive sound unit with a one symbol. The symbol may be composed of either one letter or a letter and a diacritic combination [10][7]. Again in the prototype HMM, the first and last states of the model are non-radiating while the remaining three are radiating states [12][26]. The global mean and variance of HMMs per state are calculated using the predefined prototype, in which all means are initialized to zero and variances to unity. The training files are scanned through while calculating the global mean and variance. Once an initial set of models are built, the optimal HMM parameters are re-estimated to get the acoustic model. Here we required six iterations to get the final model. The acoustic model is the backbone of a PE. A total of 30 phonetic units including a silence are used in Manipuri language PE while 34 phonetic units are used in both Assamese and Bengali PEs.

The extracted MFCC feature vectors from a test utterance are then compared with the ‘acoustic model’ to estimate the ‘acoustic likelihood’ scores. For the highest likelihood score (LS) emanating from any language PE for an unknown utterance it is considered that the utterance belongs to that language. This is illustrated in Fig. 1. below. Here the PEs of three different languages, viz L1, L2 and L3 are used to build the LID system.

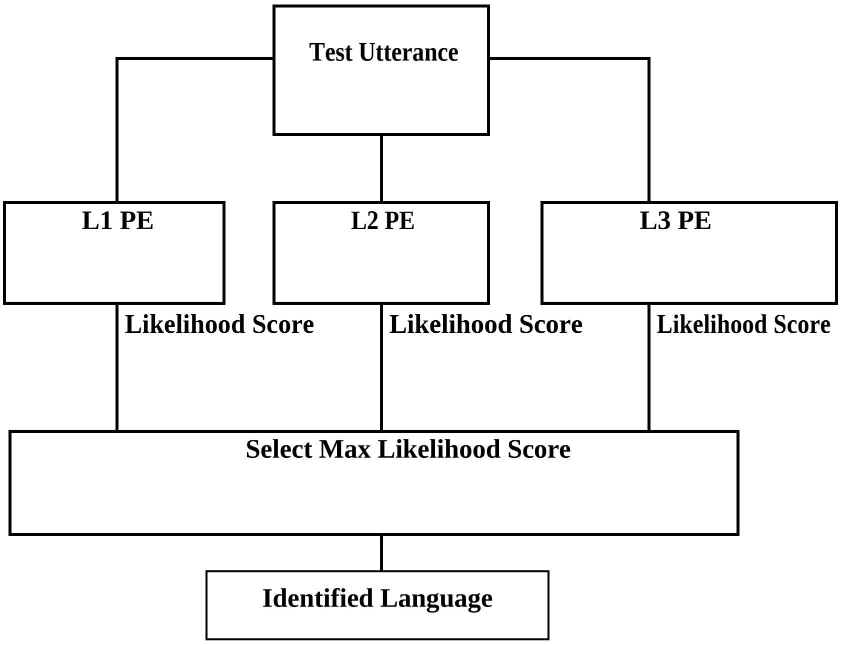

Fig. 1. PE based LID system

With the above set up, we perform the training and testing of the system. Next the performance is analyzed for the same. The formula for evaluating the performance of the PE is mentioned in below equation [17],[25]

$$PA = \frac{N-D-S-I}{N} \times 100\ \% \qquad (1)$$

Where PA is percentage accuracy, N is the number of words in test set, D is the number of deletion, S is the number of substitutions, I is the number of insertions and PA gives phone accuracy rate.

An accuracy of 62.11% is achieved while testing the data from both male and female speakers together. Figure 2 below shows the screenshot of accuracy analysis of the Manipuri PE. The implementation of PE is carried out using HMM modeling tool HTK version 3.3 [26]. The PEs for Assamese and Bengali languages is built in the similar way. The overall accuracies reported for Assamese PE is 43.28% while Bengali PE is 48.58% after analysis similar to the Manipuri PE. Now all these 3 PEs are used in the system to identify an arbitrary test utterance (in any of the languages among Manipuri, Assamese and Bengali) according to the procedure stated above.

The performance of a LID system is determined by the identification rate (IDR) [25]. The unknown test utterance which gets higher ‘Acoustic Likelihood’ score is considered as the identified language. The error rate is calculated by the number of test utterances that give false identification per total test utterances. The lower the error rate, the higher the accuracy of the LID system. For a given language L, the IDR is defined as:

$$IRD = \frac{n}{N} \qquad (2)$$

Where ‘n’ is the number of correctly identified utterances in language L. ‘N’ is the total number of utterances in language L.

V. GMM-UBM BASED LID SYSTEM

The GMM-UBM system of classification is a popular method of pattern recognition in speech processing literature, especially in the text- independent speaker verification process [14]. The method can also be adopted for LID tasks, with minor modifications. Instead of the speaker specific information, here the language specific informations are modeled when used it for LID tasks. The backbone of the present LID system is a UBM which is created by using the GMM parameters of the individual languages. The development the system can be summarized in the following steps.

1. Use the individual GMM parameters of each language to model the UBM, which is a bigger GMM.

2. Next to estimate the model parameters, i.e. Mean, Variance and Weight for UBM from development data set. This data set may be created from a section of the training data.

3. Then using the training data set UBM adaptation is done, i.e. Adapted Mean, Variance and Weight vectors are estimated. This gives a tighter coupling of UMB with the language models.

4. Maximum A-posterior (MAP) algorithm is used for approximating the adapted parameters.

In order to determine the GMM-UBM parameters, we used the Expectation Maximization (EM) algorithm. Afterwards, in the adaptation part we use the MAP algorithm. This is a well-known approach used in estimating the GMM parameters, viz, Mean, Variance and Weight vectors. In practice, in a GMM based LID system, every language has its own GMM. However, the Universal Back-ground Model based GMM system (GMM-UBM) uses a single background model for all languages. For example, in the proposed system, three GMMs had been built for Manipuri, Assamese and Bengali languages. Here, only one large GMM is required for all these languages when put together. This GMM is referred to as UBM and is considered to represent the characteristics of all the languages. From the UBM, a separate GMM for each language is derived by adaptation. The advantages of using such an adaptation over running the EM algorithm [14] to train each language model (used in GMM) separately are as follows: a tighter coupling between the language model and the UBM is achieved. Besides, all language models have the same initialization parameters as that of the UBM. In addition, MAP adaptation [14] combines the robustly estimated UBM parameters with the language model parameters. This leads to more robust estimates of language models for those languages which may have insufficient training data. Finally, training a new language model is faster here, than running the EM again and again on each language. It requires only a few adaptation iterations here.

VI. PROSODIC FEATURE BASED CLASSIFICATION SYSTEM

The details of the components used in the classification system are shown in the Fig. 2.

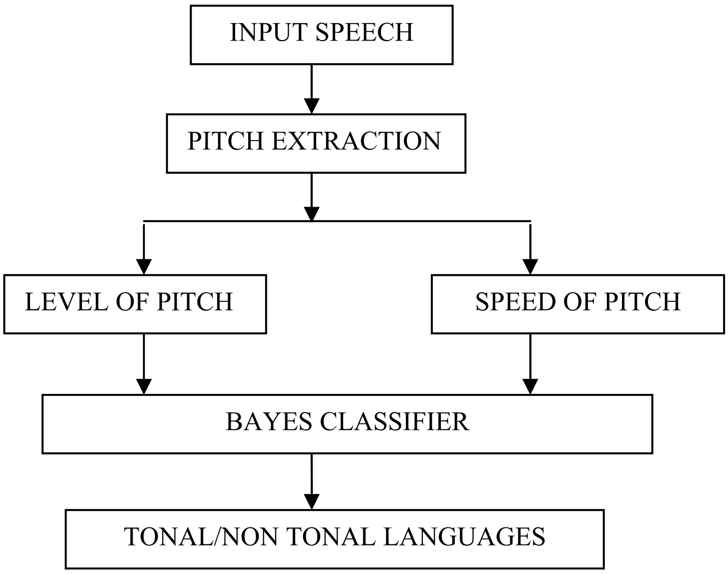

Fig. 2. Prosodic feature based classification system

The system consists of a pitch extraction module followed by a speed of pitch change module and a level of pitch change module. Any of the suitable pitch tracking algorithm can be used to extract pitch which is a representative of prosodic features in speech samples. The speed of pitch change is measured as the difference between the pitch frequencies of two consecutive frames. Again, level of pitch change is measured as the distance of the pitch frequency of a particular frame from the average of pitch frequencies of all the frames. Ideally the tonal languages have the characteristics that both the speed and the level of pitch changes are higher in them than the non-tonal languages. Now both the speed and the level of pitch change are calculated for all the voiced frames [2] from the training data set. Next these two are used as inputs to a Bayes' classifier. The classifier works on Naive Bayes' principle of class conditional probability. The test samples can then be classified as tonal or non-tonal depending upon the seed and the level of pitch changes associated with them.

VII. EXPERIMENTAL RESULTS

Here, the experimental results along with the details of the database used are presented. The implementation of HMM based PE is done using HTK-tool version 3.3 [14], while the implementation of GMM-UBM system has been carried out using MATLAB version 2011(b).The tables below discuss the various results obtained:

The description of the data used in speech corpus is mentioned in the Table I.

Table I
Description of Data Used In Speech Corpus

Language	**Training Data Used**	**Testing Data Used**	**Length of Wave Files**	**Number of Testing Files**
Assamese	2.80 hrs	33 min	4-6 sec	298
Bengali	2.90 hrs	73 min	4-6 sec	829
Manipuri	1 hr	29 min	4-6 sec	214

A. *IDR for GMM-UBM Based Lid System*

The IDR has been calculated as in equation (2).

Table II below shows the experimental results with the described data set.

Table II
Experimental Results of the GMM-UBM Based LID System

Number of Testing Files used in each language	**Total Number of Testing Files**	**Number of False Identification**	**Total Number of False Identification**	**Overall Accuracy**
Assamese- (298) Bengali- (829) Manipuri- (214)	1341	Assamese- (1) Bengali- (40) Manipuri- (Nil)	41	96.94%

Here, all the Bengali error files have been falsely identified as Assamese. This may be due to the similarity between the two languages. Both GMM and GMM-UBM are showing same level of accuracy.

B. IDR for PE Based LID System

The Table III shows the experimental results with the same data set.

Table III
Experimental Results of PE Based LID System

Sl. No.	**Language**	**Accuracy obtain using Acoustic Likelihood**
1	Assamese	99%
2	Manipuri	99%
3	Bengali	100%

The overall accuracy is 99.33%

The identification process relies on the fact that a test utterance is considered to be belonging to a particular language for which the corresponding PE produces the highest likelihood score. Thus, the above result in the Table III is obtained by manually checking all likelihood scores for the test utterances.

C. IDR of Prosodic Feature Based Classifier

The prosodic feature based classifier can classify the languages only as tonal or non-tonal languages. The Table IV shows the results obtained using this approach.

Table IV
Prosodic Feature Based Tonal/Non-Tonal Classification

Languages	**No. of Samples**	**Correct Identification**	**False Alarm**	**Identified as (Tonal/Non-Tonal)**	**Accuracy (%)**
Assamese	330	298	32	Non-tonal	90.30
Manipuri	330	313	17	Tonal	94.84
Bengali	293	219	74	Non-tonal	74.74

The above results present the performance of the classification system.

VIII. CONCLUSION

It is observed that the prosodic feature based approach can't alone be used for identification of languages. However, this method can be used to classify languages into two broad groups of tonal and non-tonal languages. So this method must be coupled with a proper language identification system such as PE or GMM-UBM for completing the identification process. Again, the PE based LID system showed better accuracy than the GMM-UBM based system. This may be attributed to the use of a phonetically labelled speech corpus in the method. The phonetic descriptions represent the speech data more properly and therefore the salient features of the data are retained. However, it is difficult to get a labelled speech corpus always, since it requires the need of language experts. The GMM-UBM method does not require any labelled corpus and can be used directly with the raw data, though accuracy is slightly below the PE. As such, it can be concluded that the price paid for accuracy is compensated with the data preparation task.

REFERENCES

[1] Y. K. Muthusamyy, E., Barnardz, and R. A. Colez, Reviewing Automatic Language Identification, (1994) *IEEE* Signal Processing Magazine, 11(4), pp 33 – 41.

[2] L. Wang, "Automatic Spoken Language Identification," Thesis, University of New South Wales, Australia, 2008.

[3] M. Zissman, Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, (1996) *IEEE Transactions on Speech and Audio Processing,* 4(1), pp 31-44.

[4] S. Furui, 50 Years of Progress in Speech and Speaker Recognition Research, (2005) *The Journal of the Acoustical Society of America,* 116(4), pp 2497-2498.

[5] H. Fujihara, and M. Goto, Three Techniques for Improving Automatic Synchronization between Music and Lyrics: Fricative Detection, Filler Model, and Novel Feature Vectors for Vocal Activity Detection, *Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing*, (pp. 69-72, 2008).

[6] P. Schwarz, "Phone Recognition Based on Long Temporal Context," Thesis, Bruno University of Technology, Czech Republic, Europe, 2008.

[7] P. Matejka, "Phonotactic and Acoustic Language Recognition," Thesis, Bruno University of Technology, Czech Republic, Europe, 2009.

[8] P. Eswar, "A Ruled-Based Approach for Spotting Characters from Continuous Speech in Indian Languages," Thesis, IIT Madras, India, 1990.

[9] S. V. Gangashetty, "Neural Network Model for Recognition of Consonant-Vowel Units of Speech in Multiple Languages," Thesis, IIT Madras, India, 2004.

[10] International Phonetic Association (1999) Handbook of the International Phonetic Association. Cambridge University Press, Edinberg Building, Cambridge.

[11] R. Rabiner, and B. H. Huang, An Introduction to Hidden Markov Models, (1986) *IEEE Acoustics and Speech Signal Processing Magazine,* vol. 3, pp 4-16.

[12] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of IEEE,* vol. 77, Issue 2, pp 257-286, 1989.

[13] P. Bhaskararao, Salient Phonetic Features of Indian Languages in Speech Technology, (2011) *Sadhana,* vol. 36, pp 587-599.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, Speaker Verification using Adapted Gaussian Mixture Models, (2000) *Digital Signal Processing,* vol. 10, pp 19-41.

[15] T. J. Hazen, and V. W. Zue, Automatic Language Identification using a Segment Based Approach. *Proceedings of Eurospeech*, vol. 2, pp 1303-1306, 1993.

[16] K. P. Li, Automatic Language Identification using Syllabic Spectral Features. *Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing*, (pp. 297-300, 1994).

[17] L. Wang, E. Ambikairajah, and E. Choi, A Novel Method for Automatic Tonal Non-Tonal Language Classification, *IEEE International Conference on Multi Media and Expo Workshop*, (pp. 352-355, 2007).

[18] K. Y. E. Wong, "Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information," Thesis, Queensland University of Technology, Australia, 2004.

[19] T. Rong, "Automatic Speaker and Language Identification," Thesis, Nanyang Technological University, Singapore, 2006.

[20] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory", Algorithm and System Development*, (Prentice Hall PTR, NJ, USA, 2001).

[21] T. Schultz, and K. Kirchhoffe, *Multilingual speech processing*, (Elsevier, Academic Press, 2006).

[22] M. Gruhne, K. Schmidt, and C. Dittmar, Phone Recognition in Popular Music, *Proceedings of 8th International Conference on Music Information Retrieval*, Austria, September 2007.

[23] A. Nagesh, and M. Sadanadam, Language Identification using Ergodic Hidden Markov Model. Int J Adv Res Comput Sci Softw Eng 2: 297-301.

[24] S. Nandakishor, L. Rahul, S. K. Dutta, and L. J. Singh, Development of Manipuri Phonetic Engine. Zonal seminar, The Institute of Electronics and Telecommunication Engineers [IETE], May 3-4, 2013.

[25] S. K. Dutta, S. Nandakishor, and L. J. Singh (2015) Development of Language Identification System Using Phonetic Engine. *Proceedings of 13th Conference on Computing and Communication Systems,* (pp. 182-186, 2015).

[26] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book,* (Cambridge University Engineering Department, Cambridge, 2009).

AUTHOR PROFILE

Sushanta Kabir Dutta received his B. E and M.Tech degrees from Dibrugarh University and Manipal University respectively. He is presently working in the Department of Electronics and Communication Engineering in North Eastern Hill University, Shillong. He is also pursuing PhD from the same University.

Dr. Lairenlakpam Joyprakash Singh received his B.Tech. degree in Electronics & Communication Engineering (ECE) from North Eastern Regional Institute of Science and Technology (NERIST), Arunachal Pradesh in 1999. He received his M.Tech. degree from Tezpur University in 2000 and Ph.D. (Engg.) from Jadavpur University in 2006. He is presently working as an Associate Professor in the Department of Electronics and Communication Engineering in North Eastern Hill University, Shillong. His area of interest is Signal Analysis and Processing. He is a member of IEEE and a life member of the Computer Society of India (CSI), Mumbai and the Indian Science Congress Association (ISCA), Kolkata.