1

Introduction
1.1 Introduction

Automatic language identification (LID) often refers to identifying languages from a spoken database by computational machines. This has become an enabling technology for a wide range of multilingual speech processing applications[1]. In general, applications of LID systems can be broadly categorized into two groups, namely front-end for human operators and front-end for machines. While the front-end for human operator LID system is useful in diverting calls to appropriate users, the front-end for machine operated LID systems find many applications.

Humans are born with the ability to discriminate among spoken languages as a part of their natural intelligence[2]. However, the quest to automate this ability continues[3, 4] and consequently many promising results have come up in the last few decades[5]. Thus LID aims to replicate human ability to identifying languages by computational means. In the state of art literature, a number of significant results are obtained in the area of language identification[6, 7]. In taking advantage of recent developments, different LID systems are developed based on Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM) etc. The important characteristics of an ideal LID system can be listed as below:

i). The system should be capable of identifying an utterance within a small interval of time.

ii). The performance degradation must not be very high if the duration of the utterance is reduced.

iii). The system should not be biased to any language or group of languages.

iv). The system should tolerate speaker variation, accent variation, channel variation etc.

v). The complexity of the system should be less, in the sense that the amount of language specific information required for the development of the system should be
small and at the same time the inclusion of new languages to the system should be easy.

1.2 Motivation

The present study has been chiefly motivated due to the wide range of applications of LID technologies available in a number of research articles\cite{8–11}. For example, a multi-lingual voice controlled travel information system accepts voice as the only input and from which it identifies the language of an unknown speaker so that he/she may be replied to the queries in his/her language. Therefore, the ultimate goal here is a multi-lingual speech dialogue system. Similarly, telephone companies may provide better services to customers speaking different languages if an LID front-end is used to route the calls to appropriate operators, viz handling emergency call, checking into hotels, arranging meeting for non-native speakers etc. In a multi-lingual country like India, this would be very reasonable because a big population in the country is residing in villages whose sole dependency is on agriculture and are able to speak only native languages. In such cases, the call centres developed for assisting the farmers would be greatly benefited if a LID front-end could be used for diverting a call of a customer to the appropriate call-center employee who speaks that native language.

Even though some of the Indian languages are included in the standard databases used in LID studies like TIMIT, OGI-TS etc, many of the languages are still left untouched. Besides, there is no significant research work reported so far dealing with the LID tasks of languages from North Eastern part of India. Since the North Eastern part of India is neighboured by countries like Bhutan, China, Bangladesh, Myanmar etc., the languages spoken in this region are also influenced by the languages of these neighbouring countries. Therefore, these languages are quite different from the main land languages. Thus, this has become as a motivation to carry out the present LID research for the selected North East Indian languages as mentioned above.
1.3 LID cues

It has been learned that human listeners discriminate among languages by using combination of phoneme and word spotting strategies along with prosodic cues. These experiments, as carried out by Muthuswami[12], showed that increased exposure to each language and longer training sessions contribute to improved performance in language identification by the subjects. Thus prior knowledge remained helpful for the participants to improve their performances. Understanding from these experiments a set of language identification cues had been developed and applied in the LID tasks. Based on the level of knowledge abstraction, these cues are categorized in the following way[1].

a). Acoustic phonetic: The speech sounds as concrete acoustic events are called phonemes[13]. The number of phonemes used in a language usually ranges from 15 and 50, with majority of them having around 30 phonemes each. Phonetic repertoires differ from language to language even though they may share some common phonemes. These contribute to differences in acoustic phonetic feature distributions[13].

b). Phonotactic: Every language has unique set of lexical phonological rules that govern the combination of different phonemes. The phonotactic constraints dictates permissible phoneme sequences. Thus, some phoneme sequences frequent in one language may be rare in another. Such phonotactic constraints may be characterized by phone n-gram model[14].

c). Prosody: Prosody refers to suprasegmental features such as stress, rhythm and intonation [15]. The set of interrelated prosodic features are the important characteristics of the spoken language. Thus, world’s languages can be grouped as stressed-timed, syllable-timed and mora-timed languages[1]. Lexical tone is another very prominent feature in tonal languages for which languages can also be divided into two broad classes as tonal and non-tonal languages. However, from the experiments conducted in human listening[5, 16], it is found the prosodic features are less informative than phonotactic features. Therefore, these features are often used
1.4. Types of LID systems

together with the other features in language classifications applications.

d). Word and syntax: Languages have their phonological systems that governs how symbols are used to form words and morphemes. Again, a syntactic system regulates how words and morphemes are combined to form phrases and utterances. Therefore, lexical approach seems to be an obvious choice in language recognition. However, the number of available languages in the world are very large and learning of all the words and syntax patterns together are also not possible. Therefore, this type of approach is seldom used.

Among all LID approaches, the phone recognition based approaches provide considerable promise as they incorporate sufficient knowledge of the language to be identified and that too without the use of any high cost like the word based approaches. There are three different types of phone based approaches, namely Phone recognition followed by language model (PRLM), parallel-PRLM and parallel phone recognizer (PPR). The novelty in all theses approaches is that more knowledge can be incorporated into the LID system[7, 17]. The front end in all these approaches requires to train the system from manually labeled data. However, in PPR the labeled data from all languages under consideration, are required to train the recognition system. Therefore, PPR is most difficult to implement although it offers a very good performance.

1.4 Types of LID systems

All spoken language identification systems are broadly classified into two types, Explicit LID and Implicit LID systems[18].

1.4.1 Explicit LID system

The LID system that requires speech recognizers of one or several languages are termed as Explicit LID system. Thus an Explicit system requires a segmented and phonetically
1.5. Overview of the Present work

labeled speech corpus.

1.4.2 Implicit LID system

The LID system that does not require segmented and labeled speech data are termed as Implicit. Thus, such a system requires only raw speech data. The language specific informations are derived from the available data in order to discriminate among the languages.

1.5 Overview of the Present work

In this work, a system tool called Phonetic Engine (PE) [19] has been developed for selected Indian languages namely Manipuri, Assamese and Bengali, which are spoken widely across the North Eastern part of India, on the basis of principles of PPR. The PE uses the acoustic phonetic information present in the speech samples and converts it into some symbolic form. The conversion from speech signal to the appropriate symbolic form requires efficient transcription in good speech recognition platform. The transcription is useful in the systematic representation of a language in written form as well. The information that needs to be captured in the transcription is essentially what the speaker has spoken and not what the speaker is intended to speak. The choice of symbol must be such that it captures all the phonetic variations in speech. Existing PEs for Indian languages use syllable like units as sub-words units. Here used a sequence of symbols based on International Phonetic Alphabet (IPA) as the sub-word units[20, 21]. The IPA is designed to represent the qualities of speech that are distinctive in oral language: phonemes, intonation and the separation of words and syllables. IPA symbols are composed of one or more elements of two basic types, letters and diacritics. In the current transcriptions, a sequence of IPA symbols are used as sub-word units since IPA provides one symbol for each distinctive sound. In a way, this is a typical sub-word based LID system, sub-words being of IPA characters and this can be thought of as an alternative to the PPR system. The system can be called parallel-sub word recognizer (PSWR). In PSWR, there are N
numbers of SWRs require for the system that has N-language tasks. Each SWR has a language dependent sub-word unit inventory, which is obtained from the training samples. The experiments in the present study were conducted in the database consisting of Read, Lecture and Conversation modes of spoken data. The identification accuracies obtained in the experiments are then compared with another dominant LID technique, the Gaussian Mixture Model based Universal Background Model (GMM-UBM). This comparison showed that percentage accuracy obtained is higher for the PSWR than GMM-UBM.

In addition to this, the prosodic feature differences between the target languages are also explored in the thesis. Although the prosodic feature differences among languages are not always very significant, but here these differences are found to be sufficient enough to distinguish the target languages into two broad categories. This has been done by measuring the changes of speed and level of pitch variations in the languages. This finding has revealed that a substantial amount of pre-assessment could be done in identifying the target languages besides being using the PSWR.

1.6 Contributions

The main contributions in the thesis are summarized as below:

i). The thesis work investigates the language specific properties from Assamese and Manipuri spoken data for the purpose of identification. A system tool called Phonetic Engine (PE) has been developed for 3 Indian languages, namely, Assamese, Manipuri and Bengali from the raw speech data. The non-target language Bengali is used in order to avoid any bias in the system. This tool explores the language specific properties from raw speech data and therefore acts as the building block for the proposed LID system.

ii). A comparative study of the PE based LID system with a GMM-UBM based LID system has been done for the same speech data to evaluate the performance of the proposed system.
1.7. Organization of the thesis

The Thesis has been organized in the following way:

a). Chapter 2 discusses the review of literature in automatic spoken language identification. The chapter begins with a note on the challenges in the area and presents the available studies for both Explicit and Implicit LID systems.

b). Chapter 3 presents the details of preparation of the speech corpus, feature extraction stage and development of the phonetic engine(PE). An analysis of selection of most suitable spectral feature for PE has also been done.

c). Chapter 4 details the development of the LID system using PEs for the selected languages. An analysis of the performance of the system with the Implicit LID system GMM- UBM has been presented.

d). Chapter 5 presents an analysis of automatic phone recognition and discusses the issues related to language identification.

e). Chapter 6 discusses the development of a prosodic feature based classification system for the selected languages. It also presents the advantage and limitations of such a system.

f). Chapter 7 presents conclusion of the work with a mention of the future possible directions.
1.8 List of publications resulting from this research


ii). Sushanta Kabir Dutta and L. Joyprakash Singh ‘A Study on Prosodic feature based Automatic Classification of languages from North Eastern India’ accepted for publication in Advances in Communication Devices and Networking as a part of the Lecture Notes in Electrical Engineering (by Springer-Scopus),


vi). Sushanta Kabir Dutta, Salam Nandakishor and L. Joyprakash Singh ‘A comparative study of feature dependency of the Manipuri language based Phonetic Engine’ in CSCITA, 2017 the IEEE conference (record no #38873) held in Kohinoor Continental, Mumbai on 7-8 April, 2017 (Published by IEEE, available online),

vii). Sushanta Kabir Dutta and L. Joyprakash Singh ‘Some Issues related to Phone Recognition and Language identification using Phonetic Engine’ in the proceedings of I3CS’16 held on November 11-13, 2016 in NEHU Shillong (Proceedings published as Lecture Notes in Network and Systems by Springer),

1.8. *List of publications resulting from this research*

communicated to a journal is under review.