Language Identification for Indian Languages usng Spectral and Prosodic features

Sudhamay Maity

Language Identification for Indian Languages using Spectral and Prosodic features

Thesis submitted to the
Indian Institute of Technology, Kharagpur
For award of the degree

Master of Science (by Research)

by

Sudhamay Maity

under the supervision of

Dr. K. Sreenivasa Rao

School of Information Technology
Indian Institute of Technology, Kharagpur

July 2012

APPROVAL OF THE VIVA-VOCE BOARD

Date: / / .

Certified that the thesis entitled **LANGUAGE IDENTIFICATION FOR INDIAN LANGUAGES USING SPECTRAL AND PROSODIC FEATURES** submitted by **SUDHAMAY MAITY** to the Indian Institute of Technology, Kharagpur, for the award of the degree Master of Science has been accepted by the external examiners and that the student has successfully defended the thesis in the viva-voce examination held today

(Member of DAC) (Member of DAC) (Member of DAC)

(Member of DAC) (Member of DAC) (Member of DAC)

(Supervisor)

(Internal Examiner) (Chairman)

CERTIFICATE

This is to certify that this thesis entitled **Language Identification for Indian Languages using Spectral and Prosodic features** submitted by **Sudhamay Maity** to Indian Institute of Technology, Kharagpur, is a record of bonafide research work carried under my supervision and I consider it worthy of consideration for award of the degree of Master of Science of the Institute.

Dated:

Dr. K. Sreenivasa Rao
Assistant Professor,
School of Information Technology,
Indian Institute of Technology, Kharagpur.

Declaration

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Sudhamay Maity

Dedicated To,

My beloved Parents,

Shri Sushil Kumar Maity

Smt. Rekha Maity

and

My Respected Mentor,

Dr. Krothapalli Sreenivasa Rao

Acknowledgements

First and foremost I offer my sincere gratitude to my supervisor, Dr K. Sreenivasa Rao, who has supported me through out my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my Masters degree to his encouragement and effort and without him this thesis, too, would not have been completed or written. It is indeed a great privilege and honour for me to have worked under his guidance which has made my research experience productive and cherish able.

I would like to thank all the faculty members of the School of Information Technology for the assistance they provided at all levels of the research project. I thank Prof. Indranil Sen Gupta and Prof. Jayanta Mukhopadhyay, the former and present Head, School of Information Technology.

My sincere thanks to speech researchers across India namely Prof. Yagnanarayana, Dr. Samudhra Vijaya, Dr. Kishore, Dr. Surya Kanth, Prof. Prasanna, Dr. Guru, Dr. Dhanu, Dr. P. Krishna Murthy, and Dr. Sri Rama Murty for sharing their valuable data, programs, ideas, support etc.

I am indebted to my colleagues for providing a stimulating and fun environment in which to learn and grow. I am especially grateful to Shashidhar G. Koolagudi, Ramu Reddy V., Krishnendu Ghosh, Narendra N. P, Anil Vuppala, Sourjya Sarkar, Dipanjan Nandi for their cooperation.

Lastly, and most importantly, I wish to thank my parents, Sushil Kumar Maity and Rekha Maity. They bore me, raised me, supported me, taught me, and loved me through my entire life. I would also like to thank my sister Prapti, uncle Sunil Kumar and best friend Aniruddha for their support and providing a loving environment for me.

Abstract

The basic goal of language identification (LID) system is to accurately identify the language from the given speech sample. In literature LID studies were carried out extensively on western and eastern languages by exploring various language-specific features and models. But, in Indian context, LID is analyzed with very few (less than six) Indian languages (ILs), and no systematic study was observed. Discrimination of ILs is a tough task due to high similarity of languages. This is mainly due to common origin (i.e., Sanskrit) for all ILs. In this work, we analyzed the LID performance on 27 Indian languages with spectral and prosodic features. In our study spectral features extracted from block processing (BP), pitch synchronous analysis (PSA) and glottal closure regions (GCR) are explored for identification of languages. It is observed that the performance of LID system is better by using the spectral features derived from PSA and GCR compared to BP. Prosodic features represented by intonation, rhythm and stress (IRS) are proposed at syllable and word levels for discriminating the language-specific information. For representing language-specific information at the global level, dynamics of fundamental frequency (F_0), duration and energy patterns are proposed in this study. Complementary information present in different features is exploited by combining the systems developed by using individual features. It is observed that the performance of LID system has been improved by combining various components of prosodic features and spectral features. The proposed features are also verified using OGI database, and observed the similar trend. The major contributions of our research work are given below :

- IITKGP-MLILSC (Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus) is developed for promoting LID research on ILs.
- Spectral features from pitch synchronous analysis and glottal closure regions are proposed for discriminating the ILs.
- Intonation, rhythm and stress features at individual syllable level and in the sequence of syllables within a word are proposed for identifying the ILs.

- Prosodic patterns at global level are proposed in terms of variation of F_0, intensity and duration patterns for recognizing the ILs.
- Language-specific information from the spectral and prosodic features is combined for improving the accuracy of LID system.

Key words:
Language Identification, Indian Languages, Pitch synchronous, Glottal closure region, prosody, Intonation, Rhythm, Stress.

Contents

List of Figures

List of Tables

List of Abbreviations

AANN - Auto-Associative Neural Network
ASR - Automatic Speech Recognition
AIR - All India Radio
BP - Block Processing
CV - Consonant-Vowel
DCT - Discrete Cosine Transform
DFT - Discrete Fourier Transform
EER - Equal Error Rate
FFNN - Feed-Forward Neural Network
FOD - First Order Difference
FOGD - First Order Gaussian Difference
GC - Glottal Closure
GCI - Glottal Closure Instants
GMM - Gaussian Mixture Models
GCR - Glottal Closure Region
HMM - Hidden Markov Model
Hz. - Hertz
IDFT - Inverse Discrete Fourier Transform
IL - Indian Language
IRS - Intonation Rhythm Stress
ISE - Instant of Significant Excitation
IITKGP-MLILSC - Indian Institute of Technology KharaGPur-
Multi-Lingual Indian Language Speech Corpus
LID - Language Identification
LP - Linear Prediction
LPCC - Linear Prediction Cepstral Coefficients
MFCC - Mel Frequency Cepstral Coefficients
ms - Milli Seconds
NIST-LRE - National Institute of Standards and Technology- Language Recognition Evaluation
OGI-MLTS - Oregon Graduate Institute database Multi-Language Telephone-based Speech
PLP - Perceptual Linear Prediction

PSA - Pitch Synchronous Analysis
SNR - Signal to Noise Ratio
SVM - Support Vector Machines
VT - Vocal Tract
VOP - Vowel Onset Point
VQ - Vector Quantization
ZFF - Zero Frequency Filter

Chapter 1

Introduction

Speech is mainly intended to convey some message. Message is conveyed through sequence of legal sound units, and the sequence should agree the constraints in certain language. Therefore, speech and language cannot be delinked. Every speaker has their own unique style of speaking and physiological characteristics of speech production. Hence, speaker-specific characteristics are also embedded in the speech signal. Thus speech signal not only contains the intended message but also the characteristics of language and speaker information.

Automatic spoken language identification (LID) is the goal of identifying the language from the short duration of speech signal uttered by an unknown speaker. In general, the applications of LID systems can be broadly divided into two categories, namely front-end for human operators and front-end for machines. As a front-end for human operators, LID system can be useful in routing calls to an appropriate human operator who is familiar to the language of the caller, whereas in the category front-end for machines, LID systems find many applications. In Speech-to-Speech translation system, in which the source language is to be converted to the target language. Here, LID system used as front-end to recognize the language of the input (source) speech signal. Another important application in this category is at public service points, where information access is enabled by voice. There are many benefits to be gained from making LID an automatic process that can be performed by a machine. Some of the benefits include the reduced costs and shorter training periods associated with using an automated system. For multiple human language identification services, several people would need to be trained to properly recognize a set of languages whereas

the LID system can be trained once and then run on multiple machines simultaneously. Therefore, it is not always possible to have trained operators who can assist in different languages, at all public service points. An automatic LID system can be extremely useful in such kind of places to provide assistance to public. In application like multi-language speech recognition system, where the language of the input signal is not known, it will be necessary to keep speech recognizers of several languages and run parallely, computationally which is very expensive. In this case, the LID system can act as front-end and is useful in reducing the number of speech recognizers to be activated, by considering the first n-best languages obtained from the LID system.

The LID system can be extremely useful in all the above mentioned applications, if the performance of the system is reasonably good. In the state-of-art literature, number of significant results are obtained in the area of language identification [1][2]. In taking advantage of recent developments, different identification systems based on Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), etc., have been widely used in the LID systems [3, 4, 5, 6]. The main features of an ideal language identification system are as given below:

- The time requirement to determine the identity of the test utterance should be small.
- The performance degradation must be graceful as the length of the utterance is reduced.
- The system should not be biased towards any language or group of languages.
- The system should tolerate speaker variation, accent variation, noisy/low SNR speech signal, channel and environment variation.
- The system should not be complex, which means that the amount of language specific information needed for developing the system should be small and inclusion of new language into existing system should be easy.

1.1 Cues for language identification

Human beings are the best known LID systems in the world today. Simply, by hearing one or two seconds of speech of a familiar language, they can easily extract the specific cues for identifying the language. Humans are endowed with the ability to integrate the knowledge from several sources to identify the language. Human beings learn a language over a period of time, and the level of language knowledge may vary and it depend on the factors like whether it is his/her native or first language, whether he/she has sufficient exposure and formal education in it. Humans use knowledge such as vocabulary, syntax, grammar and sentence structure to identify the language. It has been observed that humans often can identify the language of an utterance even they are not proficient in that language, suggesting that they are able to learn and identify language-specific patterns directly from the speech signal [7]. In the absence of higher level knowledge of a language, listener presumably relies on lower level constraints such as the prosody, phonetic repertoire and phontactics. Perceptual studies have revealed the importance of prosodic characteristics such as intonation, stress and rhythm are helpful in identifying the language, by humans [8][9]. Besides these cues, humans can also use the contextual knowledge about the speaker to identify the language.

The language part of the information contained in the speech signal is inferred using features at various levels. Even for the listener it is very difficult to describe language-specific features that he/she will be using for recognition. Thus these features are somewhat ambiguous and not unique, and they also depend on the listener. Any language follows a sequence of phones/sound units. The differences between the languages can be at several levels. Hierarchically, we can say frame level (10-30 ms), phone level, consonant-vowel (CV) level, syllable level, word level and phrase level. The possible differences among different languages at these levels are the frequency of occurrence of different units in each set, the sequence of units (Phonotactics) and their frequency of occurrence, the inventory, the acoustic signature, the duration of the same sound unit in different languages, and intonation patterns of units at higher levels. The performance of any LID system depends on the amount of information and the reliability of information extracted and how efficiently it is incorporated into the system.

There is a variety of information that humans and machines can use to distinguish one language from another. At a low level, speech features such as acoustic, phonetic, phonotactic and prosodic information are widely used in LID tasks. At a higher level, the difference between languages can be exploited, based on morphology and sentence syntax. Figure 1.1 depicts various levels of speech features from the raw speech that are available for language identification. When compared to the higher level speech features, low level acoustic features are easier to obtain, but volatile, because speaker or channel variations may be present. Higher level features, such as syntactic features, are believed to carry more language discriminative information [10], but they rely on the use of large vocabulary speech recognizers, and hence are hard to obtain. Figure 1.1 outlines commonly utilized levels of distinctions between different features spanning the range from low level to high level speech features. The different levels are elaborated below.

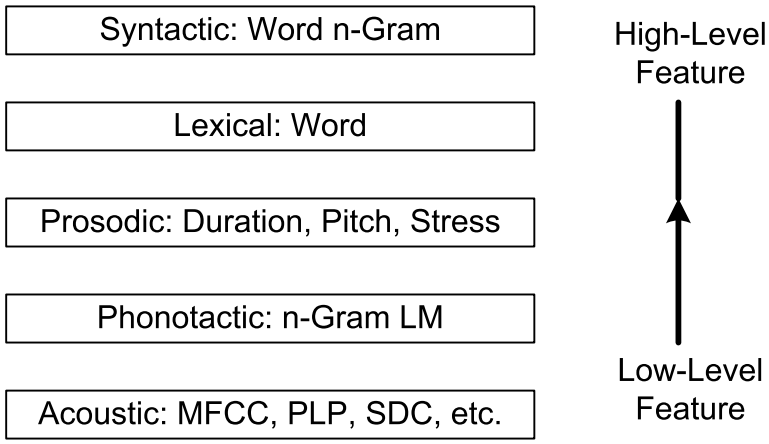

Figure 1.1: Levels of LID features

- Acoustic-phonetics: Acoustic information is generally considered as first level of analysis of speech production [11]. Human speech is a longitudinal pressure wave and different speech events can be distinguished at an acoustic level according to amplitude and frequency components of the waves [11]. Acoustic information is one of the

simplest forms of information which can be obtained during the speech parameterization process directly from raw speech. Also, higher level speech information such as phonotactic and word information can be extracted from the acoustic information. The most widely used parameterization techniques are Linear Prediction, Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) and Linear Prediction Cepstral Coefficient (LPCC) [12, 13]. Once the basic acoustic features have been obtained, additional features are appended to each feature vector with the intention of incorporating the temporal aspects of the speech signal. Some commonly utilized additional features are the delta and acceleration cepstrum and the Shifted Delta Cepstrum (SDC) [14]. Each sound unit in a language have unique articulatory configuration of the vocal tract system. However, there are some set of sound units may have significant overlap among the languages, but the same sound unit may differ across different languages due to different co articulation effects and dialects. These variations in acoustic realization of phonemes, forms the basis for the acoustic-phonetic studies.

- Phonotactics: There is a finite set of meaningful sounds that can be produced physically by humans. Not all of these sounds appear in any given language and each language has its own finite subset of meaningful sounds. Phonology is the study of the sound system of a specific language or set of languages and phonotactics is a branch of phonology that deals with the valid sound patterns in a specific language; i.e., the permissible combinations of phonemes (which are abstract sound units of a language that are capable of conveying distinctions in meaning) including consonant clusters and vowel sequences by means of phonotactical constraints [15]. Phonotactic rules, governing the way different phonemes are combined to form syllables or words, differ among languages. The sequence of allowable phonemes or syllables are different among different languages. Certain phoneme or syllable clusters common in one language may be rare in another language. There is a wide variance in phonotactic constraints across languages. For example, the phoneme cluster /st/ is very common in English, whereas it is not allowed in Japanese; the phoneme cluster /sr/ is not a legal cluster in English,

although it is very common in the Dravidian language Tamil; Japanese does not allow two adjacent consonants, but Danish and Swedish do. Hence the phonotactic information carries more language discriminative information than the phonemes themselves and therefore it is suitable for exploiting the characteristics of a language.

- Prosody: Prosody is one of the key components in human auditory perception. Tone, stress, duration, intensity and rhythm are the main aspects of prosody. The prosodic aspects vary from language to language. To utilize prosodic information, an appropriate quantitative representation is needed. Usually, pitch (or the fundamental frequency) is used for representing tone, intensity is used for indicating stress and duration sequence is used for representing rhythm. Some phonemes are shared across different languages and their duration characteristics will depend on the phonetic constraints of the language. Intonation is the variation of pitch when speaking. All languages use pitch semantically to convey surprise or irony, or to pose a question (for example in English, a sentence can be changed from a statement to a question by a rise in the tone at the end of a sentence). The pitch variations in tone are often used to identify Asian languages such as Mandarin Chinese, Vietnamese and Thai, where the intonation of a word determines meaning [16]. In some languages, a pattern of stress can determine the meaning of a word, for example in English a noun can become a verb by placing stress on different syllables. Also the stress pattern can be used to distinguish the languages with a word-final stress pattern (such as French) and the languages with a word-initial stress pattern (such as Hungarian) [15]. The manifestation of prosodic constraints of speech, conveys some important information regarding the language.

- Morphology: Morphology is the field of linguistics that studies the internal structure of words [17]. Words are generally considered to be the smallest units of syntax and in most languages words can be related to other words based on the morphology rules. The word roots and lexicons are usually different across different languages. In addition, different languages have their own sets of vocabularies and their own manner of forming words. As a result, the LID task can be performed at the word level by examining the characteristics of word forms.

- Syntax: In linguistics, syntax is the study of the principles and rules that govern the way that words in a sentence come together [18]. The sentence patterns vary across different languages. Even in the case when a single word is being shared by two languages, but the context (i.e., set of preceding and following words) may differ among languages [1]. It is known that integration of word based lexicon and grammars, exploiting morphological and syntactic information, lead to improvements in speech recognition systems and attempts to utilize such information in LID systems have been met with some success. However, constructing dictionaries and word based grammars for LID systems require a considerable extra effort when compared to the phonetic level. Systems that make use of morphological and syntactic information are currently not very common. LID systems typically consist of sub-systems that make use of some or all of the above mentioned types of information to estimate some measure of similarity to the different languages considered (such as likelihoods) and these measures from the various sub-systems are then fused/combined to make the final decision about language identity. Figure 1.2 shows a block diagram of a generic LID system that makes use of all levels of information. However, it is not necessary that an LID system do so, and in fact most LID systems do not. The most popular approach is to use acoustic and phonotactic information.

1.2 Types of language identification systems

Language identification can be text-based or speech signal-based. While text-based LID is a solved problem, the signal-based language identification remains an active area of research. All the spoken LID systems can be broadly divided into two types, explicit and implicit LID systems.

1.2.1 Explicit LID systems

The LID systems that require speech recognizers of one or several languages, in other words, the systems that require a segmented and labeled speech corpus are termed here as Explicit LID systems. A simple (but harder to implement) block diagram of an explicit LID system

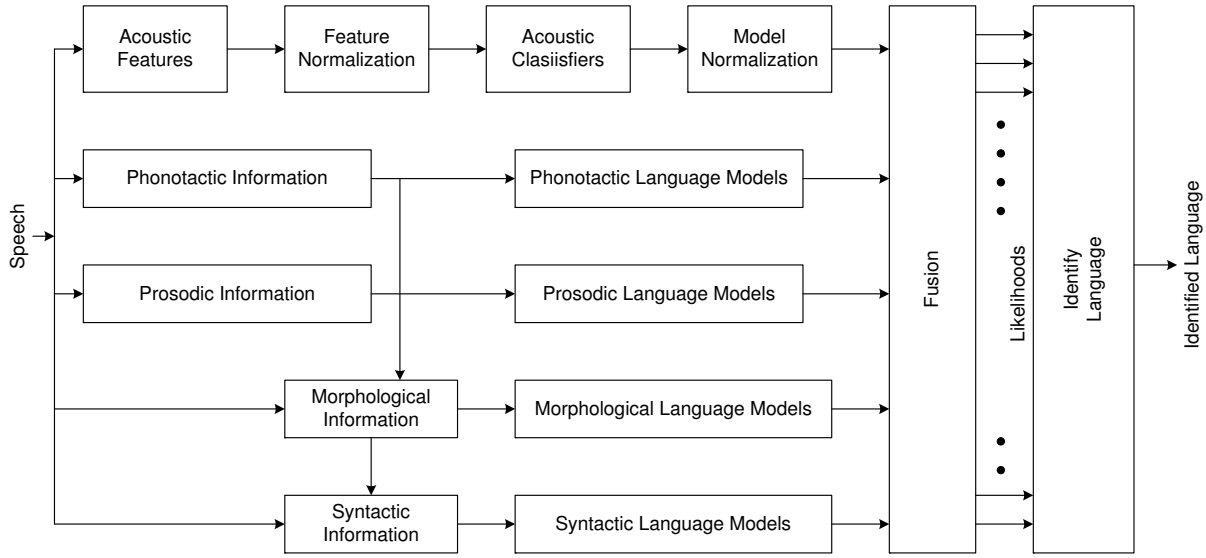

Figure 1.2: Block diagram of generic LID system using the features from various levels

is given in Figure 1.3. More details of different explicit LID systems is presented in the following chapter.

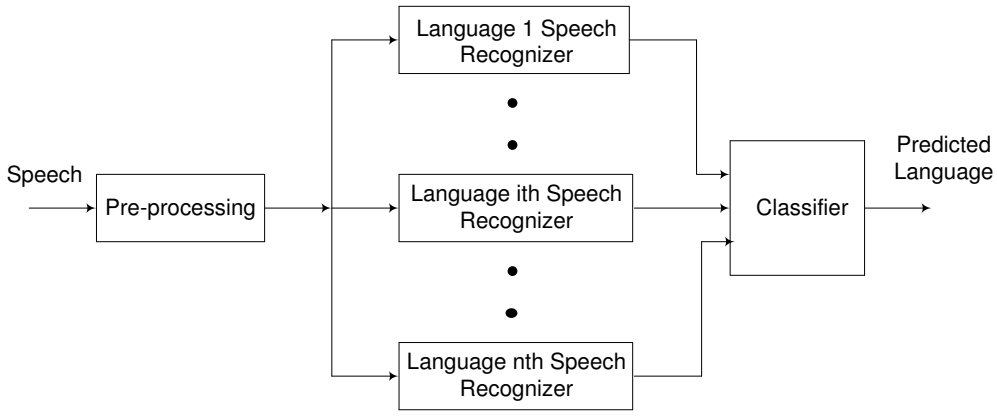

Figure 1.3: Explicit language identification system

1.2.2 Implicit LID systems

The language identification systems which do not require segmented and labeled speech data (or rather speech recognizers), are termed here as Implicit LID systems. In other words, these systems require only the raw speech data along with the true identity of the language spoken. Here, the language models or the language-specific information are derived only from the raw speech data. A generalized block diagram of an implicit LID system is shown in Figure 1.4. The existing implicit LID systems differ mainly in the feature extraction stage, since the type of feature selected for discriminating languages may be different. The details of different implicit LID systems is presented in the next chapter.

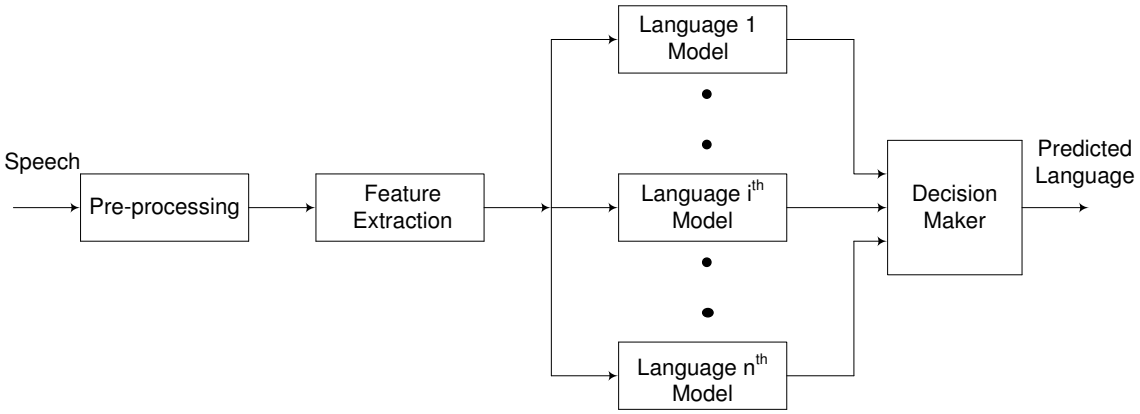

Figure 1.4: Implicit language identification system

1.3 Challenging issues in automatic language identification

Identification of language automatically without understanding the language is a challenging problem. In the language identification domain, it should be assumed that no test speaker's spectral or any other type of information is present in the reference set. The comparison between the test sample and the reference samples is always from unconstrained utterances of two different speakers. Therefore, between the two utterances there exist differences such as text, speaker, environment and language. Hence, for a reliable spoken language identification

system deriving the language differences apart from text, speaker and environment differences is the main problem. The following are some important issues in automatic language identification.

- Variation in speaker characteristics: Different speakers have different speaking styles which give rise to large amount of speaker variability, even within the same language constraints. Thus it is essential to nullify the speaker variation while doing language modeling.

- Variation in accents: Accent is mainly related to pronunciation. From the accent one can easily distinguish whether the person is native speaker or not. For example, a native of Kannada fairly fluent in Telugu may speak Telugu with Kannada accent. But, it is difficult to describe the difference in accents.

- Variation in environment and channel characteristics: Characteristics of speech signal has strong influence on the environmental conditions where the data is collected and on channel where it is transmitted. These factors can have significant effect on the features derived from the short time spectral analysis. Therefore, it is necessary to have the features which are less affected by the environment and channel to achieve the robustness needed for language identification system.

- Variation in dialects: Dialect is a regional or social variety of a language distinguished by pronunciation, grammar, or vocabulary, especially a variety of speech differing from the standard literary language or speech pattern of the culture in which it exists. For example, there is variation of the Telugu language spoken by different areas of Andhra Pradesh.

- Similarities of languages: There is a lot of similarity in Indian languages as most of the languages are derived from the common source Sanskrit. Most of the Indian languages have common set of root words and also follows similar grammatical structure. Therefore, discriminating the languages in Indian context is a great challenging task.

- Extraction and representation of language-specific prosody: Prosodic features such as intonation, duration, intensity, stress and rhythm patterns vary among different lan-

guages. But the nature of these characteristics are not well defined. For example, the rhythm of a language may be felt due to succession of syllables, vowels, sudden amplitude bursts, rise or fall nature of intonation patterns, which is not well understood. Moreover, there are no proper techniques available for speech processing to represent the high level source knowledge like prosody. Therefore, extraction and representation of language-specific prosody is difficult.

From the literature, it is observed that, identifying the language of the unknown speech utterance will be comparatively simple, if the languages are very different from each other (i.e., the phoneme set is quite different across the languages). But as far as Indian languages are concerned, even though there are few unique sounds, all the languages share a common set of phonemes, since most of the languages have the same origin. Therefore, developing a LID system for Indian languages is a challenging task.

1.4 Objective and scope of the work

The aim of the thesis is to develop an efficient language identification system for Indian languages. Building an efficient LI system for Indian languages is lacking due to unavailability of proper database covering majority of languages. In the context of Indian languages most of the works related to LI systems are limited to 4 languages. So to overcome this deficiency, we have initiated the task of developing multilingual speech corpus in Indian languages. We have collected speech corpus for 27 Indian languages and details of the database are provided in Chapter 3. The language-specific characteristics of the speech can be attributed to the characteristics of vocal tract system, behavioral characters of speaker, excitation source characteristics and suprasegmental characteristics. Language specific vocal tract information is mainly represented by spectral features such as Mel-frequency cepstral coefficients (MFCC) or linear prediction cepstral coefficients (LPCC). The parameters like pitch, duration, energy, rhythm and stress can be used as basic prosodic features to represent language specific prosodic information. Therefore, in this thesis language identification system for Indian languages is developed using above mentioned spectral and prosodic features. Since, ILs are derived from same origin, for discriminating them, finer variation in spectral characteristic

needed to be exploited in addition to conventional spectral features derived from sequence of frames. Similarly for investigating the discriminative information in prosodic features, in addition to gross statistics of prosody, the dynamics of prosody at different levels such as syllable, word and phrase need to be examined carefully. For improving the discrimination among the languages, evidences from different sources may be combined to exploit the complementary information present in various sources.

1.5 Issues addressed in the thesis

- One of the major issues for the lack of LID research in Indian languages is due to unavailability of proper database covering majority of languages. In this thesis, the database issue is addressed by developing IITKGP-MLILSC (Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus) speech database covering major 27 Indian languages and it will made available for public for promoting LID research on Indian languages.

- In implicit LID systems, the performance mainly depend on speech features representing the language-specific information. In this work, this issue is addressed by proposing spectral features from pitch synchronous analysis and glottal closure regions for discriminating the Indian languages. The proposed spectral features may represent finer variation of VT spectrum present in successive pitch cycles which may provide the efficient language discrimination. In the context of prosodic features, intonation, rhythm and stress features are proposed at syllable and word levels for discriminating the languages. For capturing the language-specific information at the global level dynamics of duration, intonation and energy patterns are exploited at phrase level.

- For improving the LID accuracy further, the evidences from the proposed features (Prosodic features at syllable, word and phrase level, proposed spectral features) may be combined. The complementary information related to language discrimination if at all present in different features will enhance the overall identification performance.

1.6 Organization of the thesis

- Chapter 1 provides brief introduction about language identification and its applications. General issues in language identification, language-specific cues in speech signal, specific issues in identification of Indian languages, scope of the work, issues addressed and organization of the thesis are discussed in this chapter.

- Chapter 2, briefly reviews about explicit and implicit LID systems present in the literature. Existing works related to LID on ILs are briefly discussed. Various speech features and models proposed in the context of LID are briefly reviewed in this chapter. Motivation for the present work from the existing literature is briefly discussed.

- Chapter 3, introduce about multilingual IL speech corpus developed in this work for analyzing the LID performance. Speaker-dependent and independent language models are also discussed in view of LID. Spectral features extracted from conventional block processing, pitch synchronous analysis, and glottal closure regions are examined for discriminating the languages.

- Chapter 4, discusses about language identification using intonation, rhythm and stress features derived from syllable and word levels. Global prosodic features are also examined for discriminating the languages. Complimentary evidences from global, word and syllable level prosodic features are exploited for improving the performance. Finally, spectral and prosodic features are combined for further improving the performance.

- Chapter 5, summarizes the thesis contributions and provides the future extensions to present work.

Chapter 2

Literature review

This chapter provides an overview of existing language identification systems. In early 1970's, research in automatic spoken language identification was started. But, the progress in this area of research was slow for almost two decades. After that, with the advent of public-domain multi-lingual corpora for speech, many researchers started showing interest in this area and lot of effort and progress have been made [19][20]. All the existing LID systems use some amount of language-specific information either explicitly or implicitly and the amount of language-specific information used in them differs in both explicit and implicit LID systems. The performance and the complexity of the system are dependent on the amount of linguistic information supplied to the system and it is proportional. While training, some systems require only the speech signal and the true identity of the language. In these systems, language models are derived only from the speech data which is supplied during training. More complicated LID systems may require segmented and labeled speech signal of all the languages under consideration. Although the performance of the more complicated LID systems is superior to others. Adding a new language into more complicated LID systems is not a trivial task. Therefore, the trade off between performance and simplicity has become inevitable if the number of languages under consideration is large. A few representative LID systems in explicit and implicit groups are described in this chapter.

This chapter is organized as follows: An overview of explicit language identification systems is presented in section 2.1. Section 2.2 describes the existing implicit language identification systems. Importance of implicit LID systems over explicit LID systems are

given in section 2.3. Motivation for carrying out the present work is discussed in section 2.4. Summary and conclusions of this chapter is presented in section 2.5

2.1 Review of explicit LID systems

Muthusamy *et al* have done the perceptual experiments on language identification task and observed that the knowledge of a particular language (i.e, linguistic and syntactic rules) will definitely help in identifying or discriminating the languages [21]. From this, it can be interpreted that even for an automatic system, if speech recognizers of all the languages to be identified are used as front-ends, the performance in classifying the languages will be better. For developing a speech recognizer for any language, the basic requirement is, the segmented and labeled speech corpus.

Lamel and Gauvain used phone recognizers as front-end for language identification task [22][23]. Phone recognizers for the languages French and English are built and used in parallel. The unknown speech signal from any of these two languages is processed by the two phone recognizers in parallel. The language associated with the model having the highest likelihood is declared as the language of the unknown speech signal.

Berkling *et al.* [24], have considered a superset of phonemes of three different languages like English, Japanese, and German. They have explored the possibility of finding and using only those phones that best discriminate between language pairs.

Hazen and Zue [25] pursued a single multi-language front-end phone recognizer instead of language-dependent phone recognizer and incorporated the phonetic, acoustic, and prosodic information derived from speech within a probabilistic framework.

Andersen *et al.* [26] have grouped the total inventory of phonemes into a number of groups, one which contains language-independent phones and three language-dependent phone inventories for the three languages under consideration and tried to classify the languages.

Tucker *et al* [27] have utilized a single language phone recognizer to label multi-lingual training speech corpora, which have then been used to train language-dependent phone recognizers for language identification. They used three techniques such as the acoustic

difference between phonemes of each language, the relative frequencies of the phonemes of each language and the combination of above two sources of information for classifying the languages.

Zissman and Singer [28] used a single English phoneme recognizer and proved that it was feasible to model phonotactic constraints with the information of phoneme inventory from one language. Zissman [1] used a single language-dependent phone recognizer to convert the input speech signal to a sequence of phones and used the statistics of the resulting symbol sequences for language identification, which is termed as Phone Recognition followed by Language Modeling (PRLM). He has further extended it using multiple language-dependent phone recognizers in parallel (Parallel-PRLM) and achieved a reasonable improvement in the language identification performance.

Kadambe and Hieronymus [29] demonstrated that the performance of an LID system which is based only on acoustic models can be improved by incorporating higher level linguistic knowledge in the form of trigram and lexical matching. This system is also based on the Parallel Phone Recognition (PPR) approach.

Yan and Bernard [30] extended the Parallel Phone Recognition (PPR) approach which was discussed in [1] for six languages with refined bigram models and context-dependent duration models. For combining the evidence derived from the scores of acoustic model, language model, and duration model, they proposed a neural network based approach.

Navratil and Zhulke [31] have also used a single language-independent phone recognizer but have used improved language models instead of standard bigram models. The aim of using a single language-independent phone recognizer is to reduce the computation complexity introduced by the parallel phoneme tokenizers. Navratil improved the PPRLM LID system by using the binary-tree (BT) structures and acoustic pronunciation models instead of the traditional N-gram language models [32]. Two approaches of BT estimation are proposed - building the whole tree for each class in one case, and adapting from a universal background model (UBM) in the other case. The resulting system serves for language identification as well as for unknown language rejection, and achieved the error rates of 9.7% and 1.9% on the 1995 NIST (based on OGI-TS corpus) six-language identification task and 14.9% and 5.1% on the nine-language task for 10-sec and 45-sec test utterances respectively.

Hazen and Zue [33] proposed a phonotactic based LID system with a single decoder with a multilingual phoneme repertory and a variable number of phoneme units. The phonotactic classifiers use multiple phone recognizers as the front-end to derive phonotactic statistics of a language. Since the individual phone recognizers are trained on different languages, they capture different acoustic characteristics from the speech data. From a broader perspective, characterization of the spoken language is possible by combining these recognizers to form a parallel phone recognizer (PPR) front-end. GMM was applied to model the phonetic class in a segment-based approach. They achieved LID accuracy rates of about 50% measured on the 1994 NIST evaluation dataset, compared to about 70% achieved by a phonotactic component on the same data.

Kirchhoff and Parandekar [34] made an interesting study for modeling the cross-stream dependencies for the phonotactic based LID systems. In their approach, a multi-stream system was used to model the phonotactic constraints within as well as across multiple streams.

Prasad [35] developed a language-independent syllable recognizer for a two language task and tried using it as a front-end, in which the syllable statistics are used to determine language identity. Ramasubramanian *et al* [36] have studied the PPR system in detail for a 6-language task. They made a study on three different classifiers such as the maximum likelihood classifier, Gaussian classifier, and K-nearest neighbor classifier. For each classifier they have explored with three different scores, namely, acoustic score, language-model score, and joint acoustic-language model score and concluded that maximum likelihood classifier with acoustic likelihood score gives best LID accuracy.

Gauvain *et al.* [37] proposed another approach of generating a multitude of streams with the use of phoneme lattices. The use of phoneme lattices has been shown to significantly improve the performance of PPRLM systems when compared to the 1-best approach of considering only one phoneme (token) sequence [37, 38]. Gleason and Zissman [39] described two enhancements to the Parallel PRLM (PPRLM) system by using the composite background (CBG) modeling and score standardization.

In order to model reliably a longer time span than the traditional PPRLM (to model 5-gram instead of trigram), Cordoba *et al.* [40] presented an approach for language identi-

fication based on the text categorization technique. With the parallel phoneme recognizer as the front-end, the N-gram frequency ranking technique was used instead of the language model. The resulting LID system is capable of modeling the long span dependencies (4-gram or even 5-gram), which could not be modeled appropriately by the traditional N-gram language model, probably due to insufficient training data. The proposed parallel phoneme recognition followed by n-gram frequency ranking achieved a 6% relative improvement compared to the PPRLM LID system.

Li *et al.* [41] proposed to use a "universal phoneme recognizer", which was trained to recognize 258 phonemes from 6 languages (English, German, Hindi, Japanese, Mandarin and Spanish). For the back-end, both the N-gram models and the vector space modeling (VSM) were adopted to make a pair-wise decision. This PPR-VSM LID system achieved an EER of 2.75% and 4.02% on 30-sec test utterances for 1996 NIST LRE and 2003 NIST LRE respectively.

Sim and Li [42] improved the PPRLM based LID system by using the acoustic diversification as an alternative acoustic modeling technique. Unlike the standard PPRLM systems where the subsystems are derived using language dependent phoneme sets to provide phonetic diversification, the proposed method aims at improving the acoustic diversification among the parallel subsystems by using multiple acoustic models. By combining the phonetic and acoustic diversification (PAD), the resulting LID system achieved EERs of 4.71% and 8.61% on the 2003 and 2005 NIST LRE data sets respectively.

Tong *et al.* [43] proposed a target-oriented phone tokenizers (TOPT) that uses the same phone recognizer for different target languages in the PPR front end. For example, Arabic-oriented English phone tokenizer, Mandarin- oriented English phone tokenizer, as Arabic and Mandarin each is believed to have its unique phonotactic features to an English listener. Note that not all the phones and their phonotactics in the target language may provide equally discriminative information to the listener, it is therefore desirable that the phones in each of the TOPTs can be those identified from the entire phone inventory, and having the highest discriminative ability in telling the target language from other languages.

You *et al* [44] have used morphological information, including letter or letter-chunk N-grams, to enhance the performance of language identification in conjunction with web-based

page counts. Six languages, namely, English, German, French, Portuguese, Chinese, and Japanese are tested. Experiments show that when classifying four Latin languages, including English, German, French, and Portuguese, which are written in Latin alphabets, features from different information sources yield substantial performance improvements in the classification accuracy over a letter 4-gram-based baseline system. The accuracy increases from 75.0% to 86.3%, or a 45.2% relative error reduction.

Botha and Barnard [45] have exploited effects and the relations of different factors such as size of text fragment, amount of training data, classification features and algorithm employed, and similarity of the languages which affect the accuracy of text-based language identification. They have used 11 official languages of South Africa. Within these languages distinct language families can be found. They found that it is much more difficult to discriminate languages within languages families than languages in different families. They have used n-gram statistics as features for classification The relationship between the amount of training data and the accuracy achieved is found to depend on the windowsize: for the largest window (300 characters) about 400000 characters are sufficient to achieve close-to-optimal accuracy,whereas improvements in accuracy are found even beyond 1.6million characters of training data for smaller windows.

2.2 Review of implicit LID systems

The NIST language recognition evaluation conducted in the years 1996 [46] and 2003 [47], shows that the approaches based on a bank of parallel-phone recognizers of multiple languages were the best performing systems. But, some problems can be anticipated with these systems when the number of languages to be identified is increased. In [1], it is shown that, parallel-PRLM performs better than PRLM, in which a single phone recognizer is used as a front-end. Further, it is shown that reducing the number of channels (phone recognizers) has a strong negative effect in the performance. From this one can conclude that the performance of the Parallel-PRLM system is proportional to the number of speech recognizers used in parallel in the system. When the number of languages to be recognized is increased, the number of phone recognizers may need to be increased further. But developing a single phone recognizer

with reasonable performance itself is not a trivial task. This clearly shows that, attention should be given to developing language identification systems which do not require phone recognizers. The existing implicit LID systems differ mainly in the feature extraction stage, since the type of feature selected for discriminating languages may be different. Some of the representative implicit LID systems in the literature are discussed below.

The earliest sustained effort in automatic spoken LID systems were reported by Leonard and Doddington [48] at Texas Instruments (TI) labs. House and Neuburg [49] have grouped all the speech samples into five broad linguistic categories, with the assumption that it is possible to identify gross linguistic categories with great accuracy. They proposed the first HMM based language identification system. Cimarusti and Eves [50] showed that a pattern classifier approach can be used for language identification. They ran the experiment using a 100 dimensional LPC derived feature vector. Eady [51] performed a two language identification task (English and Mandarin Chinese) by examining the differences in the F_0 patterns.

Ives [52] extended the previous study by developing a rule-based LID system using an extended multi-lingual corpus. Foil [53] examined both formant and prosodic feature vectors and found that formant features were generally superior. The formant vector based language identification system used k-means clustering. Goodman and *et al* [54] have extended Foil's work by refining the formant feature vector.

Muthusamy *et al* [55] have suggested a segment based approach to identify the language, where the speech is first segmented into seven broad phonetic categories using a multi-language neural network based system with the basic idea that the acoustic structure of any language can be estimated by segmenting the speech into broad phonetic categories.

Sugiyama [56] performed vector quantization classification on acoustic features such as LPC coefficients, autocorrelation coefficients and delta cepstral coefficients. The difference between using one VQ code book per language versus one common code book was explored in [56]. In the latter case, languages were classified according to their VQ histograms. Riek [57], Nakagawa [4] and Zissman [58] used a Gaussian mixture classifier for language identification based on the observation that different languages have different sounds and sound frequencies.

Itahashi *et al.* [59], and Itahashi and Liang [60] proposed LID systems based on fundamental frequency and energy contours with the modeling using a piecewise-linear function. Li [61] proposed a LID system which is based on features extracted at syllable level. In this system, the syllable nuclei (vowels) for each speech utterance are located automatically. Next, feature vectors containing spectral information are computed for regions near the syllable nuclei. Each of these vectors consists of spectral sub-vectors computed on neighboring frames of speech data. Rather than collecting and modeling these vectors over all training speech, Li keeps separate collections of feature vectors for each training speaker. During testing, syllable nuclei of the test utterance are located and feature vector extraction is performed. Each speaker-dependent set of training feature vectors is compared to feature vectors of the test utterance, and most similar speaker-dependent set of training vectors is found.

Pellegrino and Andre-Abrecht [62] proposed an unsupervised approach to LID, in which each language vowel system is modeled by GMM trained with automatically detected vowels. Even though GMM for language identification is well experimented in [1] [63][64], the difference here is, the language models are generated using vowels alone. Corredor *et al.* [65], Dalsgaard *et al.* [66], Lamel and Gauvain [23], Pellegrino *et al.* [67], Ueda and Nakagawa [68], and Zissman [1] did extensive studies on LID using HMMs. Due to its abilities to capture temporal information in human speech, HMMs represent a natural bridge between the purely acoustic approaches and the phonotactic approaches [32, 1].

Cole *et al.* [3] applied ANN in the form of a multilayer perceptron trained by the PLP features. Braun and Levkowitz [69] described the use of the recurrent neural networks (RNNs) for the LID task. The GMM-UBM method was proposed for language identification by Wong *et al.* [5] and has gained momentum and become one of the dominant techniques for acoustic based language identification. Campbell *et al.* [70], Zhai *et al.* [6] and Castaldo *et al.* [71] applied SVMs for the language identification task and showed improved results compared to the GMM based approach. In a more recent development Noor and Aronowitz [72] combined the anchor models with the SVM.

In [63], Gaussian Mixture Model (GMM) is used to tokenize the speech signal and language models are derived from the sequence of tokens. Similar to phone recognizers in the

PRLM system, the GMM tokenizer is trained on one language but is used to decode information for any candidate language. The decoded sequence is then used to train bigram models. Further, a parallel GMM tokenization system has been tried, where for each language, a separate GMM tokenizer is used. It is shown that as the number of GMM tokenizers is increased beyond a level, there is a degradation in the performance. In this work Callfriend [20] corpus is used for both training and testing.

Lin and Wang [73] proposed the use of a dynamic model in ergodic topology with the input of the temporal information of prosodic features. Rouas *et al.* developed a LID system with only prosodic features, where a set of rhythmic parameters and fundamental frequency parameters were extracted. They later improved this with a modified algorithm of rhythm extraction and several prosodic parameters were extracted (consonantal and vowel durations, cluster complexity) and were modeled by the GMM [74]. The resulting LID system achieved a language identification rate of 67% on a 7-language task.

Chung-Hsien [75] *et al* have done segmenting and identification of mixed languages. A delta Bayesian information criterion (delta-BIC) is firstly applied to segment the input speech utterance into a sequence of language-dependent segments using acoustic features. A VQ-based bi-gram model is used to characterize the acoustic-phonetic dynamics of two consecutive codewords in a language. A Gaussian mixture model (GMM) is used to model codeword occurrence vectors orthogonally transformed using latent semantic analysis (LSA) for each language-dependent segment. They have achieved language identification accuracies of 92.1% and 74.9% for single-language and mixed-language speech, respectively.

Rouas [76] also implemented an LID system based on the modeling of the prosodic variations, which was achieved by the separation of phrase and accentual components of intonation. An independent coding of phrase and accentual components is proposed on differentiated scales of duration. Short-term and long-term language-dependent sequences of labels are modeled by n-gram models. The performance of the system is demonstrated by experiments on read speech and evaluated by experiments on spontaneous speech. An experiment is described on the discrimination of Arabic dialects, for which there is a lack of linguistic studies, notably on prosodic comparisons.

Siu *et al* [77] have used discriminative GMMs for language identification using boosting

methods. The effectiveness of boosting variation comes from the emphasis on working with the misclassified data to achieve discriminatively trained models. The discriminative GMMs aproach is applied on the 12-language NIST 2003 language identification task.

Sangwan *et al* [78] have done analysis and language classification based on speech production knowledge. The speech utterance is first parsed into consonant and vowel clusters. Subsequently, the production traits for each cluster is represented by the corresponding temporal evolution of speech articulatory states. Evaluation is carried out on South Indian Languages, namely, Kannada, Tamil, Telegu, Malayalam, and Marathi which are closely related. Accuracy rate of 65% was achieved with about 4 sec of train and test speech data per utterance.

Martnez *etal* [79] have extracted prosodic features such as intonation, rhythm and stress and classified the language with a generative classifier based on iVectors.

In the context of Indian languages, very few attempts are reported in the area of language identification. First attempt has been made on Indian languages by Jyotsna *et al.*, [80]. As Sugiyama, they also performed VQ classification on four Indian languages using 17 dimensional Mel Frequency Cepstral Coeffcients (MFCC). They observed that the acoustic realization of the same sound unit from different Indian languages are different and some sound units (key sounds) contribute to the performance of the LID system. Nagarajan et. al., [81] have explored different code book methods for building LID system. Later using automated segmentation of speech into syllable like units and parallel syllable like unit recognition are used to build implicit language identification system. In [2], Sai Jayaram and *et al* have used a parallel sub-word unit recognition (PSWR) approach for LID, where the sub-word unit models are trained without using segmented and labeled speech data. In this work, first, the training utterances are segmented into acoustic segments automatically and clustered using K-means algorithm. After clustering, HMMs for each class are generated. The rest of the work is similar to the PPR approach discussed in [1] and [36]. It is claimed that the language identification performance for this system is almost the same as that of the system which uses phone recognizers in parallel [36]. The resulting PSWR LID system achieved an LID accuracy of 70% on a six-language task (based on OGI-TS corpus) for 45-sec test utterances. Mary et. al., [82][83] have explored spectral features with autoassociative

neural network models for language identification with varying durations of test speech samples.In their later work [84], they focused on prosodic (intonation, rhythm and stress) and syllabic features for language recognition.

2.3 Reasons for attraction towards implicit LID systems

The most successful approach to LID uses the phone recognizer of one language or several languages as a front-end. Zissman [1] has compared the performance of four approaches for LID and argued that LID systems using Parallel Phone Recognition (PPR) approach will outperform other systems. But the basic requirement for this approach is the availability of phone recognizers of all the languages to be identified. To develop a phone recognizer for any language, a segmented and labeled speech corpus is necessary. Building segmented and labeled speech corpora for all the languages to be recognized, is both time consuming and expensive, requiring trained human annotators and substantial amount of supervision [85]. The other approaches tried in [1], do not require segmented and labeled speech corpora for all the languages to be recognized. Phone recognition followed by language modeling (PRLM) system explained in [1] requires only phone recognizer for any one language. Since all the sounds in the languages to be identified do not always occur in a single language used in the PRLM approach, it seems natural to look for a way to incorporate phones from more than one language into a PRLM system [1]. This approach, which uses more than one language speech recognizers, is referred to as Parallel-PRLM. Parallel-PRLM approach requires that labeled speech be available in more than one language, although the labeled training speech does not need to be available for all, or even any, of the languages to be identified. The performance of the Parallel-PRLM approach seems to be better than PRLM approach. But, as mentioned in the previous Section, the performance of the Parallel-PRLM system for language identification is proportional to the number of phone recognizers used in parallel.

Even though the performance of the PPR approach or the Parallel-PRLM approach is

impressive, non availability of segmented and labeled speech corpora of all the languages to be recognized, makes the implementation of PPR approach, harder. In all the above mentioned approaches (PPR, Parallel-PRLM), phone recognizers of several languages need to be used as a front-end for LID. The speech recognizer which will be used as a front-end for LID, should be capable of handling two mismatches; the channel mismatch and the language mismatch. Even though both of these mismatches can be handled to certain extent, the performance may not be optimal. The worse scenario is the non availability of speech recognizers of any language. The difficulties in implementing LID systems which rely on speech recognizers as the front-end, or the non availability of any of the speech recognizers, makes implicit LID systems more attractive, even though the performance is slightly inferior to that of explicit LID systems.

2.4 Motivation for the present work

The phenomena of globalization has brought people together from different parts of India. However, one of the barriers to global communication is due to existence of several languages and different people speaks in different languages. There are several hundreds of mother tongues exist in India. Almost 30 languages are spoken by more than a million native speakers. For an effective communication among the people across the states, appropriate speech interface is required. Here, the basic goal of the speech interface is to automatically transform the speech from source language to the desired target language without loss of any information. For example, a person from Tamilnadu can communicate with a person in Kashmir in their respective mother tongues. Here, speech interface has to identify the source and target languages based on initial speech samples, and then it has to transform the speech from source to target language. Speech transformation may be carried out with the following sequence of speech tasks : (i) speech to text transformation, (ii) source text to target text conversion and (iii) synthesizing the speech from target text. Therefore for speech to speech transformation language identification has to be performed first.

In Indian context, there is no systematic study on identification of ILs. The existing LID studies dealt with only 4 ILs. At present, there is no standard speech corpus covering

majority of ILs. But, for initiating LID task on ILs, standard speech corpus covering all ILs is essential. Therefore, in this work we have developed the IL speech corpus covering 27 ILs spoken by majority community. Due to complexity in accessing the resources for all 27 ILs, we have attempted implicit LID on ILs. For exploring the features to represent language-specific information, most of the existing works are based on spectral features extracted using conventional block processing and prosodic features extracted from phrases. Since, all the ILs are originated from Sanskrit, they have lot of similar characteristics in various aspects and hence the difficulty in discriminating them using conventional features. Therefore, in this work we have explored spectral features extracted from PSA and GCR for representing the language-specific information. Similarly in view of prosodic features we have explored intonation, rhythm and stress (IRS) features at syllable and word levels in addition to global level prosodic features. For minimizing the confuseability among the languages, complementary evidences from various features are combined at various phases.

2.5 Summary and conclusions

This chapter summarizes the existing philosophies in LID. Existing works on explicit and implicit LID are discussed in terms of features and models for capturing the language-specific information and fusion techniques for improving the performance. Tendency of present research on LID towards implicit LID task is briefly discussed. Existing works on LID in Indian context are briefly discussed. Finally the motivation for the present work is described.

Chapter 3

Language identification using spectral features

From the review of language identification discussed in Chapter 2, it is clear that the choice between explicit and implicit systems for language identification (LID) is a compromise between complexity and performance. The explicit LID systems that make use of phonotactics and word level knowledge perform better than implicit systems which rely on acoustic-phonetics and prosody. But the higher performance of explicit LID systems is achieved at the cost of additional complexity of using a subword unit recognizer at the front end. For building such a subword unit recognizer, invariably one requires manually labeled corpora for number of languages. Such a corpora may not be available in many languages. This is especially true for Indian languages, where a large number of languages are spoken even within a small geographical area. Therefore, the LID system that operates on features derived directly from the speech signal is appropriate in Indian context.

Language identification task involve three stages namely, feature extraction, modeling and evaluation. Feature extraction deals with extraction of language-specific features from the speech signal. Appropriate models are developed using features obtained from the training data. The models are evaluated using the features derived from a test utterance for identifying the language. The performance of language identification systems are influenced by all the three stages, namely, feature extraction, model building and evaluation strategies. An LID system must exploit the primary differences which exist among the languages, while

still being robust for variation due to speaker, channel and vocabulary. The system also needs to be computationally efficient. Thus, it is desirable to determine language discriminating characteristics which are easy to extract from the acoustic signal and are relatively free from variations due to speaker, channel and vocabulary. Each language has a set of phonemes. Phonemes are combined to form syllables. As the vocal apparatus used for the production of languages is universal, there will be considerable overlap of the phoneme and syllable sets. Also there are differences in the way the same phoneme or syllable is pronounced in different languages. Such variations can be represented using *acoustic-phonetic* features represented by spectral features. In this chapter, language identification is carried out using spectral features such as linear prediction cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC). The language-specific information from the extracted features is captured using Gaussian mixture models (GMMs). For extracting the spectral features, in addition to conventional block processing approach (frame level features using 20 ms as frame size and 10 ms frame shift), pitch synchronous and glottal closure based approach are explored for deriving the finer variations of spectral characteristics present in successive pitch cycles.

The chapter is organized as follows: The details of speech databases used in this work for LID task are presented in section 3.1. Section 3.2 discuss the features used for automatic language identification. Development of language models using Gaussian mixture models is explained in section 3.3. Performance of LID systems in speaker dependant and independent models is dicussed in Section 3.4. Section. 3.5 presents language identification performance using the proposed spectral features. Performance evaluation of proposed PSA and GCR based spectral features on OGI-MLTS database is presented in section 3.6. Final section discuss the summary and conclusions of this chapter.

3.1 Speech databases

In this work, two speech databases namely, Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus (IITKGP-MLILSC) and Oregon Graduate Institute (OGI) multi-language telephone-based speech (MLTS) corpus are used for identification of

languages. The details of the databases are given in the following subsections.

3.1.1 Indian Institute of Technology Kharagpur Multi-lingual Indian Language Speech Corpus (IITKGP-MLILSC)

The proposed IITKGP-MLILSC database is recorded from broadcast television channels using DISH-TV direct to home connection. For some languages, where TV broadcast channels are not available, broadcast radio channels are used for collecting the speech. The speech corpus is recorded from news bulletins, interviews, live shows and talk shows. Around 80% of the database is collected from news bulletins. The Indian Institute of Technology Kharagpur multi-lingual Indian language speech corpus (IITKGP-MLILSC) covers 27 Indian languages, namely, Arunachali, Assamese, Bengali, Bhojpuri, Chhattisgarhi, Dogri, Gojri, Gujrati, Hindi, Indian English, Kannada, Kashmiri, Konkani, Manipuri, Mizo, Malayalam, Marathi, Nagamese, Neplai, Oriya, Punjabi, Rajasthani, Sanskrit, Sindhi, Tamil, Telugu and Urdu. Every language in the database contains at least 10 speakers. From each speaker, about 5-10 minutes of speech is collected. On the whole, each language contains minimum of 1 hour speech data. Details of the database are described in Table 3.1. Pixelview TV tuner card is used to access the required TV channels through the computer, using VentiTV software. Audacity software is used to record the speech from TV channels. Speech data from broadcast radio channels are collected from archives of Prasar Bharati, All India Radio. The main reasons for choosing TV and Radio channels for collecting the proposed database are: (1) It is difficult to get sufficient number of native speakers for each of the above mentioned languages, in a single place, (2) It is highly time consuming to record the required speech data from different speakers at different places, (3) The quality of speech data used in broadcast TV channels is observed to be clean and noise free and (4) Speakers from TV channels are more professional and matured. Since, most of the speakers in the database are well matured and professional, hence the collected speech corpus ensures standard quality in terms of articulation, speaking rate and pronunciation. The speech signal is recorded with the sampling rate of 16 kHz, and each sample is stored as 16 bit number. The speech recorded through TV channels may contain some inherent problems such as (1) News bul-

letins contain background music during headlines, background videos and commercials, and (2) During talk shows and interviews, there are chances of overlapping of speech of different speakers. Therefore, while recording and editing the database proper care has been taken to minimize the above mentioned problems.

Table 3.1: Details of Indian Institute of Technology Kharagpur multi-lingual Indian language speech corpus (IITKGP-MLILSC)

Iden-tity	Language	Region	Population (Million)	Speakers		Duration (Minutes)
				Female	Male	
1	Arunachali	Arunachal Pradesh	0.41	6	15	72.00
2	Assamese	Assam	13.17	6	8	67.33
3	Bengali	West Bengal	83.37	14	10	69.78
4	Bhojpuri	Bihar	38.55*	5	7	59.82
5	Chhattisgarhi	Chhattisgarh	11.50*	9	11	70.00
6	Dogri	Jammu and Kashmir	2.28	8	12	70.00
7	Gojri	Jammu and Kashmir	20.00*	3	12	44.00
8	Gujrati	Gujarat	46.09	7	6	48.96
9	Hindi	Uttar Pradesh	422.05	14	24	134.70
10	Indian English	All over India	125.23	12	13	81.66
11	Kannada	Karnataka	37.92	4	8	69.33
12	Kashmiri	Jammu and Kashmir	5.53	2	19	59.64
13	Konkani	Goa and Karnataka	2.49	5	15	50.00
14	Manipuri	Manipur	1.47	11	11	64.00
15	Mizo	Mizoram	0.67	3	8	48.00
16	Malyalam	Kerala	33.07	7	12	81.09
17	Marathi	Maharashtra	71.94	7	9	74.33
18	Nagamese	Nagaland	0.03	11	9	60.00
19	Neplai	West Bengal	2.87	7	6	54.19
20	Oriya	Orissa	33.02	10	4	59.87
21	Punjabi	Punjab	29.10	7	10	80.91
22	Rajasthani	Rajasthan	50.00*	10	10	60.00
23	Sanskrit	Uttar Pradesh	0.014	0	20	70.00
24	Sindhi	Gujarat and Maharashtra	2.54	14	6	50.00
25	Tamil	Tamil Nadu	60.79	7	10	70.96
26	Telugu	Andhra Pradesh	74.00	7	8	73.72
27	Urdu	Uttar Pradesh	51.54	5	16	86.49

3.1.2 Oregon Graduate Institute database Multi-language Telephone-based Speech (OGI-MLTS)

The Oregon Graduate Institute (OGI) multi-language telephone-based speech (MLTS) corpus consists of telephone speech from 11 languages. This data was collected by Yeshwant Muthusamy for his PhD research, included 90 telephone calls in each of the 10 languages [19]. The languages are: English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese. This corpus was used by the National Institute of Standards and Technology (NIST) for evaluation of automatic language identification in 1996. Later the corpus was extended with additional recordings for each of the ten above, and 200 Hindi calls were added, making a total of 11 languages.

Speech utterances in each call were spoken by a unique speaker over the telephone channel and the speech was sampled at 8kHz. For collecting the data, each caller was asked a series of questions designed to elicit:

1. Fixed vocabulary speech (e.g. days of the week).
2. Domain-specific vocabulary speech.
3. Unrestricted vocabulary speech.

The unrestricted vocabulary speech was obtained by asking callers to speak on any topic of their choice. The "story-before-tone" (maximum duration 50 sec) and the "story-after-tone" (maximum duration 10-sec) utterances together form the 1 minute unrestricted vocabulary speech portion of each call. This corpus was collected and developed in 1992, and the latest version was released in 2002 which includes recorded utterances from about 2052 speakers, for a total of about 38.5 hours of speech. The OGI 22-Language Corpus [86] was also developed by Oregon Graduate Institute. The current version of the OGI 22 Language corpus consists of telephone speech from 22 languages: Arabic, Cantonese, Czech, English, Farsi, German, Hindi, Hungarian, Japanese, Korean, Indonesian, Mandarin Chinese, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil and Vietnamese. The corpus contains fixed vocabulary utterances (e.g. days of the week) as well as fluent continuous speech. Approximately 20,000 utterances in 16 languages have corresponding orthographic

transcriptions. Table 3.2 shows the comparison of OGI database and the IITKGP-MLILSC used in LID studies.

Table 3.2: Comparison of OGI-MLTS and IITKGP-MLILSC speech databases.

Sl. no.	Features	OGI-MLTS database	IITKGP-MLILSC database
1.	No. of languages	11	27
2.	No. of speakers/language	Average of 90	Average of 10
3.	Type of speech	Casual conversation	Well articulated read speech
4.	Environment of recording	Realistic	Studio quality
5.	Noise	Background noise	No background noise
6.	Channel characteristics	Different	Similar
7.	Language characteristics	Different	Similar

3.2 Features used for automatic language identification

In this chapter we explore various frame level features for analyzing the language-specific information. As the linguistic content of the speech influences the vocal tract shape, the distribution of vocal tract system features may be unique for a language. The vocal tract features are represented using spectral features. The distribution of the spectral feature vectors for each language is captured to model the language. Gaussian mixture models (GMMs) are used to capture the language-specific distributions from the derived spectral features.

The language-specific information is present in speech at different levels such as excitation source level, vocal tract system level and prosodic level. In this work, language-specific characteristics are explored at vocal tract system and prosodic levels. At the vocal tract system level, language-specific information can be viewed as unique sequence of vocal tract shapes for the given sound unit. Each sound unit corresponds to a particular articulatory configuration of the vocal tract. For different languages, the articulatory configuration corresponding to even same sound unit may slightly vary due to the differences in pronunciation and co articulation. With this effect, each sound unit has certain uniqueness with respect to each language due to the constraints associated to that particular language. This acoustic-

phonetic variations among languages can be represented using short-time spectral features. Due to the non-stationary nature of speech production mechanism, the spectral features are extracted over short (typically 20 msec) quasi stationary segments of speech data. We study the presence of language-specific information in spectral characteristics using features derived from short spans of speech signal. For representing these characteristics, we have explored two popular spectral features namely, linear prediction cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC) [87]. MFCCs and LPCCs are determined from speech using the following steps.

1. Pre-emphasize the speech signal.
2. Divide the speech signal into sequence of frames with a frame size of 20 ms and a shift of 10 ms. Apply the Hamming window over each of the frames.
3. Compute magnitude spectrum for each windowed frame by applying DFT.
4. Mel spectrum is computed by passing the DFT signal through mel filter bank.
5. DCT is applied to the log mel frequency coefficients (log mel spectrum) to derive the desired MFCCs.
6. For deriving LPCCs, LP spectrum is computed for the windowed speech frame obtained from step 2.
7. LPCCs are determined by applying the DFT over log-magnitude LP spectrum.

In this work, 13 MFCC and 13 LPCC features are derived from a speech frame of 20 ms with a frame shift of 10 ms. For deriving the MFCCs, 24 filter bands are used. LPCCs are derived using 10^{th} order LP filter. Detailed description of LPCC and MFCC is given in Appendix A and Appendix B respectively.

3.3 Development of language models

In this work, Gaussian mixture models are used for developing the LID systems using spectral features. GMMs are well known to capture the distribution of data in the feature space. The

accuracy in capturing the true distribution of data depends on various parameters such as dimension of feature vectors, number of feature vectors and number of mixture components. In this work, GMMs are assumed to capture the language-specific information from the given spectral features.

For developing the language identification system using GMMs, we need to develop a specific GMM for each of the language. These GMMs are developed using the spectral vectors derived from the speech corresponding to the languages considered. In this work, 27 Indian languages (IL) are considered for analyzing the LID performance using spectral features. Therefore, the proposed LID system consists of 27 GMMs (language models) developed using speech corresponding to 27 languages. General block diagram of an LID system is shown in Figure 3.1. For evaluating the developed LID system, feature vectors derived from test speech samples are given to all the language models. The evidence in response to test speech samples from all the models are analyzed, and the highest evidence among the models is hypothesized as the language identity corresponding to the given speech sample. More details about training and testing of GMM models are provided in Appendix C.

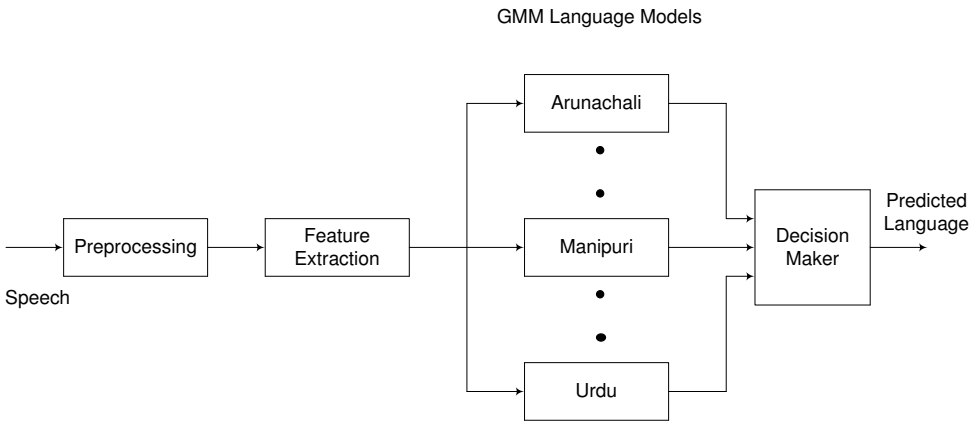

Figure 3.1: Block diagram of LID system

3.4 LID performance on Indian language database (IITKGP-MLILSC)

In this study, LID systems are developed in speaker dependent and speaker independent environments. In speaker dependent LID system, speech from all the speakers of a given language is used for developing and evaluating the models. Whereas for speaker independent LID system, speaker set used for developing the models or LID system is different from the set of speakers used for evaluating the models. The following subsections discuss more details about speaker dependent/independent LID systems.

3.4.1 Speaker dependent LID system

In this study, 80% of speech from all the speakers of a given language is used for developing the specific language model. For evaluating the developed speaker dependent LID system, 20% of speech from each speaker (which is not used for developing the models) is used. Separate LID systems are developed using LPCCs and MFCCs. For analyzing the effect of number of mixture components of GMMs on the identification performance, various LID systems are developed by varying the number of mixture components of GMMs from 32 to 512. Similarly, for analyzing the effect of length of test speech utterance on identification performance, three different lengths (5, 10 and 20 secs) of test speech utterances are analyzed. The performance of speaker dependent LID system with different spectral features, different mixture components and different lengths of test speech utterances is given in Table 3.3. In Table 3.3, first column shows the number of mixture components used for developing the LID system. Columns 2-4, indicate the performance of LID system developed using MFCC features for different durations of test speech utterances. Columns 5-7, indicate the performance of LID system developed using LPCC features for different durations of test speech utterances. From the results, it is observed that the identification performance has improved significantly, for both LPCC and MFCC based systems while increasing the number of mixture components from 32 to 256, and beyond 256 mixture components a slight improvement is observed. The improvement in identification rate by increasing the number

of mixtures from 256 to 512 is not worth in view of time complexity associated to building the models. LID performance using MFCC features seems to be slightly better compared to LPCC features. But, it is observed that LPCC features seems to perform better compared to MFCC features for the models built with lower number of mixture components (32 to 256). In view of test utterance duration, the performance seems to be better at 10 secs, compared to 5 and 20 secs. The performance of speaker dependent LID system for 256

Table 3.3: Performance of speaker dependent LID system for different Gaussian mixture components

No. of mixture components	Average Recognition Performance (%)					
	MFCC			LPCC		
	5 sec	10 sec	20 sec	5 sec	10 sec	20 sec
32	73.19	77.34	78.12	79.37	78.46	80.67
64	80.35	82.23	83.92	86.79	84.12	85.76
128	87.29	86.33	87.04	88.96	86.32	87.97
256	90.12	90.19	90.17	90.19	**90.16**	90.16
512	90.78	**91.83**	90.65	88.64	89.22	91.16

Gaussian mixture components with different spectral features and different lengths of test speech utterances of each language is given in Table 3.4.

3.4.2 Speaker independent LID system

In this study, we have conducted the language identification using leave-one-speaker-out approach. In this approach language models are developed using *(n-1)* speakers of a particular language and the testing/validation is carried out using the speech utterances of a speaker, who has not participated in training the models. Like-wise in each iteration, one speaker will be left for testing and other speakers speech is used for developing the models. The average recognition across all the speakers is considered as recognition performance of the LID system. In speaker dependent LID system, all the speakers present during training stage are also present during testing stage. Hence, we have observed the better recognition performance. But, in practice LID system should perform well even under speaker independent environment. Therefore, to analyze the LID performance in speaker independent environ-

Table 3.4: Performance of speaker dependent LID system with different test durations

Language	Average Recognition Performance (%)					
	LPCC			MFCC		
	5 Sec	10 Sec	20 Sec	5 Sec	10 Sec	20 Sec
Arunachali	100.00	100.00	100.00	100.00	100.00	100.00
Assamese	89.78	89.59	90.02	90.35	90.54	89.29
Bengali	84.33	85.17	84.66	85.29	84.41	84.85
Bhojpuri	95.59	95.59	95.11	95.24	95.05	94.65
Chhattisgarhi	94.32	94.14	93.35	93.99	93.22	94.22
Dogri	89.30	89.85	89.17	90.14	89.71	90.19
Gojri	100.00	99.92	99.80	100.00	100.00	99.86
Gujrati	97.73	98.27	98.61	97.40	97.38	98.64
Hindi	99.40	99.31	99.58	99.87	99.67	98.69
Indian English	93.85	92.98	93.99	93.56	93.17	93.64
Kannada	98.94	98.24	98.33	98.60	99.02	99.20
Kashmiri	95.00	95.68	95.82	95.93	95.07	95.28
Konkani	90.66	90.49	91.35	90.55	91.12	91.48
Manipuri	100.00	100.00	100.00	100.00	100.00	100.00
Mizo	90.36	90.70	90.39	90.62	91.04	90.65
Malyalam	83.60	84.35	83.46	83.64	84.22	83.29
Marathi	92.13	92.07	92.40	91.55	92.39	91.65
Nagamese	95.72	95.66	95.00	94.72	95.82	94.93
Neplai	95.80	95.43	95.30	95.80	95.75	95.27
Oriya	84.97	85.70	85.44	84.63	85.23	84.52
Punjabi	81.27	81.86	82.04	82.12	81.91	81.96
Rajasthani	79.35	80.27	80.32	79.84	80.06	80.64
Sanskrit	82.87	82.51	83.36	83.64	82.85	82.96
Sindhi	87.79	88.48	87.94	88.38	87.70	87.92
Tamil	87.87	88.00	87.50	87.75	87.58	88.33
Telugu	72.70	71.53	72.50	71.61	71.68	72.68
Urdu	69.99	69.42	69.12	70.02	69.81	69.63
Average	**90.12**	**90.19**	**90.17**	**90.19**	**90.16**	**90.16**

ment, we have developed and evaluated the LID system with non-overlapping speaker sets. For testing the speaker independent LID system, speech utterances of a speaker (who is not involved during training) from each language are used. Similar to speaker dependent LID system, speaker independent LID system is also evaluated with varying number of Gaussian mixture components, varying test utterance durations and different spectral features. The performance of speaker independent LID system for 64 Gaussian mixture components is

given in Table 3.5. From the results presented in Table 3.5, it is observed that the performance of LID system in speaker independent environment is drastically reduced compared to speaker dependent environment (see Tables 3.4 and 3.5). From this observation, it may be noted that LID system in speaker dependent environment has captured speaker specific information in addition to language specific information. In speaker independent environment, LID system with MFCC features has achieved the highest recognition performance of 48.46% with 10 secs test utterance speech duration and 64 Gaussian mixture components. From the results present in Tables 3.4 and 3.5, it is observed that there exists a strong speaker bias in the developed language models. Therefore, the identification performance is superior (about 90%) in the case of speaker dependent environment compared to (about 48%) speaker independent environment.

3.4.3 Speaker independent LID system with speaker specific language models

For improving the performance of LID system in speaker independent environment, we have proposed speaker-specific language models in this work. In this framework, each language model is represented by set of speaker models associated to that particular language. The block diagram of the LID system with proposed speaker-specific language models is given in Figure 3.2. For analyzing the language identification performance in speaker independent environment, language models are developed with multiple speaker models by omitting one speaker from each language. The process of building language models is similar to speaker independent LID system discussed above. The basic idea behind the proposed speaker specific language models is that, the multiple evidence from each language model (see Figure 3.2) can be exploited in a better way compared to single evidence from each language model (see Figure 3.1). In speaker independent case, it is known that speakers during training and testing are mutually exclusive, and the common information may present in both training and testing is only language specific information. Hence, our hypothesis is that the evidences from the matched language is high compared to non-matched languages. The proposed speaker specific language models are analyzed with respect to various spectral features and

Table 3.5: Performance of speaker independent LID system with different test durations

Language	Average Recognition Performance (%)					
	LPCC			MFCC		
	5 Sec	10 Sec	20 Sec	5 Sec	10 Sec	20 Sec
Arunachali	96.84	96.99	96.57	62.58	73.39	88.55
Assamese	8.71	22.85	4.68	0.00	0.00	0.00
Bengali	42.50	41.57	37.53	46.35	59.78	55.38
Bhojpuri	10.53	14.71	16.32	28.66	33.65	38.97
Chhattisgarhi	0.00	0.00	0.00	96.36	96.60	96.60
Dogri	78.30	88.24	94.36	75.66	95.13	94.54
Gojri	19.60	21.37	25.43	16.53	20.01	23.07
Gujrati	48.23	71.52	51.85	2.16	21.48	7.30
Hindi	6.04	9.26	6.76	1.84	12.36	12.91
Indian English	16.38	10.77	11.16	26.57	27.79	24.77
Kannada	55.57	64.28	69.72	60.72	72.41	76.78
Kashmiri	39.46	50.37	59.70	75.72	70.92	69.89
Konkani	5.96	32.13	45.52	41.33	38.13	45.03
Manipuri	13.39	23.92	15.56	4.33	11.87	0.52
Mizo	96.30	96.39	96.23	96.93	96.75	96.33
Malyalam	0.00	0.00	0.00	0.00	0.00	1.82
Marathi	0.00	0.00	0.00	65.78	75.41	69.63
Nagamese	91.58	91.34	94.66	96.63	96.63	96.23
Neplai	63.00	73.86	75.11	57.04	68.14	70.50
Oriya	49.09	56.72	53.86	50.29	56.17	53.83
Punjabi	67.62	92.23	92.19	33.89	52.83	41.46
Rajasthani	76.22	79.41	86.38	91.07	91.94	96.48
Sanskrit	2.99	0.00	0.00	36.03	20.94	33.61
Sindhi	22.28	43.57	46.45	57.04	64.08	86.41
Tamil	34.85	40.86	42.56	49.27	51.98	53.98
Telugu	0.00	0.00	0.00	0.00	0.00	0.00
Urdu	0.00	0.00	0.00	0.00	0.00	0.00
Average	**35.02**	**41.57**	**41.58**	**43.44**	**48.46**	**49.43**

varying Gaussian mixture components. In this study, the length of test speech utterances is chosen as 10 secs. The evidence from multiple speaker models of each language are explored in various ways such as (a) maximum of the evidence, (b) mean of the evidence, (c) sum of best 2, 3, 4 and 5 evidences and (d) maximum plus mean. It is observed, among the various strategies to exploit the multiple evidences, maximum evidence from multiple speaker models seems to be out performed compared to other strategies. The identification performance

Figure 3.2: LID system using speaker specific models.

of the speaker independent LID system developed using speaker specific language models with maximum evidence criteria is given as confusion matrix in Table 3.6. Here, language identities are mentioned with numbers from 1-27, indicating the languages in alphabetical sequence mentioned in Table 3.1. Each row corresponds to classification accuracy of a particular language test samples. For example, first row indicates the classification accuracy of "Arunachali". It shows that 77% of Arunachali test samples are correctly classified, and the remaining 23% samples are misclassified as Dogri (with 1%) and Sanskrit (with 22%). From the confusion matrix it is observed that each language is mostly misclassified (confused) into 4-6 other languages. Hence, the most of the entries in the confusion matrix are found

to be zero. By properly analyzing the confusion matrix one can design the multi-level LID system for improving the performance, where initial classification is performed on broader groups, and later the finer classification is performed within the groups. From the results, it is observed that the average identification accuracy of the speaker independent LID system using the proposed speaker specific models with MFCC features is about 51%.

Table 3.6: Performance of speaker independent LID system with speaker specific language models

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	**77**	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0
2	0	**0**	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	23	0	0	0	0	0	0	0
3	0	0	**64**	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23
4	0	0	0	**37**	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	34	26	0	0	0	0	0	0
5	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	**99**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	**24**	0	24	0	0	35	0	0	0	0	0	0	0	10	0	0	7	0	0	0	0
8	0	0	25	0	0	0	0	**25**	0	0	0	8	0	0	0	0	0	0	0	25	0	0	0	0	0	0	17
9	0	9	7	1	0	0	0	0	**16**	0	0	33	0	1	0	1	0	0	0	10	7	0	0	0	3	0	11
10	0	0	24	0	0	0	2	0	26	**31**	0	5	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
11	0	0	12	0	0	0	0	0	0	0	**76**	4	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0
12	0	10	0	0	0	0	0	0	10	0	0	**74**	0	0	0	0	0	0	0	0	3	0	0	0	0	3	0
13	0	0	0	0	0	0	0	0	2	0	0	0	**41**	0	0	0	0	33	0	0	0	0	0	24	0	0	0
14	0	9	0	0	0	0	0	0	3	0	0	9	0	**16**	0	0	0	0	0	9	3	0	0	0	50	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	18	0	0	0	0	4	0	0	14	0	0	0	0	**0**	0	0	0	39	0	0	0	0	0	0	25
17	0	0	12	0	0	0	0	0	0	0	0	0	0	3	0	0	**79**	0	0	0	0	0	0	4	1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0
19	0	9	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**72**	0	0	0	0	0	9	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	**60**	0	0	0	0	0	0	0
21	0	4	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	**57**	0	0	0	0	0	0
22	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**96**	0	0	0	0	0
23	76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**24**	0	0	0	0
24	0	0	0	0	0	29	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	**67**	0	0	0
25	0	0	0	2	0	0	0	0	4	0	0	32	0	6	0	0	0	0	0	0	0	0	0	0	**55**	0	0
26	0	20	40	7	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	7	13	0	0	0	0	**0**	0
27	0	2	4	0	0	0	0	0	6	0	0	35	0	0	0	0	0	0	0	16	37	0	0	0	0	0	**0**

For more detailed analysis of the recognition results, we have also examined the recognition performance of individual languages. Table 3.7 shows the recognition performance of the 27 Indian languages, individually. We have also examined the recognition performance by considering the desired language in either rank 1 or rank 2 or rank 3 position. Column 2, 3 and 4 indicate the recognition performance of individual languages within rank 1, rank 2 and rank 3 positions respectively. From the results, it is observed that the average recognition performance has been improved from 51.42% to 65.60% by considering the top 3 ranks. Among the 27 languages, Assamese, Hindi, Manipuri, Malyalam, Telugu, and Urdu have achieved the recognition performance less than 20%, whereas Bhojpuri, Gojri, Gujrati, Indian English, Konkani, and Sanskrit have achieved the recognition performance in between 23-45% and rest have achieved more than 50% recognition performance based on rank 1 (see column 2). From column 4 of Table 3.7, it can be noted that the recognition performance of Malayalam, Sanskrit, Telugu, and Urdu has still less than 30% even after considering top 3 choices. The cause for the poor performance of the above mentioned 4 languages has to be thoroughly investigated by examining the quality of test speech samples.

In this work, pair-wise testing is also performed among the languages to analyze the pair-wise language discrimination. The details of the pair-wise testing results on Indian language database are given in Table 3.8. The total number of pairs formed from 27 languages are 351 (27_{C2}). Table 3.8 represent the results in 27x27 matrix form where the performance of 351 pairs of languages is represented by upper diagonal elements in the matrix. The basic goal of the pair-wise testing is to determine the most confusable languages with respect to each language considered for the identification task. In Table 3.8, first row indicates the results of pair-wise testing for the languages with respect to languages 2-27. In this case, a pair-wise models are considered and they are evaluated with the test samples of the corresponding pair of languages. In this manner the entries in the first row of Table 3.8 represent the recognition performance of pairs of languages correspond to (1,2), (1,3), (1,4) ...(1,27). Similarly, the entries in the second row represent recognition performance corresponding to pair of languages (2,3), (2,4), ...(2,27).

From the study, we can observe that if the recognition accuracy is less than 75%, it indicates that the particular pair of languages have similar acoustic characteristics, and

Table 3.7: Performance of LID system using speaker specific models. Columns 2-4 represent the identification performance by considering the genuine model within (i) First, (ii) Two and (iii) Three ranks respectively.

Language	k-Best performance		
	k=1	k=2	k=3
Arunachali	76.92	88.34	92.45
Assamese	0	31.82	41.52
Bengali	63.64	68.18	75
Bhojpuri	37.14	54.58	58.26
Chhattisgarhi	100	100	100
Dogri	98.8	100	100
Gojri	23.75	32.12	40.43
Gujrati	25	41.67	66.67
Hindi	15.71	44.29	44.17
Indian English	30.95	58.21	65.7
Kannada	76	80.25	82.45
Kashmiri	74.19	81.19	81.19
Konkani	41.18	65.66	67.52
Manipuri	15.63	25	28.12
Mizo	100	100	100
Malyalam	0	9.88	14.65
Marathi	79.1	87.16	89.15
Nagamese	100	100	100
Neplai	71.88	78.15	79.85
Oriya	59.57	59.57	59.57
Punjabi	56.52	65.85	67.25
Rajasthani	95.65	100	100
Sanskrit	24.39	28.12	28.12
Sindhi	67.11	72.15	77.11
Tamil	55.32	65.83	85.25
Telugu	0	13.33	13.33
Urdu	0	13.33	13.33
Average	**51.42**	**61.65**	**65.60**

hence they are more confusable and discrimination accuracy is less. For example, as per the entries in first row, languages 1 and 23 (Arunachali and Sanskrit) are more similar and confusable to each other and hence the recognition performance is 72% between them. Whereas with other languages, language 1 is highly discriminative. In this manner, if we analyze the discrimination between each pair of languages, we can determine the confusable

languages. Table 3.9 shows the list of confusable languages for each language based on the threshold of 75% recognition accuracy. By analyzing the pair-wise performance, we can design the LID system using hierarchical approach. At the first level languages are divided into few groups based on their similarity characteristics observed in pair-wise testing.

Table 3.8: Performance of pair-wise language discrimination on Indian language database

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	-	100	100	100	100	100	100	100	100	100	100	100	98	100	100	100	99	100	100	100	100	99	72	100	100	100	100
2	-	-	75	100	100	100	100	100	67	100	89	87	84	72	100	90	100	100	70	59	73	91	95	100	96	59	52
3	-	-	-	79	100	100	100	68	59	75	72	83	88	76	99	50	85	100	93	74	100	100	100	100	90	62	38
4	-	-	-	-	100	100	100	83	92	100	92	100	100	96	100	100	100	100	97	54	71	100	100	100	96	96	99
5	-	-	-	-	-	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
6	-	-	-	-	-	-	100	100	100	100	97	100	85	100	100	100	100	100	100	100	100	100	96	91	100	100	100
7	-	-	-	-	-	-	100	100	57	100	100	45	100	100	100	100	100	100	100	85	100	100	97	100	100	100	100
8	-	-	-	-	-	-	-	-	66	97	78	93	95	70	99	78	100	100	100	83	89	100	100	100	95	81	48
9	-	-	-	-	-	-	-	-	-	75	85	70	81	69	100	76	99	100	93	70	81	100	100	100	78	68	41
10	-	-	-	-	-	-	-	-	-	-	100	85	100	100	100	91	100	100	100	71	100	100	100	100	100	100	100
11	-	-	-	-	-	-	-	-	-	-	-	95	97	72	97	89	100	100	91	93	90	100	100	99	88	88	93
12	-	-	-	-	-	-	-	-	-	-	-	-	100	83	100	100	100	100	98	96	69	100	100	100	78	76	50
13	-	-	-	-	-	-	-	-	-	-	-	-	-	77	89	95	98	76	95	99	100	100	98	77	89	92	94
14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	57	92	100	98	87	84	100	100	100	63	74	77
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	100	100	100	100	100	100	99	100	100	100
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	99	100	98	59	67	100	100	100	89	60	70
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	99	100	100	100	97	91	100	100
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	100	100	100	99	100	100	100
19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76	100	100	100	100	97	94	99
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76	100	100	100	91	55	60
21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	100	87	74	42
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	99	100	100	100
23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91	100	100	100
24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100	100	100
25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	98	85
26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72
27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.9: List of confusable languages obtained from pair-wise test.

Language	Most confusable pairs (less than 75% accuracy)
1	23
2	3, 9, 14, 19, 20, 21, 26, 27
3	2, 8, 9, 11, 16, 20, 26, 27
4	20, 21
5	-
6	-
7	9, 12
8	3, 9, 14, 27
9	2, 3, 7, 8, 12, 14, 20, 26, 27
10	20
11	3, 14
12	7, 21, 27
13	-
14	2, 8, 11, 16, 25, 26
15	-
16	3, 20, 21, 26, 27
17	-
18	-
19	2
20	2, 3, 4, 9, 10, 16, 26, 27
21	2, 4, 12, 16, 26, 27
22	-
23	1
24	-
25	14
26	2, 3, 9, 14, 16, 20, 21, 27
27	2, 3, 8, 9, 12, 16, 20, 21, 26

3.5 LID system using spectral features from pitch synchronous analysis (PSA) and glottal closure regions (GCRs)

The motivation to use spectral features from pitch synchronous analysis (PSA) and glottal closure regions (GCR) is that in case of conventional block processing, 20-30 ms of speech frame is considered for analysis, and hence it provides the average spectral characteristics

corresponding to multiple pitch cycles. In the block processing approach, the basic assumption is that speech signal is stationary during 20-30 ms time interval. But, in strict sense (in practice) shape of the vocal tract (VT) will vary continuously. At the gross level, we may not perceive the change, but at the finer level there exists variations in vocal tract system characteristics within a speech frame. For analyzing the variations in VT characteristics within speech frame, it is divided into pitch cycles, and each pitch cycle is further divided into glottal closure and glottal open phases. The analysis of each pitch cycle is viewed as pitch synchronous analysis. Within a pitch cycle, glottal closure (GC) phase is more important compared to glottal open phase. In general, speech segment in a GC region contain high signal strength, and hence it has high signal to noise ratio (SNR) compared to other regions. This is due to presence of significant (impulse like) excitation at the instant of glottal closure. Because of this high SNR property, speech in GC regions is less susceptible to noise. Another important characteristic of GC phase is that the vocal tract resonances present during this phase are more accurate, and they represent true resonances of VT. This is because during GC phase oral cavity (cavity from glottis to lips) is completely decoupled from laryngeal and lungs cavities. Therefore, in this work pitch synchronous and GC based spectral features are investigated in addition to conventional frame level spectral features to examine the presence of language specific information.

In case of conventional block processing (BP) approach, 13 MFCC and 13 LPCC features are derived from a speech frame of 20 ms with a frame shift of 10 ms. For deriving the spectral features from pitch synchronous analysis (PSA) and glottal closure regions (GCR), first we need to determine the instants of significant excitation (ISE). In this work, ISE are determined by using zero frequency filter (ZFF) based method [88]. These ISE indicate instants of glottal closure (epochs or pitch markers) during voiced speech. After determining the ISE, the sequence of pitch cycles can be located by using the knowledge of ISE. Speech segment between two successive epochs is considered as one pitch cycle. Figure 3.3 illustrates different approches for extracting spectral features from speech signal using BP, PSA and GCR with the help of ISE. In Figure 3.3(a) successive stems of speech signal are considered for BP. Whereas, in Figure 3.3(b) successive stems (pitch cycle) are considered for PSA and 30% of a pitch cycle from a ISE is considered for GCR.

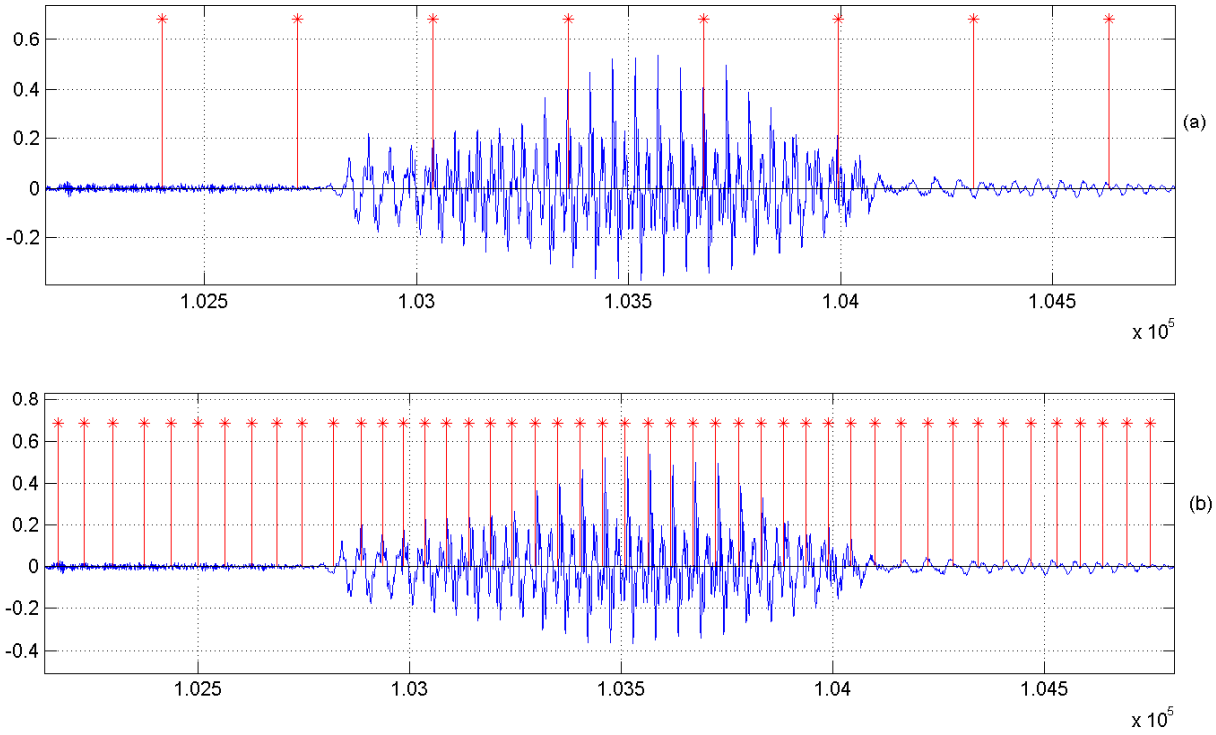

Figure 3.3: Illustration of feature extraction using (a)BP approach. (b) PSA approach.

The details of ZFF method for extracting the ISEs are given below.

3.5.1 Epoch extraction using zero frequency filter method

Among the existing epoch extraction methods, ZFF method determines the epoch locations with highest accuracy [88]. ZFF exploits the discontinuities due to impulse excitation reflected across all the frequencies including the zero frequency. The influence of vocal tract system is negligible at zero frequency. Therefore, zero frequency filtered speech signal carries excitation source information, which is used for extracting the epoch locations. The sequence of steps needed in ZFF method for extracting the ISEs are as follows:

1. Difference the input speech signal to remove any time-varying low frequency bias in the signal

$$x(n) = s(n) - s(n-1)$$

2. Compute the output of cascade of two ideal digital resonators at 0 Hz i.e.,

$$y(n) = \sum_{k=1}^{4} a_k y(n-k) + x(n)$$

where $a_1 = +4$, $a_2 = 6$, $a_3 = +4$, $a_4 = 1$. Note that this is equivalent to passing the signal $x(n)$ through a digital filter given by

$$H(z) = \frac{1}{(1 - z^{-1})^{-1}}$$

3. Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n)$$

where

$$\bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n)$$

Here $2N+1$ corresponds to the size of the window used for computing the local mean, which is typically the average pitch period computed over a long segment of speech.

4. The trend removed signal $\hat{y}(n)$ is termed as *zero frequency filtered* (ZFF) signal. Positive zero-crossings in the ZFF signal correspond to the epoch locations.

Epoch extraction for the segment of voiced speech using ZFF method is shown in Figure 3.4. Figure 3.4(a) shows the differenced electro-glotto-graph (EGG) signal of voiced speech segment shown in Figure 3.4(b). ZFF signal and the derived epoch locations are shown in Figures 3.4(c) and 3.4(d), respectively. From the Figures 3.4(a) and 3.4(d), it is evident that the epochs extracted using ZFF method are almost coincide with the negative peaks of differenced EGG signal, which indicate the instants of glottal closure.

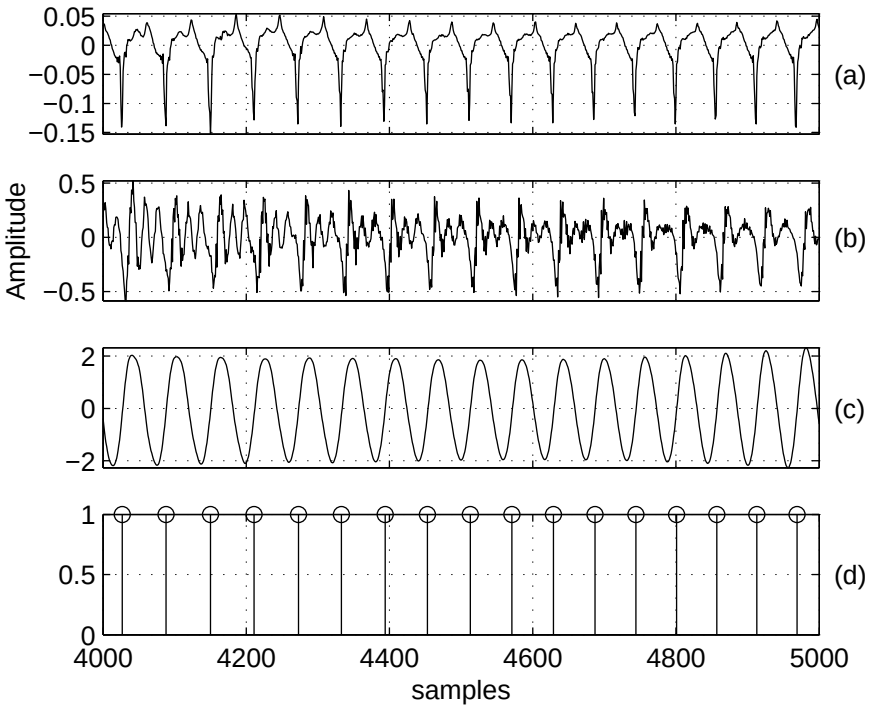

Figure 3.4: Epoch (GCI) extraction using zero frequency filtering method. (a) differenced EGG signal. (b) speech signal. (c) zero frequency filtering signal, and (d) epochs (GCIs) derived from zero frequency filtered signal.

3.5.2 Extraction of the spectral features from PSA and GCRs

In the case of pitch synchronous analysis, LPCC and MFCC are derived from each pitch cycle. Glottal closure instant (GCI) indicates the beginning of the GC region (GCR). But, determining GCR precisely within each glottal cycle (pitch cycle) is difficult. Therefore, we have analyzed various frame durations varying from 10% to 60% of pitch period for choosing the appropriate frame size to represent the glottal closure region. After determining the GCR in each pitch cycle, MFCC and LPCC features are extracted from each of the GC regions.

For developing the language recognition system using GMMs, we need to develop a specific GMM for each of the language. These GMMs are developed using the spectral vectors derived from the speech corresponding to the languages considered. In this work, 27 Indian languages are considered for analyzing the language recognition performance using

spectral features derived from block processing, pitch synchronous analysis and GC regions. The proposed LID system consists of 27 GMMs (language models) developed using speech corresponding to 27 languages.

For analyzing the language discrimination capability of the proposed spectral features, separate LID systems are developed with respect to each group of features. For choosing the optimal number of mixture components, we have examined various LID systems developed with mixture components varying from 2 to 512.

In this work, language models were trained and tested using leave-one-speaker-out cross-validation on the proposed feature sets. Each of the test utterance duration is about 10 s. For deriving the spectral features from GC regions, we have explored various frame durations ranging from 10% to 60% of glottal cycle starting from the GCI for choosing the suitable frame size to represent the GC regions. Language recognition systems are developed separately, with each of the speech frames (10%, 20%, 30%, 40%, 50% and 60% of pitch period). The recognition performance is shown in Table 3.10. From the recognition results, it is observed that 30% glottal cycle seems to be optimal for representing the GC region. In our further studies, GC region is represented with 30% of glottal cycle starting from the GCI.

3.5.3 Performance evaluation

Table 3.10: Performance of LID system with different durations of speech frames starting from the GCI.

% Pitch period	Average recognition performance (%)	
	LPCC	MFCC
10	34.82	35.41
20	38.67	39.19
30	**41.23**	**43.82**
40	40.11	42.53
50	39.36	40.27
60	38.54	39.08

Performance of LID system developed using MFCC features extracted from BP, PSA and GCR with 64 Gaussian mixture components of each language is given in Table 3.11.

Table 3.11: Performance of language identification system developed using MFCC features derived from block processing (BP), pitch synchronous analysis (PSA) and glottal closure regions (GCR).

Language	Recognition performance (%)		
	BP	PSA	GCR
Arunachali	76.92	81.08	85.12
Assamese	0	4.24	7.80
Bengali	63.64	67.53	70.86
Bhojpuri	37.14	41.23	45.03
Chhattisgarhi	100	100.00	100.00
Dogri	98.8	100.00	100.00
Gojri	23.75	27.31	31.53
Gujrati	25	28.19	32.26
Hindi	15.71	19.11	22.43
Indian English	30.95	35.09	38.25
Kannada	76	80.14	84.05
Kashmiri	74.19	78.55	82.10
Konkani	41.18	45.13	49.57
Manipuri	15.63	18.86	23.36
Mizo	100	100.00	100.00
Malyalam	0	4.25	8.50
Marathi	79.1	82.63	87.01
Nagamese	100	100.00	100.00
Neplai	71.88	76.39	80.76
Oriya	59.57	63.70	68.15
Punjabi	56.52	60.82	64.72
Rajasthani	95.65	99.39	103.54
Sanskrit	24.39	28.18	32.12
Sindhi	67.11	71.02	74.19
Tamil	55.32	58.83	62.58
Telugu	0	4.18	8.21
Urdu	0	3.84	7.70
Average	**51.42**	**54.80**	**58.14**

From the results, it is observed that among various feature extraction methods, spectral features derived from PSA and GCR have achieved better recognition performance compared to conventional block processing approach. Spectral vectors derived from PSA and GCR have shown the improvement of 3% and 6% in LID performance, compared to spectral features derived from BP. This indicates that PSA effectively captures the finer variations in the

spectral characteristics from the adjacent pitch cycles, and these finer variations may be more language specific, and it leads to better language discrimination. In the case of GCR based spectral features, LID performance is still better compared to PSA. It is known that, GCR based approach may also viewed as PSA. But, in GCR approach, in each pitch cycle only 30% of the pitch period of speech segment starting from the GCI is used for extracting the spectral features. With this, in each pitch cycle only high SNR portion representing the GC region is used for feature extraction.

For analyzing the robustness of the proposed feature extraction methods, LID systems developed based on speaker specific language models are evaluated using noisy test cases. Here, noisy test utterances are generated by adding the white noise taken from NOISEX-92 database. The performance of LID systems developed using MFCC features derived from BP, PSA and GCR, with noisy test utterances is given in Table 3.12. From the performance results, it is observed that spectral features derived from GCR are robust to noise. This is mainly due to the presence of high SNR speech in GCR. With this fact, the spectral features derived exclusively from GCR are robust to background noise compared to PSA and BP based features.

Table 3.12: Performance of LID systems with noisy test utterances

Noise level (dB)	Average recognition performance (%)		
	BP	PSA	GCR
Clean	51.42	54.80	58.14
30 (dB)	46.17	50.73	57.94
20 (dB)	40.26	42.55	54.79
10 (dB)	26.37	29.45	48.53

3.6 Performance of proposed spectral features on OGI-MLTS database

The language-specific information represented by the proposed spectral features is also analyzed by performing the LID task on OGI-MLTS database. For conducting these studies 10 speakers speech data from each language is considered. Evaluation is carried out using

leave-one-speaker-out approach. LID is carried out using MFCC features derived from BP, PSA, and GCRs. Training and testing conditions are similar to the LID systems developed using IL database. The LID performance using the proposed spectral features on OGI-MLTS database is given in Table 3.13. From the results, it is observed that among various feature extraction methods, spectral features derived from PSA and GCR have achieved better recognition performance compared to conventional block processing approach. Spectral vectors derived from PSA and GCR have shown the improvement of 5% and 7% in LID performance, compared to spectral features derived from BP. The overall identification accuracy of the proposed spectral features seems to be better for OGI-MLTS database compared to IL database. This may be due to (i) less number of languages in OGI-MLTS database, (ii) discrimination among the languages may be more, because the languages may be originated from different language families. Whereas all ILs are originated from Sanskrit.

Table 3.13: Performance of language identification system developed using MFCC features derived from block processing (BP), pitch synchronous analysis (PSA) and glottal closure regions (GCR) on OGI-MLTS database.

Language	Average recognition performance (%)		
	BP	PSA	GCR
English	89.75	91.85	92.05
Farsi	65.12	65.12	69.20
French	70.08	72.25	74.08
German	50.45	54.15	55.30
Hindi	47.10	50.20	51.22
Japanese	82.09	87.34	89.12
Korean	75.37	77.10	80.24
Mandarin	24.15	35.48	36.07
Spanish	47.21	52.24	55.06
Tamil	60.24	67.25	72.76
Vietnam	38.08	48.54	49.05
Average	**59.05**	**63.77**	**65.83**

3.7 Summary and conclusions

In this thesis we have proposed IITKGP-MLILSC (Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus) speech database for language identification task in the context of Indian languages. The proposed speech database consists of 27 Indian languages. The developed speech database is analyzed in view of language identification by using spectral features. The language identification system is analyzed with speaker dependent and independent environments. The performance of LID system in speaker independent environment is observed to be very low compared to speaker dependent environment. The dominant performance in speaker dependent case is mainly due to speaker bias in the developed language models. The performance of speaker independent LID system has been improved by proposing speaker specific models for each language. Results have indicated that spectral features derived from PSA and GCR have better language discrimination ability compared to conventional block processing method. This can be attributed to finer spectral variations captured by PSA and high SNR property of GCR. Compared to PSA, spectral features from GCR were observed to have more language specific information. In the presence of background noise, spectral features derived from GCR were found to be more robust compared to PSA and BP based spectral features. LID studies were also performed on OGI database using spectral features derived from BP, PSA and GCRs and the recognition trend is observed to be similar to ILs. The recognition performance may be improved further by exploiting the language specific prosodic and source information in addition to spectral information. Development of language identification system using prosodic features is discussed in the following chapter.

Chapter 4

Language identification using prosodic features

In the previous chapter frame level spectral features are explored for capturing the implicit language-specific knowledge present in speech. Features at frame level represent limited span of speech, and hence do not directly represent variations of sequence of sound units among languages. To represent variations of sequence of sound units, features need to be derived from the larger spans of speech signal corresponding to these sound units. Therefore, in this chapter we examine the usefulness of prosodic features at syllable, multi-syllable and global (phrase) levels for language identification. Human beings impose some constraints on the sequence of sound units while producing speech [89]. These constraints that lend naturalness to speech. Therefore, speech can not be merely characterized as a sequence of sound units. The variation of the pitch provides some melodic properties to speech, and this controlled modulation of pitch is referred as *intonation*. The duration of sound units are varied (shortened or lengthened) in accordance to some underlying pattern, giving some *rhythm* to speech. Some syllables or words are made more prominent than others, resulting in *stress*. The information gleaned from the melody, timing and stress in speech increases the intelligibility of spoken message, enabling the listener to segment continuous speech into phrases and words with ease [90]. These properties are also capable of conveying many more lexical and non-lexical information such as lexical tone, accent and emotion. The characteristics that make us perceive these effects are collectively referred to as *prosody*.

Much of the LID research so far has placed its emphasis on spectral information, mainly using the acoustic features of sound units (referred as acoustic phonetics), and their alignment (referred as phonotactics). Such systems may perform well in similar acoustic conditions [32, 1]. But their performance degrade due to noise and channel mismatch. Prosodic features derived from pitch contour, amplitude contour and duration pattern are relatively less affected by channel variations and noise. Therefore in this chapter we focused on developing the LID systems using prosodic features. Prosodic features such as intonation, rhythm and stress are extracted from syllables and trisyllables (word), whereas dynamics of changes in fundamental frequencies (F_0), durations of syllables, changes in energies are used at global level for discriminating the languages. Though the systems based on spectral features outperform the prosody-based LID systems, their combined performance may provide the needed robustness.

The chapter is organized as follows: Section 4.1 presents the extraction of consonant-vowel (CV) portions from the continuous speech. The similarities and differences of languages with respect to prosodic aspects is discussed in section 4.2. Development of LID systems using prosodic features from syllables and words is explained in section 4.3. Performance results of LID system at syllable and word levels is given in section 4.4. Section 4.5 presents the LID system developed using prosodic features at global level. The details of performance of LID system at global level is given in section 4.6. Performance evaluation of prosodic features on OGI-MLTS database is presented in section 4.7. Performance of LID system with various combination of features is presented in section 4.8. Final section discuss the summary and conclusions of this chapter.

4.1 Extraction of CV units from continuous speech

To make use of the syllable level differences among languages, ideally we should have separate models for all the syllables in a language, and these models should be speaker independent. In order to train such syllable models, sufficient number of occurrences of each of the syllable from different speakers should be available. This may require several hours of manually labeled speech data, which may not be available in practice. Syllables in general can be of

the form C^mVC^n, where $m, n >= 0$, indicating that a syllable invariably contains a vowel (V) unit with zero or more consonants (C) preceding and/or succeeding the vowel. The possible syllable structures includes CCV, CV, V, CVC,CVCC, VC, VCC etc. A study conducted on an Indian language database [91] revealed that a major proportion of the syllables were of CV types. A cross-lingual study found that CVs are the most common type of syllables in the world languages [92]. The articulatory movements for the vowel in CV starts at the same time as the movements for the initial consonant [93]. The characteristics of the consonant part is influenced by the succeeding vowel, and therefore CV units capture significant co articulation effects. In CV-based language identification, the first step is to identify CV regions directly from speech signal. CV units consist of three main regions, region before the onset of vowel, the region of transition and the region of vowel as shown in Figure 4.1. Studies have demonstrated that a CV type of syllable can be represented using features corresponding to a fixed region around an important event called vowel onset point (VOP) [94, 95, 96].

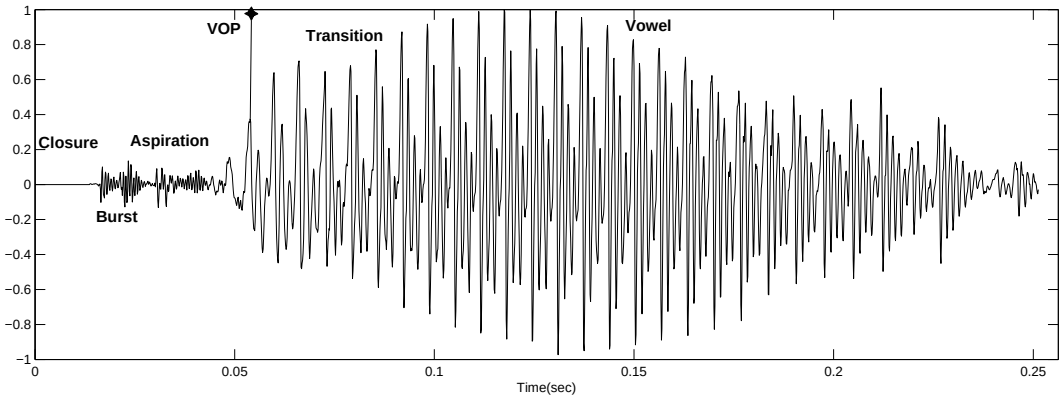

Figure 4.1: Regions of significant events in the production of the CV unit */ka/*.

In implicit LID, transcribed corpora will not be available, and hence it is important to select tokens from speech signal automatically. It is difficult to have a language-independent algorithm for segmentation of continuous speech. On the other hand, if certain events of significance for each unit can be identified and detected, feature extraction can be anchored around such events. Vowel onset point is one such event helpful for locating the CV patterns automatically from continuous speech. In this work vowel onset points are determined using

the spectral energy within a frequency band of 500-2500 Hz extracted from GCRs of speech [97]. The sequence of steps for identifying the VOPs from the continuous speech is as follows:

1. Determine the epoch locations (glottal closure instants) by using ZFF method.

2. Compute Discrete Fourier Transform (DFT) for the speech samples present in 30% of glottal cycle starting from the GCI.

3. Determine the spectral energy within the frequency band of 500-2500 Hz. Here, spectral energy in 500-2500 Hz band is considered, where energy of the vowel is much higher than the consonant.

4. Spectral energy is plotted as a function of time. Fluctuations in the spectral energy contour are smoothed by using mean smoothing with 50 ms window.

5. The change at the VOP present in the smoothed spectral energy of the speech signal is enhanced by computing its slope using First-Order Difference (FOD). FOD of $x(n)$ is given by

$$x_d(n) = x(n) - x(n-1)$$

The finer details involved in the enhancement of VOP evidence are illustrated by using Figure 4.2.

Figure 4.2(a) shows the speech utterance. Smoothed spectral energy in 500–2500 Hz band around each epoch is shown in Figure 4.2(b). The FOD signal of smoothed spectral energy is shown in Figure 4.2(c). Since FOD values corresponding to slopes, positive to negative zero crossings of slopes correspond to local peaks in the smoothed spectral energy signal. These local peaks are shown by star (*) symbols in Figure 4.2(b). The unwanted peaks in Figure 4.2(b) are eliminated by using the sum of slope values within 10 ms window centered at each peak. Figure 4.2(d) shows the sum of slope values within 10 ms around each peak. The peaks with the lower slope values are eliminated with a threshold set to 0.5 times the mean value of the slopes. This threshold has been considered after experimenting with huge data. Further, if two successive peaks present within 50 ms, then the lower peak among the two will be eliminated, based

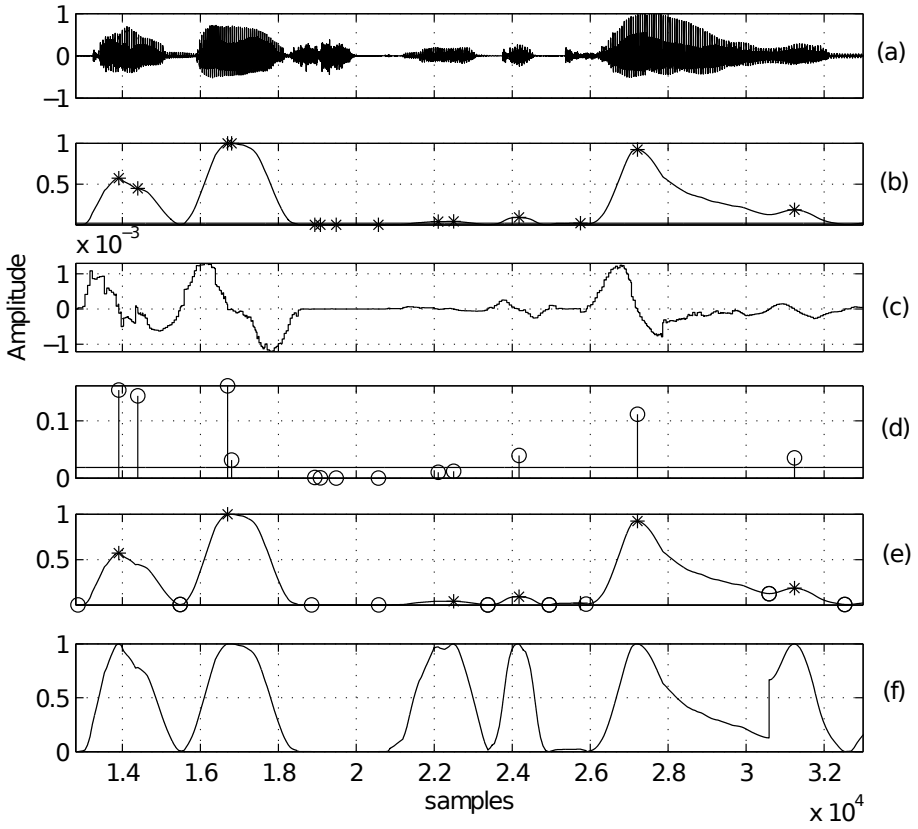

Figure 4.2: Enhancement of VOP evidence for a speech utterance /"Don't ask me to carry an"/. (a) speech signal, (b) smoothed spectral energy in 5002500 Hz band around each epoch, (c) FOD values, (d) slope values computed at each peak locations, (e) smoothed spectral energy plot with peak locations, and (f) enhanced values.

on the assumption that two VOPs won't present within 50 ms interval. The desired peak locations are shown in Figure 4.2(e) with star (*) symbol after eliminating the unwanted peaks. At each local peak location, the nearest negative to positive zero crossing points (see Figure 4.2(c)) on either side are identified and marked by circles on Figure 4.2(e). The regions bounded by negative to positive zero crossing points are enhanced by normalization process shown in Figure 4.2(f).

6. Significant changes in spectral characteristics present in the enhanced version of the smoothed spectral energy are detected by convolving with First Order Gaussian Difference (FOGD) operator of length 100 ms. A Gaussian window $g(n)$ of length L is

given by

$$g(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-n^2}{2\sigma^2}}$$

where σ is standard deviation. In this work, σ value of 200 is considered. The choice of length of Gaussian window (L) is based on assumption that the VOP occurs as gross level changes at intervals of about 100 ms [16, 19]. FOGD is given by $g_d(n)$, and it is shown in Figure 4.3.

$$g_d(n) = g(n) - g(n-1)$$

.

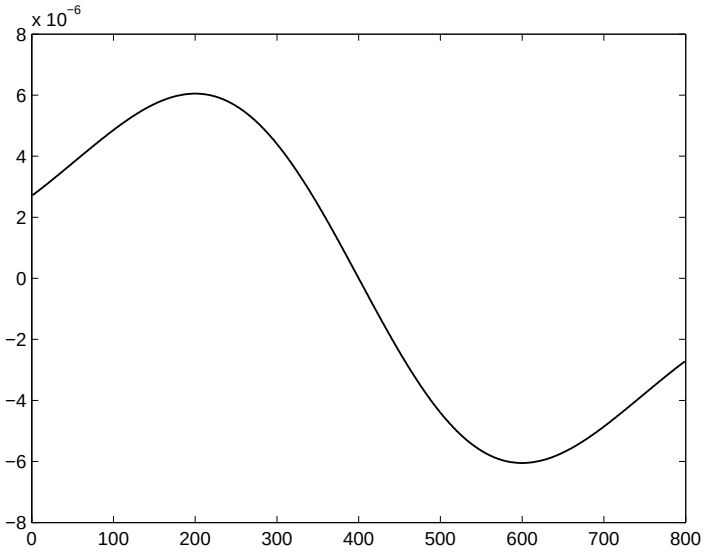

Figure 4.3: FOGD operator with L = 800, and σ = 200.

The convolved output is the VOP evidence plot.

7. Positive peaks in the VOP evidence plot represent the VOP locations. The flow diagram of the proposed VOP detection method is shown in Figure 4.4.

   Output of each step in the proposed method is shown in Figure 4.5 by using speech utterance /"Don't ask me to carry an"/. Figure 4.5(a) shows the speech signal with manually marked VOPs. Spectral energy in 500-2500 Hz band, and its smoothed signal are shown in Figures 4.5(b) and (c) respectively. Figure 4.5(d) shows the enhanced signal correspond to the signal present in Figure 4.5(c). Figure 4.5(e) shows

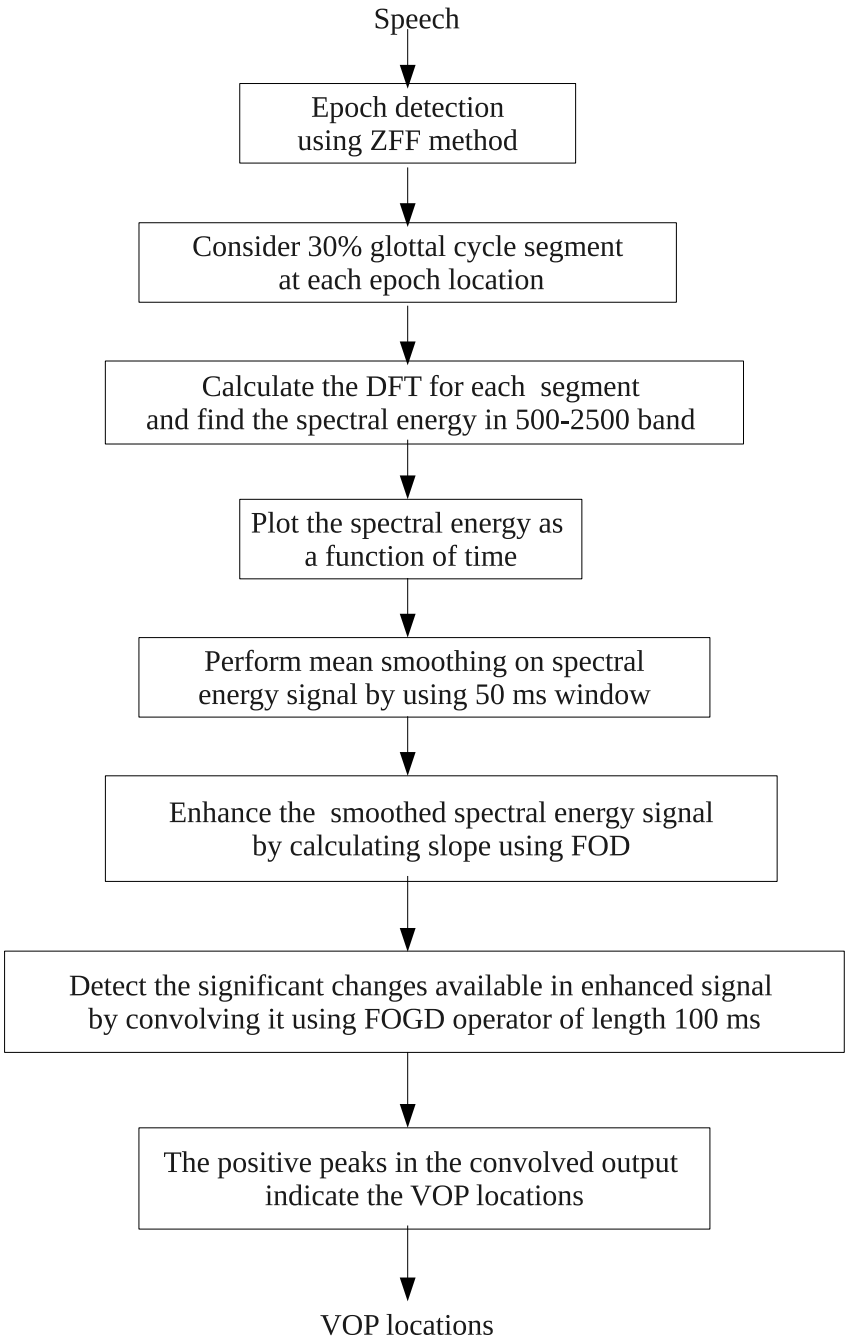

Figure 4.4: Flow diagram of the VOP detection method.

the VOP evidence signal obtained by convolving the enhanced spectral energy signal with FOGD. We can observe that manual VOPs marked in Figure 4.5(a) and detected VOPs marked in 4.5(e) are close to each other.

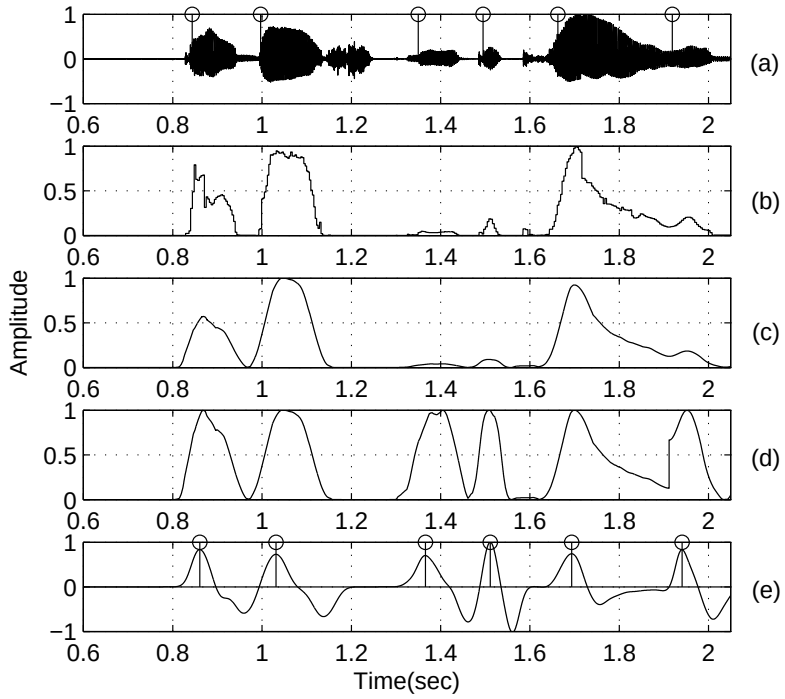

Figure 4.5: VOP detection for a speech utterance /"*Dont ask me to carry an*"/. (a) speech signal with manually marked VOPs, (b) spectral energy in 500–2500 Hz band around each epoch, (c) mean smoothed spectral energy, (d) enhanced spectral energy signal, and (e) VOP evidence signal.

The above mentioned VOP detection method is thoroughly evaluated on labeled speech databases such as TIMIT and IITM-Doordarshan news bulletin. From the result, it is observed that about 98% of VOPs are detected within 40 ms deviation, 2% of the VOPs are missed and certain 1% of VOPs are spurious. Among the detected VOPs about 85% are located within 10 ms deviation. Therefore, in the present study this VOP method will accurately determine the locations of CV units.

4.2 Prosodic differences among languages

The similarities in the prosodic aspects of neutral sentences in different languages are mostly due to identical constraints of the production and perception apparatus. There are similarities in the nature and position of pauses, and fundamental frequency (F_0) variations at sentence/phrase levels. The similarity in F_0 variations at sentence/phrase levels include the

tendency of F_0 values to fluctuate between two abstract lines, declination tendency of F_0 range, resetting of base line and the tendencies to repeat the succession of F_0 rises and falls [90, 98, 99]. But inspite of these natural tendencies, there are some prosodic characteristics that make a particular language different from others.

Languages can be broadly categorized as stress-timed and syllable-timed, based on their timing/rhythmic properties. In stress-timed languages like English and German, duration of the syllables are mainly controlled by the presence of stressed syllables which may occur at random. In stress-timed languages, roughly constant separation (in terms of time) is maintained between two stressed syllables. Syllables that occur in between two stressed syllables are shortened to accommodate this property. In syllable-timed languages such as French and Spanish, the durations of syllables remain almost constant. Languages are also classified as stress-accented and pitch-accented, based on the realization of prominence. In pitch-accented languages like Japanese, prominence of a syllable is achieved through pitch variations, whereas in stress-accented language, pitch variation is only one factor that helps to assign prominence. There is yet another categorization of languages as tonal and non tonal, based on the tonal properties of a language. We can identify languages which employ lexical tone such as Mandarin Chinese or Zulu (tonal languages), those which use lexically based pitch accents like Swedish or Japanese (pitch accented languages), and stress accented languages such as English or German [100]. But, there is no systematic study on ILs to discuss their implicit characteristics such as stress timed or syllable timed. There are many other languages which strictly do not follow the rules of a class, which means that these classifications are rather a continuum. Therefore languages may differ in terms of intonation, rhythm and stress.

4.3 Extraction of intonation, rhythm and stress (IRS) features from syllable and word levels

The term prosody refers to certain properties of speech signal such as audible changes in pitch, loudness and syllable length which makes human speech sound natural. Prosodic

characteristics are acquired over a period of time and prosodic events appear to be time-aligned with syllables or group of syllables [101]. The acoustic manifestation of prosody can be measured from F_0 contour, energy and duration. At the perceptual level, these acoustic measurement correspond to pitch, energy and length [101]. At the linguistic level, prosody of an utterance is represented at the level of sequences of syllables. We hypothesize that prosody is governed by the underlying syllable sequence, and measurement of prosodic characteristics involve segmentation into syllables. In this work, the locations of VOP are used for segmenting speech into syllable-like units. The locations of VOP are then associated with pitch contour for extracting the prosodic features.

In spoken communication, we use and interpret prosody without conscious effort. But it is difficult to describe them. Prosodic characteristics such as intonation, rhythm and stress (IRS) vary among languages. The variations in prosodic characteristics are represented using features derived from duration, fundamental frequency (F_0) contour and amplitude contour. Features corresponding to the syllable-like regions are derived to represent syllable-based intonation, rhythm, and stress. Therefore, in this chapter, prosodic features represented by intonation, rhythm and stress are extracted for syllables and words (tri-syllables) for evaluating the performance of LID system developed using Indian language and OGI- MLTS databases. Gaussian mixture models (GMMs) are used to capture the language specific distributions from the derived prosodic features. The details of prosodic features and its representation are given in the following subsections.

4.3.1 Intonation

Pitch is a perceptual attribute of sound which can be described as a sensation of the relative "altitude" of sound. The physical correlate of pitch is the fundamental frequency (F_0). The direction of F_0 change, either rising or falling, is determined by the phonological patterns of the constituent words, which are language-specific. The difference in F_0 contour between languages is illustrated for the case of three Indian languages, namely Assamese, Punjabi and Tamil in Figure 4.6. It can be observed that in general Punjabi has large variations in F_0 values compared to other languages, inspite of the variations in speaker characteristics.

In this study, our goal is to represent these pitch contour with suitable features to bring

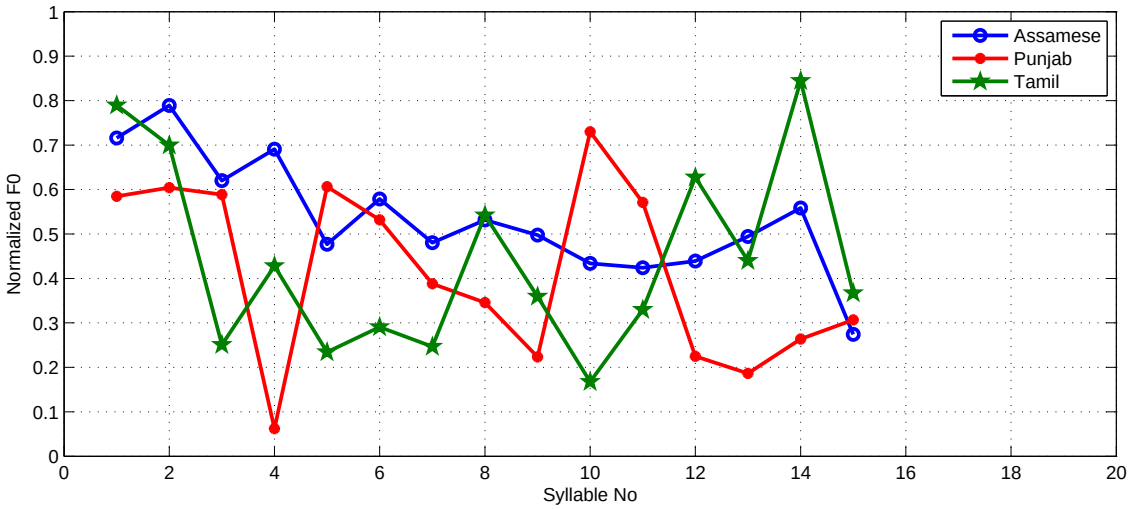

Figure 4.6: Variation in dynamics of F_0 contour for utterances in Assamese, Punjabi and Tamil.

out the language-specific information present in it. It has been observed that certain F_0 events, such as F_0 peaks and valleys, maintain a relatively stable alignment with the onset or offset of a syllable. In English, Greek and Dutch, it is found to occur quite regularly at the onset of the accented syllable. In Mandarin, peaks of F_0 are found to be consistently aligned with the offset of the tone-bearing syllable in certain situations [102].

Pitch analysis generates F_0 curves with finer variations. But the finer variations referred as micro prosody cannot be perceived and have no function in intonation. Therefore, F_0 contour obtained using zero frequency filter method [88] is smoothed to remove the finer variations. Since it is unnatural to have an abrupt variation of F_0, a simple median filtering is sufficient for smoothing the F_0 contour.

As illustrated in Figure 4.6, the dynamics of the F_0 contour reflects language-specific characteristics. Therefore it is important to represent it using suitable parameters. The F_0 contour between two consecutive VOPs (as shown in Figure 4.7) corresponds to the F_0 movement in a syllable-like region, and it is treated as a segment of F_0 contour. The nature of F_0 variations for such a segment may be a rise, a fall, or a rise followed by a fall in most of the cases. We assume that more complex F_0 variations are unlikely within a segment. To represent the dynamics of the F_0 contour segment, some parameters are derived.

With reference to Figure 4.7, tilt parameters [103], namely amplitude tilt (A_t) and du-

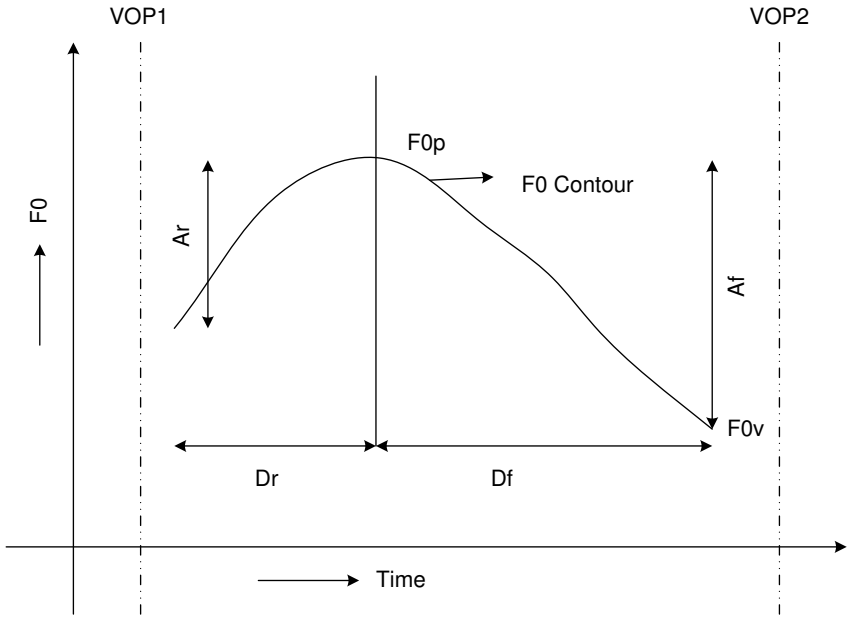

Figure 4.7: A segment of F_0 contour. Tilt parameters A_t and D_t defined in terms of A_r, A_f, D_r and D_f represent the dynamics of a segment of F_0 contour.

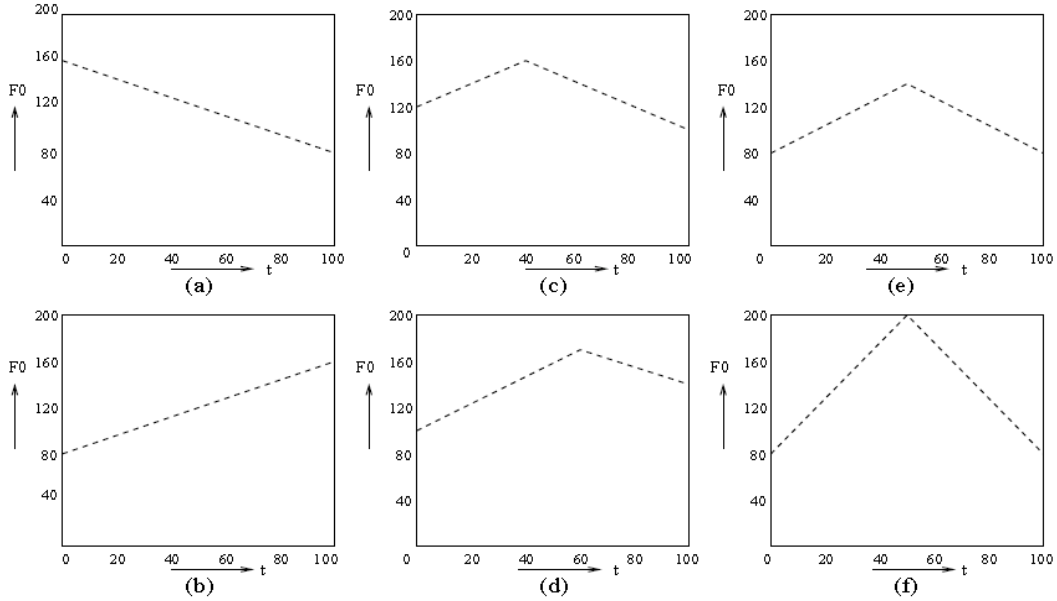

Figure 4.8: Illustration of F_0 contours with various tlt parameters (a) A_t=-1, D_t=-1; (b) A_t=1, D_t=1; (c) A_t=-0.2, D_t=-0.2; (d) A_t=0.4, D_t=0.2; (e) A_t=0, D_t=0; (f) A_t=0, D_t=0;.

ration tilt (D_t) for a segment of F_0 contour are defined as follows:

$$A_t = \frac{|A_r| - |A_f|}{|A_r| - |A_f|}$$

$$D_t = \frac{|D_r| - |D_f|}{|D_r| - |D_f|}$$

where A_r, and A_f represent the rise and fall in F_0 amplitude, respectively, with respect to peak value of fundamental frequency F_{o_p}. Similarly D_r and D_f represent the duration taken for rise and fall respectively. It can be observed from Figures 4.8(e) and (f) that the tilt parameters are not sufficient to reflect the swing of the F_0 values. Studies have shown that, speakers can vary the prominence of pitch accents by varying the height of the fundamental frequency peak, to express different degrees of emphasis. Likewise, the listeners judgment of prominence reflect the role of F_0 variation in relation to variation in prominence [104]. To express the height of the F_0 peak, the difference between peak and valley fundamental frequency ($\Delta F_0 = F_{0_p} \;\; F_{o_v}$) is used in this study. It has been observed that the length of the F_0 peak (length of onset) has a role in the perceptual prominence [104]. In this study, this is represented using the distance of F_0 peak location with respect to VOP (D_r).

In summary, the intonation features used for this language identification study are the following:

1. Change in F_0 (ΔF_0)
2. Distance of F_0 peak with respect to VOP (D_r)
3. Amplitude tilt (A_t)
4. Duration tilt (D_t)

Absolute values of the frame level F_0 are dependent on the physiological constraints, and hence are more speaker-dependent. Therefore absolute F_0 values are not included in the feature set for language discrimination studies. Positional details of syllables are not used in this study, as it is difficult to segment conversational speech into phrases.

4.3.2 Rhythm

Rhythmic properties of speech are felt when speech in different languages are contrasted. The ability to distinguish languages based on rhythm has been documented in infants as well as in adults [8]. According to the frame/content theory of speech production [105], all

spoken utterances are superimposed on successive syllables which constitute a "continual rhythmic alternation between an open and a closed mouth (a frame) on the production process". In [106], a consonant-vowel (CV) type of syllable is characterized as the basic rhythmic unit, beginning with a tight constriction and ending with an open vocal tract, resulting in a kind of rhythm. Two (correlated) variables defined over an utterance, namely the proportion of vocalic intervals and the standard deviation of the duration of consonantal intervals, are identified as correlates of linguistic rhythm [107]. Both these measures will be directly influenced by segmental inventory and the phone tactic regularities of a specific language.

In this work, we hypothesize that rhythm is perceived due to closing and opening of the vocal tract in the succession of syllables. The proportion of voiced intervals within each syllable region gives a measure of this transition. Segmenting continuous speech into syllable-like units enables representation of the rhythmic characteristics. We use the duration of syllable (D_s) (approximated to the distance between successive VOPs) and the duration of voiced region (D_v), to represent rhythm.

The voiced portion of syllables is obtained from the epochs which are extracted from the zero frequency method. It is clearly evident in the voiced speech, where significant excitation within a pitch period coincides with the glottal closure (GC) event. The instants of significant excitation show periodic nature in the voiced regions, and this is not present in unvoiced regions. This periodicity property along with the strength of excitation at the instants of glottal closure (strength of instants) is used for detecting the voiced regions.

The following features are used to represent rhythm:

1. Syllable duration (D_s)

2. Duration of voiced region (D_v) within each syllable

4.3.3 Stress

In all languages, some syllables are in some sense perceptually stronger than other syllables, and they are described as stressed syllables. The way stress manifests itself in the speech stream is highly language-dependent. The difference between strong and weak syllables is

of some linguistic importance in every language. However, languages differ in the linguistic function of such differences. It is necessary to consider what factors make a syllable count as stressed. It seems likely that stressed syllables are produced with greater effort than unstressed. This effort is manifested in the air pressure generated in the lungs for producing the syllable, and also in the articulatory movements in the vocal tract. A stressed syllable can produce the following audible changes:

1. Pitch prominence, in which the stressed syllable stand out from its context. Often a pitch glide such as a fall or rise is used for pitch prominence.

2. Stressed syllable tend to be longer. The length of the vowel in stressed syllable is longer than that of unstressed syllable. This syllable lengthening effect is noticeable in languages like English, and it is less in certain other languages.

3. Stressed syllable is powerful, intensive and loud in pronunciation than unstressed.

In most of the languages, higher intensity, larger pitch variation and longer duration help to assign prominence to stressed syllables. But the position of stressed syllable in a word varies from language to language. English is a stress-timed language, where stressed syllables appear roughly at a constant rate, and unstressed syllables are shortened to accommodate this. In some languages, stress is always placed on a given syllable, as in French, where the words are always stressed in the last syllable. In English and French, a longer duration syllable carries more pitch movements. But such a correlation may not hold equally well for all languages. Therefore, it is possible that, the specific interaction between the suprasegmental features, and relation between suprasegmental and segmental aspects, are the most salient characteristics that differentiate languages [108]. The syllable carrying stress is prominent with respect to the surrounding syllables, due to its loudness, large movement of F_0 and/or longer duration [101]. Therefore, along with F_0 and duration features mentioned above, we use change in log energy (ΔE) within voiced region to represent the stress.

It has been observed that tones of adjacent syllables influence both the shape and height of the F_0 contour of a particular syllable [102], and prominence of a syllable is estimated based on the pitch characteristics of the contour around it [104]. Similarly, rhythm is formed by a

sequence of syllables, and a syllable in isolation cannot be associated with rhythm. Therefore temporal dynamics of these parameters are important while representing the prosodic variations among languages. The context of a syllable, i.e., the characteristics of preceding and succeeding syllable is used to represent up-down movement of F_0 curve. When the distance of separation between two successive VOPs exceeds certain threshold, the region is hypothesized as a probable word/phrase boundary or as a long pause. In this work the language identification at syllabic and word (trisyllabic) levels is carried out using prosodic features represented by intonation, rhythm and stress.

4.4 Performance evaluation using syllable and word level prosodic features

In this work, speaker independent LID system is developed using prosodic features such as intonation, rhythm and stress (IRS) from syllable and word levels. The dimension of the feature vector representing IRS extracted from syllable is 7 (ΔF_0, D_r, A_t, D_t, D_s, D_v, ΔE), whereas the dimension of IRS feature vector at word level is 21 (present syllable(7), previous syllable(7), following syllable(7)). Thus word level feature vectors consists of concatenation of IRS features of previous, present and following syllables. In this manner, word level features are extracted for every syllable. In this study, LID systems at syllable and word levels are developed using Gaussian mixture models (GMMs), which can capture the language-specific information from the given prosodic features. Here language models are trained and tested using leave-one-speaker-out approach. For evaluating the developed LID systems, feature vectors derived from test speech samples are given to all the language models. The evidence in response to test speech samples from all the models are analyzed, and the highest evidence among the models is hypothesized as the language identity corresponding to the given speech sample. Here, test speech samples of 10 sec duration are considered for evaluating the performance of LID systems. The experiments are carried out for different Gaussian mixtures (32, 64, 128, 256, 512) and it is observed that overage recognition performance is better for 64 mixture components. The performance of the LID systems at syllable and word levels

using intonation, rhythm and stress features for 64 mixtures are given in Table 4.1. Column 1 of Table 4.1 indicates the languages used for discrimination. Columns 2 and 3 indicates the recognition performance of each language using IRS features derived from syllable and word levels respectively.

Table 4.1: Performance of LID system using intonation, rhythm and stress features at syllable and word levels

Language	Average recognition performance using IRS features (%)	
	Syllable level	Word level
Arunachali	15.35	20.06
Assamese	32.43	36.63
Bengali	53.19	56.49
Bhojpuri	25.32	28.62
Chhattisgarhi	36.34	38.33
Dogri	20.07	25.18
Gojri	43.81	46.85
Gujrati	21.48	25.34
Hindi	32.77	36.35
Indian English	41.46	44.95
Kannada	50.09	51.05
Kashmiri	53.17	56.40
Konkani	34.62	37.62
Manipuri	16.24	21.96
Mizo	16.72	17.68
Malyalam	21.59	25.82
Marathi	47.83	51.91
Nagamese	21.44	24.42
Neplai	46.64	48.38
Oriya	20.96	23.08
Punjabi	51.82	55.26
Rajasthani	25.75	29.35
Sanskrit	18.85	22.04
Sindhi	21.30	24.80
Tamil	37.72	40.44
Telugu	31.30	33.32
Urdu	25.82	28.58
Average	**32.00**	**35.22**

From the Table 4.1, it is observed that there is an improvement in the recognition performance for each language using IRS features at the word level, compared to syllable level. The overall language discrimination performance is improved by 3% with word level IRS

features compared to syllable level features.This improvement may be due to presence of dynamics of IRS features at word level due to concatenation of IRS features of 3 consecutive syllable. Whereas at syllable level the static characteristic of IRS features will be reflected.

Whereas spectral features discriminate some languages (e.g. Chhattisgarhi, Dogri, Mizo, Nagamese and Rajasthani) with very high accuracy and confuse with some other languages (e.g Assamese, Gojri, Gujarati, Hindi, Manipuri, Malyalam, Telugu and Urdu). It is also observed that the deviation in recognition accuracy is very high. From this, we can infer that, spectral features alone may not be good enough to discriminate the languages in uniform way. The performance results indicate that the proposed IRS features discriminate the language in uniform manner. For most of the languages, the recognition performance is close to overall (mean performance) recognition accuracy. None of the languages have recognized either very high or very low accurately. With this, we can say that the proposed IRS are balanced and stable across all the languages.

4.5 Extraction of prosodic features from global level

In this work, we have considered the sequence of 15 syllables to analyze the language-specific prosodic patterns at global level. In this study, the IRS features are not appropriate for representing the language-specific prosodic information. The main reason for not using IRS features is to avoid the redundant information, which is already captured using syllable and word level IRS features. For each syllable prosodic features such as ΔF_0, duration, and ΔE are extracted. In this study, absolute F_0s and energies are not considered in the feature set for discrimination of languages as they are dependent on the physiological constraints, and hence are more speaker-dependent. Therefore, the dynamics of ΔF_0s, durations, and ΔEs corresponding to sequence of 15 syllables are used to develop the LID system at global level. In this study feature vectors are derived by shifting one syllable at a time. With this the number of feature vector is equivalent to the number of syllable present in training and testing speech corpus. The dimension of feature vector is 45 (ΔF_0 (15), duration (15), and ΔE (15). The extraction and representation of the features are given in the following subsections.

4.5.1 ΔF_0 contour

The steps followed for deriving the ΔF_0 contour are as follows:

1. The sequence of fundamental frequency values (F_0) of each syllable is obtained using zero frequency filter method.
2. The successive difference of F_0 values of a syllable represent the sequence of ΔF_0s.
3. The average of ΔF_0s present within a syllable represent ΔF_0 of that particular syllable.
4. The sequence of ΔF_0s corresponds to sequence of 15 syllables represents ΔF_0 contour.

4.5.2 Duration contour

1. The duration of the syllable is determined as the time difference between the successive vowel onset points.
2. The sequence of 15 duration values corresponds to sequence of 15 syllables represents the duration contour.

4.5.3 ΔE contour

The steps followed for calculation of ΔE contour are as follows:

1. The frame energies within a syllable are calculated as

$$E = \sum_{i=1}^{N} x_i^2$$

2. The successive difference of E values of syllable represent the sequence of ΔEs.
3. The average of ΔEs present within a syllable represent ΔE of that particular syllable.
4. The sequence of ΔEs corresponds to sequence of 15 syllables represents ΔE contour.

4.6 Performance evaluation using global level prosodic features

In this work, speaker independent LID systems are developed separately using the following global features (i) ΔF_0 contour, (ii) duration contour, (iii) ΔE contour and (iv) ΔF_0 + duration + ΔE contour. Gaussian mixture models (GMMs) are used to capture the language-specific information from the given prosodic features. For evaluating the developed LID systems, feature vectors derived from test speech samples are given to all the language models. The evidence in response to test speech samples from all the models are analyzed, and the highest evidence among the models is hypothesized as the language identity corresponding to the given speech sample. Here, 10 sec duration of test samples are considered for evaluating the performance of LID systems. The experiments are carried out for different Gaussian mixtures (32, 64, 128, 256, 512) and it is observed that average performance is better for 64 mixture components. The performance of the LID system at global level with 64 mixtures is given in Table 4.2. Column 1 of Table 4.2 indicates the languages considered for discrimination. Columns 2-4 indicates the recognition performance of using ΔF_0 contour, duration contour and ΔE contour respectively. The last column indicates the recognition performance of the combination of above mentioned global prosodic features at feature level. The average recognition performance of the LID system is given in the last row of Table 4.2. From the results, it is observed that the trend in recognition performance is similar to IRS features in view of uniform discrimination. But the discrimination of individual languages is different for IRS features. Among the above feature sets, ΔF_0 contours have highest discrimination and ΔE contours have least discrimination. The combination of 3 features has improved the overall accuracy by 5%. This shows that the above feature sets contain some complementary language-specific information.

Table 4.2: Performance of LID system using prosodic features at global level

Language	Recognition performance (%)			
	ΔF_0 contour	duration contour	ΔE contour	ΔF_0 + duration + ΔE
Arunachali	12.13	8.71	5.33	16.96
Assamese	28.55	25.65	21.64	33.99
Bengali	49.57	46.68	42.54	54.80
Bhojpuri	21.73	18.78	14.93	27.12
Chhattisgarhi	33.11	29.07	26.00	38.03
Dogri	16.20	12.79	9.72	22.31
Gojri	40.15	36.87	33.58	45.61
Gujrati	18.06	15.01	11.07	23.06
Hindi	29.21	26.14	22.36	34.99
Indian English	38.00	34.72	30.77	43.75
Kannada	46.92	42.93	39.42	51.89
Kashmiri	49.85	46.74	42.64	54.68
Konkani	31.41	28.16	24.33	36.26
Manipuri	12.97	9.67	5.56	18.02
Mizo	13.40	9.72	6.29	18.67
Malyalam	18.11	14.51	11.32	23.24
Marathi	44.69	40.83	37.03	49.78
Nagamese	17.53	14.61	10.70	23.49
Neplai	42.69	39.73	36.19	48.25
Oriya	17.42	14.27	10.45	22.48
Punjabi	48.28	44.73	41.34	53.50
Rajasthani	22.35	19.16	15.61	27.45
Sanskrit	14.94	11.81	8.63	20.64
Sindhi	17.87	14.71	10.93	23.17
Tamil	34.52	30.97	27.56	39.21
Telugu	27.50	24.32	20.59	32.95
Urdu	22.37	18.70	15.69	27.95
Average	**28.50**	**25.18**	**21.57**	**33.79**

4.7 Performance evaluation using prosodic features on OGI-MLTS database

The proposed prosodic features extracted from syllable, word and global levels are also evaluated on OGI-MLTS database. Training and testing the models are carried out similar to the models developed using ILs speech database. The performance of the LID system using the prosodic at different levels is given in Table 4.3. From the Table 4.3, it is observed that

the results show similar phenomena as that of Indian language database. But the average performance of LID system using prosodic features on OGI-MLTS database is better than Indian database. This is mainly due to less number of languages and the languages which are present in the OGI database are distinct from each other. The set of 11 languages present in the database may be divided into 4-5 groups based on geographical regions where they use. French, German and Spanish may be grouped as European languages, Hindi, Tamil and Vietnamese may be grouped as South Asian languages, Japanese, Korean and Mandarin may be considered as Eastern Asian languages. In this way we can view the languages are very different to each other, and hence they can be discriminated with less difficulty.

Table 4.3: Language identification performance using prosodic features at syllable, word and global levels on OGI-MLTS database

Language	Recognition performance (%)		
	Syllable	Word	Global
English	45.12	48.24	48.08
Farsi	35.21	39.21	37.25
French	42.09	46.77	45.12
German	55.89	59.22	58.56
Hindi	47.15	51.35	48.23
Japanese	30.45	33.65	33.15
Korean	35.05	39.10	37.52
Mandarin	31.78	35.09	33.23
Spanish	55.45	60.27	58.45
Tamil	52.25	56.62	53.75
Vietnam	37.77	40.25	39.87
Average	**42.56**	**46.34**	**44.84**

4.8 LID using combination of features

In this work the LID study is carried out using spectral features and prosodic features. The prosodic features are explored at syllable, word and global levels for discriminating the languages. By analyzing the result, we have observed that all the proposed features carry the language-specific information. It is also noted that the language-specific information is distinct for each of the features considered. We can observe that some set of languages are recognized at high accuracy with some specific features. Therefore by combining the evi-

dences appropriately from different systems developed using individual features may improve the recognition accuracy by exploiting the complimentary evidences.

In this work we have explored various features and score level fusion techniques for improving the recognition performance. From the prosodic features prospective, we have explored individually by developing the LID systems separately with each of the following features (i) IRS features at syllable level, (ii) IRS features at word level, (iii) ΔF_0 contour, (iv) duration contour, (v) ΔE contour and (vi) ΔF_0 + duration + ΔE contour. In the context of global features, we have also examined the performance of score level fusion of individual global level prosodic features. But the recognition performance of global level prosodic features using score level fusion seems to be slightly inferior compared to feature level fusion. Therefore for future combinations, LID system with all global features together has considered for representing global prosodic features. In this work we have explored the LID task in four phases.

1. Phase 1: LID systems are developed separately using individual features.
2. Phase 2: LID systems developed using IRS features at syllable and word levels are combined using score level fusion.
3. Phase 3: LID systems developed at Phase 2 (i.e, syllable and word level) and LID system developed using all global prosodic features are combined.
4. Phase 4: LID system developed using spectral features and LID system developed using syllable, word and global prosodic features (Phase 3) are combined.

All four phases of combining individual LID systems are shown in 4.9.

In this work, LID systems of different modalities are combined by summing the weighted confidence scores (evidences). The weighting rule for combining the confidence scores of individual modalities is as follows:

$$c^m = \frac{1}{m}\sum_{i=1}^{m} w_i c_i$$

, where c^m is the confidence score, w_i and c_i are weighting factor and confidence score of the i^{th} modality, and m indicates number of modalities used for combining the scores. In this

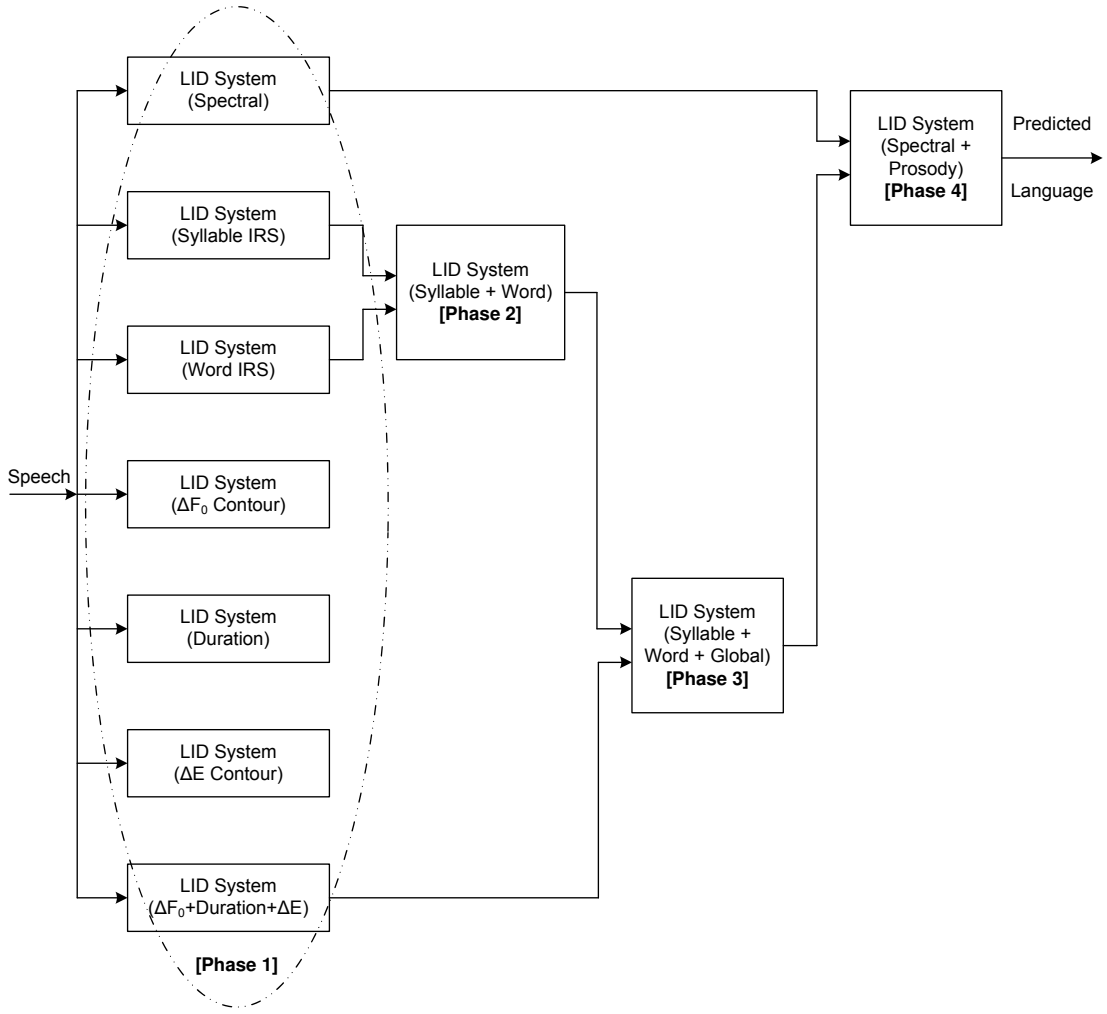

Figure 4.9: LID systems and their combinations in defferent phases.

study, each of the weights (w_i) is varied in steps of 0.1 from 0 to 1, and the sum of weights should equal unity ($\sum_{i=1}^{m} w_i = 1$).

4.8.1 Performance of LID system using IRS features from syllable and word levels

Performance of the combined LID system developed by combining the evidences from the LID systems developed using syllable and word level IRS features is given in second column of Table 4.4. Here, the scores of the individual systems are combined, using the weighting rule mentioned above. The recognition performance for various combinations of weighting

factors is shown in Figure 4.10. It is observed that the best recognition performance is about

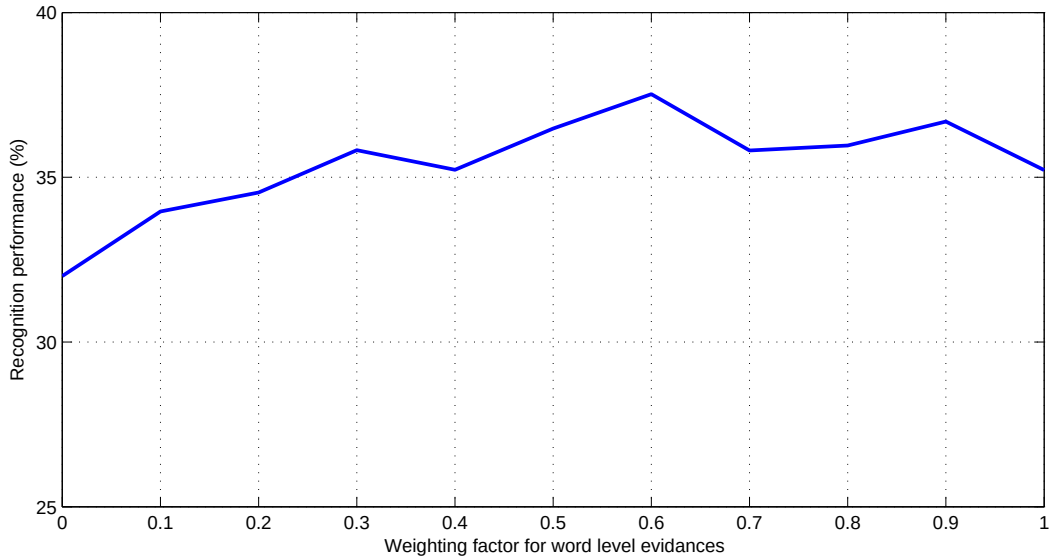

Figure 4.10: Variation in language recognition performance with different weighting factors, for the combination of prosodic features at syllable and word levels.

37.52%, and the corresponding weighting factors associated to LID systems developed using IRS features derived from syllable and word level features are 0.4 and 0.6 respectively. From the results, it is observed that the recognition performance improved in the combined system, compared to individual systems (see Table 4.4). This may be due to presence of some non overlapping language-specific information in the individual systems.

4.8.2 Performance of LID system using prosodic features from syllable, word and global level

Language-specific information represented by local and global prosodic features are combined in Phase 3. Here, the language-specific cues at syllable, word and global (phrase) levels are exploited for enhancing the recognition performance. It is known that syllable and word level prosodic features represent the language-specific information in the local level, and global prosodic features represents the language-specific information at a global level. Hence, combining the evidences may improve the performance. The recognition performance of the combined LID system developed by combining evidences from from LID systems developed

in Phase 2 and LID system developed using all global prosodic features is given in third column of Table 4.4. Here, the scores of the individual systems are combined using the weighting rule we mentioned above. The recognition performance for various combination of weighting factors is shown in Figure 4.11.

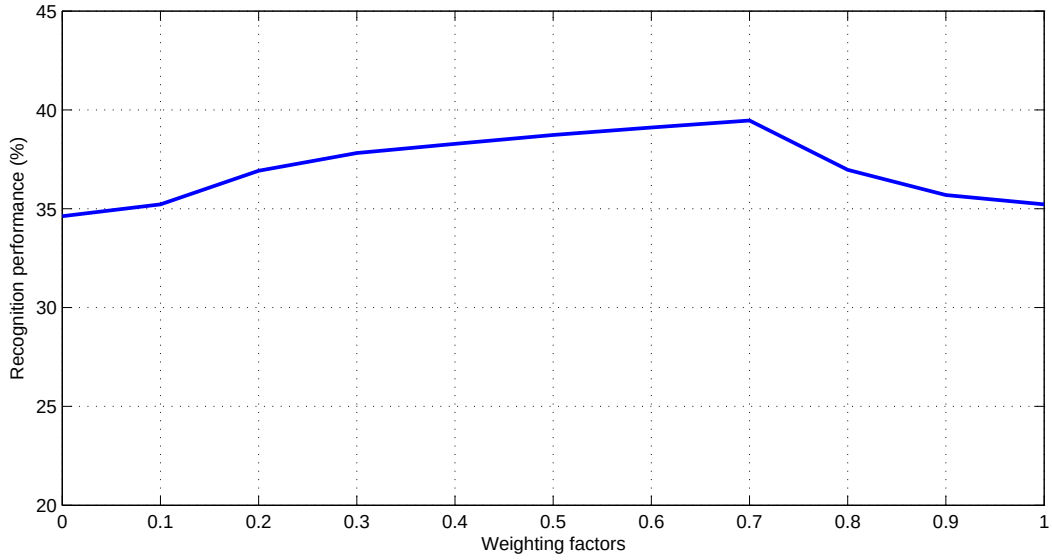

Figure 4.11: Variation in language recognition performance with different weighting factors for the combination of prosodic features at syllable, word and global level.

It is observed that the best recognition performance is about 39.46%, and the corresponding weighting factors associated to LID systems developed Phase 2 and all global level prosodic features are 0.7 and 0.3 respectively. From the results, it is observed that the recognition performance improved in the combined system, compared to individual systems (see table 4.4). This may be due to presence of some nonoverlapping language-specific information in the Phase 2 system and system developed by global prosodic features.

4.8.3 Performance of LID system using spectral and prosodic features

The evidences of LID systems based on spectral and prosodic features can be viewed as complementary to each other. This is because, spectral features (i.e, MFCCs) are extracted by processing the speech frames at segmental level, whereas prosodic features are extracted

Table 4.4: Language identification performance using the combination of prosodic features at (i) syllable and word levels (i) syllable, word and global levels.

Language	Recognition performance (%)	
	Syllable + Word	Syllable + Word + Global
Arunachali	21.01	21.81
Assamese	37.82	40.36
Bengali	58.59	60.90
Bhojpuri	31.25	32.70
Chhattisgarhi	41.59	43.51
Dogri	26.29	27.81
Gojri	49.59	50.86
Gujrati	26.70	28.68
Hindi	38.80	40.61
Indian English	47.12	49.92
Kannada	55.24	58.02
Kashmiri	58.46	60.54
Konkani	40.26	41.64
Manipuri	22.16	24.18
Mizo	22.09	24.33
Malyalam	27.05	28.57
Marathi	53.50	56.17
Nagamese	26.80	29.77
Neplai	51.85	53.94
Oriya	25.93	28.30
Punjabi	57.51	59.26
Rajasthani	31.03	32.52
Sanskrit	24.42	25.88
Sindhi	26.62	28.72
Tamil	43.05	44.32
Telugu	36.49	39.17
Urdu	31.84	33.00
Average	**37.52**	**39.46**

from the supra-segmental level of speech signal. Another basic difference between these two features is that spectral features characterizes the languages in form of distinct shapes of the vocal tract, whereas prosodic features characterizes the languages in terms of distinct duration, intonation and stress patterns. Hence, combining the evidences of these two systems may enhance the recognition performance further.

From the studies carried out so far, it is evident that spectral features and prosodic features have certain language discriminative power. It is also observed that the languages

such as Arunachali, Bengali, Chhattisgarhi, Dogri, Kannada, Kashmiri, Mizo, Marathi, Nagamese, Neplai are well discriminated and languages like Assamese, Gojri, Gujarati, Hindi, Indian English, Manipuri, Malyalam, Sanskrit, Telugu, Urdu are poorly discriminated using spectral information. With prosodic features languages like Assamese, Bengali, Bhojpuri, Chhattisgarhi, Gojri, Hindi, Indian English, Manipuri, Kannada, Kashmiri, Konkani, Marathi, Neplai, Punjabi, Tamil, Telugu are well classified and Arunachali, Mizo, Nagamese are having poor performance. From this we can hypothesize that spectral and prosodic features in discriminating the languages are contrary to each other. Combining the evidences from the LID systems developed using these features (spectral and prosodic) may improve the performance.The recognition performance of the combined LID system developed by combining evidences from LID systems developed in Phase 3 and LID system developed using spectral features in Phase 1 is given in Table 4.5. Here, the scores of the individual systems are combined using the weighting rule mentioned above. The recognition performance for various combination of weighting factors is shown in Figure 4.12.

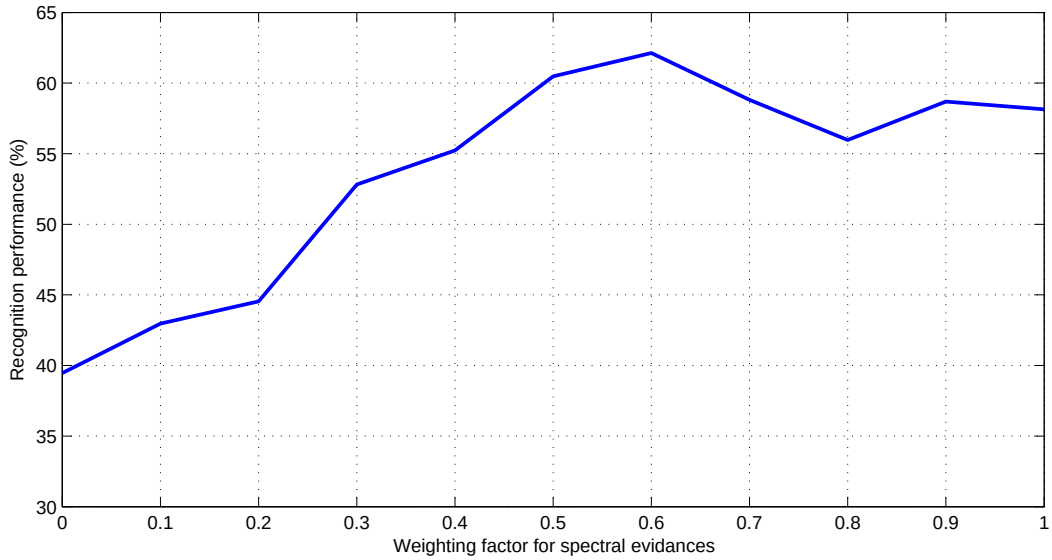

Figure 4.12: Variation in language recognition performance with different weighting factors, for the combination of spectral and prosodic features.

It is observed that, the best recognition performance is about 62.13% for the weighting factors of 0.6 and 0.4 to the confidence scores of spectral and prosodic features respectively. From the results (see Table 4.5), it is observed that recognition performance of languages

are improved in the combined system. The recognition performance of the combined system is observed to be increased by around 11% compared to the system developed using spectral features alone. The improvement in the performance may be due to the supplementary information provided by the prosodic features.

Table 4.5: Language identification performance using the combination of spectral and prosodic features.

Language	Recognition performance (%)
	Spectral + Prosodic
Arunachali	86.19
Assamese	21.49
Bengali	72.73
Bhojpuri	45.31
Chhattisgarhi	100.00
Dogri	100.00
Gojri	31.57
Gujrati	32.95
Hindi	36.77
Indian English	39.84
Kannada	84.69
Kashmiri	83.17
Konkani	49.92
Manipuri	33.96
Mizo	100.00
Malyalam	24.78
Marathi	88.39
Nagamese	100.00
Neplai	81.67
Oriya	68.34
Punjabi	65.19
Rajasthani	100.00
Sanskrit	38.84
Sindhi	75.85
Tamil	63.67
Telugu	23.20
Urdu	28.87
Average	**62.13**

4.9 Summary and conclusions

In this work, the developed Indian language speech database is analyzed in view of language identification by using prosodic features. In this work IRS features are used to represent the language-specific prosodic information at syllable and word levels. F_0, duration and energy contours are used for representing the language-specific information at global (phrase) level. Results have indicated that prosodic features derived from words (tri-syllables) have better language discrimination ability compared to syllable and global levels. The recognition performance is further improved by fusion of evidences from the prosodic features derived from syllable, word and global levels. The complementary nature of spectral and prosodic features is exploited for further improvement in the performance of LID system by combining the evidences of spectral and prosodic features. Experiments were also performed on OGI-MLTS database using prosodic features at different levels.

Chapter 5

Summary and conclusions

5.1 Summary of the thesis

In this thesis, we have focused on the extraction and representation of language-specific features at different levels of speech signal, for the purpose of language identification. For recognizing a language, different cues are available in speech. Languages differ in acoustic-phonetics, prosody, phonotactics, morphology and syntax. These characteristics are manifested at different time spans of the speech signal. To represent these characteristics, features should be derived from multiple levels of speech signal.

In this work, the problem of language identification has been formulated using a probabilistic approach. The observations from speech signal has been split into different components based on the level of manifestation. These components represented in terms of different feature vectors were modeled separately using Gaussian mixture models. The evidence obtained from different levels was later combined to arrive at a decision.

For automatic language identification, language-specific features at four different levels of speech signal, namely, frame, syllabic, tri-syllabic (word) and global levels are explored. At the frame level, spectral features derived from lower span of speech signal ($<$ 20 msec) were used for language identification. Features such as linear prediction cepstral coefficients and mel-frequency cepstral coefficients are examined. The presence of language-specific information in these features was demonstrated using twenty-seven Indian languages. It was observed that the frame level features are dominated by the physiological characteristics of

the speaker. Speaker specific models with k-best scoring approaches are explored to reduce the effect of speaker variability on language identification. The similarities and dissimilarities among the languages are examined using pair-wise testing. To improve the classification accuracy further by exploiting teh finer variation in spectral characteristic, we proposed spectral features using pitch synchronous analysis (PSA) and glottal closure region (GCR) approaches. In the case of PSA, spectral features are derived from each pitch cycle, whereas in case of GCR approaches, spectral features are derived from 30% of pitch cycle starting from the glottal closure instants. It was observed that the spectral features from PSA effectively captures the finer variations in the spectral characteristics from the adjacent pitch cycles, and these finer variations may be more language specific, and hence it leads to better language discrimination. In the case of GCR based spectral features, LID performance is still better compared to PSA. For analyzing the robustness of the proposed feature extraction methods, LID systems developed based on speaker specific language models are evaluated under noisy condition. It is observed that spectral features derived from GCR are more robust to noise compared to spectral features derived from PSA and BP approach. This is mainly due to the presence of high SNR speech in GCR. The effectiveness of the proposed spectral features is also demonstrated using OGI-MLTS database.

To extract prosodic features from syllable, word and global levels, VOPs are used as anchor points. In this work, VOPs are determined using spectral energy within 500–2500 Hz frequency band from the GCR. The prosodic characteristics of languages differ in terms of rhythm, intonation and stress (IRS). These characteristics are manifested in acoustic speech signal in terms of duration, F_0 (pitch) and energy. The IRS features extracted from syllable and word levels are examined separately for discriminating the languages. It was observed that IRS features from word level were slightly better performed over syllable level. This may be due to presence of IRS interactions among adjacent syllables. At the global level, sequence of 15 syllables were considered to represent the language-specific prosodic characteristics. Prosodic features at global level are represented by ΔF_0 contour, duration contour and ΔE contour associated in the sequence of 15 syllables. It is observed that the performance of LID system is better for ΔF_0 contour compared to other global level features. The performance of the language identification system using prosodic features is improved by combining the

evidences from syllable, word and global levels. The contrary language-specific information present in the spectral and prosodic features is further exploited by combining the evidences of LID systems developed using spectral and prosodic features with different weights. It is observed that the language discrimination is enhanced with the combination of spectral and prosodic features. The effectiveness of prosodic features for language discrimination is also demonstrated on OGI-MLTS database.

5.2 Major contributions of the thesis

- IITKGP-MLILSC (Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus) is developed for promoting LID research on ILs.
- Spectral features from pitch synchronous analysis and glottal closure regions are proposed for discriminating the ILs.
- Intonation, rhythm and stress features at individual syllable level and in the sequence of syllables within a word are proposed for identifying the ILs.
- Prosodic patterns at global level are proposed in terms of variation of F_0, intensity and duration patterns for recognizing the ILs.
- Language-specific information from the spectral and prosodic features is combined for improving the accuracy of LID system.

5.3 Scope for future work

- In addition to spectral and prosodic features, one can explore language specific source features for identifying the languages. The recognition performance may be further improved by combining spectral, prosodic and source information effectively. Other nonlinear models such as neural networks, support vector machines and hidden Markov models can be explored in future for further improvement in the performance.

- This work has attempted the use of implicit acoustic-phonetics and prosody for language identification. Implicit phonotactic features are not explored in this work. It may be possible to represent the phonotactics of a language without explicitly identifying the subword units. There is scope for exploring the representation of phonotactics (alignment of subword units) by implicit means, in a manner useful for language identification.

- Prosodic features give rise to grouping of syllables and words into larger chunks. There are prosodic features that indicate the relation between such groups, indicating that two or more groups of syllables are linked in some way. This linguistic aspect of intonation is a part of the structure of language and specific to any given language [101]. It may be possible to employ prosodic features corresponding to groups of syllables, instead of using features of a syllable and its context for language identification.

- For a language, there can be different accents. Accent variations are felt due to variations in pronunciations and prosodic gestures. Therefore, the difference in accents may be represented using acoustic-phonetic and prosodic features. There is scope for extending the approach used in this study for accent identification studies as well.

- Multi-stage approaches and hybrid models (linear and non-linear) can be explored in future to improve the performance of the LID systems.

Appendix A

LPCC Features

The cepstral coefficients derived from either linear prediction (LP) analysis or a filter bank approach are almost treated as standard front end features. Speech systems developed based on these features have achieved a very high level of accuracy, for speech recorded in a clean environment. Basically, spectral features represent phonetic information, as they are derived directly from spectra. The features extracted from spectra, using the energy values of linearly arranged filter banks, equally emphasize the contribution of all frequency components of a speech signal. In this context, LPCCs are used to capture emotion-specific information manifested through vocal tract features. In this work, the 10^{th} order LP analysis has been performed, on the speech signal, to obtain 13 LPCCs per speech frame of 20 ms using a frame shift of 10 ms. The human way of emotion recognition depends equally on two factors, namely: its expression by the speaker as well as its perception by a listener. The purpose of using LPCCs is to consider vocal tract characteristics of the speaker, while performing automatic emotion recognition.

Cepstrum may be obtained using linear prediction analysis of a speech signal. The basic idea behind linear predictive analysis is that the n^{th} speech sample can be estimated by a linear combination of its previous p samples as shown in the following equation.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p)$$

where a_1,a_2, a_3 $\cdots$ are assumed to be constants over a speech analysis frame. These are

known as predictor coefficients or linear predictive coefficients. These coefficients are used to predict the speech samples. The difference of actual and predicted speech samples is known as an error. It is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

where $e(n)$ is the error in prediction, $s(n)$ is the original speech signal, $\hat{s}(n)$ is a predicted speech signal, $a_k s$ are the predictor coefficients.

To compute a unique set of predictor coefficients, the sum of squared differences between the actual and predicted speech samples has been minimized (error minimization) as shown in the equation below

$$E_n = \sum_{m} \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$

where m is the number of samples in an analysis frame. To solve the above equation for LP coefficients, E_n has to be differentiated with respect to each a_k and the result is equated to zero as shown below

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad \text{for } k = 1, 2, 3, \cdots, p$$

After finding the $a_k s$, one may find cepstral coefficients using the following recursion.

$$C_0 = \log_e p$$

$$C_m = a_m + \textstyle\sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad \text{for } 1 < m < p \text{ and}$$

$$C_m = \textstyle\sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad \text{for } m > p$$

Appendix B

MFCC Features

The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. The detailed description of various steps involved in the MFCC feature extraction is explained below.

1. **Pre-emphasis:** Pre-emphasis refers to filtering that emphasizes the higher frequencies. Its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region. For voiced sounds, the glottal source has an approximately -12 dB/octave slope [109]. However, when the acoustic energy radiates from the lips, this causes a roughly +6 dB/octave boost to the spectrum. As a result, a speech signal when recorded with a microphone from a distance has approximately a -6 dB/octave slope downward compared to the true spectrum of the vocal tract. Therefore, pre-emphasis removes some of the glottal effects from the vocal tract parameters. The most commonly used pre-emphasis filter is given by the following transfer function

$$H(z) = 1 - bz^{-1} \tag{B.1}$$

   where the value of b controls the slope of the filter and is usually between 0.4 to 1.0 [109].

2. **Frame blocking and windowing:** The speech signal is a slowly time-varying or

quasi-stationary signal. For stable acoustic characteristics, speech needs to be examined over a sufficiently short period of time. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over 20 ms windows, and advanced every 10 ms [110, 111]. Advancing the time window every 10 ms enables the temporal characteristics of individual speech sounds to be tracked and the 20 ms analysis window is usually sufficient to provide good spectral resolution of these sounds, and at the same time short enough to resolve significant temporal characteristics. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frame. On each frame a window is applied to taper the signal towards the frame boundaries. Generally, Hanning or Hamming windows are used [109]. This is done to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the DFT on the signal.

3. **DFT spectrum:** Each windowed frame is converted into magnitude spectrum by applying DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \qquad 0 \leq k \leq N-1 \tag{B.2}$$

where N is the number of points used to compute the DFT.

4. **Mel-spectrum:** Mel-Spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as mel-filter bank. A mel is a unit of measure based on the human ears perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly. The mel scale is approximately a linear frequency spacing below 1 kHz, and a logarithmic spacing above 1 kHz [112]. The approximation of mel from physical frequency can be expressed as

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{B.3}$$

where f denotes the physical frequency in Hz, and f_{mel} denotes the perceived frequency [110].

Filter banks can be implemented in both time domain and frequency domain. For MFCC computation, filter banks are generally implemented in frequency domain. The center frequencies of the filters are normally evenly spaced on the frequency axis. However, in order to mimic the human ears perception, the warped axis according to the non-linear function given in Eqn. (B.3), is implemented. The most commonly used filter shaper is triangular, and in some cases the Hanning filter can be found [109]. The triangular filter banks with mel-frequency warping is given in Figure B.1.

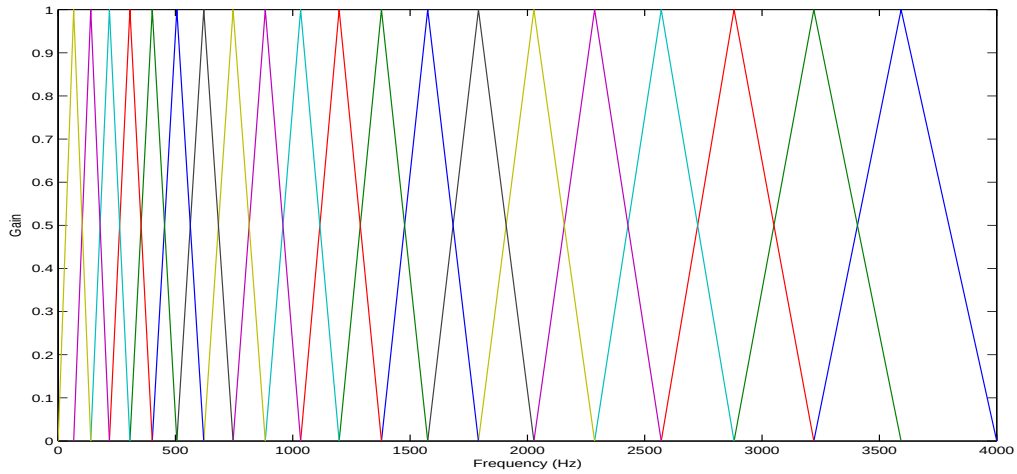

Figure B.1: Mel-filter bank

The mel spectrum of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the of the triangular mel weighting filters.

$$s(m) = \sum_{k=0}^{N-1} \left[|X(k)|^2 H_m(k) \right]; \qquad 0 \leq m \leq M-1 \tag{B.4}$$

where M is total number of triangular mel weighting filters [113, 114]. $H_m(k)$ is the weight given to the k^{th} energy spectrum bin contributing to the m^{th} output band and

is expressed as :

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{B.5}$$

with m ranging from 0 to $M-1$.

5. **Discrete Cosine Transform (DCT):** Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients produces a set of cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low quefrequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components [109]. Finally, MFCC is calculated as [109]

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0, 1, 2, ..., C-1 \tag{B.6}$$

where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. Traditional MFCC systems use only 8 to 13 cepstral coefficients. The zeroth coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information.

6. **Dynamic MFCC features:** The cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first and second derivatives of cepstral coefficients [115, 116]. The first order derivative is called delta coefficients, and the second order derivative is called delta-delta coefficients. Delta coefficients tell about the speech rate, and delta-delta coefficients provide information

similar to acceleration of speech. The commonly used definition for computing dynamic parameter is

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{B.7}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight and T is the number of successive frames used for computation. Generally T is taken as 2. The delta-delta coefficients are computed by taking the first order derivative of the delta coefficients.

Appendix C

Gaussian Mixture Model (GMM)

In the speech and speaker recognition the acoustic events are usually modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix. However unimodel PDF with only one mean and covariance are unsuitable to model all variations of a single event in speech signals. Therefore, a mixture of single densities is used to model the complex structure of the density probability. For a D-dimensional feature vector denoted as x_t, the mixture density for speaker Ω is defined as weighted sum of M component Gaussian densities as given by the following [117]

$$P(x_t|\Omega) = \sum_{i=1}^{M} w_i P_i(x_t) \tag{C.1}$$

where w_i are the weights and $P_i(x_t)$ are the component densities. Each component density is a D-variate Gaussian function of the form

$$P_i(x_t) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(x_t-\mu_i)'\Sigma_i^{-1}(x_t-\mu_i)\right]} \tag{C.2}$$

where μ_i is a mean vector and Σ_i covariance matrix for i^{th} component. The mixture weights have to satisfy the constraint [117]

$$\sum_{i=1}^{M} w_i = 1. \tag{C.3}$$

The complete Gaussian mixture density is parameterized by the mean vector, the covariance matrix and the mixture weight from all component densities. These parameters are collectively represented by

$$\Omega = \{w_i, \mu_i, \Sigma_i\}; \qquad i = 1, 2, ...M. \tag{C.4}$$

C.1 Training the GMMs

To determine the model parameters of GMM of the speaker, the GMM has to be trained. In the training process, the maximum likelihood (ML) procedure is adopted to estimate model parameters. For a sequence of training vectors $X = \{x_1, x_2, .., x_T\}$, the GMM likelihood can be written as (assuming observations independence) [117]

$$P(X|\Omega) = \prod_{t=1}^{T} P(x_t|\Omega). \tag{C.5}$$

Usually this is done by taking the logarithm and is commonly named as log-likelihood function. From Eqns. (C.1) and (C.5), the log-likelihood function can be written as

$$\log\left[P(X|\Omega)\right] = \sum_{t=1}^{T} \log\left[\sum_{i=1}^{M} w_i P_i(x_t)\right]. \tag{C.6}$$

Often, the average log-likelihood is used value is used by dividing $\log\left[P(X|\Omega)\right]$ by T. This is done to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor [118]. The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. The likelihood value is, however, a highly nonlinear function in the model parameters and direct maximization is not possible. Instead, maximization is done through iterative procedures. Of the many techniques developed to maximize the likelihood value, the most popular is the iterative expectation maximization

(EM) algorithm [119].

C.1.1 Expectation Maximization (EM) Algorithm

The EM algorithm begins with an initial model Ω and tends to estimate a new model such that the likelihood of the model increasing with each iteration. This new model is considered to be an initial model in the next iteration and the entire process is repeated until a certain convergence threshold is obtained or a certain predetermined number of iterations have been made. A summary of the various steps followed in the EM algorithm are described below.

1. **Initialization:** In this step an initial estimate of the parameters is obtained. The performance of the EM algorithm depends on this initialization. Generally, LBG [120] or K-means algorithm [121, 122] is used to initialize the GMM parameters.

2. **Likelihood Computation:** In each iteration the posterior probabilities for the i^{th} mixture is computed as [117]:

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{C.7}$$

3. **Parameter Update:** Having the posterior probabilities, the model parameters are updated according to the following expressions [117].

   Mixture weight update:

$$\overline{w_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t)}{T}. \tag{C.8}$$

   Mean vector update:

$$\overline{\mu_i} = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{C.9}$$

Covariance matrix update:

$$\overline{\sigma}_i^2 = \frac{\sum_{i=1}^{T} \Pr(i|x_t) \left|x_t - \overline{\mu_i}\right|^2}{\sum_{i=1}^{T} \Pr(i|x_t)}. \tag{C.10}$$

In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. It is more common to use diagonal covariance matrices for GMM, since linear combination of diagonal covariance Gaussians has the same model capability with full matrices [123]. Another reason is that speech utterances are usually parameterized with cepstral features. Cepstral features are more compactable, discriminative, and most important, they are nearly uncorrelated, which allows diagonal covariance to be used by the GMMs [117, 124]. The iterative process is normally carried out 10 times, at which point the model is assumed to converge to a local maximum [117].

C.1.2 Maximum *a posteriori* (MAP) Adaptation

Gaussian mixture models for a speaker can be trained using the modeling described earlier. For this, it is necessary that sufficient training data is available in order to create a model of the speaker. Another way of estimating a statistical model, which is especially useful when the training data available is of short duration, is by using maximum *a posteriori* adaptation (MAP) of a background model trained on the speech data of several other speakers [125]. This background model is a large GMM that is trained with a large amount of data which encompasses the different kinds of speech that may be encountered by the system during training. These different kinds may include different channel conditions, composition of speakers, acoustic conditions, etc. A summary of MAP adaptation steps are given below.

For each mixture i from the background model, $Pr(i|x_t)$ is calculated as [126]

$$\Pr(i|x_t) = \frac{w_i P_i(x_t)}{\sum_{j=1}^{M} w_j P_j(x_t)}. \tag{C.11}$$

Using $Pr(i|x_t)$, the statistics of the weight, mean and variance are calculated as follows [126]

$$n_i = \sum_{i=1}^{T} \Pr(i|x_t) \tag{C.12}$$

$$E_i(x_t) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t}{n_i} \tag{C.13}$$

$$E_i(x_t^2) = \frac{\sum_{i=1}^{T} \Pr(i|x_t) x_t^2}{n_i}. \tag{C.14}$$

These new statistics calculated from the training data are then used adapt the background model, and the new weights ($\hat{w}_i$), means ($\hat{\mu}_i$) and variances ($\hat{\sigma}_i^2$) are given by [126]

$$\hat{w}_i = \left[\frac{\alpha_i n_i}{T} + (1-\alpha_i) w_i\right] \gamma \tag{C.15}$$

$$\hat{\mu}_i = \alpha_i E_i(x_t) + (1-\alpha_i)\mu_i \tag{C.16}$$

$$\hat{\sigma}_i^2 = \alpha_i E_i(x_t^2) + (1-\alpha_i)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2. \tag{C.17}$$

A scale factor γ is used, which ensures that all the new mixture weights sum to 1. α_i is the adaptation coefficient which controls the balance between the old and new model parameter estimates. α_i is defined as [126]

$$\alpha_i = \frac{n_i}{n_i + r} \tag{C.18}$$

where r is a fixed relevance factor, which determines the extent of mixing of the old and new estimates of the parameters. Low values for α_i ($\alpha_i \to 0$), will result in new parameter estimates from the data to be de-emphasized, while higher values ($\alpha_i \to 1$) will emphasize the use of the new training data-dependent parameters. Generally only mean values are adapted [118]. It is experimentally shown that mean adaptation gives slightly higher performance than adapting all three parameters [126].

C.2 Testing

In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with maximum likelihood is selected as identified speaker. For example, if S speaker models $\{\Omega_{1,}\Omega_{2,...,}\Omega_S\}$ are available after the training, speaker identification can be done based on a new speech data set. First, the sequence of feature vectors $X = \{x_1, x_2, .., x_T\}$ is calculated. Then the speaker model $\hat{s}$ is determined which maximizes the a posteriori probability $P(\Omega_S|X)$. That is, according to the Bayes rule [117]

$$\hat{s} = \max_{1 \leq s \leq S} P(\Omega_S|X) = \max_{1 \leq s \leq S} \frac{P(X|\Omega_S)}{P(X)} P(\Omega_S). \tag{C.19}$$

Assuming equal probability of all speakers and the statistical independence of the observations, the decision rule for the most probable speaker can be redefined as

$$\hat{s} = \max_{1 \leq s \leq S} \sum_{t=1}^{T} \log P(x_t|\Omega_s) \tag{C.20}$$

with T the number of feature vectors of the speech data set under test and $P(x_t|\Omega_s)$ given by Eqn. (C.1).

Decision in verification is obtained by comparing the score computed using the model for the claimed speaker Ω_S given by $P(\Omega_S|X)$ to a predefined threshold θ. The claim is accepted if $P(\Omega_S|X) > \theta$, and rejected otherwise [118].

Publications

1. Sudhamay Maity, Anil Kumar Vuppala, K. Sreenivasa Rao and Dipanjan Nandi,"IITKGP-MLILSC Speech Database for Language Identification, *Eighteenth National Conference on Communications(NCC)*, 2012.

2. K. Sreenivasa Rao and Sudhamay Maity,"Pitch Synchronous and Glottal Closure based Speech Analysis for Language Recognition",*Computer Speech and Language.*(Communicated)

Bibliography

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech, Audio Processing*, vol. 4, pp. 31–44, 1996.

[2] A. K. V. S. Jayaram, V. Ramasubramanian, and T. V. Sreenivas, "Language identification using parallel sub-word recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. I, pp. 32–35, 2003.

[3] R. A. Cole, J. W. T. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, "Language identification with neural networks: A feasibility study," in *Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, pp. 525–529, 1989.

[4] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text independent language identification by HMM," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-1992),*, pp. 1011–1014, 1992.

[5] E. Wong and S. Sridharan, "Gaussian mixture model based language identification system," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-2002)*, pp. 93–96, 2002.

[6] Z. Lu-Feng, S. Man-hung, Y. Xi, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," in *Proc. Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006*, pp. 1–6, 2006.

[7] Y. K. Muthusamy and R. A. Cole, "Automatic segmentation and identification of ten languages using telephone speech," in *Proc. Int. Conf. Spoken Language Processing*, (Banff, Alberta, Canada), pp. 1007–1010, Oct. 1992.

[8] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *J. Acoust. Soc. Am.*, vol. 105, pp. 512–521, Jan. 1999.

[9] K. Mori, N. Toba, T. Harada, T. Arai, M. Kometsu, M. Aoyagi, and Y. Murahara, "Human language identification with reduced spectral information," in *Proc. EUROSPEECH*, vol. 1, (Budapest, Hungary), pp. 391–394, Sep. 1999.

[10] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, vol. 2, pp. 781–784, 1996.

[11] J. Laver, *Principles of Phonetics*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[12] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall, 2 ed., 2008.

[13] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall, 1993.

[14] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Proc. 14th Annual Speech Research Symp.*, 1994.

[15] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. New York: Academic, 2006.

[16] M. Yip, *Tone*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[17] L. Bauer, *Introducing Linguistic Morphology*. Georgetown Univ. Press, 2003.

[18] A. Carnie, *Syntax: A Generative Introduction*. New York: Wiley-Blackwell, 2 ed., 2006.

[19] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of Int. Conf. Spoken Language Processing*, pp. 895–898, Oct. 1992.

[20] LDC,Philadelphia,PA,http://www.ldc.upenn.edu/Catalog, 1996. LDC96S46 - LDC96S60.

[21] Y. K. Muthusamy, N. Jain, and R. A. Cole, "Perceptual benchmarks for automatic language identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 333–336, Apr. 1994.

[22] L. F. Lamel and J. L. Gauvain, "Cross lingual experiments with phone recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 507–510, Apr. 1993.

[23] L. F. Lamel and J. L. Gauvain, "Language identification using phonebased acoustic likelihoods," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 293–296, Apr. 1994.

[24] K. M. Berkling, T. Arai, and E. Bernard, "Analysis of phoneme based features for langugae identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 289–292, Apr. 1994.

[25] T. J. Hazen and V. W. Zue, "Recent improvements in an approach to segement-based automatic language identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 1883–1886, Sep. 1994.

[26] O. Andersen, P. Dalsgaard, and W. Barry, "On the use of datadriven clustering technique for identification of poly and mono-phonemes for four European languages," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 121–124, Apr. 1994.

[27] R. C. F. Tucker, M. J. Carey, and E. S. Paris, "Automatic language identification using sub-words models," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 301–304, Apr. 1994.

[28] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-94)*, vol. 1, pp. I/305–I/308, 1994.

[29] S. Kadambe and J. L. Hieronymus, "Language identification with phonological and lexical models," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 3507–3510, May 1995.

[30] Y. Yan and E. Barnard, "Analysis approach to automatic langauge identification based on language-dependent phone recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 5, pp. 3511–3514, May 1995.

[31] J. Navratil and W. Zuhlke, "Phonetic-context mapping in language identification," in *Proceedings of EUROSPEECH*, vol. 1, (Greece), pp. 71–74, Sep. 1997.

[32] J. Navratil, "Spoken language recognition a step toward multilinguality in speech processing," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 678–685, Sep. 2001.

[33] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification," *J. Acoust. Soc. Amer.*, vol. 101, pp. 2323–2331, 1997.

[34] K. Kirchhoff and S. Parandekar, "Multi-stream statistical N-gram modeling with application to automatic language identification," in *Proc. EUROSPEECH-2001*, pp. 803–806, 2001.

[35] V. K. Prasad, *Segmentation and Recognition of Continuous Speech*. PhD thesis, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 2003.

[36] V. Ramasubramanian, A. K. V. S. Jayaram, and T. V. Sreenivas, "Language identification using parallel phone recognition," in *WSLP, TIFR*, (Mumbai), pp. 109–116, Jan. 2003.

[37] J. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proc. INTERSPEECH-2004*, pp. 25–28, 2004.

[38] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," in *Proc. IEEE Odyssey 2006: Speaker and Language Recognition Workshop*, pp. 1–6, 2006.

[39] T. P. Gleason and M. A. Zissman, "Composite background models and score standardization for language identification systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP01)*, vol. 1, pp. 529–532, 2001.

[40] R. Cordoba, L. Dharo, F. Fernandez-Martinez, J. Macias-Guarasa, and J. Ferreiros, "Language identification based on n-gram frequency ranking," in *Proc. EUROSPEECH-2007*, pp. 2137–2140, 2007.

[41] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 271–284, Jan. 2007.

[42] S. K. Chai and L. Haizhou, "On acoustic diversification front-end for spoken language identification," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, pp. 1029–1037, 2008.

[43] R. Tong, B. Ma, H. Li, and E. Chng, "Target-oriented phone selection from universal phone set for spoken language recognition," in *Proc. INTERSPEECH- 2008*, 2008.

[44] J.-L. You, Y.-N. Chen, M. Chu, F. K. Soong, and J.-L. Wang, "Identifying Language Origin of Named Entity With Multiple Information Sources," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, pp. 1077–1086, Aug. 2008.

[45] G. R. Botha and E. Barnard, "Factors that affect the accuracy of text-based language identification," *Computer Speech and Language*, vol. 26, p. 307320, 2012.

[46] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, pp. 115–124, 2001.

[47] A. F. Martin and M. A. Przybocki, "NIST 2003 language recognition evaluation," in *Proceedings of EUROSPEECH*, (Geneva, Switzerland), pp. 1341–1344, Sep. 2003.

[48] R. G. Leonard and G. R. Doddington, "Automatic language identification," tech. rep., A.F.R.A.D. Centre Tech. Rep. RADC-TR-74-200, 1974.

[49] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance," *J. Acoust. Soc. Amer.*, vol. 62, pp. 708–713, 1977.

[50] D. Cimarusti and R. B. Eves, "Development of an automatic identification system of spoken languages : Phase I," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 1661–1663, May 1982.

[51] S. Eady, "Differences in F0 patterns of speech: Tone languages versus stress language," *Lang. Speech*, vol. 25, pp. 29–42, 1982.

[52] R. Ives, "A minimal rule AI expert system for real-time classification of natural spoken languages," in *Proc. 2nd Annual Artificial Intelligence and Advanced Computer Technology Conf*, pp. 337–340, 1986.

[53] J. T. Foil, "Language identification using noisy speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 861–864, Apr. 1986.

[54] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 528–531, May 1989.

[55] Y. K. Muthusamy, R. A. Cole, and M. Gopalakrishnan, "A segment-based approach to automatic language identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 353–356, Apr. 1991.

[56] M. Sugiyama, "Automatic language recognition using acoustic features," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 813–816, May 1991.

[57] L. Riek, W. Mistreta, and D. Morgan, "Experiments in language identification," tech. rep., Lockheed Sanders Tech. Rep. SPCOT-91-002, 1991.

[58] M. A. Zissman, "Automatic langauge identification using Gaussian mixture and hidden Markov models," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 399–402, Apr. 1993.

[59] S. Itahashi, J. Zhou, and K. Tanaka, "Spoken language discrimination using speech fundamental frequency," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-1994)*, pp. 1899–1902, 1994.

[60] I. Shuichi and D. Liang, "Language identification based on speech fundamental frequency," in *Proc. EUROSPEECH-1995*, pp. 1359–1362, 1995.

[61] K. Li, "Automatic language identification using syllabic features," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 297–300, 1994.

[62] F. Pellegrino and R. Andre-Abrecht, "An unsupervised approach to language identification," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 833–836, 1999.

[63] P. A. T. Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using Gaussian mixture model tokenization," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. I, pp. 757–760, 2002.

[64] P. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. Reynolds, and J. J. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-2002)*, 2002.

[65] C. Corredor-Ardoy, J. Gauvain, M. Adda-Decker, and L. Lamel, "Language identification with language-independent acoustic models," in *Proc. EUROSPEECH-1997*, pp. 55–58, 1997.

[66] P. Dalsgaard and O. Andersen, "Identification of mono- and polyphonemes using acoustic-phonetic features derived by a self-organising neural network,," in *Proc. Int. Conf. Spoken Language Processing (ICSLP- 1992)*, pp. 547–550, 1992.

[67] F. Pellegrino, J. Farinas, and R. Andr-Obrecht, "Comparison of two phonetic approaches to language identification," in *Proc. EUROSPEECH99,*, pp. 399–402, 1999.

[68] Y. Ueda and S. Nakagawa, "Diction for phoneme/syllable/word-category and identification of language using HMM," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-1990)*, pp. 1209–1212, 1990.

[69] J. Braun and H. Levkowitz, "Automatic language identification with perceptually guided training and recurrent neural networks," in *Proc. Int. Conf. Spoken Language Processing (ICSLP-1998)*, 1998.

[70] W. Campbell, E. Singera, P. Torres-Carrasquillo, and D. Reynolds, "Language recognition with support vector machines," in *Proc. ODYSSEY-2004*, 2004.

[71] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, and C. Vair, "Language identification using acoustic models and speaker compensated cepstral-time matrices," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2007)*, pp. IV–1013IV–1016, 2007.

[72] E. Noor and H. Aronowitz, "Efficient language identification using anchor models and support vector machines," in *Proc. IEEE Odyssey 2006 Speaker and Language Recognition Workshop*, pp. 1–6, 2006.

[73] C. Lin and H. Wang, "Language identification using pitch contour information in the ergodic Markov model," in *Proc. 2006 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2006)*, pp. I–I, 2006.

[74] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. Andr-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, pp. 436–456, 2005.

[75] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 266–276, Jan. 2006.

[76] J. L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 15, pp. 1904–1911, 2007.

[77] M.-H. Siu, X. Yang, and H. Gish, "Discriminatively Trained GMMs for Language Classification Using Boosting Methods," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17, pp. 187–197, Jan. 2009.

[78] A. Sangwan, M. Mehrabani, and J. H. L. Hansen, "Automatic Language Analysis and Identification based on Speech Production Knowledge," in *ICASSP*, 2010.

[79] D. Martnez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-based Prosodic System for Language Identification," in *ICASSP*, 2012.

[80] B. Jyotsna, H. A. Murthy, and T. Nagarajan, "Language identification from short segments of speech," in *Proceedings of Int. Conf. Spoken Language Processing*, (Beijing, China), pp. 1033–1036, Oct. 2000.

[81] T. Nagarajan and H. A. Murthy, "Language identification using spectral vector distribution across languages," in *Proc. International Conference on Natural Language Processing*, pp. 327–335, 2002.

[82] L. Mary and B. Yegnanarayana, "Autoassociative neural network models for language identification," in *Proc. Int. Conf. Intelligent Sensing and Information Processing*, (Chennai, India), pp. 317–320, 2004.

[83] L. Mary, K. S. Rao, and B. Yegnanarayana, "Neural Network Classifiers for Language Identification using Syntactic and Prosodic features," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing*, (Chennai, India), pp. 404–408, Jan. 2005.

[84] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Communication*, vol. 50, pp. 782–796, 2008.

[85] S. Greenberg, "Speaking in short hand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[86] T. Lander, R. Cole, B. Oshika, and M. Noel, "The OGI 22 language telephone speech corpus," in *Proc. EUROSPEECH-1995*, pp. 817–820, 1995.

[87] G. Z. Fang Zheng and Z. Song, "Comparison of Different Implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 16, pp. 582–589, 2001.

[88] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEEASLP*, vol. 16, pp. 1602–1613, 2008.

[89] K. S. Rao and B. Yegnanarayana, "Modeling durations of syllables using neural networks," *Computer Speech and Language*, vol. 21, pp. 282–295, 2007.

[90] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[91] A. N. Khan, S. V. Gangashetty, and B. Yegnanarayana, "Syllabic properties of three Indian languages: Implications for speech recognition and language identification," in *Proc. Int. Conf. Natural Language Processing*, (Mysore, India), pp. 125–134, Dec. 2003.

[92] M. Ember and C. R. Ember, "Cross-language predictors of consonant-vowel syllables," *American Anthropologist*, vol. 101, pp. 730–742, Dec. 1999.

[93] S. E. G. Ohman, "Coarticulation in VCV utterances:spectrographic measurements," *J. Acoust. Soc. Am.*, vol. 39, pp. 151–168, Jan. 1966.

[94] C. C. Sekhar, *Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) Segments in Continuous Speech.* PhD thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Chennai, India, 1996.

[95] S. R. M. Prasanna, S. V. Gangashetty, and B. Yegnanarayana, "Significance of vowel onset point for speech analysis," in *Proc. Int. Conf. Signal Processing and Communication*, vol. 1, (Bangalore, India), pp. 81–86, Jul. 2001.

[96] S. V. Gangashetty, *Neural Network Models for Recognition of Consonant-Vowel Units of Speech in Multiple Languages.* PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Feb. 2005.

[97] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel Onset Point Detection for Low Bit Rate Coded Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 1894–1903, 2012.

[98] A. S. Madhukumar, *Intonation knowledge for Speech Systems for an Indian Language*. PhD thesis, Department of Computer Science and Engg., Indian Institute of Technology Madras, Chennai-600 036, India, 1993.

[99] A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of text-to-speech system for Hindi," *Computer, Speech and Language*, vol. 7, pp. 283–301, 1993.

[100] F. Cummins, F. Gers, and J. Schmidhuber, "Comparing prosody across languages," Tech. Rep. I. D. S. I. A. Technical Report IDSIA-07-99, Istituto Molle di Studie sull'Intelligenza Artificiale, CH6900 Lugano, Switzerland, 1999.

[101] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers, 1997.

[102] Y. Xu, "Consistency of tone-syllable alignment across different syllable structures and speaking rates," *Phonetica*, vol. 55, pp. 179–203, 1998.

[103] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *J. Acoust. Soc. Am.*, vol. 107, pp. 1697–1714, Mar. 2000.

[104] C. Gussenhoven, B. H. Reepp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *J. Acoust. Soc. Am.*, vol. 102, pp. 3009–3022, Nov. 1997.

[105] P. F. MacNeilage, "The frame/content theory of evolution of speech production," *Behavial and Brain Sciences*, vol. 21, pp. 499–546, 1998.

[106] R. A. Krakow, "Physiological organization of syllables: a review," *Journal of Phonetics*, vol. 27, pp. 23–54, 1999.

[107] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[108] A. Cutler and D. R. Ladd, *Prosody: models and measurements*. Berlin Heidelberg New York Tokyo: Springer-Verlag, 1983.

[109] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of IEEE*, vol. 81, pp. 1215–1247, Sep 1993.

[110] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[111] J. Benesty, M. M. Sondhi, and Y. A. Huang, *Springer Handbook of Speech Processing*. Springer-Verlag New York, Inc., 2008.

[112] J. Volkmann, S. Stevens, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, pp. 185–190, Jan. 1937.

[113] Z. Fang, Z. Guoliang, and S. Zhanjiang, "Comparison of different implementations of MFCC," *J. Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[114] G. K. T. Ganchev and N. Fakotakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. of Int. Conf. on Speech and Computer*, (Patras, Greece), pp. 191–194, 2005.

[115] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 342–350, 1981.

[116] J. S. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Toronto, Canada), pp. 3673–3676, Apr. 1991.

[117] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, Aug. 1995.

[118] B. F., B. J. F., F. C., G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal process.*, vol. 2004, no. 4, pp. 430–451.

[119] Dempster, A.P., Laird, N.M., and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[120] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, pp. 84–95, Jan. 1980.

[121] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. M. L. Cam and J. Neyman, eds.), vol. 1, pp. 281–297, University of California Press, 1967.

[122] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[123] Q. Y. Hong and S. Kwong, "A discriminative training approach for text-independent speaker recognition," *Signal process.*, vol. 85, no. 7, pp. 1449–1463, 2005.

[124] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio processeing*, vol. 3, pp. 72–83, Jan. 1995.

[125] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio process.*, vol. 2, pp. 291–298, Apr. 1994.

[126] D. A. Reynolds, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, Jan. 2000.