**Indian Journal of Science and Technology,** *Vol 10(23), DOI: 10.17485/ijst/2017/v10i23/114011, June 2017*
*ISSN (Print) : 0974-6846*
*ISSN (Online) : 0974-5645*

# An Analysis of Automatic Phone Recognition and Identification of a Few Languages from North Eastern India

**Sushanta Kabir Dutta* and Lairenlapam Joyprakash Singh**

Department of Electronics and Communication Engineering, North Eastern Hill University, Shillong – 793022, Meghalaya, India; sushantatzp@gmail.com, jplairen@gmail.com

## Abstract

**Objective**: Phones are basic sound units available in the spoken data. Languages differ among themselves due to use of different phone sets. This paper analyzes some aspects of automatic phone recognition and subsequent identification of a few languages from North Eastern region of India using Phonetic Engine (PE). **Methods and Statistical Analysis**: PE is a system which converts a speech sample into some symbolic form so that these symbols are capable of capturing all the information carried by the speech sample. In the development of PEs, the International Phonetic Alphabet (IPA) symbols are used in the data transcription process. In modelling the phonetic units Hidden Markov Models (HMM) have been used in the training phase. These trained HMMs are then used in phone recognitions leading to the identification of language(s) of unknown test utterances. **Findings:** PEs are built for three Indian Languages and one dialect, namely Manipuri, Assamese and Bengali and the Kakching dialect of Manipuri. These languages are widely spoken across the North Eastern region of India. The overall phone recognition accuracies reported by the PEs for the above selected languages are 62:11% for standard Manipuri language, 59:0% for Kakching dialect of Manipuri, 43:28% for standard Assamese and 48:58% for Bengali language. **Application**: Automatic LID is possible using a set of PEs in testing unknown utterances due to the language bias of these systems. Various level of identification rates reported in some LID tasks carried out with PEs are discussed here to make an analysis of the issues belonging to it.

**Keywords:** Hidden Markov Models (HMM), International Phonetic Alphabet (IPA), Language Dependency, Language Identification (LID), Phone Recognition, Phonetic Engine (PE)

## 1. Introduction

The present paper shows an analysis of automatic phone recognition and subsequent identification of languages using phonetic engine. The term Phonetic Engine (PE) was used to transform acoustic phonetic information present in a speech sample into some symbolic form[1,2]. Therefore, it produces a sequence of symbols without using any language constraints like lexicon, syntax etc. The detail of PE has been described in some literatures[3]. The symbols here are chosen in a way so that they are capable of capturing all kinds of phonetic variations in the speech signal. A few PEs implemented for selected Indian languages produce syllable-like units as the output which are found as the most basic units in sound production system. The International Phonetic Alphabet (IPA) symbols are used as the underlying sub-word units in the process. These sub-word units work as one level of abstraction for continuous information found in the speech data. Therefore, these symbols may be considered as some form of the quantized versions of the information present in the speech sample[4].

The PE finds its applications as a front end module in speech recognition (SR), information retrieval (IR) and language identification (LID). One distinct advantage of PE is that it uses an open vocabulary set. It incurs a low amount of error for out of vocabulary (OOV) words. Another advantage is the possibility to make a PE for a

**Author for correspondence*

new language with similar sounds using minimum speech data from the language since it uses very low level of representation like the phonetic units of the speech data. Besides, the use of PE is more suitable for speech data from Indian languages because they use phonetic nature of characters during production and writing[3]. Thus the PEs developed for Indian languages use syllable like sub-word units coupled with Hidden Markov Model (HMM) to model each of them[3,5]. The PEs use IPA symbols where, IPA provides one unique symbol for each distinctive sound unit. A symbol usually comprises of one or more elements of two basic types, letters and diacritic marks[6]. While letters represent basic sound units, the diacritics are small markings placed around an IPA letter that show a certain alteration or specific description of sound.

The present study analyzes the PEs available for the three different languages Manipuri, Assamese and Bengali widely spoken in the North Eastern part of India. In Manipuri, there are two PEs, one for the standard Manipuri language (the officially used version) and another for a dialect known as 'Kakching' spoken by the people in Kakching area of South Eastern part of Manipur. However, for Assamese and Bengali languages, the PEs is for the standard versions of languages (being officially used). The phone recognition accuracies as implemented and reported by the Manipuri PEs are 62:11% for the standard version[7] and 59:0% for Kakching dialect. Similarly the accuracies reported by Assamese and Bengali PEs are 43:28% and 48:58% respectively[7]. An analysis of these variations in accuracies with overall effect of them in automatic identification of the languages has been done here.

The other sections of the paper are organized in the following way: The section 2 provides an overview of the phone recognition and language identification using PEs. The section 3 describes the database and addresses various issues with phone recognition and language identification. The section 4 summarizes and concludes the work.

## 2. Overview of Phone Recognition and Language Identification using PE

The basic idea in PEs is that they are being developed using mono-phone HMMs[8,9]. In order to develop any PE in a particular language, first thing is to collect a set of spoken data. Although there is no limitation for data requirement, it is always better to collect data in a way such that all the phonetic units in the language are available in the data. The collected data are then separated into training and testing data sets. Transcription is done for the training data set using IPA. Next, the speech features are extracted from the samples by a suitable feature extraction technique. Ideally, a set of 13 to 39 dimensional feature vector is used to represent a speech frame to 20-25msec duration. The speech frames are prepared as overlapped frames of around 10 msec. Now, each phonetic unit is represented by using a 5 state left-to right context independent mono-phone HMM with one Gaussian per state. The implementation of PE is carried out using the HTK toolkit[10]. Initially a prototype model is defined with all means set at zero and variances set at unity. After this, the global mean and variances are estimated scanning through the training files. A new prototype model with all means set at global mean and all variances set at global variance is thus found. This model is used for creating flat start HMMs. These HMMs are re-estimated using training data and after some iteration the estimation process is completed. This process produces one model for each phone from the considered phone set.

Feature vectors are extracted from the test data in a similar manner as of the training data. Next, the Viterbi decoding is used for decoding the test samples to sequence of phone. Viterbi decoding finds the hidden sequence of states within a phone. Thus these states are the most likely states to have produced the observed sequence of feature vectors corresponding to the phone. The automatic phone recognition accuracy[11] is measured as below

$$PA = (N\text{-}S\text{-}D\text{-}I)/N \times 100 \quad (1)$$

Where N is the number of phones, S is the number of substitutions, D is the number of deletions and I is the number of insertions.

While building an automatic LID system[12,13] a set of PEs are put together. Each individual PE corresponds to one language. Now, the extracted feature vectors from a test utterance are compared with the PEs to get an estimate of the probability of the set of feature vector coming from each PE is illustrated in Figure 1. Here the PEs of three different languages, viz L1, L2 and L3 are shown to build the LID system[11].

The highest likelihood score emanating from any particular PE is used to identify a language. The system can be extended to contain any number of PEs as required.

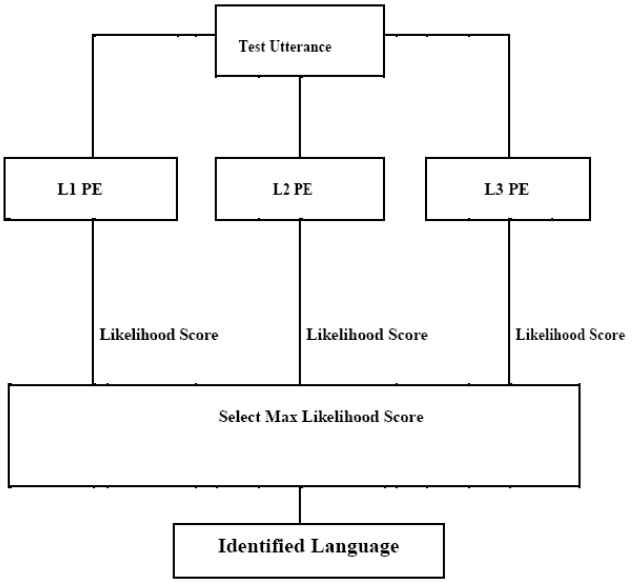


**Figure 1.** LID system.

While using in LID tasks, the identification rate (IDR) of the overall system is estimated as

$$IDR = \frac{n}{N} \quad (2)$$

Where 'n' is the number of correctly identified utterances and 'N' is the total number of utterances used in a particular language L.

## 3 Analysis of Phone Recognition and LID Accuracies

Here some experiments were carried out to test the automatic phone recognition accuracies of the PEs.

In the first level the accuracy of one PE is tested with data being collected in various modes. For this purpose the speech data are collected in 'Read', 'Lecture' and 'Conversation' modes. Here, the standard Manipuri language based PE is selected for the experiments. The 'Read' mode data are collected from AIR Imphal (capital city of the State of Manipur). The 'Lecture' mode data are recorded in the studio with a Zoom H4n high quality recorder from professional teachers. The 'Conversation' mode data are recorded by the same device while allowing a group of people to discuss on a topic in the studio. Table 1 shows the data set. Here, a total number of 31 speakers including both male and female were available to offer the data and all the speakers had used 30 distinct phones including a silence. The PE was built with these phones using the HTK toolkit.

**Table 1.** Durations of training and testing data for Manipuri PE

| Mode | Training Data (Hr:Min:Sec) | Testing Data (Hr:Min:Sec) | Total Data (Hr:Min:Sec) |
|---|---|---|---|
| Read | 2:23:53 | 1:26:48 | 3:50:41 |
| Lecture | 1:36:57 | 00:53:41 | 2:30:38 |
| Conversation | 1:41:42 | 00:49:35 | 2:31:17 |

The automatic phone recognition accuracy was available from the confusion matrix[10]. The accuracies reported by the PE were compared for different set of feature vectors[12,14] used. Here, the PE was tested with LPCC, PLP and MFCC feature vectors with 13, 26 and 39 dimensions. The various levels of accuracies reported by the PE are listed in Table 2.

From this, it can be inferred that the accuracy of PE depends upon the type of data collected and at the same time acoustic representation of the speech samples. From Table 2 it can be noticed that percentage accuracy (PA) of the PE is high with 'Read' and 'Conversation' mode of data in all cases. This is due to these modes of data being the more formal mode of speech as compared with 'Conversation' mode. Again, the PA is also dependent on the set of feature vector. It is observed that the PA is high with both PLP and MFCC compared to LPCC. This is due to PLP and MFCC both are being capable of incorporating some level of the hearing aspects of human ear. Further, the feature dimension also affects the percentage accuracy. Thus, the same Table 2 shows that the PA of the PE is highest in all cases when the feature set is represented by 39 dimensional feature vectors.

Another observation is that the phone recognition rate is higher when the data base contains only one type of speaker. Table 3 lists details of this. However, the overall accuracy was 62:11% when all types of speakers and all modes of data were used. Thus the accuracy depends on the type of speakers as well as the mode of data collected.

The accuracy of transcription also plays a vital role in the performance of the PE. In order to distinguish between two varieties of the Manipuri language - the PE with the standard Manipuri and the PE with the Kakching dialect are used in phone durations measurements.

This provides significant information regarding variations in the two languages. Figure 2 shows a comparison of phone durations measured in units of $10^{-7}$sec. Only vowels are considered as consonant durations are not significant. The Kakching dialect has longer durations

**Table 2.** Performance of Manipuri PE in Read, Lecture and conversation speech data using 13 static coefficients, 26 coefficients (13 static and 13 delta coefficients) and 39 coefficients (13 static, 13 delta and 13 acceleration coefficients) of LPCC, PLP, MFCC

| Feature used | Percentage accuracy in different modes of data with selected coefficient dimensions | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Read Mode | | | Lecture Mode | | | Conversation Mode | | |
| | [13] | [26] | [39] | [13] | [26] | [39] | [13] | [26] | [39] |
| LPCC | 42.32 % | 56.21 % | 64.26 % | 43.36 % | 57.27 % | 62.71 % | 37.07 % | 48.10 % | 53.29 % |
| PLP | 49.09 % | 65.11 % | 70.55 % | 49.35 % | 64.28 % | 68.00 % | 44.83 % | 58.44 % | 63.17 % |
| MFCC | 48.85 % | 63.40 % | 70.66 % | 49.03 % | 64.50 % | 68.66 % | 44.54 % | 57.88 % | 63.11 % |

**Table 3.** Performance analysis of Manipuri language based PE

| Sl No | Mode | Type of Speaker | Accuracy (%) |
|---|---|---|---|
| 1 | Conversation | Female | 64.71 |
| 2 | Lecture | Male | 69.71 |
| 3 | Lecture | Female | 72.50 |
| 4 | Lecture | Both male & female | 68.62 |
| 5 | Read | Female | 74.21 |
| 6 | Read | Male | 72.07 |
| 7 | Read | Both male & female | 70.49 |

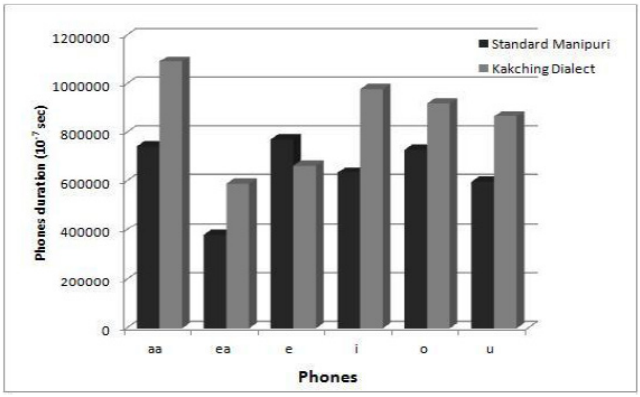


**Figure 2.** Comparison of phone duration measurements of Manipuri and Kakching.

for vowel pronunciations as compared to the standard version of Manipuri language except for the vowel 'e'. In case of inaccurate transcriptions, the phone identification results would be affected.

On the other hand, the number of speakers used in the data-base also influences the phone recognition tasks. With 16 speakers using 30 phones the Manipuri PE has overall accuracy 62:11%, while with 31 speakers using

**Table 4.** IDR by LID system with 3 PEs

| Language | IDR (%) |
|---|---|
| Maniprui | 99 |
| Assamese | 99 |
| Bengali | 100 |

34 different phones the Assamese PE overall accuracy is found to be 43:28%. Similarly, with 43 speakers using 34 phones, the Bengali PE has overall accuracy of 48:58%. Again with 8 speakers using 30 phones the Kakching dialect PE has an overall accuracy of 59:0%. These variations may be attributed to the speaking variations among the speakers which influence the transcription process. The data collection procedure for all these languages were similar to the one adopted for the Manipuri languages and the amount of data collected were nearly of same durations as in the Table 1.

However, when used with LID tasks, the system comprising of a set of PEs the IDR is much higher even though individual PE's PA is less. This is a good sign that PEs are worth using in LID tasks. Table 4 shows the IDR for the LID system built using Manipuri, Assamese and Bengali PEs together. The reason behind this may be due to the finding the maximum likelihood estimates of an unknown test utterance being coming from any of the three PEs. The overall IDR reported by the LID system is found to be 99:33%, when tested with a total of 300 test samples, each of around 10 sec duration from Assamese, Manipuri and Bengali languages.

## 4. Conclusion

Thus an analysis of the automatic phone recognition as well as language identification by PEs has been done

here. The primary task of a PE is phone recognition and the higher the level of accuracy, better the phone recognizer it would be. While using in language identification, a set of Phone Recognizers (PR) with high accuracy levels are usually expected, since the common understanding is that it will result in an LID system with better performance. However, from the experiments with the present LID system it is found that the IDR of the system does not much depend on the accuracies of individual PRs (here the PEs). This is due to considering the maximum likelihood estimates in case of the identification of an unknown test utterance. Another point is that different languages often use different phone sets. Therefore, such an LID system can always result with sufficient IDRs.

However, the IDR can be tested with one PE for different varieties of a single language. That is the PE of a single language can be used to observe how it identifies different dialects of the language. For this the transcribed data from all the dialects would be required to train the system and this may be carried out in future.

## 5. References


1. Eswar P. A rule-based approach for spotting characters from continuous speech in Indian languages [Ph.D. thesis]. IIT Madras: 1990.
2. Gangashetty SV. Neural network model for recognition of consonent-vowel units in multiple languages [Ph.D. thesis]. IIT Madras: 2004.
3. Yegnanarayana B, Gangashetty SV, Rajendran S, Murty KSR, Dhananjaya N, Guruprasad S. A Phonetic Engine for Indian languages. 7th International Conference on Natural Language Processing (ICON-2009); Hyderabad, India. 2009, pp. 8391.
4. Sarma BD, Sarma M, Prasanna SRM. Development of Assamese Phonetic Engine: some issues. IEEE India Conference (INDICON-2013); Mumbai, India. 2013, pp. 1-6. Crossref
5. Yegnanarayana B, Gangashetty SV. Machine learning for speech recognition-an illustartion of phonetic engine using Hidden Markov Models. International Conference on Frontiers of Interface between Statistics and Sciences, Hyderabad: 2009. PMid:19813815
6. International Phonetic Association. Handbook of International Phonetic Association. UK: Cambridge University Press; 1999.
7. Dutta SK, Salam NK, Singh LJ. Development of Manipuri Phonetic Engine and its application in language identification. International Journal of Engineering and Technical Research. 2015; 3(8):200–3.
8. Schwarz P. Phone Recognition based on long temporal context [Ph.D. thesis]. Faculty of Information technology, Bruno University of Technology; 2008.
9. Nagesh A, Sadanadam M. Language identification using Ergodic Hidden Markov Model. International Journal of Advanced research in Computer Science and software Engineering. 2012; 2(11):297–301.
10. Young S. HTK book. Cambridge, UK: Cambridge University Engineering Department; 2005.
11. Dutta SK, Salam NK, Singh LJ. Development of language identification system using Phonetic Engine. International Conference on Computing and Communication Systems (I3CS'15); Shillong: 2015.
12. Huang X, Accero A, Hon HW. Spoken language processing: A guide to theory algorithm and system development. USA: Prentice Hall; 2001. p. 1008.
13. Rong T. Automatic speaker and language identification [Ph.D. thesis] Singapore: Nayang Technical University; 2006.
14. Rabiner L, Juang B. Fundamentals of speech processing. USA: Prentice Hall; 1993. p. 1–3.