International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-3, Issue-8, August 2015

Development of Manipuri Phonetic Engine and its application in language identification

Sushanta Kabir Dutta, Salam Nandakishor, L. Joyprakash Singh

***Abstract*— This paper discusses the development of a phonetic engine which can be used for building an automatic language identification system. Since a phonetic engine can best extract the acoustic information of a speech signal and converts that into symbolic form, it can suitably be used to improve the performance of existing phone recognizers. A detailed discussion is presented for the phonetic engine in the Manipuri language. The two other phonetic engines for Assamese and Bengali languages are built with similar ideas and only an overview is stated here. Around 5 hours of 'read speech' data have been collected separately for each of Manipuri, Assamese and Bengali languages for training and testing purposes. The collected speech data of Manipuri, Assamese and Bengali consisted of 16, 31 and 43 speakers respectively. Symbols of IPA (International Phonetic Alphabet) revised in 2005 have been used in transcription of the data. A 5-state left to right Hidden Markov Model (HMM) with 32 continuous density diagonal covariance Gaussian Mixture Model (GMM) per state is used to build a model for each phonetic unit. The Manipuri phonetic unit is trained with 30 phonetic units including a silence symbol, which is used to indicate break between two words. After training and testing we analysed the performance of the system. An overall accuracy of 62.11% has been achieved with the engine. Assamese and Bengali phonetic engines are built with 34 phonetic units each, which results in accuracies of 43.28% and 48.58% respectively. These together with the Manipuri phonetic engine are then used to build a system for automatic language identification. It has been observed that the system is well capable of identifying a target language. The Identification Rates (IDR) of the Manipuri and Assamese languages are 99% and the IDR for Bengali Language is 100%.**

***Index Terms*—Manipuri phonetic engine, Mel-frequency Cepstral Coefficients (MFCC), Hidden Markov Model (HMM), Language Identification (LID).**

I. INTRODUCTION

Phonetic engine (PE) is the signal to symbol transformation module which uses the acoustic phonetic information present in the speech signal to convert it into symbolic form [1], [2]. The engine produces a sequence of symbols without using any language constraints in the form of lexical, syntactic and higher level knowledge source. The choice of symbol should be such that it can capture all the phonetic variations in speech. Existing PE implemented for Indian languages produces syllable like units as the output where constraint at the syllable level are used, as syllable-like units are most basic in the production of speech. PE is the front end module for both speech recognition system and information retrieval system. In automatic speech recognition of continuous speech, the speech signal is first converted to the sub-word units of speech which in turn is converted to text. The first part of converting speech to sub-word units is done by a PE. Existing PE implemented for Indian languages uses syllable like units as the sub-word units. Here we will use sequence of International Phonetic Alphabet (IPA) as the sub-word units as IPA provides one symbol for each distinctive sound (speech segment) [3]. These symbols are composed of one or more elements of two basic types, Letters and Diacritics. Letters represent basic sound units while Diacritics are small markings which are placed around the IPA letter in order to show a certain alteration or more specific description in the Letter's pronunciation. Since IPA symbols capture all distinctive acoustic phonetic characteristics of speech, they can be called as acoustic phonetic sequence (APS) [4]. In an automatic language identification task, the PE can be used to recognize the phone units and also to calculate the 'Acoustic (AC) Likelihood' of an unknown test utterance. As we get the highest 'Acoustic Likelihood' score in a particular PE for an unknown test utterance, it is then considered as the identified language. The PE can be used in various other applications too, for example keyword detection [5] language recognition [6], speaker identification [7], music identification and translation [8], [9].

The rest of the paper is organized as follows: Section-II briefly describes an automatic language identification system cues and characteristics. Section-III explains the corpus and in depth development of Manipuri Phonetic Engine. Section-IV details about the proposed language identification system that uses the Manipuri PE together with Assamese and Bengali PEs developed similarly. Experimental Results are discussed in Section-V. The Conclusion is presented in Section-VI along with a possible direction for the future work.

II. BRIEF REVIEW OF AUTOMATIC LANGUAGE IDENTIFICATION CUES AND SYSTEMS

The task of automatic language identification (LID) is the process of identifying a particular language from a set of languages. In literature, LID systems are broadly classified into two main categories, namely Explicit LID system and Implicit LID systems [10]. This is done on the basis of how the languages are modeled within the system. An Explicit LID system requires segmented and labeled speech corpus while an Implicit system uses digitized speech samples with corresponding true identities of the language. Our proposed LID system using phonetic engine is an Explicit LID system [11]. A few LID cues are listed below.

A. Acoustic

It is a physical characteristic of the speech signal described by frequency, time and intensity information of the speech. Typically, the acoustic information of a spoken utterance is

Sushanta Kabir Dutta, Department of Electronics and Communication Engineering, NEHU, Shillong, India, +91-9436333097.
Salam Nandakishor, Department of Electronics and Communication Engineering, NEHU, Shillong, India, +91-9863132849.
Lairenlakpam Joyprakash Singh, Department of Electronics and Communication Engineering, NEHU, Shillong, India, +91-9436349192.

represented as a sequence of feature vectors where each individual vector corresponds to acoustic information for a particular time frame. Acoustic information is one of the most primitive forms of information which can be obtained by a process called Speech Parameterizations process (also Acoustic Analysis) directly from raw speech [12]-[15]. The most widely used parameterizations techniques are Linear Prediction Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) and Linear Prediction Cepstral Coefficient (LPCC).

B. Phonotactic

There are various phonological factors which govern the distinctiveness of a particular language. Some of these factors include the phone set and the Phonotactic constraints of the language. 'Phonotactic' refers to the rules that govern the combinations of the different phones in a language. There is a wide variation in Phonotactics across languages world over. Different languages may have different rules for describing how sequences of phonemes may be constructed. These Phonotactic constraints may result in having appeared some phonetic sequences similar in some languages while very different to many others. For example, Japanese language has strict Phonotactic constraints which generally prohibit consonants from following consonants. English, on the other hand, has looser constraints which allow for the possibilities of multiple consonants occur in succession. Hence Phonotactic information can be useful in capturing some of the dynamic nature of speech lost during feature extraction [16].

C. Vocabulary

Conceptually, the most important difference among the languages which is that they use different word sets. So their vocabularies differ [16].

D. Prosodic

The stress, intonation (pitch contour), and rhythm (the duration of phones, speech rate) are all the important elements being used within the prosodic structure of a spoken utterance. The manner in which these elements are incorporated into the prosodic structure varies across the languages. The differences among the languages can often be observed in the realization to their prosodic features which in turn determine the tones or the stress contained throughout an utterance. For example, tonal languages such as Mandarin have very different intonation characteristics compared with that of the stress languages such as English [16].

A few researches [17]-[19] have exposed a significant importance of acoustic and Phonotactic information in language identification work. In order to get an accurate estimation of these information sources, a detailed modeling is found to be necessary [17], [25]. In this paper, an approach to automatic language identification based on language-dependent phone recognition has been proposed. Continuous HMMs are used to build the language-dependent phone recognizers (PRs). An acoustic model is created by using audio recordings of speech with their corresponding transcriptions which later on are compiled to get statistical representations of the phone units.

III. Corpus and Development of the Manipuri Phonetic Engine

A. Data collection and transcription

A good quality data of about 5 hours in read speech have been collected from the recording studio as well as from the AIR Imphal. This data consists of speech read by male and female speakers. The H4n recording devices have been used during recording in the studio. The device is maintained at a sampling frequency of 48 kHz and 44.1 kHz, 16 bit per sample size and WAV format.

The broadcast data acquired has been sliced into smaller parts proportionate to the length of a sentence. Each chunk of data is listened and analyzed carefully to obtain higher accuracy in transcription. The Read mode data has been collected from 5 males and 11 female native speakers of Manipuri language. Each of these male speakers used about 35 phones while among female speakers, three have used 35 phones each while remaining speakers used 34 phones only. A total of 36 phones have been used by the speakers altogether. The transcription of collected read data have been done using the IPA chart (revision 2005) to build up the database of this system. Table-I below shows the list of phonetic units in the Manipuri language and the reduced units after merging the similar sounds.

B. Data preparation and Task definition

In building the Manipuri Phonetic Engine [23], we merged some phonetic units having similar sounds, as explained in Table-I. After merging these phonetic units we are left with a total of 30 distinct phonetic units including a silence unit and these are finally used in the development of the Manipuri Phonetic Engine. Then, we assigned the 29 phonetic units equal number of ASCII codes, while the silence symbol is denoted by 'sil'. Using these 29 ASCII codes altogether with 'sil' and we create the basic architecture of the PE.

C. Acoustic Analysis

The system tools cannot process directly on speech waveforms and hence these have to be represented in a more compact and efficient way which is achieved through the acoustical analysis. During the analysis, the signal is segmented in successive frames of 25 ms with a frame shift of 10 ms. Each frame is then multiplied by a Hamming window. A vector of acoustic coefficients giving a compact representation of the spectral properties is extracted from each windowed frame. Here, each feature vector consists of an energy coefficient, 12 MFCCs, 13 delta coefficients and another 13 acceleration coefficients respectively. These 39 coefficients give the vocal tract information of the speaker.

D. Training phase

For each of the phonemes including the silence, a HMM is designed. Each model consists of 5 states. The first and the last states are non-emitting states and the remaining 3 states are active state. The pre-defined prototype along with acoustic vectors and transcription of training data are first initialized. Then it calculated the global speech mean and variance of the HMMs per state. In the next phase of the development process, the flat start mono-phones calculated thus far are re-estimated. In our system implementation, re-estimation iteration is repeated up to six times as convergence is achieved.

E. Testing Phase

The data to be tested are first transformed into a series of acoustic vectors (MFCCs) in the same way as being done during acoustic analysis in the training phase. The acoustic vectors with HMMs definition, task network, and dictionary and HMM lists are processed in order to produce the transcription of the test data.

Table-I: List of Phonetic units used in Manipuri Phonetic Engine and the Reduced set after merging similar units

Sl. no.	Phonetic units	Reduced Phonetic units	Name in ASCII
1	i	i	i
2	a	a	aa
3	ə	ə	ea
4	o, ɔ	o	o
5	e	e	ee
6	u	u	u
7	n	n	n
8	m	m	m
9	ŋ	ŋ	ng
10	P	p	p
11	b	b	b
12	t	t	t
13	d	d	d
14	k	k	k
15	g	g	g
16	p^h, f	p^h	ph
17	b^h, v	b^h	bh
18	t^h	t^h	th
19	d^h	d^h	dh
20	k^h	k^h	kh
21	g^h	g^h	gh
22	z, d͡z	z	z
23	s, ʃ	s	s
24	h	h	h
25	w, ʊ	w	w
26	ɹ, ɾ	ɹ	r
27	y	y	y
28	l	l	l
29	t͡s	t͡s	ts

IV. Proposed Language Identification System

Acoustic analysis is used as the first step in the development of the LID system. This step is same as discussed in the subsection- C above. A 5-state prototype left to right HMM is used for each phonetic unit of a language [20]. The first and last states of the HMM are non-emitting states while remaining 3 states are the emitting states [21], [22]. We calculated the global mean and variance of HMMs per state using the predefined prototype along with the Acoustic vectors and transcriptions of the training data set [15, 16]. Once an initial set of models has been created, the optimal values for the HMM parameters (transition probability, mean and variance vectors for each observation function) are re-estimated. Thus we get the Acoustic Model.

Then extracted the feature vectors from a test utterance during the `Identification phase' and compared the same with the `acoustic model' to estimate the `Acoustic likelihood' score of the utterance. Prior to this step is the development of Assamese and Bengali phonetic engines that uses 34 phonetic units for each language. The units have been extracted from an equal amount of data as that of the Manipuri language. The techniques adopted in building these PEs are same as discussed for the Manipuri PE above. Then Manipuri, Assamese and Bengali PEs are run over a randomly chosen test utterance spoken in any of these Languages. The highest likelihood score emanating from one of the PEs is used to identify a particular test utterance belonging to that language. Thus the identified language is the one for which the PE of that language yields the highest likelihood score. The Fig-1 below shows a block diagram of the proposed LID system.

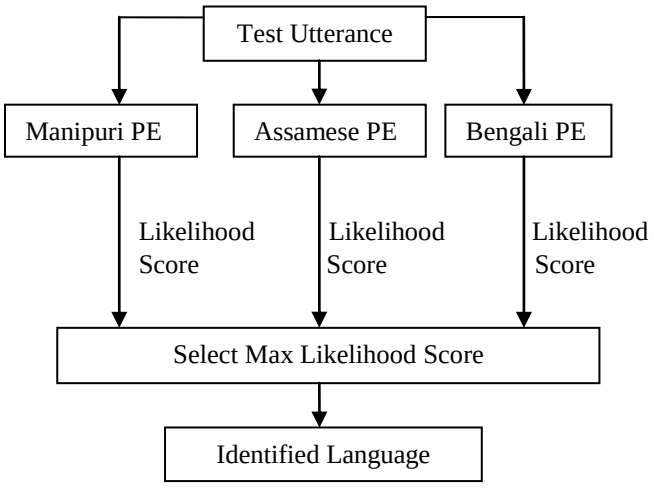

Fig-1: Proposed LID System

V. Experimental Result

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

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

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 and 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 speaker together. Similarly we developed the PEs for Assamese and Bengali languages. The overall accuracies reported for Assamese PE is 43.28% while Bengali PE is 48.58%. Now these PEs are used to identify an arbitrary test utterance (in any of the languages among Manipuri, Assamese and Bengali) according to the procedure stated above.

We used 100 utterances for each of the languages for testing the performance of LID system. The performance of a LID system is determined by the identification rate (IDR). 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.

Table-II: Experimental Results of LID system

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

VI. Conclusion and Future Work

The above results shown in Table-II reveal that the performance of this system is good enough to identify a target language. However, this type of LID system requires phonetic transcriptions and its dictionaries. Producing phonetic transcriptions and dictionaries is an expensive, time consuming process that usually requires a skilled linguistic fluent in the language of interest. As a part of future work, this can be extended with an increasing the number of languages for a few more number of test utterances.

Acknowledgment

This work is a part of the consortium project "Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages" headed by IIIT, Hyderabad and funded by the Department of Information Technology, Govt. of India. We also acknowledge the assistance received from EMST lab IIT, Guwahati for the Assamese and Bengali data.

References

[1] P. Eswar, "*A Ruled-based Approach for Spotting Characters from Continuous Speech in Indian Languages*," Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras, July 1990.

[2] S. V. Gangashetty, "*Neural Network Model for Recognition of Consonant-Vowel units of speech in multiple languages*," Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras, October, 2004.

[3] International Phonetic Association, "*Handbook of the International Phonetic Association*, Cambridge University Press," the Edinberg Building, Cambridge CB2 2RU, UK 1999.

[4] Peri Bhaskararao, "*Salient phonetic features of Indian languages in speech technology*," Sadhana, vol. 36, part. 5, pp. 587599, 2011.

[5] P. Schwarz, "*Phone recognition based on long temporal context*," Ph.D. thesis, Faculty of Information Technology, Bruno University of Technology, 2008.

[6] P. Matejka, *"Phonotactic and Acoustic Language Recognition*," Ph.D. thesis, Faculty of Electrical Engineering and Communication, Bruno University of Technology, 2009.

[7] S. Furui, "*50 Years of Progress in Speech and Speaker Recognition Research*", Proceedings of ECTI Transactions on Computer and Information Technology, vol. 1, no. 2, 2005.

[8] 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, Speech, and Signal Processing (ICASSP), pp. 69-72, USA, April 2008.

[9] M. Gruhne, K. Schmidt, and C. Dittmar, "*Phone recognition in popular music"*, Proceedings of 8th International Conference on Music Information Retrieval, Austria, September 2007.

[10] A. Nagesh and M. Sadanadam, "*Language Identification Using Ergodic Hidden Markov Model*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 11, November 2012.

[11] S.K.Dutta, Salam Nandakishor and L.Joyprakash Singh "*Development of Language Identification System using Phonetic Engine*", I3CS'15, Vol.1, pp.182-186, 2015

[12] Liang Wang, "*Automatic Spoken Language Identification"*, Ph.D. thesis, School of Electrical Engineering and Telecommunications Faculty of Engineering, The University of New South Wales, 2008.

[13] Schultz, T. and K. Kirchhoffe, "*Multilingual Speech Processing*", Elsevier, 2006.

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

[15] Kim-Yung Eddie Wong, "*Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information*", Ph.D. thesis, School of Electrical and Electronics Systems Engineering, Queensland University of Technology.

[16] Tong Rong, "*Automatic Speaker and Language Identification*", Ph.D. report, Nanyang Technological University, March, 2006.

[17] M. Zissman, "Comparison of four approaches to Automatic Language Identification of telephone speech", IEEE Trans, Speech and Audio Processing, Vol. 4, pp. 33-44, 1996.

[18] T.J. Hazen and V.W. Zue, *"Automatic language identification using a segment based approach"*, 3rd European Conference on Speech Communication and Technology (Eurospeech '93). Vol. 2, pages 1303-1306, September, 1993.

[19] K.P. Li, *"Automatic Language Identification Using Syllabic Spectral Features"*, International Conference on Acoustic, Speech, and Signal Processing Proceedings. Vol.1, pages 297-300, April, 1994.

[20] Steve Young et. al., "*The HTK Book (for HTK Version 3.4)"*, Cambridge University Engineering Department, Cambridge, 2009.

[21] L.R. Rabiner , "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*", Proc.of the IEEE Vol. 77, Issue 2, pp. 257286, 1989.

[22] R. Rabiner, and B. H. Huang, "*An introduction to hidden markov models*", IEEE Acoustics Speech Signal Processing, Mag., pp. 4-16, 1986.

[23] Salam Nandakishor, Laishram Rahul, S.K Dutta and L. Joyprakash Singh, "Development of Manipuri Phonetic Engine", Zonal Seminar, The Institute of Electronics and Telecommunication Engineers [IETE], May 3-4, 2013.

[24] Muthusamy et. al, "*Automatic Language Identification: A Review/Tutorial*", IEEE Signal Processing Magazine, October, 1994.

Sushanta Kabir Dutta received his B. E and M.Tech degrees from Dibrugarh University and Manipal University. 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.

Salam Nandakishor received his B. E degree from North Maharashtra University in Electronics and Communication Engineering and pursuing M. Tech degree in Electronics and Communication Engineering from North Eastern Hill University, Shillong. He is currently working as Assistant Laboratory Engineer in Speech and Image Processing Laboratory, Department of Electronics and Communication Engineering, NEHU, Shillong.

Lairenlakpam Joyprakash Singh received his B.Tech. 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.