MULTILEVEL IMPLICIT FEATURES FOR LANGUAGE AND SPEAKER RECOGNITION

A THESIS

submitted by

LEENA MARY

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JUNE 2006

To my parents

Mrs. and Mr. K. S. Kuriakose

and

parents-in-law

Mrs. and Mr. P. S. Koshy

THESIS CERTIFICATE

This is to certify that the thesis entitled **Multilevel Implicit Features for Language and Speaker Recognition** submitted by **Leena Mary** to the Indian Institute of Technology, Madras for the award of the degree of Doctor of Philosophy is a bonafide record of research work carried out by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chennai - 600 036 Prof. B. Yegnanarayana

Date: Department of Computer Science and Engineering

ACKNOWLEDGEMENTS

I would like to express my deep sense of gratitude to my guide Prof. B. Yegnanarayana for his constant guidance and support throughout this research work. I am indebted to him for selecting me as one of his research students. His discipline, hard work and attitude towards research, are inspiring. The excellent research environment he has created in Speech and Vision Laboratory is appreciable, and I consider myself fortunate to be associated with this laboratory.

I thank Prof. C. Pandurangan, Prof. S. Raman, and Prof. T. A. Gonsalves, Chairpersons of the department of Computer Science and Engineering during the period of this research, for providing constant support and excellent facilities to carry out my research. I thank Dr. C. Chandrasekhar for his timely advise and suggestions at various stages of my research. I thank my doctoral committee members, Dr. Deepak Khemani, Dr. C. Chandrasekhar, Dr. S. Mohan, and Dr. K. S. Swaroop, for sparing their valuable time to evaluate the progress of my research work. I would like to mention the concern and encouragement shown to me by Dr. Hema A. Murty and Dr. Darsanambika. I am thankful to all the teaching and non-teaching staff of the department for their help. I am also grateful to the QIP office, IIT Madras for their co-operation.

I would like to thank the Director of Technical Education, Government of Kerala for providing deputation for the period 2002 - 2005 to carry out this research. I also thank the principal and staff members of Government Engineering College, Palakkad for the support given to me during the period of research. I take this opportunity to thank all my teachers right from my school days who have helped me to come up.

I am grateful to Dr. Suryakanth V. Gangashetty, Dr. S. R. Mahadeva Prasanna and Dr. Sreenivasa Rao, my seniors in speech and vision lab for giving initiative for this

work. Now let me express my sincere gratitude to all members of Speech and Vision Lab, especially to Dr. Palanivel, Dr. Rajendran, Dr. Nayeem, Jinu, Gupta, Anitha, Shahina, Guru, Dhanu, Anil, Satish, Sriram, Chaitanya, Suresh, Subitha, Kartik, Krishnamohan, Kumaraswamy, Gomathi, Sangeetha, Balavamsi, Anand, Chandrakala, Swapna, Lakshmi Narayan, Sheetal, Jayasree, Dileep, Veena and Rajeev for their co-operation and support. My special thanks to Guru and Surya for giving corrections to my technical writings, Sriram, Dhanu, Satish, Chaitanya and Anand for their help on various occasions.

I thank Anitha and Sujatha for making my stay in IIT Campus pleasant and memorable. I thank Prema for the concern and care shown to me and my family. I am grateful to my brother-in-law Dr. Reebu Z. Koshy, and sister-in-law Dr. Anitha Joseph, through whom I came to know about the activities of Speech and Vision group in IIT Madras. I was lucky to have my children, Linda and Alan, with me in IIT campus. The moments I spend with them make my life worthwhile. I do not have words to express my gratitude to my husband Bino who has been a constant source of encouragement and motivation to do better things in life. The blessings and steadfast support of my parents and parents-in-law always strengthened me and I dedicate this thesis to them for their affection.

I praise God almighty for all the blessings I received from Him.

Leena Mary

ABSTRACT

Keywords: *Language identification; speaker verification; vowel onset point; consonant-vowel units; autoassociative neural network models; multilayer feedforward neural networks; prosody; intonation; stress; rhythm; tilt.*

Human beings recognize language and speaker using features in the speech at various levels. It is difficult to identify and describe these features. Hence these features can be viewed as *implicit* features. Speech signal carries characteristics of both the language and the speaker, and it is difficult to distinguish features specific to language and speaker separately. Human beings seem to combine evidence obtained from features at *multiple* levels to arrive at a decision. In general, the levels, and the combination of evidence at these levels, are difficult to articulate. This thesis is an attempt to identify, extract and represent some implicit features at multiple levels of speech signal for language and speaker recognition.

In this work, we explore language-specific features at three different levels of speech signal, namely, frame, syllabic and multisyllabic levels, for automatic language identification (LID). At the frame level, the features explored for LID are weighted linear prediction cepstral coefficients (WLPCC), linear prediction (LP) residual and phase of the LP residual. At the syllable level, language-specific variations in the realization of syllables are represented using spectral features. At the multisyllabic level, prosodic characteristics such as rhythm, intonation and stress are used. Prosodic characteristics are represented using features derived from duration, F_0 contour and energy. The effectiveness of these multilevel features for language identification are discussed (a) for an Indian language database and (b) for the Oregon Graduate Institute (OGI) multi-language telephone-based speech (MLTS) corpus.

For speaker verification, speaker differences in excitation source, vocal tract dimensions and prosodic characteristics are examined. These characteristics are represented using residual, spectral and prosodic features derived from subsegmental, segmental and suprasegmental levels, respectively. The vulnerability of spectral features due to the effects of channel characteristics and noise, motivated us to explore the use of prosodic features for speaker verification. We propose a method for modeling speaker-specific prosody, which is based on capturing the distribution of prosodic features. The effectiveness of the proposed prosody-based speaker verification system is evaluated using National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) 2003 extended data.

The major contributions of this work are:

- An approach to language identification based on features derived from speech signal at different levels
- A method based on vowel onset point (VOP) for extraction and representation of the prosodic features directly from the speech signal
- A prosody-based approach for language discrimination and speaker verification
- A speaker verification system combining evidence from spectral features and prosodic features

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AANN	- AutoAssociative Neural Network
ANN	- Artificial Neural Network
ASR	- Automatic Speech Recognizer
CCV	- Consonant-Consonant-Vowel
CV	- Consonant-Vowel
CVC	- Consonant-Vowel-Consonant
DET	- Detection Error Tradeoff
DFT	- Discrete Fourier Transform
DTW	- Dynamic Time Warping
EER	- Equal Error Rate
En	- English
FFT	- Fast Fourier Transform
FA	- False Acceptance
Fa	- Farsi
FR	- False Rejection
Fr	- French
GC	- Glottal Closure
Ge	- German
GMM	- Gaussian Mixture Model
HE	- Hilbert Envelope
Hi	- Hindi
HMM	- Hidden Markov Model

IDFT	- Inverse Discrete Fourier Transform
IITM	- Indian Institute of Technology Madras
ITRANS	- Indian language TRANSliteration
Ja	- Japanese
JHU	- John Hopkins University
k-NN	- k-Nearest Neighbor
Ko	- Korean
LVCR	- Large Vocabulary Speech Recognizer
LID	- Language IDentification
LF	- Liljencrants-Fant
LM	- Language Model
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
LPCC	- Linear Prediction Cepstral Coefficient
LSA	- Latent Semantic Analysis
Ma	- Mandarin
MFCC	- Mel-Frequency Cepstral Coefficients
MLFFNN	- MultiLayer FeedForward Neural Network
NIST	- National Institute of Standards and Technology
NN	- Neural Network
OGI	- Oregon Graduate Institute
OGI MLTS	- Oregon Graduate Institute Multi-Language Telephone Speech
PCA	- Principal Component Analysis
PDF	- Probability Density Function
PPR	- Parallel Phone Recognition
PRLM	- Phone Recognition followed by Language Modeling
PPRLM	- Parallel Phone Recognition followed by Language Modeling

SNR	- Signal-to-Noise ratio
Sp	- Spanish
SRI	- Stanford Research Institute
SVM	- Support Vector Machine
Ta	- Tamil
TIMIT	- Texas Instruments and Massachusetts Institute of Technology
Vi	- Vietnamese
VOP	- Vowel Onset Point
VQ	- Vector Quantization
WLPCC	- Weighted Linear Prediction Cepstral Coefficients

NOTATION

$P(C_i\|O)$	–	a posteriori probability of the class C_i for given O
$P(O\|C_i)$	–	likelihood probability of O to the class C_i
$P(C_i)$	–	a priori probability of the class C_i
a_k	–	k^{th} linear prediction coefficient
c_n	–	n^{th}th linear prediction cepstral coefficient
p	–	linear prediction order
n	–	discrete time index
$s(n)$	–	speech signal as a function of time index n
$\hat{s}(n)$	–	predicted speech signal as a function of time index n
$x(n)$	–	signal variable
ω	–	frequency variable in radians for a discrete time signal
$X(\omega)$	–	Fourier transform of $x(n)$
$r(n)$	–	residual signal as a function of time index
$r_h(n)$	–	Hilbert transform of $r(n)$
$R(\omega)$	–	discrete Fourier transform of $r(n)$
$R_h(\omega)$	–	discrete Fourier transform of $r_h(n)$
$h_e(n)$	–	Hilbert envelope of $r(n)$
$cos\theta$	–	phase of residual signal
L	–	linear unit
N	–	nonlinear unit
$g(n)$	–	Gabor filter of length n
F_0	–	fundamental frequency
ΔF_0	–	change in F_0
ΔE	–	change in log energy
A_r	–	amplitude of rise
A_f	–	amplitude of fall
D_r	–	duration of rise

D_f	–	duration of fall
A_t	–	amplitude tilt
D_t	–	duration tilt
D_s	–	duration of syllable
D_v	–	duration of voiced region
D_p	–	distance of F_0 peak with reference to VOP
μ	–	mean
σ	–	standard deviation
F_{0_p}	–	F_0 peak
F_{0_v}	–	F_0 valley
F_{0_μ}	–	F_0 mean

CHAPTER 1

MULTILEVEL IMPLICIT FEATURES FOR LANGUAGE AND SPEAKER RECOGNITION - AN INTRODUCTION

1.1 OBJECTIVES OF THE THESIS

Speech is primarily intended to convey some message. It is conveyed through a sequence of legal sound units in certain language, and this sequence has to obey the constraints imposed by the language. Hence speech and language can not be delinked. Since each speaker has unique physiological characteristics of speech production and speaking style, speaker-specific characteristics are also embedded in the speech signal. Thus speech signal contains not only the intended message, but also the characteristics of the language and speaker.

The message part of speech is mostly conveyed as a sequence of legal sound units, where each unit corresponds to a particular manner and place of speech production. Extracting the message part of information in speech constitutes *speech recognition*. Typically in speech recognition, the speech signal is represented as a sequence of acoustic-phonetic features derived from analysis frames. These sequences of features are used to determine the boundaries and the class of each of the sound units present in the sequence. Here the feature extraction and representation is based on the characteristics of speech signal rather than on the characteristics of production. These features may be called *explicit* features, in the sense that they can be derived from the signal and can be interpreted in terms of production process. There will be some ambiguity in the recognition of the class of a sound unit from these features. This ambiguity is resolved most of the time by the production and linguistic constraints on

the sequence of units.

The language and speaker part of the information contained in the speech signal is inferred using features at several levels. It is difficult even for a listener to describe the language-specific and speaker-specific features that he/she will be using for recognition. Thus these features are somewhat ambiguous and not unique. They also depend on the listener. But the class label (language or speaker) is unique. These features, spread over different levels typically referring to the different time span of the speech, are *implicit* rather than explicit as for the case of speech recognition. It is difficult to distinguish the language-specific and speaker-specific parts in these implicit features. Usually in human listening, the evidence from *multiple* levels are combined to arrive at a decision. But in general, the levels, and the combination of evidence at these levels, are difficult to articulate. It is a challenging task to identify and extract the implicit features for a language or speaker from the speech signal. As the features derived from the acoustic speech signal are both language and speaker-specific, their representation is crucial to highlight the language or speaker-specific part depending on the task. This thesis is an attempt to identify, extract and represent some implicit features at multiple levels for language and speaker recognition.

The existing language and speaker recognition systems rely on features derived through short-time spectral analysis. But spectral features are affected by channel characteristics and noise. This motivated us to explore the use of additional features, for the presence of language or speaker-specific information. These features may provide complementary evidence to the spectral-based systems. To make use of the language-specific or speaker-specific information present at larger span of speech, most of the existing language/speaker recognition systems use segment boundaries and text labels obtained using speech recognizers. For building a speech recognizer, several man hours are needed for preparing a manually labeled corpora. We aim for language and speaker recognition using features derived directly from the acoustic speech signal, without the use of transcription of the signal.

The general area of language or speaker recognition can be classified as identification and verification. Language identification or speaker identification is the task of determining the language or speaker from a set of known languages or speakers. Verification is the task of determining whether the claim regarding the language or speaker is valid or not (a yes/no decision). In this thesis, we address the issues related to *language identification* and *speaker verification*, focusing on features extracted from multiple levels of speech signal.

This chapter is organized as follows: Next section discusses the significance of automatic language and speaker recognition systems. Section 1.3 describes the cues used by the human beings for recognizing a language or a speaker. In Section 1.4, language-specific and speaker-specific aspects of speech is discussed. Representation of language-specific and speaker-specific characteristics, embedded at multiple levels of speech signal, used in this study are discussed in Section 1.5. In Section 1.6, issues in automatic language identification and speaker verification are discussed. Section 1.7 describes the issues addressed in this thesis. Section 1.8 gives an overview of the thesis.

1.2 SIGNIFICANCE OF LANGUAGE AND SPEAKER RECOGNITION

The research area of speech processing emerged as a result of one's desire to implement an intelligent machine that can recognize the speech and comprehend the meaning in it. In spite of the research effort in this direction, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by any speaker in a language of his or her preference.

Automatic language identification (LID) is the task of identifying the language of a given utterance of speech using a machine [1,2]. Applications of LID fall in two main categories [1]: Pre-processing for machines and pre-processing for human listeners. A multilingual voice-controlled information retrieval system is an example of the first

category. Language identification system used to route an incoming telephone call to a human operator at a switchboard, fluent in the corresponding language, is an example of the second category [1, 2]. Multilingual interoperability is an important issue for many applications of modern speech technology [3]. In a multi-lingual country like India, the development of multi-lingual speech recognizer and spoken dialog systems is very important. Applications such as spoken dialog systems, database search and retrieve systems, automatic call routing, and language translation need to address the possible presence of multiple languages in the input [4]. For such multilingual applications, the machine should be capable of distinguishing among languages. Identification of language of the input speech is a tool to adapt a recognizer suitable to a specific type of speech. An approach for a multilingual system is to integrate several monolingual recognizers with a front-end for identifying the language [5].

Person authentication or verification systems are useful in applications where access to a facility need to be controlled. Biometrics, which bases the person authentication on the intrinsic aspect of a human being, appears as a viable alternative to more traditional approaches (such as alphanumeric codes or passwords). It could be done with various modalities, such as face, voice, iris, gait and fingerprints, among others. A person's identity is embedded in his or her voice, and can be recognized using automatic speaker recognition systems. Voice-based access control systems are attractive, since speech is inexpensive to collect and analyze, and hard to mimic. Automatic speaker verification systems are useful for applications such as transaction authentication, access control to systems, monitoring of telephone usage and voice matching for forensics. In the next section, the human way of recognizing language and speaker are discussed.

1.3 LANGUAGE AND SPEAKER RECOGNITION - THE HUMAN WAY

An insight into the ability of human beings to identify language or speaker from speech may offer clues for developing an automatic recognition system. Having certain degree

of familiarity with a given language or a speaker, human beings can extract specific cues for identifying the language or speaker. Human beings are endowed with the ability to integrate knowledge from various sources for recognizing a language or a speaker.

1.3.1 Language Recognition by Humans

Human beings learn a language over a period of time, and the level of his language knowledge may vary depending on whether it is his native/first language, whether he has sufficient exposure and formal education in it. He uses knowledge of vocabulary, syntax, grammar and sentence structure to identify a language, in which he is proficient. It has been observed that humans often can identify the language of an utterance even when they have no working linguistic knowledge of that language, suggesting that they are able to learn and recognize language-specific patterns directly from the signal [6]. In the absence of higher level knowledge of a language, listener presumably relies on lower level constraints such as phonetic repertoire, prosody and phonotactics. Perceptual studies have revealed the importance of prosodic characteristics such as rhythm, intonation and stress, for language recognition by humans [7,8]. Besides these cues from speech, human beings also use contextual knowledge about the speaker, to identify the language spoken.

1.3.2 Speaker Recognition by Humans

Human beings use several levels of perceptual cues for speaker recognition, ranging from high-level cues such as semantics, pronunciations, idiosyncrasies and prosody to low-level cues such as acoustic aspects of speech [9]. The high-level features such as prosody and idiolect are the behavioral attributes of the speaker, different from physiological characteristics of the speech production system. Human beings derive evidence regarding the identity of a speaker from certain prosodic cues such as pitch gestures, accents, and speech rate. It is generally recognized that human listeners can better

recognize those speakers who are familiar to them, than those who are relatively less familiar. This increased ability is due to speaker-specific prosody and idiosyncrasies that are recognized by the listener, either consciously or otherwise [10]. "Familiar-speaker" differences, however, surely relate to long term speech characteristics, such as the usage of certain words and phrases, and to the features such as intonation, stress and timing.

Speaker recognition is one area of artificial intelligence where machine can exceed human performance [9]. Using short test utterance and a large number of speakers, machine accuracy often exceeds that of humans. This is especially true for unfamiliar speakers, where the training time for humans to learn a new voice well is very long compared with that of machines [11]. The language-specific and speaker-specific cues present in the speech are examined in the next section.

1.4 LANGUAGE-SPECIFIC AND SPEAKER-SPECIFIC ASPECT OF SPEECH

Speech signal contains characteristics of sound units, speaker, language and channel. For recognizing the language or speaker, the differences among languages or speakers need to be identified and these differences should be brought out using appropriate features.

1.4.1 Language-specific Aspects of Speech

The following aspects of speech may differ among languages:

(a) Acoustic-phonetics: Each sound corresponds to a unique articulatory configuration of the vocal tract. Even though there is significant overlap in the set of sound units in languages, the same sound unit may differ across different languages due to coarticulation effects and dialects. This variations in the acoustic realization of phoneme, forms the basis for the acoustic-phonetic studies.

(b) Phonotactics: 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 from one language to another. Certain phoneme or syllable clusters common in one language may be rare or illegal in some other language.

(c) Prosody: Prosodic features such as rhythm, stress, and intonation vary among languages. The manifestation of prosodic constraints in speech, conveys some important information regarding the language.

(d) Vocabulary and lexical structure: The word roots and lexicon are usually different between languages. Each language has its own vocabulary, and its own manner of forming words. Even when two languages share a word, the set of words that may precede or follow the word may be different. At higher levels, the sentence pattern and grammar are different between languages.

1.4.2 Speaker-specific Aspects of Speech

Speaker characteristics vary due to difference in:

(a) Physiological characteristics of speech production organs

(b) Acquired or learned habits

Physiological difference include the differences in the shape and size of oral tract, nasal tract, vocal folds and trachea. This can lead to difference in the vocal tract dynamics and excitation characteristics. The acquired habits are characteristics that are learned over a period of time, mostly influenced by the social environment and also by the characteristics of the first or native language in the 'critical period' (lasting roughly from infancy until puberty) of learning. The way prosodic characteristics are manifested in speech give important information regarding the identity of a speaker. Idiosyncrasies of a speaker are reflected in the usage of certain words and phrases and it is present even at the semantic level.

Differences in speaker characteristics may be summarized as follows:

(a) Vocal tract size and shape

(b) Excitation characteristics

(c) Prosody

(d) Idiolect

(e) Semantic

Fig. 1.1 illustrates various language and speaker-specific cues and their levels of manifestation in speech. Language and speaker-specific cues are present at low level as well as high level of speech. Low level cues are directly derivable from the speech signal whereas high level cues are present in the textual content. Therefore transcription of speech is required for representing high level cues.

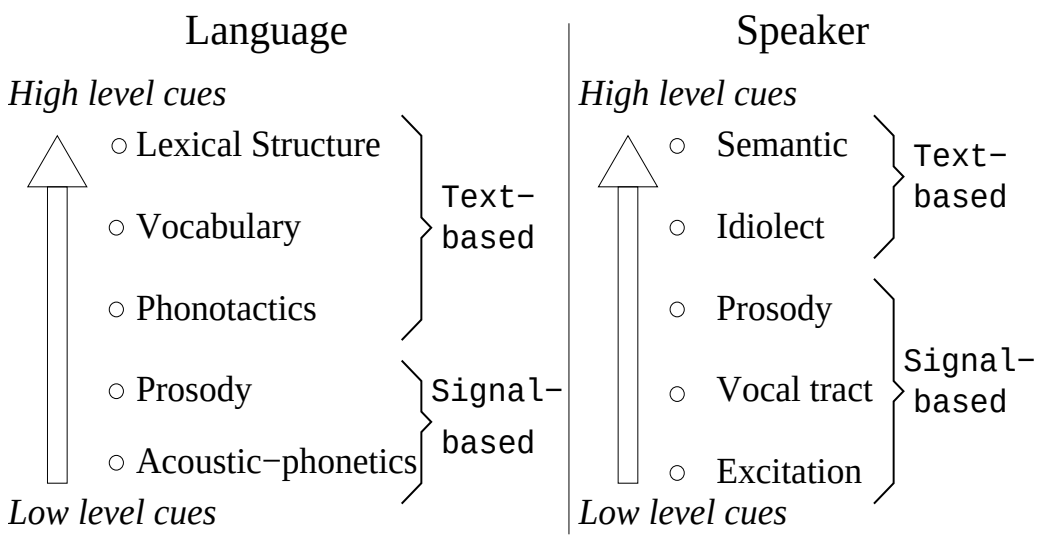


Fig. 1.1: Various language and speaker-specific cues and their levels of manifestation.

1.4.3 Signal-based Versus Text-based Recognition

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. Researchers have approached spoken language identification in two different ways. In one approach, phoneme recognizers are used as the front-end, and the transcription generated by the phoneme recognizers are used for modeling phonotactic constraints of a language using statistical language models. This approach requires large amount of labeled speech data in different languages for training the phone recognizers, which may not be available for many languages. Therefore, developing an LID system using this approach can be as complex as multilingual speech recognition systems. On the other hand, there is another approach which rely on features derived from speech signals, and do not need labeled speech data at any stage of identification. Similarly most of the existing speaker verification systems make use of segment boundaries and transcriptions for the extraction and representation of prosody and idiolect. Systems which use text transcriptions directly or indirectly for automatic recognition systems are referred as explicit systems, in this thesis. Systems which rely on features derived from speech signal, and do not use any text-based information, are referred as implicit systems. The present study focuses on the use of signal-based features for developing an implicit language and speaker recognition system.

In order to represent language and speaker-specific aspects of speech, features need to be extracted and represented from different time spans of speech signal. Next we discuss the levels of representation of implicit features used in this study.

1.5 MULTILEVEL IMPLICIT FEATURES FOR LANGUAGE AND SPEAKER RECOGNITION

Characteristics of sound units, language, speaker and that of channel are embedded in the speech signal. In language or speaker recognition, the task is to extract language-specific or speaker-specific features from speech. In order to represent the acoustic variations among languages or speakers, features should be derived from short windows of speech typically a few ($\leq$ 20 msec) milliseconds duration. But to represent

pronunciation variation of subword units like syllables, features should be represented at the syllabic level. Variations in prosody and phonotactics should be represented at a level higher than a syllable.

1.5.1 Language Identification

In this work, we focus on lower level cues such as acoustic-phonetics and prosody directly derived from speech signal for identifying a language. We do not try to model subword units separately and therefore do not use any labeled corpora for the development of LID system. This precludes us from looking into the higher level features such as vocabulary, grammar, and also phonotactics up to some extent. In this study, we represent the variations due to acoustic-phonetics and prosody using language-specific features derived at different levels of speech signal. Even though there is less clarity regarding the levels of representation of features, it is approximated as the following:

(a) Frame Level: A short window of speech, $\leq$ 20 msec duration, is referred to as frame. In the context of language identification, the frame level features mainly represent production level constraints imposed by the language. Since production level constraints may be sound-specific and speaker-specific, the language-specific part should be emphasized by reducing the effect of others.

(b) Syllabic Level: Features derived from subword units such as syllables are referred to as syllable level features. Syllables are often represented by a collection of frame level features derived from corresponding speech regions. These features can be used to develop models of language. Syllabic level features aim to model the acoustic variations in the realization of syllables among languages.

(c) Multisyllabic Level: Features corresponding to a larger span of speech which go beyond syllables are referred to as multisyllabic features. To study the role of phonotactic constraints and prosody of a language, features should be represented at a multisyllabic level.

1.5.2 Speaker Verification

In order to represent differences among speakers in terms of characteristics of excitation source, vocal tract system and prosody, features should be extracted from levels of speech at which these characteristics are manifested. In our study, the approximated levels of representation of speaker-specific features are the following:

(a) Subsegmental: The main source of excitation for production of speech is the glottal vibration. In each glottal cycle, the instant of glottal closure is the instant at which significant excitation of vocal tract takes place. Hence a small region around the instant of glottal closure contains significant information about the speaker. In order to represent the excitation source characteristics of a speaker, excitation sequence corresponding to a duration of 1-5 msec, which is less than one pitch period, can be considered.

(b) Segmental: Vocal tract system characteristics are extracted from a window of speech signal that contains a few pitch periods (10-30 msec). Time varying nature of the vocal tract system is taken into account by sliding the window by about 5-10 msec.

(c) Suprasegmental: Features corresponding to a larger span of speech (> 100 msec) which go beyond segments are referred as suprasegmental features. Prosodic characteristics such as intonation, duration and stress are visible only for a large span of speech. Therefore to model the prosodic characteristics of a speaker, features are represented at the suprasegmental level.

1.6 ISSUES IN LANGUAGE AND SPEAKER RECOGNITION

1.6.1 Issues in Automatic Language Identification

The LID system should identify the language of a speech utterance independent of the message spoken, speaker characteristics and channel characteristics. The challenge

in language identification is the identification and extraction of features which will bring out the difference among languages. The following are some issues in automatic language identification:

(a) Variability in Speaker Characteristics: Within the constraints of a language, speakers can have their own speaking styles giving rise to a large amount of speaker variability. Therefore it is necessary to capture the speaker variability while modeling a language.

(b) Variability in Accents: Accent is primarily about pronunciation. A person's accent is a good indicator whether he is a native speaker of a given language or not. For example, a native of English fairly fluent in Spanish may speak Spanish with an English accent. But difference in accents are difficult to describe.

(c) Variability in Environment and Channel Characteristics: Characteristics of speech signal are influenced by the environment in which it is collected and channel through which it is transmitted. These factors can significantly change the features derived from short time spectrum analysis. It is important to have features which are less affected by channel and environment characteristics to achieve the robustness needed for a language identification system.

(d) Extraction and Representation of Language-specific Prosody: Prosodic features such as rhythm, stress, and intonation vary among languages. But the nature of these characteristics is not well-defined. For example, the rhythm of a language may be felt due to succession of syllables, vowels, amplitude bursts, or rise or fall in pitch, which is not well-understood. Moreover, the techniques available for speech processing are not adequate to represent higher sources of knowledge such as prosody. Therefore, extraction and representation of language-specific prosody is difficult.

1.6.2 Issues in Automatic Speaker Verification

(a) Variability in Environment and Channel Characteristics: The anatomy and physiology of speech production organs that is manifested in speech can be well represented using features derived through spectral analysis. But the spectral features are also influenced by the channel characteristics which may adversely affect the speaker recognition performance. For example, a speaker model trained using speech collected over a microphone may not give correct result for a genuine test utterance collected over land line or cellular environment.

(b) Variability in Language: In a multi-lingual environment, it is natural for a human being to use more than one language. Since the spectral characteristics are also language-specific, speaker models should be trained to capture this variability.

(c) Extraction and Representation of speaker-specific Prosody: The exact nature of the speaker-dependent prosodic cues are not fully understood. Also there is less clarity regarding the level at which it should represented.

1.7 ISSUES ADDRESSED IN THIS THESIS

In this thesis, we explore language-specific features at three levels of speech, namely, frame, syllabic and multisyllabic levels, for language identification. The issue of speaker and channel variability in spectral-based LID is addressed using separate models of speakers within the same language. For using syllable level variations among languages, fixed regions around the vowel onset points (VOP) are used for representing the consonant-vowel (CV) type of syllables.

The existing techniques of feature extraction are not sufficient to represent information such as prosody, which should be represented for a larger span of speech. We propose a method for extraction and representation of prosodic features with syllable

as the basic unit. Continuous speech is segmented into syllable-like regions using the knowledge of VOPs. Prosodic features are derived for each syllable-like regions.

Another issue addressed is the representation of prosodic features for language identification. Language-specific prosodic characteristics such as intonation, rhythm and stress are linked to the syllable sequence rather than individual syllables. Therefore prosodic features are represented at a multisyllabic level. The systems proposed for language identification do not require labeled corpora for training and testing.

Finally, we address the issues in the development of prosody-based speaker verification system. Prosodic features are less affected by channel characteristics and noise. Evidence from prosodic features and spectral features are combined to improve the accuracy of speaker verification system.

1.8 ORGANIZATION OF THE THESIS

The evolution of ideas reported in this thesis is given in Table 1.1.

The thesis is organized as follows: In Chapter 2, a review of automatic language and speaker recognition is given. Chapter 3 gives a probabilistic formulation to the problem of language and speaker recognition. In Chapter 4, features at the frame level are examined for possible presence of language-specific characteristics. The details of language database, and the results of experimental studies based on frame level features are described in this chapter. Chapter 5 describes the use of features derived at the level of syllables for language identification. The goal is to make use of the acoustic variation in the realization of syllables, which is language-specific. In Chapter 6, the role of phonotactics and prosody are studied using manual segmentation and labeling information obtained from an Indian language database. Later, a method is proposed for the extraction and representation of prosodic features from speech signal, and its effectiveness is demonstrated in case of Oregon Graduate Institute (OGI) multi-language telephone speech corpus. Chapter 7 describes multilevel

features examined for speaker verification, and explains the need for incorporation of the prosodic features. The prosodic characteristics of speakers are captured using distribution capturing techniques. The effectiveness of prosodic features for speaker verification is illustrated using National Institute of Standards and Technology (NIST) 2003 extended data task. We also demonstrate that, by combining evidence from spectral and prosodic features, the performance of speaker verification system can be improved. Finally, we summarize the contributions of this research work, and discuss some directions for further work in Chapter 8.

Table 1.1: Evolution of ideas presented in the thesis.

1. **Features for language and speaker recognition**
   - Implicit and are at multiple levels
2. **Language-specific and speaker-specific cues**
   - Language-specific: acoustic-phonetics, prosody, phonotactics and vocabulary
   - Speaker-specific : excitation, vocal tract, prosody and idiolect
   - Issue: Extraction and representation at appropriate levels
3. **Multilevel implicit features for LID**
   - Frame level and syllabic level for capturing acoustic-phonetic variations
   - Multisyllabic level for capturing prosody and phonotactics
   - Issue: Speaker variability within a language
4. **Prosodic features in the speech signal**
   - Hypothesis: Prosody is linked to the underlying syllable sequence
   - Segmentation of speech into syllable-like regions using VOPs
   - Issue: Representation of F_0 contour, energy and syllable durations
5. **Prosodic features for LID**
   - Rhythm, intonation and stress
   - Issue: Representation of language-specific prosody
6. **Multilevel implicit features for speaker verification**
   - Subsegmental level for excitation source characteristics
   - Segmental level for vocal tract characteristics
   - Suprasegmental level for prosodic characteristics
   - Issue: Identification and extraction of implicit features
7. **Prosodic features for speaker verification**
   - Physiological : Distribution of F_0 values
   - Acquired habits: Dynamics of F_0 contour
   - Issue: Representation of speaker-specific prosody
8. **Combining evidence from multiple levels for improving robustness in speaker verification**

CHAPTER 2

REVIEW OF AUTOMATIC LANGUAGE AND SPEAKER RECOGNITION

2.1 INTRODUCTION

This chapter provides a review of existing approaches to language and speaker recognition, categorized according to the characteristics and features used for these tasks. Approaches to automatic language identification (LID) systems range from simple systems based on acoustic-phonetics to complex systems based on large vocabulary continuous speech recognition. It has been shown that, by using higher level information such as phonotactics and vocabulary, better performance can be achieved. But better performance in such systems is achieved with the transcriptions obtained using an automatic speech recognizer (ASR). A large manually labeled corpora is required for building such an ASR. Implicit LID systems which use features extracted directly from the speech signal are attractive in many practical applications as they do not require any labeled corpora at any stage.

Current speaker recognition systems are dominated by the vocal tract characteristics represented using spectral features such as Mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) derived through short-time spectral analysis. Systems based on spectral features perform well in acoustically matched and noise-free conditions. However, they fail to model information about the speaker at many other level that might contribute to speaker recognition. It has been shown that spectral features are affected by channel and noise. Therefore researchers try additional features to capture the speaker-specific characteristics of

excitation source, prosody and idiolect.

This chapter is organized as follows: Review of automatic language identification is discussed in Section 2.2. Approaches for LID, categorized as systems based on spectral similarity, prosody, phoneme, vocabulary and continuous speech recognition are described. In Section 2.3, we discuss approaches to speaker recognition, categorized as systems based on vocal tract, excitation source, prosody, phone pronunciation and idiolect characteristics for modeling. Finally in Section 2.4, we discuss the challenges in language and speaker recognition that motivated the present work.

2.2 APPROACHES TO AUTOMATIC LANGUAGE IDENTIFICATION

An LID system must exploit the primary differences which exist among languages, while still being robust for variations 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.

An understanding of the characteristics of spoken language is essential to the development of an LID system. 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 is 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.

The frequency of occurrence of phonemes differ significantly among languages. The rules govern the way different phonemes are combined to form larger units are referred as *phonotactics*. The sequence of allowable phonemes are different from one language to another. Certain syllables common in one particular language may be rare in some other language. These variations in phoneme statistics and phonotactics are normally

represented using the sequence of phoneme labels obtained using phoneme recognizers.

Prosody refers to a collection of characteristics which makes human speech sound natural. Prosodic characteristics are acquired over a period of time. In spoken communication, we use and interpret prosody without conscious effort. But it is difficult to describe them. Prosodic characteristics such as rhythm, intonation and stress vary among languages. The variations in prosodic characteristics are represented using features derived from duration, fundamental frequency (F_0), F_0 contour and amplitude contour.

Each language has its own *vocabulary*, and its own manner of forming words. At higher levels, the *sentence structure* and grammar vary among languages. Even when two languages share a word, the set of words that may precede or follow the word will be different. Table 2.1 lists some language discriminating cues and their representation normally used by the researchers.

Table 2.1: Language discriminating cues and their representations for LID.

Cues for LID	**Representation**
Acoustic-phonetics	Spectral features (MFCC, LPCC)
Prosody	Features from duration, F_0 and amplitude
Phonotactics	Sequences of subword labels
Vocabulary and lexical structure	Sequences of word transcriptions

It has been observed that human beings often can identify the language of an utterance even when they have no strong linguistic knowledge of that language [6]. This suggest that they are able to learn and recognize language-specific patterns directly from the signal [6]. In the absence of vocabulary and sentence level knowledge of a language, a listener presumably relies on characteristics such as acoustic-phonetics, phonotactics and prosody [12]. Automatic LID can make use of any one or a combination of the above cues [1]. The LID approaches reviewed in this chapter are categorized as systems based on the following:

(a) Spectral similarity

(b) Prosody

(c) Phonemes or syllables

(d) Words

(e) Continuous speech

We review some research efforts in language identification according to this categorization.

2.2.1 Systems Based on Spectral Similarity

Languages differ from each other with respect to their typical short-time acoustic features. This is caused by the difference in phoneme inventory, and also due to differences in the realization of similar phones. Modeling of languages using short-time spectral vectors was implemented by several researchers using a variety of techniques and approaches as summarized in Table 2.2 [13–16].

Table 2.2: Summary of major LID efforts based on spectral similarity.

Feature	**Technique**	**Reference**
Linear prediction coefficients (LPC)	Vector quantization (VQ)	Sugiyama 1991 [13]
Mel-frequency cepstral coefficients (MFCC)	Gaussian mixture models (GMM)	Nakagawa 1992 [14] Zissman 1993 [15]
LP features for syllable nuclei	Speaker similarity using nearest neighbor algorithm	Li 1994 [17]
Shifted delta cepstral features	Support vector machines (SVM)	Torress -Carassquillo 2003 [16]

2.2.2 Systems Based on Prosody

The prosody offers an enhancement to spectral, phoneme or word-based LID system by being an additional source of information, robust to noise [18]. Language-specific prosodic cues include stress, rhythm and intonation. Each cue is a complex language dependent perceptual entity expressed primarily as a combination of three measurable parameters: fundamental frequency, amplitude and duration.

LID researchers in the early era found that inclusion of prosodic features such as speech rate, F_0 and syllable timing offered little to improve the performance of their systems [2,12]. The most direct study on the utility of prosodic features is attempted to derive parameters to capture F_0 and amplitude contours on a syllable-by-syllable basis [18]. It computes inter-syllable (timing related) relationships in the F_0 and amplitude information, collects histogram of various features or feature-pairs. Then log likelihood ratio functions of histograms are computed to evaluate the unknown utterances in a pairwise discrimination task. The results showed that prosodic parameters are useful in discriminating one language from another.

Comparison of 10 languages in OGI database was done using prosody features namely ΔF_0 (first differenced pitch estimate) and ΔEnv (first differenced amplitude envelope of band-limited speech) [19]. In another effort, the rhythmic characteristics of languages are represented using syllable structure and durations of consonants and vowels [20]. It also used stress-related features in terms of pitch and energy [21]. In another approach [22], the stylized F_0 trajectories are quantized and labeled into a small set of classes that describes the dynamics of pitch and energy. The n-gram models based on these labels are formed to capture the prosodic characteristics of a language. F_0 contour represented using coefficients of Legendre polynomial is also shown to be useful for language discrimination [23].

Table 2.3: Summary of various efforts toward modeling prosody for LID.

Feature	**Modeling technique**	**Reference**
F_0 and duration	Neural network	Muthusamy 1993 [2]
F_0 values	HMM	Hazen 1993 [12]
F_0 and amplitude on a syllable-syllable basis	Histogram	Thyme-Gobbel and Hutchins 1996 [18]
Differenced F_0 and amplitude envelope	Recurrent neural network	Cummins 1999 [19]
Rhythmic and stress	GMM	Rouas 2003 [20, 21]
F_0 and energy contour	n-gram model	Adami 2003 [22]
F_0 contour	GMM	Lin 2005 [23]

2.2.3 Systems Based on Phones or Syllables

Given that different languages have different phone inventories, researchers have built LID systems based on phoneme/syllable recognizers that hypothesize the phoneme or syllable as a function of time. The likelihood scores emanating from language dependent phone recognizers can be used to discriminate languages [24]. Also, by labeling the input speech to sequence of phones, the phonotactics of the resulting phone sequence can be used to perform language identification [4, 25]. The phonotactics are modeled using statistical language models such as bigrams or trigrams. Systems which try to model phones, phone frequencies and phonotactics perform better than models based only on the acoustic information.

In speech recognition systems, language models in the form of stochastic grammar will give the likelihood of certain words or subword units appearing together. This will help to reduce many of the errors of the word recognizer or subword unit recognizer. For text independent language identification, it is generally not feasible to construct word

models for each of the target languages, as it would be difficult to obtain dictionaries with sufficient coverage in each of the languages. However it is possible to create models which represents the sequential statistics of more basic units such as phonemes, or broad categories of phonemes in each of the languages.

In one approach, a neural network was used for classifying speech to seven broad phonetic categories such as silence, fricative, pre-vocalic-stop, vowel, pre-vocalic sonorant, inter vocalic sonorant and post-vocalic sonorant [2]. The main reason for the choice of broad categories is that they are language-independent, and thus can be used even for languages for which labeled training data is not available. A stochastic grammar is then used to compute the likelihoods of co-occurrence of those units, which will capture some of the so-called phonotactic regularities in the target languages.

Bigram models can be employed to capture the likelihood that a phoneme is followed by any other phoneme. The language of an utterance is determined by successively decoding the test utterance using phone recognizer and bigram model of each of the target languages. The decoding with the highest likelihood is taken to indicate the language to which the utterance belongs. Since the likelihood computed during the decoding process is a product of both acoustic and phonotactic components, this score actually incorporates both acoustic and phonotactic information [12]. There are three different variations for this phone-based approach [15].

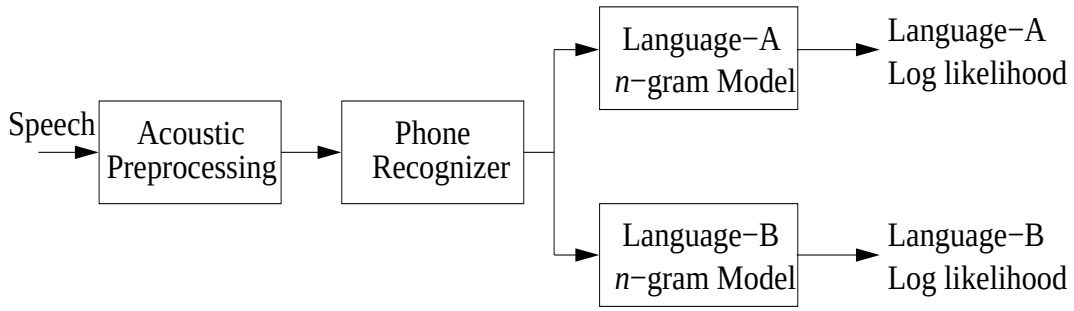


Fig. 2.1: Block schematic of PRLM - Single language phone recognition followed by language-dependent n-gram modeling.

(a) PRLM: As shown in Fig. 2.1, single language phone recognition followed by

language-dependent n-gram language modeling is abbreviated as PRLM. In this system, phones present in training data are recognized using a front end phone recognizer. The resulting symbol sequence is analyzed, and an n-gram language model is obtained for each language.

(b) Parallel PRLM: Although PRLM is an effective means of identifying the language of a speech, it is difficult to train a phone recognizer which works well for all the languages. Alternatively, parallel PRLM uses multiple single language phone recognizers as shown in Fig. 2.2. Separate phone recognizers are trained for languages in which labeled data is available. This is followed by phonotactic analysis using language models.

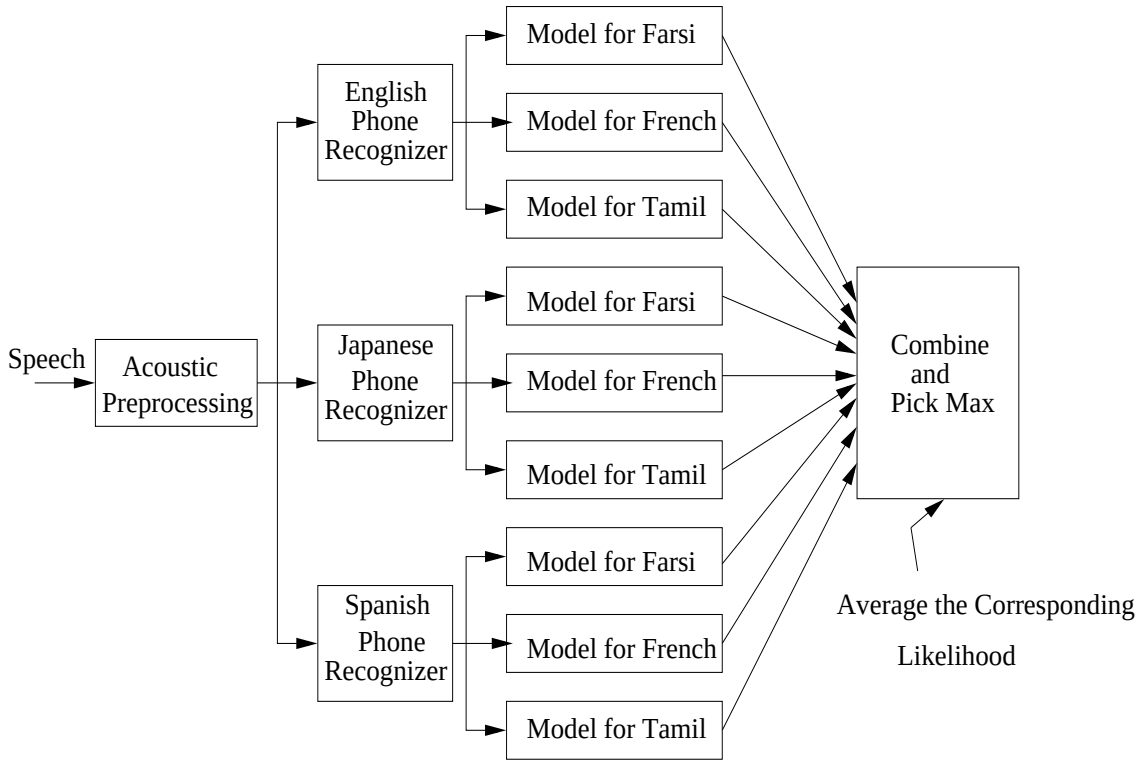


Fig. 2.2: Block schematic of parallel PRLM - Multiple phone recognizer followed by n-gram modeling.

(c) PPR: Language-dependent parallel phone recognition (PPR) is preferred when labeled training speech is available in each language to be identified. In PPR

several language-dependent phone recognition front-ends are used in parallel as shown in Fig. 2.3. The likelihood of the parallel phone recognizers are compared, from which language is hypothesized [15]. The prior knowledge of the language in terms of n-gram models is in-built in each of the language-dependent phone recognizer.

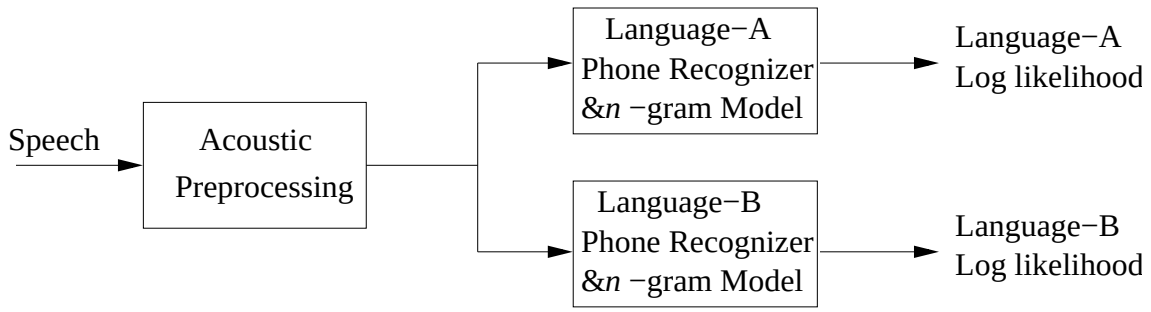


Fig. 2.3: Block schematic of PPR - Language dependent phone recognizers running in parallel.

Researchers have also focused on the problem of identifying and processing only those phones or syllables that carry the most language discriminating information [26]. These language dependent phones are called mono phones in the literature, and such phones need to be identified in each language. It is possible to use pair-wise contrastive monophones alone for language identification [26].

In a parallel subword recognition (PSWR) approach, speech is segmented in terms of subword units. This is followed by segment clustering using k-means clustering and modeling using hidden Markov models (HMM) [27]. This PSWR based LID system operates in a PPR framework, but without requiring manually labeled phonetic data in any of the languages. An input utterance is classified by maximum likelihood decision obtained by the front-end subword recognizer or by the back-end language model score, or jointly by both.

In another parallel syllable-like unit recognition approach, speech is segmented into syllable-like units. Similar syllables are grouped together to train HMM-based syllable

Table 2.4: Summary of phoneme or syllable-based LID systems

Unit recognized	Method	Reference
Broad class of phoneme	Broad phonetic statistics	Muthusamy 1991 [2]
Phoneme	PRLM	Zissman 1993 [15], Hazen 1993 [25]
	Parallel PRLM	Zissman 1993 [15], Jiri 2001 [4]
	PPR	Zissman 1993 [15], Lamel 1994 [24]
Monophone	Language-specific phones	Berkling 1998 [26]
Subword unit	Parallel subword recognition	Sai Jayaraman 2003 [27]
Syllable-like unit	Parallel syllable recognition	Nagarajan 2004 [28]

models [28]. After the initialization of the selected syllable models, the parameters of the models are re-estimated by a process called incremental training. These language-dependent syllable models are then used for identifying the language of the unknown test utterance. Table 2.4 summarizes various phoneme or syllable-based approaches.

2.2.4 Systems Based on Words

LID systems based on words employ sequence modeling at word level, but do not use full speech-to-text systems. This is an approach to the LID problem where phones are recognized first, followed by words, and eventually language. It uses lexical modeling for language identification [29]. An incoming utterance is processed by parallel lan-

guage dependent phone recognizers. Hypothesized language-specific word occurrences are identified from the resulting phone sequences.

2.2.5 Systems Based on Continuous Speech

LID systems based on continuous speech recognition use more language-specific knowledge to obtain better performance [30,31]. It employs one continuous speech recognizer per language. While testing, all these recognizers are run in parallel. The language of the recognizer which yields the highest likelihood is hypothesized as the language of the test utterance. Such systems hold the promise of high-quality language identification, because they use higher-level knowledge (words and word sequences) to make the LID decision. Furthermore, one obtains a transcription of the utterance as a byproduct of the LID. On the other hand, they require several hours of labeled training data for each language to create separate continuous speech recognizers.

Next we review systems used for automatic speaker recognition. The systems are broadly categorized based on the characteristics that are used for modeling the speaker.

2.3 REVIEW OF AUTOMATIC SPEAKER RECOGNITION

The objective of speaker recognition is to recognize a person from his or her voice. Speaker characteristics are manifested in speech signal as a result of anatomical differences inherent in the speech production organs, and differences in the learned speaking habits of individuals [32]. Research on automatic speaker recognition has been undertaken for well over four decades, and it continues to be an active area [33, 34].

The general area of speaker recognition can be classified as speaker identification and speaker verification. Speaker identification is the task of determining who is talking from a set of speakers. The unknown person makes no identity claim, and so the system must perform a $1 : N$ classification. Generally it is assumed that the unknown voice must come from a fixed set of known speakers, thus the task is often

referred to as closed-set identification. Speaker verification (also known as speaker authentication or detection) is the task of determining whether a person is who he or she claims to be (a yes or no decision) [35]. Since it is generally assumed that impostors are not known to the system, this is referred to as an open-set task. In general, most applications of speaker recognition technology use open-set speaker verification [36].

The speaker recognition is also classified as text-dependent and text-independent. In text-dependent speaker recognition, a speaker speaks the same text during enrollment and verification, and the recognition system has prior knowledge of the text. The prior knowledge and constraint of the text can greatly improve the performance of the recognition system. In text-independent system, there is no prior knowledge by the system of the text to be spoken.

In speaker recognition systems, the first step is to identify the possible characteristic variation among different speakers. Speech is generated by the excitation of vocal tract system with a time varying input. Speaker-specific information is present in both vocal tract system as well as excitation source components of speech production mechanism. Characteristics of speakers vary due to difference in physiology of vocal tract and excitation source. Speakers also have difference in certain acquired speaking habits learned over a period of time. The learned characteristics are reflected in prosody and idiolect of a speaker. Speaker recognition systems are now reviewed, based on the speaker-specific characteristics used for recognition.

2.3.1 Systems Based on Vocal Tract Characteristics

The vocal tract system can be considered as a cascade of cavities of varying cross sections. These cavities assume varying size and shape depending on the sound unit that is uttered. Also there are some variations in the physiology of vocal tract among speakers. The resonant frequencies or formants of vocal tract vary in frequency, bandwidth and relative amplitude depending on both the sound unit and the speaker. The magnitude of the frequency spectrum encodes this information and is useful for repre-

senting the physiology of speaker's vocal tract system. Short-time analysis, typically with 20 msec frame size and 5 to 10 msec frame shift, is used to compute a sequence of magnitude spectra using either linear prediction (LP) analysis or discrete Fourier transform (DFT) analysis.

First, some form of speech activity detection is performed to remove non-speech portions from the signal. Next, features conveying speaker information are extracted from the speech. Linear prediction analysis of speech [37] provides an approximation to short-time spectrum of the vocal tract filter. Different parametric representation of speech derived from the LP analysis were investigated for their effectiveness for automatic speaker recognition [38]. Linear prediction cepstral coefficients (LPCC) were shown to be effective for speaker recognition [39]. Baseline speaker recognition system of Indian Institute of Technology (IIT) Madras uses weighted linear prediction cepstral coefficients (WLPCC) for modeling speakers [40]. The Mel-frequency cepstral coefficients (MFCC) have been used for speaker recognition [35, 41]. The MFCCs are obtained by warping the frequency scale in such a way to resolve the magnitude spectrum finely at lower frequencies and relatively coarsely at higher frequencies [42]. The magnitude spectra are then converted to cepstral features, and time-differential (delta) cepstrals are appended. Finally, some form of channel compensation technique, most commonly cepstral mean subtraction, are applied to the features.

2.3.2 Systems Based on Excitation Source Characteristics

During the production of speech, the vibration of vocal folds provides quasi-periodic impulse-like excitation to the vocal tract system. Inverse filtering in LP analysis results in LP residual, which is an approximation to the excitation signal [43, 44]. It has been shown that combining evidence from spectral features and excitation source features improves the overall performance of the IITM speaker recognition system [45].

Liljencrants-Fant (LF) model has been used as a parametric model to characterize glottal flow derivative [46]. Estimate of glottal flow derivative was obtained using LF

model to capture the coarse structure, while fine structure was represented by energy and perturbation measures [47]. Both coarse and fine structure resulted in reduction of error when used with MFCCs.

2.3.3 Systems Based on Prosody

Pitch is a perceptual attribute of sound. The physical correlate of pitch is the fundamental frequency (F_0) of vibration of vocal folds. Fundamental frequency reflects speaker-specific characteristics due to the differences in physical structure of the vocal folds among speakers. Variation of pitch as a function of time is called intonation, and is represented by the F_0 contour. The dynamics of the F_0 contour can be different among speakers due to different speaking style and accent.

The distribution of F_0 values is speaker-specific. The global statistics of F_0 values of a speaker is captured using appropriate distributions for speaker verification task [48]. It has been shown that the dynamics of F_0 contour which reflect the speaking style of a person can also contribute to speaker verification task. In [49], the speaker's F_0 movements are modeled by fitting a piecewise linear model to the F_0 track to obtain a stylized F_0 contour. Each linear F_0 segment is represented using median F_0, slope, and duration. These features are modeled by log-normal, normal and shifted exponential distributions, respectively.

In order to incorporate prosodic features, speaker recognition systems generally require significantly more data for training. In 2001, in response to the growing interest in the use of prosody and idiolect for speaker recognition, NIST introduced the extended data task, based on switchboard corpus of conversational telephone speech [50]. Unlike the traditional speaker recognition tasks, the extended data task provide multiple whole conversation sides for speaker training (4-side /8-side /16-side), where each conversation side contains approximately 2.5 minutes of speech, and tested on one side conversation.

A workshop was conducted at the John Hopkins University (JHU), USA to explore a wide range of features for speaker recognition using the extended data task as its testbed [51]. The usage of various prosodic features were explored in this workshop [52, 53]. The dynamics of linear stylized F_0 and energy trajectories of each speaker are modeled using bigram models [52]. In [53], the focus is on investigations of a diverse collection of prosodic features based on F_0, segment duration, and pause duration.

In [54], duration, pitch and energy features are computed for each estimated syllable regions. The syllable boundaries are obtained from automatic speech recognizer (ASR) output. These features are quantized and it is used to form N-grams which is referred as SNERF-grams (N-grams of Syllable-based Non-uniform Extraction Region Features). Duration characteristics, namely, word features (sequence of phone durations in the word), duration of phones, and sequence of HMM state durations have been used for modeling duration [55, 56].

2.3.4 Systems Based on Phone Pronunciations

The phone-based system is a text-independent speaker recognition system based on difference among speakers in the dynamic realization of phonetic features [57]. Pronunciation variation among speakers are used here, rather than the difference in the distribution of spectral features. System based on phone pronunciations uses the time sequence of phones obtained from a bank of open-loop phone recognizers to capture some information about the speaker-dependent pronunciations. This is followed by the "bag-of-n-grams" classifier [57]. A binary tree is also shown to be useful instead of an n-gram model [58]. Speaker-dependent pronunciations are learned by comparing constrained ASR phone streams with open-loop phone streams. The conditional probabilities of phones are computed for each speaker [59].

2.3.5 Systems Based on Idiolect

It is generally recognized that human listeners can distinguish between speakers who are familiar to them better than those who are unfamiliar. This increased ability is due to speaker idiosyncrasies that are recognized by the listener, either consciously or subconsciously [10]. "Familiar-speaker" differences, however, surely relate to long term speech patterns, such as the usage of certain words and phrases, and to the features tied to these patterns, such as intonation, stress and timing. The use of such patterns helps to improve the performance of speaker recognition systems.

Idiolect refers to the way a particular person uses language. It has been proved that n-gram models have the potential to capture idiolect of a speaker [10]. The n-gram language models are conventionally used for improving the performance of speech recognition system, and these models are made speaker-independent by including data from various speakers. While using it for speaker recognition, language model is trained for a specific speaker, assuming that sufficient data is available. In this approach, transcriptions of speech data at levels of words (or phrases) are used as input for training the language models. Table 2.5 summarizes the speaker discriminating cues and the features used by the researchers.

Researchers also use output from a speech recognizer to allow application of classic text-dependent template matching techniques to the text-independent speech. For example, for F_0 contour matching, the F_0 contour from an enrollment phrase is used as the reference template for a speaker. This is matched to the F_0 contour of the same phrase using dynamic time warping (DTW). Dynamic modeling using HMM has been suggested for modeling F_0 contours of selected words and phrases [9].

The review of approaches for language and speaker recognition discussed here is from the perspective of characteristics and features. Motivation for the present work is discussed in the next section.

Table 2.5: Summary of speaker discriminating cues and features for speaker recognition.

Cue	Feature	Reference
Vocal tract system	Spectral features (MFCC, LPCC)	Atal 1974 [38] Furui 1981 [39] Reynolds 1995 [35, 41]
Excitation source	LP residual	IITM system 2001 [45]
	Glottal flow derivative	Plumpe 1999 [47]
Prosody	F_0 values	Sonmez 1998 [48]
	F_0 dynamics	Sonmez 1998 [49] Adami 2003 [52] superSID 2003 [51, 53]
	Prosody (using ASR)	Shriberg 2005 [54]
Phone pronunciations	Spectral features	Andrews 2002 [57]
Idiolect	Text transcription	Doddington 2001 [10]

2.4 SUMMARY AND MOTIVATION FOR THE PRESENT WORK

This chapter gives a brief review of language and speaker recognition. The LID system range from a single acoustic model per language, to a complex multilingual speech recognizer. Better recognition performance is obtained by using language-specific knowledge at the subword and word levels. Automatic speaker recognition systems range from systems which use physiological difference to systems that try to model the acquired speaking habits.

This review on language and speaker recognition systems reveals the similarity in approaches for language and speaker recognition. Spectral features encodes information regarding the physiology of the vocal tract system, and hence useful for speaker modeling. As the vocal tract size and shape is also decided by the sound unit spoken, the spectral features are also language-specific. Prosodic features convey information regarding the physiological constraints as well as acquired speaking habits of a person.

It also conveys the prosodic constraints of the language. As spectral and prosodic features are both language and speaker-specific, their representation should be in a manner that highlights the language-specific part or speaker-specific part, depending on the task. We hypothesize that both language and speaker characteristics are present in speech signal and it is difficult to separate the language part and speaker part from features derived from speech signal. The focus of this work is on the extraction and representation of language-specific and speaker-specific features for recognizing language or speaker.

To use vocabulary and phonotactics for LID, a multilingual recognizer is needed. Building a multilingual recognizer may be impractical due to several reasons. It requires knowledge about the acoustic-phonetic, lexical and linguistic rules for each language of interest [12], and hours of manually labeled database is required for this. Developing such systems for a number of languages would be laborious and time consuming. Preparing manually labeled database for all languages to be identified requires human experts in those languages and substantial amount of supervision. As labeled corpora is not available for most of the Indian languages, identifying language using features derived from speech signal is important. Though the explicit LID systems outperform implicit systems, the nonavailability of labeled database in many languages make the implicit systems attractive for many practical applications. Similarly for speaker recognition, to make use of prosody, some of the existing systems use segment boundaries obtained using an automatic speech recognizer. Extracting features directly from the speech signal is important for tasks like language and speaker recognition. In this work, we focus on the extraction of features directly from the speech signal, and its representation suitable language and speaker recognition.

A language or speaker recognition system based on acoustic-phonetics uses spectral vectors for modeling. The accuracy of recognition systems using spectral features reaches a saturation for a fairly long test speech. Spectral-based systems may perform well in acoustically matched and noise-free conditions. The current speaker recognition systems are dominated by the short-time spectral features. These systems ignore

speaker-specific patterns in prosody and idiolect. Due to the susceptibility of spectral features to channel, noise and environment, researchers have realized the importance of non-spectral features for increasing the reliability of recognition systems. The prosodic features derived from pitch contour, amplitude contour, duration etc. are less affected by channel characteristics, environment and noise. Therefore systems which combine evidence from prosodic features along with spectral features may be robust than systems using spectral features alone. This motivated us to consider prosody as an additional feature for language and speaker recognition.

In the next chapter, we discuss formulation of the problem of language and speaker recognition using probabilistic approach.

CHAPTER 3

ANN MODELS FOR LANGUAGE AND SPEAKER RECOGNITION

3.1 INTRODUCTION

In this chapter, we discuss formulation of the problem of language and speaker recognition using probabilistic approach, and the implementation of various stages using artificial neural network (ANN) models. Probabilistic approaches have received attention in various domains such as speech processing and image processing. According to this approach, language/speaker recognition problem is treated as the estimation of a posteriori probability. It is further simplified to the estimation of likelihood probability, assuming equal a priori probabilities of occurrence to all languages or speakers included in the task. We propose neural network models for obtaining an estimate similar to these probabilities.

Language or speaker recognition task mainly involve three stages namely, feature extraction, modeling and evaluation. Feature extraction deals with extraction of language or speaker-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 recognizing the language or speaker. The performance of language or speaker recognition systems are influenced by all the three stages, namely, feature extraction, model building and evaluation strategies. This chapter discusses the modeling strategies used in this work.

This chapter is organized as follows: The next section describes the probabilistic formulation for language and speaker recognition. Section 3.3 describes neural network models, namely, multilayer feedforward neural network and autoassociative neural net-

work used for modeling. In Section 3.4, we discuss features derived at various levels of speech, namely frame level, syllable level and multisyllabic level, useful for language identification. In Section 3.5, we discuss the probabilistic approach for speaker recognition using features derived at three different levels, namely, subsegmental, segmental and suprasegmental. Section 3.6 summarizes the chapter.

3.2 PROBABILISTIC APPROACH FOR LANGUAGE AND SPEAKER RECOGNITION

Language (or speaker) recognition can be expressed as a problem of finding the most likely language (or speaker) C^* of the input speech, from a set of known languages (or speakers). Let $\{C_i\}$, $1 \leq i \leq M$ denote the set of classes representing languages (or speakers), and let O denotes the observations derived from input speech. The recognition problem can be formulated in probabilistic terms as follows:

$$C^* = \arg\max_i P(C_i|O) \tag{3.1}$$

where $P(C_i|O)$ is the a posteriori probability of the class C_i for the given speech utterance expressed in terms of O. Let us assume that the observation O belongs to one of M classes $C_i, 1 \leq i \leq M$. According to the rule given in (3.1), the objective is to choose the class C^* for which the a posteriori probability $P(C_i|O)$ is maximum for a given O. By Bayes rule,

$$P(C_i|O) = \frac{P(O|C_i)P(C_i)}{P(O)} \tag{3.2}$$

where $P(O|C_i)$ represents the likelihood probability of O corresponding to the class C_i, and $P(C_i)$ denotes the a priori probability of the class C_i. The problem can be reformulated as follows:

$$C^* = \arg\max_i \frac{P(O|C_i)P(C_i)}{P(O)} \tag{3.3}$$

If there is no reason to prefer one class over another, $P(C_i)$ can be assumed equal for all the classes [60]. $P(O)$ being a common term for all the classes, the problem can be

simplified to:

$$C^* = \arg\max_i P(O|C_i) \tag{3.4}$$

The probabilities of different classes can be compared by evaluating $P(O|C_i)$, $1 \leq i \leq M$, and it can be used for selecting a particular class for which the probability is the largest. Thus the task of language (or speaker) recognition is treated as the estimation of a posteriori probability, and can be simplified to the estimation of likelihood probability under certain assumptions.

In our work, we use neural network models for obtaining estimates similar to these probabilities. The types of neural network models used in this study are explained in the next section.

3.3 NEURAL NETWORK MODELS FOR LANGUAGE AND SPEAKER RECOGNITION

Artificial neural network (ANN) models with different topologies can perform different pattern recognition tasks [61, 62]. A multilayer feedforward network can be designed to perform the task of pattern classification. A special class of feedforward neural networks called autoassociative neural network (AANN) is useful for capturing the distribution of the feature vectors from the given training data [63].

3.3.1 Multilayer Feedforward Neural Network Classifier

The main objective in pattern classification is to assign a label to a given pattern, often represented as a feature vector. Let us represent the output of a multilayer feedforward neural network (MLFFNN) by the function $f(x, \theta)$ where x is the input vector, and θ represents the values of all parameters that define the network. For simplicity of notation, consider a two class problem for which the desired output of the neural network takes on the value a, if x corresponds to class C_1, or the value b, if x corresponds to class C_2. Typical structure of MLFFNN for a two class problem is shown in Fig. 3.1. The performance of the network is measured using the mean

squared error defined as:

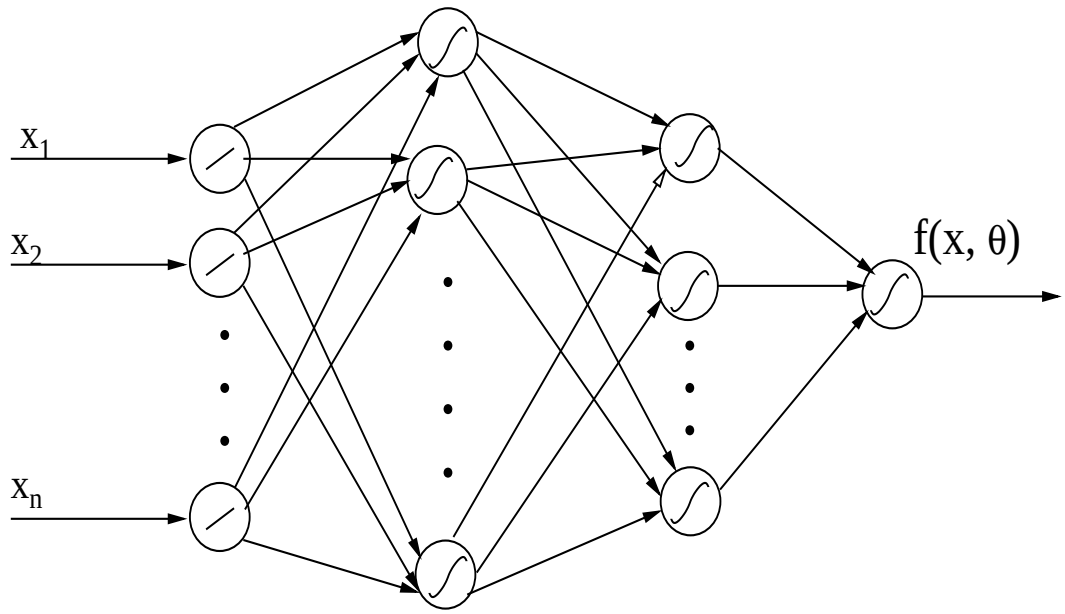


Fig. 3.1: Structure of a multilayer feedforward neural network with single output.

$$E = \frac{1}{N}\left(\sum_{x \in C_1}(f(x,\theta) - a)^2 + \sum_{x \in C_2}(f(x,\theta) - b)^2\right) \tag{3.5}$$

where N is the total number of training samples. If we assume that N is large, and the number of samples from each of the classes is in proportion to the a priori probabilities of the two classes, we can approximate the above summation by integrals [64] as follows:

$$E = \int (f(x,\theta) - a)^2 P(x, C_1)dx + \int (f(x,\theta) - b)^2 P(x, C_2)dx \tag{3.6}$$

where $P(x, C_i)$, $i = 1, 2$, is the joint probability density function of the observations x and the class C_i. Equation (3.6) can be rewritten as

$$\begin{aligned} E = & \int f^2(x,\theta)(P(x,C_1) + P(x,C_2))dx \\ & -2\int f(x,\theta)(aP(x,C_1) + bP(x,C_2))dx \\ & +a^2\int P(x,C_1)dx + b^2\int P(x,C_2)dx \end{aligned} \tag{3.7}$$

Let $P(x) = P(x, C_1) + P(x, C_2)$ denote the unconditional probability of an observation. Defining the term

$$d(x) = \frac{aP(x, C_1) + bP(x, C_2)}{P(x)} = aP(C_1|x) + bP(C_2|x)$$

Substituting $d(x)$ error E becomes

$$E = \int f^2(x, \theta)p(x)dx - 2\int f(x, \theta)d(x)p(x)dx \tag{3.8}$$
$$+a^2 \int P(x, C_1)dx + b^2 \int P(x, C_2)dx$$

This can be written as follows:

$$E = \int [f(x, \theta) - d(x)]^2 p(x)dx + a^2 P(C_1) + b^2 P(C_2) - \int d^2(x)p(x)dx \tag{3.9}$$

Only the first term in the above equation depends on the parameters of the network. Therefore adjusting the network parameters θ to minimize E is equivalent to minimizing the mean square error between the network output $f(x, \theta)$ and $d(x)$. When $x \in C_1$, we choose $a = 1$ and $b = 0$ as desired output and then $d(x) = P(C1|x)$. Therefore when $x \in C_1$, the network parameters are adjusted to minimize the first term in equation (3.9), the network output is expected to be

$$f(x, \theta) = P(C_1|x) \tag{3.10}$$

where $P(C_1|x)$ is the a posteriori probability of class C_1 , the probability that the class C_1 has occurred given that x has been observed. The Eqn. (3.10) indicates that the network is trained to approximate the a posteriori probability in a mean square sense [64].

In many applications a neural network is designed to discriminate between M classes ($M > 2$). In this case network will have M outputs $f_j(x, \theta)$, $j = 1, 2, ..., M$. The desired output of the neural network will be 1 for the class to which the input training class belongs and 0 for all the other outputs. It can be proved that minimizing the squared error criterion in this case is equivalent to minimizing the term

$$\sum_{j=1}^{M} \int [f(x, \theta) - P(C_j|x)]^2 p(x)dx$$

This shows that the parameters of network are being used to simultaneously approximate M different functions, such that average of the squared errors is minimized.

3.3.2 Autoassociative Neural Networks

For a pattern classification problem, it is necessary to capture the characteristics of each class from the features derived from the training data for that class. This involves estimation of probability density function of the feature vectors for each class. Conventionally parametric models such as Gaussian mixture models (GMM) have been used to capture the distribution of the feature vectors for each class. While using GMM, the components of the distribution are assumed to be Gaussian and the number of mixtures is generally fixed a priori [65]. In applications such as speech processing, feature vectors can have any arbtrary distribution and hence can not be adequately be described by a GMM. Autoassociative neural network (AANN) models provide an alternate modeling technique which does not impose any constraint on the shape of the distribution of the feature vectors. AANN models can capture arbitrary shape of distribution, including the patterns that can be captured by GMM [63]. Since AANN is more general, we have adopted it for capturing the distribution of data.

When a feedforward neural network is trained with an output equal to the input feature vector, then it is said to operate in an *autoassociation* mode. The network is called an *autoassociative* neural network, because it is trained (or expected) to reproduce its input [61, 62]. It consists of an input layer, an output layer and one or more hidden layers. A typical structure of a five layer AANN model is shown in Fig. 3.2. The number of units in the input and output layers is equal to the size of the input vectors. The number of units in the middle hidden layer is less than the number of units in the input and output layers, and this layer is called the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in the hidden layers can be either linear or nonlinear. The AANN model, with a dimension compression layer

in the middle, is used primarily for capturing the distribution of input features in the feature space. The ability of AANN models to estimate arbitrary densities has been demonstrated [63]. It was illustrated experimentally that a network can be designed to capture the distribution of the given data, depending on the constraints imposed by the structure of the network [63, 66]. The details of this study on AANN models is given in Appendix A.

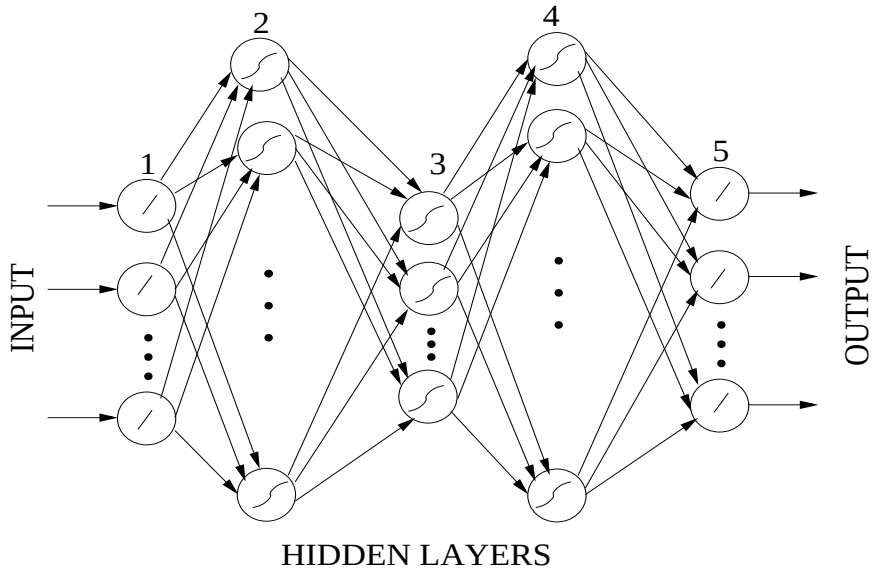


Fig. 3.2: Structure of five layer AANN model.

To capture the distribution of the input feature vectors in the feature space, the feature vectors are extracted from the signal, and are presented in a random order to the AANN. The weights of the network are adjusted using the backpropagation training algorithm. As the mean square error between the actual and desired outputs are minimized, the clusters of points in the input space determine the shape of the hypersurface obtained by the projection onto the lower dimensional space.

While testing, the output of the AANN model is computed with input test vector, and the squared error with respect to the output vector is calculated for each input test vector. The error E_i for i^{th} test vector is transformed into confidence value by using

$C_i = exp(-E_i)$. If the error is small, the resulting C_i will be close to 1, and when the error is large, C_i will be small. When the error is zero, C_i will take the maximum value of 1. Though C_i is not strictly a probability value, it may be interpreted as similar to the log likelihood probability. The average confidence value for a test utterance is computed as $C_\mu = 1/N \sum_{i=1}^{N} C_i$, where N is the total number of feature vectors in the test utterance [40].

3.4 MULTILEVEL FEATURES FOR LANGUAGE IDENTIFICATION

Let $L = \{L_1, L_2, L_3, ..., L_M\}$ represent the set of M languages. When a speech utterance is given, the LID system must derive features represented by O from the utterance to decide which of the M languages in L was spoken. The problem can be expressed as:

$$L^* = \arg\max_i P(L_i|O) \tag{3.11}$$

In this study, we use three levels of feature representation, namely, frame, syllabic and multisyllabic levels. In order to represent variations in acoustic-phonetics, features are derived at the frame level. Features are represented at the level of syllable, to use the difference in the realization of syllables in various languages. Difference among languages in terms of phonotactics and prosody extend beyond the level of syllables and hence should be represented at multisyllabic level. Let F, S and T be the sequence of feature vectors derived from the test utterance at frame, syllabic and multisyllabic levels, respectively. With this information, the most likely language L^* is found using the following expression:

$$L^* = \arg\max_i P(L_i|F, S, T) \tag{3.12}$$

$P(L_i|F, S, T)$ can be written as:

$$P(L_i|F,S,T) = \frac{P(F,S,T,L_i)}{P(F,S,T)} \tag{3.13}$$

Assuming that F, S and T are independent [12], this gets simplified to

$$P(L_i|F,S,T) = \frac{P(F|L_i)P(S|L_i)P(T|L_i)P(L_i)}{P(F)P(S)P(T)} \tag{3.14}$$

Assuming equal a priori probability $P(L_i)$ for all the languages and assigning constant values for $P(F)$, $P(S)$ and $P(T)$, the language recognition problem can be represented as

$$L^* = \arg\max_i P(F|L_i)P(S|L_i)P(T|L_i) \tag{3.15}$$

The information obtained from frame, syllabic and multisyllabic levels may be modeled separately, and the three likelihood components in Eqn. (3.15) may be computed independently, where $P(F|L_i)$, $P(S|L_i)$ and $P(T|L_i)$ refers to the likelihood probabilities obtained using features represented at the frame, syllabic and multisyllabic levels, respectively.

3.4.1 Frame Level Features

Let $F = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, ..., \mathbf{f}_k\}$ represent the sequence of feature vectors derived at the frame level. The likelihood probability $P(F|L_i)$ is given by

$$P(F|L_i) = P(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, ..., \mathbf{f}_k|L_i) \tag{3.16}$$

Assuming that each frame is independent, it can be represented as

$$P(F|L_i) = \prod_{j=1}^{k} P(\mathbf{f}_j|L_i) \tag{3.17}$$

This is equivalent to accumulating the log likelihood probability

$$logP(F|L_i) = \sum_{j=1}^{k} logP(\mathbf{f}_j|L_i) \tag{3.18}$$

3.4.2 Syllabic Features

Let $S = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \cdots, \mathbf{s}_n\}$ represent syllable level features derived from the test utterance consisting of n syllables.

Considering each syllable as an independent unit, the $p(S|L_i)$ is computed as

$$P(S|L_i) = \prod_{j=1}^{n} p(\mathbf{s}_j|L_i) \tag{3.19}$$

which is equivalent to accumulating the log-likelihood probability

$$log[P(S|L_i)] = \sum_{j=1}^{n} log[p(\mathbf{s}_j|L_i)] \tag{3.20}$$

Autoassociative neural network models are used for capturing the distribution of both the frame level and syllable level features. The average confidence score obtained for a test utterance using AANN model is used in place of log likelihood probability $log[P(S|L_i)]$.

3.4.3 Multisyllabic Features

We hypothesize that, in order to capture prosodic characteristics such as stress, rhythm and intonation of a language, features need to be extracted from a level higher than syllable. At a normal rate of conversation, it is not possible to give stress to two successive syllables. The stress assigned to one syllable is achieved at the expense of the syllable immediately preceding or following. In words with more than two syllables, it is possible to contrast the extreme syllables relative to the intermediate syllable [67]. Also the tones of adjacent syllables influence the shape and height of the pitch contour of a particular syllable. To model the phonotactic variations among languages, the co-occurrence of subword units need to be captured. As an approximation, features derived from three consecutive syllables are chosen to form a basic unit for representing prosodic and phonotactic features.

Let $T = \{\mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_q\}$ represent feature vectors obtained from the test sequence containing q trisyllabic units. Assuming that each trisyllabic unit is independent,

$P(T|L_i)$ can be represented as

$$P(T|L_i) = \prod_{j=1}^{q} P(\mathbf{t}_j|L_i) \tag{3.21}$$

3.5 MULTILEVEL FEATURES FOR SPEAKER RECOGNITION

The problem of speaker recognition can be expressed as:

$$S^* = \arg\max_i P(S_i|O) \tag{3.22}$$

where $P(S_i|O)$ is the a posteriori probability of a speaker S_i and O denotes the observations/features derived from the given speech utterance. Here the objective is to choose the speaker S^* for which the posterior probability $P(S_i|O)$ is maximum for a given O.

In this work, the observation O is represented using three different components. The speaker-specific characteristics present in excitation signal is represented at subsegmental (1-5 msec) level, which is less than a pitch period. The vocal tract system characteristics are obtained using short-time spectral analysis of windows or segment containing a few pitch periods (20-30 msec). The prosodic characteristics are represented at suprasegmental (>100 msec) level. Let $R = \{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_m\}$, $F = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_k\}$ and $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ denotes the subsegmental, segmental and suprasegmental features, respectively, for a given test utterance.

$$P(S_i|O) = P(S_i|R, F, X) \tag{3.23}$$

According to the definition of class conditional probability

$$P(S_i|R, F, X) = \frac{P(R, F, X, S_i)}{P(R, F, X)} \tag{3.24}$$

Assuming that R, F and X are independent, we get

$$P(S_i|R, F, X) = \frac{P(R|S_i)P(F|S_i)P(X|S_i)P(S_i)}{P(R)P(F)P(X)} \tag{3.25}$$

Assuming equal a priori probability $P(S_i)$ for all the speakers, and assigning constant values for $P(R)$, $P(F)$ and $P(X)$, the speaker recognition problem can be represented as

$$S^* = \arg\max_i P(R|S_i)P(F|S_i)P(X|S_i) \tag{3.26}$$

Effectively, there are three likelihood probability components which can can contribute for speaker recognition. Assuming that observations R, F and X are independent, each likelihood components can be computed separately. Assuming independence among values of R, F and X, the individual likelihood components can be computed as follows:

$$P(R|S_i) = \prod_{j=1}^{m} p(\mathbf{r}_j|S_i) \tag{3.27}$$

$$P(F|S_i) = \prod_{j=1}^{k} p(\mathbf{f}_j|S_i) \tag{3.28}$$

$$P(X|S_i) = \prod_{j=1}^{n} p(\mathbf{x}_j|S_i) \tag{3.29}$$

These likelihood components may be combined by adding them in the logaithmic domain, to get improved speaker recognition performance.

$$S^* = \arg\max_i \left[\sum_{j=1}^{m} logP(\mathbf{r}_j|S_i) + \sum_{j=1}^{k} logP(\mathbf{f}_j|S_i) + \sum_{j=1}^{n} logP(\mathbf{x}_j|S_i) \right] \tag{3.30}$$

3.6 SUMMARY

In this chapter, we have presented a probabilistic formulation for language and speaker recognition. Here the recognition problem is interpreted as the estimation of a posteriori probability, which is again simplified to estimation of likelihood probability. For capturing various language and speaker-specific characteristics, features from multiple levels of speech are modeled separately. The likelihood from various levels may be combined to improve the recognition performance.

In the next chapter, we explore various features derived at frame level for language recognition.

CHAPTER 4

FRAME LEVEL FEATURES FOR LANGUAGE IDENTIFICATION

4.1 INTRODUCTION

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 useful.

In this chapter we explore various frame ($\leq$ 20 msec) level features for the possible presence of 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. From the review of the existing approaches for LID, we notice that no specific attempt has been made so far in exploring the language-specific information present in excitation source signal. Excitation source characteristics are represented

using the residual signal obtained using LP analysis. As the second order relations are removed in the LP analysis, we hypothesize that language-specific information is present in the higher (> 2) order relations in the samples of the LP residual. This higher order relation need to be captured by the models. The complementary evidence obtained by the spectral and source features may be combined to improve the performance of the combined LID system.

This chapter is organized as follows: In the next section, we describe three different frame level features and their representation used in this study. In Section 4.3, we discuss the performance of the spectral-based LID system for a database of Indian languages. In this section, the issue of speaker variability within the same language is addressed. A method using on multiple speaker models is suggested to deal with speaker variability within the language and the effectiveness of the method is demonstrated using OGI database. In Section 4.4, LID systems based on LP residual and phase of the LP residual are discussed. The performance of different approaches and the performance of the combined system for the database of Indian languages is discussed in Section 4.5. Section 4.6 summarizes the work on LID using frame level features.

4.2 FRAME LEVEL FEATURES FOR LANGUAGE IDENTIFICATION

Each sound unit corresponds to a particular articulatory configuration of the vocal tract. For different languages, the articulatory configuration corresponding to even same sound units may slightly vary due to the difference in pronunciation and effects of coarticulation. This acoustic-phonetic variations among languages can be represented using short-time spectral features. Due to the nonstationary nature of speech production mechanism, the spectral features are extracted over short (typically 20 msec) quasistationary segments of speech data. The excitation corresponding to sound units also may be sound-specific and hence may contain language-specific information. We study the presence of language-specific information in spectral and excitation charac-

teristics using features derived from short spans ($\leq$ 20 msec) of speech signal. For representing these characteristics, features are derived using linear prediction (LP) analysis [37].

4.2.1 Weighted Linear Prediction Cepstral Coefficients

Speech is produced as a result of excitation of time-varying vocal tract system with time-varying excitation. The information corresponding to the vocal tract system and the excitation source may be separated approximately from the speech signal using LP analysis [37]. In the LP analysis each sample is predicted as a linear combination of the past p samples, where p is the order of the prediction. The predicted sample of $\hat{s}(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k). \tag{4.1}$$

In Eqn.(4.1), $\{a_k\}$ are termed as LP coefficients, and they are obtained by minimizing the squared error between the actual and predicted samples. This leads to solving a set of normal equations given by

$$\sum_{k=1}^{p} a_k R(n-k) = -R(n), \quad n = 1, 2, \ldots, p \tag{4.2}$$

$$R(k) = \sum_{n=0}^{N-(p-1)} s(n)s(n-k), \quad k = 1, 2, \ldots, p \tag{4.3}$$

The cepstral coefficients may be derived from the LPCs using the following relations [42]:

$$c_0 = ln\sigma^2 \tag{4.4}$$

$$c_n = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k a_{n-k}, \quad 1 \leq n \leq p \tag{4.5}$$

$$c_n = \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c_k a_{n-k}, \quad n > p \tag{4.6}$$

where σ^2 is the gain term in the LP analysis. The weighted linear prediction cepstral coefficients (WLPCC) are used for representing the spectral features for language

identification tasks.The algorithm used for the extraction of WLPCCs is described in Appendix B.

4.2.2 LP Residual

It is known that the excitation signals of speech are unique for each sound unit and hence it may contain language specific information. However, this information is usually ignored in most of the applications like language identification. The prediction error referred as the LP residual is given by

$$r(n) = s(n) - \hat{s}(n). \tag{4.7}$$

The residual signal obtained after removing the short-time spectral features, contains significant amount of information about the excitation source [43, 44] both at the micro (1-3 msec), and at the macro level (>100 msec). As shown in Fig. 4.1, at the macro level, the LP residual contains the intonation and duration information. The significance of these prosodic characteristics for language identification will be discussed in Chapter 6. At the microlevel, excitation source characteristics represent the sequence information present in excitation signal.

Figure 4.2 shows a 25 msec segment of speech, its 10th order LP spectrum and corresponding LP residual. The LP spectrum and the residual signal can be viewed as decomposition of the signal into approximate vocal tract system and excitation source components. The excitation represented using samples of LP residual may be specific to a particular sound and hence may contain language-specific characteristics. Since the second order correlations in the speech signal are already captured by the LP spectrum, excitation source information is expected to retain the higher (> 2) order relations among the samples. From Fig. 4.2 (c), it can be seen that prediction error is large around the instant of glottal closure. So the LP residual around the region of glottal closure has higher signal-to-noise ratio (SNR), and thus may contain useful excitation source characteristics. We use the LP residual corresponding to high SNR regions to capture the information present in the sequence of excitation.

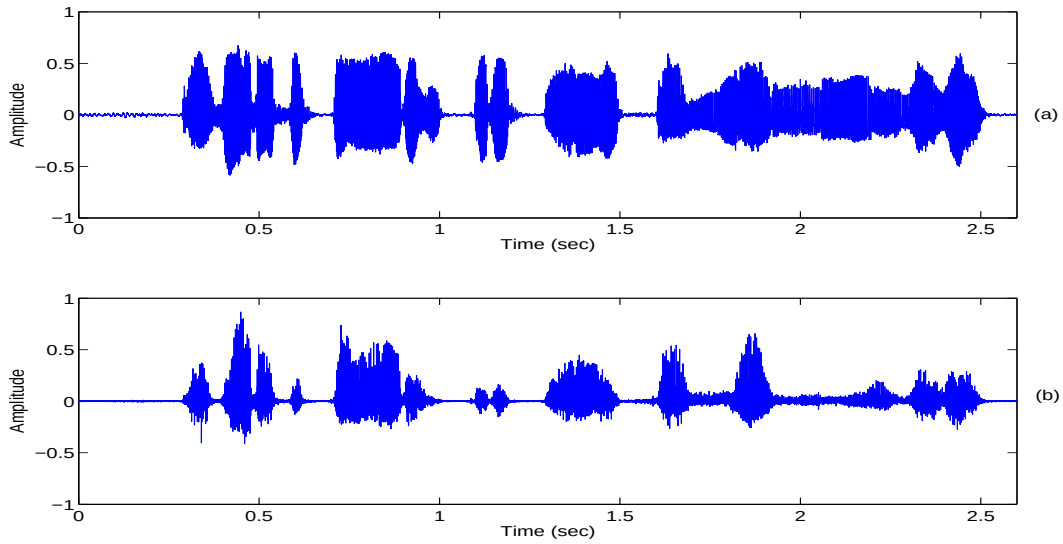


Fig. 4.1: (a) A speech signal and its (b) LP residual.

4.2.3 Phase of LP Residual

Intuitively we believe that the phase of the analytic signal derived from the linear prediction (LP) residual of speech may also contain information about the language. In this work we show experimentally that language-specific information is present even in the phase of the LP residual of speech. But, extraction of the phase information is a difficult task due to phase warping problem. Also the phase of the speech signal may be degraded due to various factors like background noise and channel effects. The knowledge of the Hilbert envelope of the LP residual is used to derive the phase information [68, 69].

The phase information can be obtained from the LP residual ($r(n)$) using the knowledge of the Hilbert transform ($r_h(n)$), which is a 90° phase-shifted version of $r(n)$. The Hilbert envelope of the LP residual is computed from $r(n)$ and $r_h(n)$ as

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)}, \tag{4.8}$$

where $r(n)$ is the LP residual of the speech signal, and $r_h(n)$ is the Hilbert transform

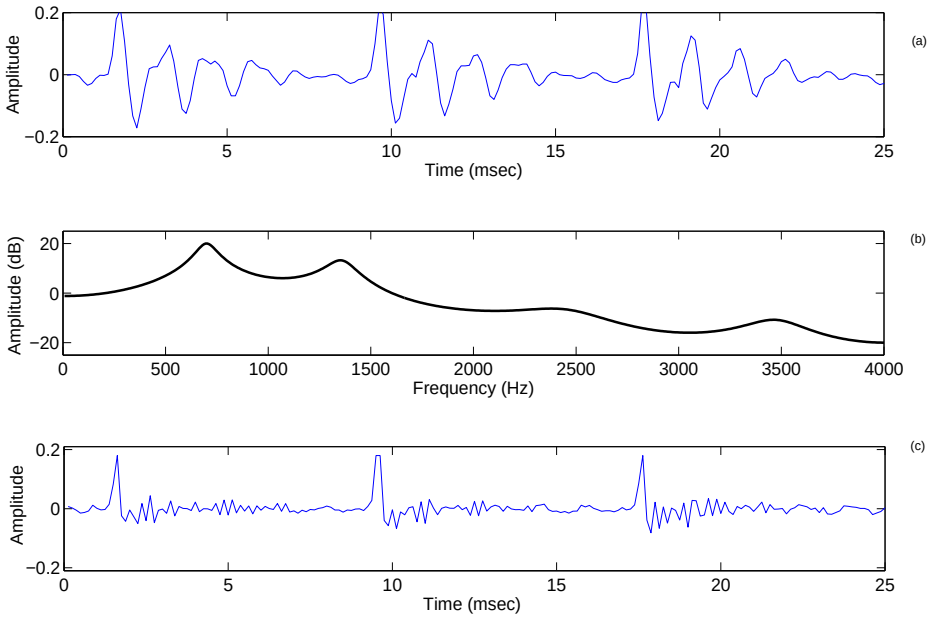


Fig. 4.2: (a) A segment of voiced speech, its (b) 10th order LP spectrum and (c) LP residual.

of $r(n)$ where the Hilbert transform is defined as

$$r_h(n) = IDFT[R_h(\omega)], \tag{4.9}$$

where

$$R_h(\omega) = \begin{cases} jR(\omega), & -\pi \leq \omega < 0 \\ -jR(\omega), & 0 \leq \omega < \pi \end{cases} \tag{4.10}$$

where IDFT refers to the inverse discrete Fourier transform; $R(\omega)$ is the discrete Fourier transform of $r(n)$. Since the Hilbert envelope $h_e(n)$ represents the magnitude information of the LP residual signal, we can obtain the cosine of the phase from $r(n)$ by dividing it with $h_e(n)$. Therefore the phase information ($\theta(n)$) is given by

$$cos\theta(n) = r(n)/h_e(n). \tag{4.11}$$

The phase information extracted from the LP residual signal is represented through $cos\theta(n)$, and it is used to represent the language-specific information [69].

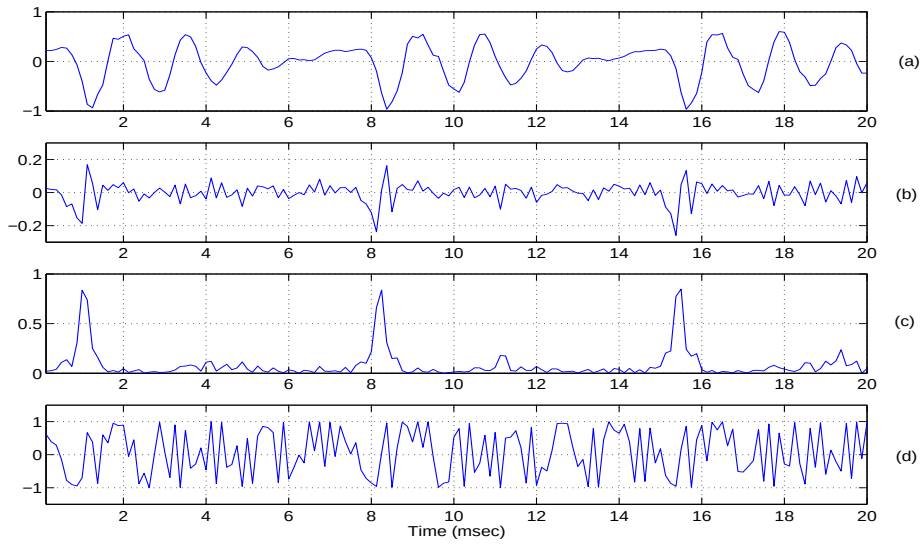


Fig. 4.3: (a) A segment of voiced speech and its (b) LP residual, (c) Hilbert envelope of the LP residual and (d) $cos\theta(n)$ derived from the LP residual.

A segment of voiced speech, the corresponding LP residual, the Hilbert envelope of the LP residual and the $cos\theta(n)$ of the LP residual are shown in Fig. 4.3. Since the LP residual and the phase of the LP residual do not contain any significant second order correlations, we conjecture that the language-specific information may be present in some higher order relations among the samples of the LP residual, and among the samples of $cos\theta(n)$ of the LP residual.

4.3 SPECTRAL FEATURES FOR LANGUAGE IDENTIFICATION

The spectral similarity approach for language identification concentrates on the acoustic-phonetic variations among languages. This exploits the fact that languages have different phonetic repertoire. The acoustic realization of even the same phoneme may be slightly different across various languages. Therefore the distribution of short-time spectral feature vectors in the feature space is considered to be unique for speech of a given language.

The preprocessing of speech signal for deriving WLPCC is as shown in Fig. 4.4.

The silence frames are removed using an amplitude threshold. The differenced speech signal is segmented into frames of 20 msec with a shift of 5 msec, and samples in each frame is multiplied with Hamming window. An 8th order LP analysis is used to capture the properties of the signal spectrum [37]. The recursive relation between the predictor coefficients and cepstral coefficients is used to convert the 8 LP coefficients into 12 cepstral coefficients. The cepstral coefficients for each frame are linearly weighted, and cepstral mean subtraction is done to reduce the effect of channel variability.

Fig. 4.4: Block schematic showing the steps involved in deriving spectral features for language identification.

The weighted linear prediction cepstral coefficients (WLPCCs) extracted from the training data of a language are used to train an AANN model. Separate AANN models are used to capture the distribution of feature vectors of each language. The structure of the AANN model used in the present studies is *12L 38N 4N 38N 12L*, where L denotes linear units, and N denotes nonlinear units. The activation function of the nonlinear units is a hyperbolic tangent function. The network is trained using error backpropagation learning algorithm for 200 epochs [62]. The number of epochs for training was chosen using cross-validation for verification, to obtain the best performance. The learning algorithm adjusts the weights of the network to minimize the mean squared error obtained for each feature vector. Once the AANN model is trained, it is used as a model for the language.

The choice of prediction order p is very important from language identification point of view. Fig. 4.5 compares LP spectra for two different orders of prediction along with short-time DFT spectra. It is clear that lower order (say p=8) LP spectrum captures the gross features of the envelope of speech spectrum. Due to the absence of higher formants, speaker information may be lost in such a representation,

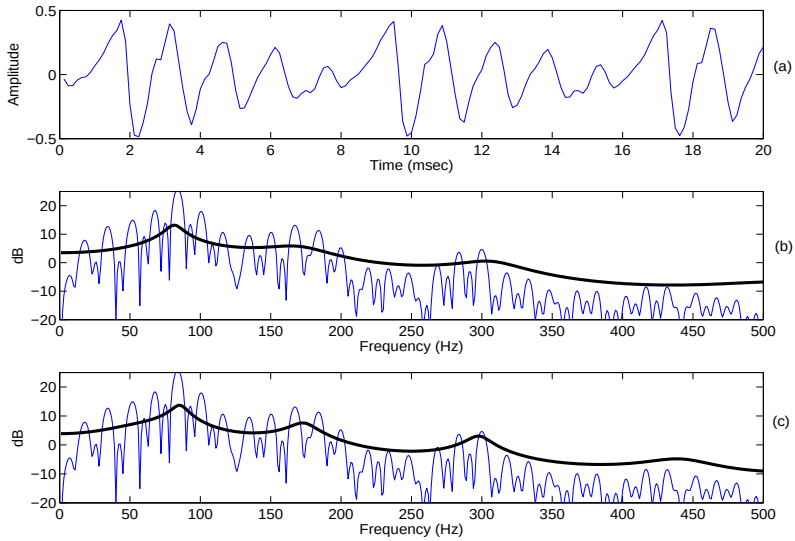


Fig. 4.5: Comparison of linear prediction analysis with different order of prediction. (a) A voiced region of speech. (b) Short-time DFT spectrum and LP log-spectrum for p=8. (c) Short-time spectrum and LP log-spectrum for p=14.

but linguistic information may be preserved. In contrast, a higher order (p=14) LP analysis captures both the gross and finer details of the envelope of the spectrum, thus preserving both linguistic and speaker-specific information. Hence for implementing a speaker independent LID system, lower order analysis is preferable. An 8th order LP analysis is used to capture the properties of the signal spectrum.

4.3.1 Performance Evaluation on Indian Language Database

The database used in this study consists of speech segments excised from continuous speech in broadcast TV news bulletins for Indian languages. It contains four different languages, namely, Hindi, Kannada, Tamil and Telugu. Training data for each language are obtained by concatenating speech data of different male and female speakers to make the model speaker-independent. For each language, speech data of duration 200 sec is used for training. Forty test utterances in each language are used for evaluating the performance.

The block diagram of the LID system used for this experimental study is shown in Fig. 4.6. The LID system consists of one AANN model per language. For identification, features extracted from the test utterance are given as input to all the AANN models. The output of each of the model is computed with its input to calculate the squared error for each frame and it is transformed into confidence value. A given test utterance is passed through each of the language models and average confidence value is obtained [40]. The model which gives the highest confidence value is hypothesized as the language of the test utterance.

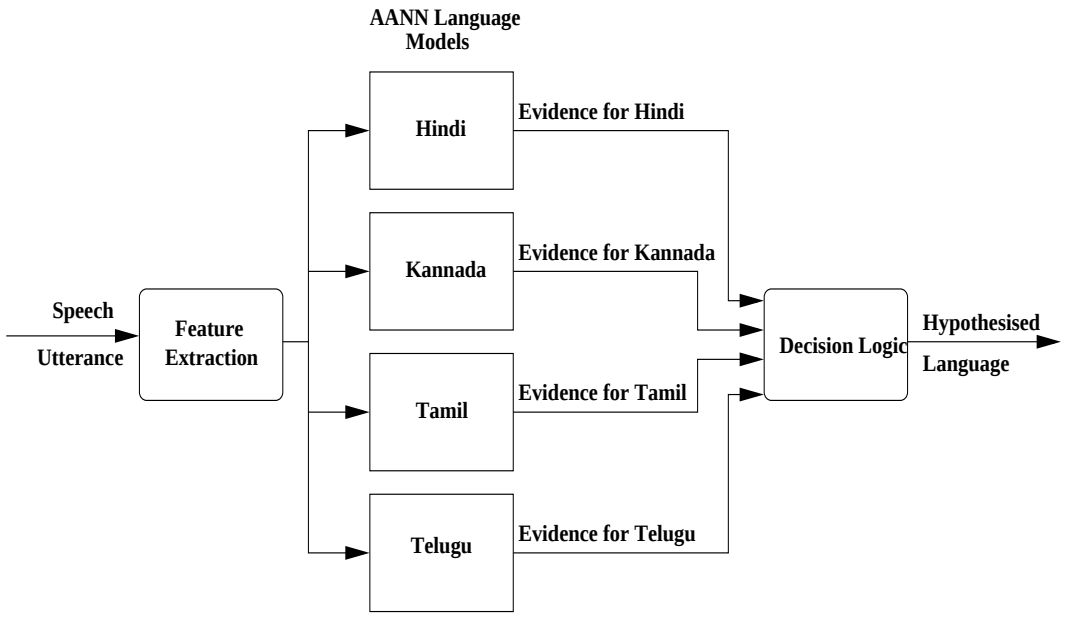


Fig. 4.6: Block diagram of the LID system used for Indian language database.

The performance of the AANN based LID system for varying test duration is given in Table 4.1. It is seen that the LID system gives better performance when the duration of the test utterance is of 10 sec duration. But even for test utterances of 5 sec and 1 sec duration, identification accuracy is reasonably good. This experimental study shows that the duration of speech required for training and testing is less for frame level features, compared to larger span features. The lower performance for Tamil and Kannada may be due to the large number of speakers present in the database,

Table 4.1: Performance of AANN based LID system for four languages. The entries from columns 2 to 4 represent the percentage of language identification.

Language	**Test duration**		
	10 sec	**5 sec**	**1 sec**
Hindi	100	100	95
Kannada	80	77.5	57.5
Tamil	77.5	75	60
Telugu	95	95	75

indicating that the framework used for this database may not be sufficient to capture the variability of large number of speakers within a language.

4.3.2 OGI Database - Issue of Speaker and Channel Variability

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 each in 10 languages [70]. The languages are: English, Farsi, French, German, Japanese, Korean, Mandarin, 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.

For collecting the data, each caller was asked a series of questions designed to elicit:

(a) Fixed vocabulary speech

(b) Domain-specific vocabulary speech

(c) Unrestricted vocabulary speech

The unrestricted vocabulary speech was obtained by asking callers to speak on any

Table 4.2: Comparison of OGI MLTS database and the Indian language database used in LID studies.

Sl. no.	Property	OGI MLTS database	Indian language database
1.	No. of languages	11	4
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

topic of their choice. Table 4.2 shows the comparison of OGI database and the Indian language database used in LID studies.

While working with the OGI database, the framework which worked well for Indian language database with limited number of speakers was found ineffective. It was observed that one AANN model per language is not sufficient to capture the variability due to large number of speakers and channels for the same language. Considering the influence of speaker characteristics on spectral features, it is necessary to reduce the effect of this variability. To address this issue we propose multiple models corresponding to different speakers for a single language.

4.3.3 Proposed Method for Handling Speaker and Channel Variability

As in the case of OGI database, when the number of speakers is large and the variability among speakers is high, it is difficult for a single AANN model to capture all the variability of a language. Since each call is collected over a unique telephone circuit, channel characteristics also vary. Considering the influence of speaker and channel characteristics on spectral features, it is necessary to capture this variability within a

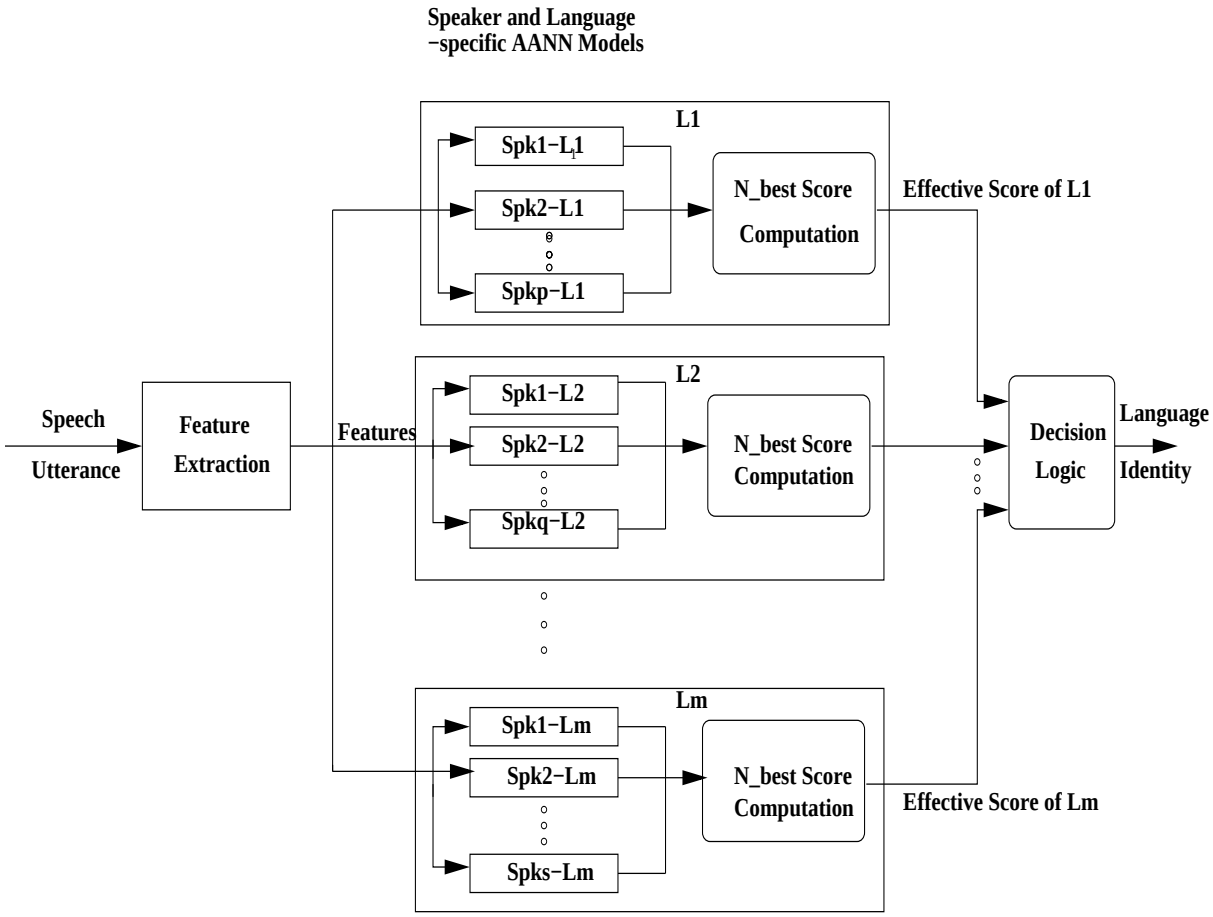


Fig. 4.7: Framework of the proposed LID system using frame level spectral features for OGI database.

language. This is done by building speaker-specific AANN models for each language. Training file for each speaker is used to build a separate AANN model, and these models are grouped according to the language as shown in Fig. 4.7. For identification, the spectral features derived from a test utterance are applied to all the models, and average scores are computed. Majority of the models in the genuine language are likely to give high scores, though a few models may give low scores due to poor match in speaker and channel characteristics. The other languages may have some models giving high scores due to similar speaker and channel characteristics, but it is unlikely to give large number of high scores.

Table 4.3: Performance of AANN based LID system using various scoring strategies for a six language subset of OGI database. Entries in columns 2 to 5 represent the identification accuracy in percentage for various scoring strategies.

Language	**Identification accuracy (in %)**			
	Average	**Maximum**	**Average+ Maximum**	**20-best**
English	95	15	55	95
French	100	45	70	90
Hindi	37	21	42	58
Japanese	12	41	30	41
Korean	6.25	50	50	44
Mandarin	16.7	25	42	17

To illustrate the effectiveness of the proposed method, initially we have chosen a subset of 6 languages from OGI database. The unrestricted vocabulary speech of each speaker with average duration of 45 sec alone was used for training as well as testing in this study. For all languages, 40 speech files corresponding to 40 different speakers were used for training. An average of 20 speech files from different speakers were tested using the proposed LID system. The training and testing speaker sets were different. Various scoring strategies, namely average (average of all the 40 model scores), maximum (maximum among the 40 model scores), average+maximum (sum of average and maximum scores) and N-best score (average of N-best scores), were experimented for identifying the language of the test utterance. The results are given in Table 4.3.

Best results are obtained through the N-best scoring. In N-best scoring, we consider the average of 20-best scores among the 40 model scores to compute the effective score of a particular language. This approach resulted in an overall identification accuracy of 61.5%. The effect of LP order on the identification accuracy in this study is shown in Fig. 4.8. The best results are obtained for an LP order of 10. Similar

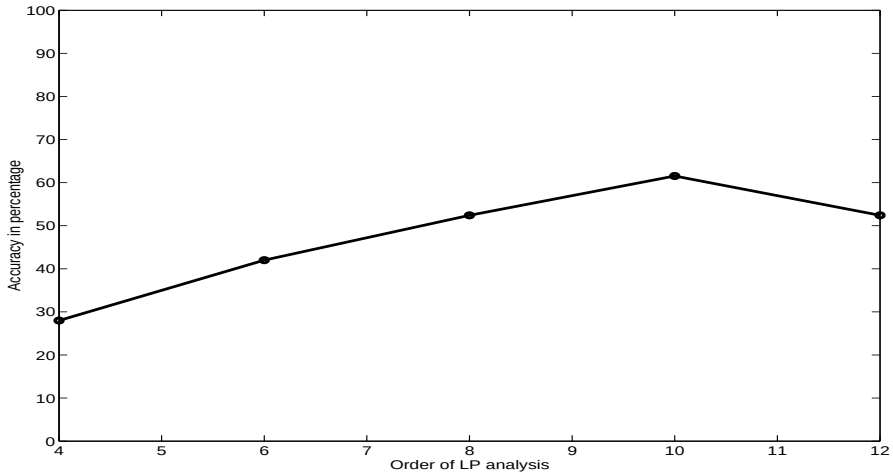


Fig. 4.8: Performance of frame level spectral-based LID system for varying LP order (using N-best scoring, evaluated on six language subset of OGI database)

study for all eleven languages in the OGI database resulted in an overall identification accuracy of 40%.

4.4 RESIDUAL FEATURES FOR LANGUAGE IDENTIFICATION

For developing models based on source features, our hypothesis is that the characteristics of a language may be present in the higher (>2) order relations among the samples of the speech data. Extraction and representation of these relations may be difficult due to nonlinear nature of the relations among the samples in the residual, although these can be perceived by the human listener. Nonlinear models may be required to extract features for language identification task from the residual. The AANN models may provide better choice compared to linear parametric models. When an AANN is presented with raw data such as samples of speech or LP residual, the explanation of the behavior of AANN is different from distribution capturing. This is because adjacent frames in the residual signal are not feature vectors. The frames of the residual may have similar features reflected in some higher order relations among samples of

the residual signal. Speech signal contains both the second order and higher order relations among samples. If the signal is directly given to the network, then, the system features may dominate the training of the network. If the second order relations are removed using the LP analysis, then the network is expected to capture the implicit higher order relations in the LP residual signal.

LP residual of the speech signal with sampling frequency of 8 kHz is computed using 12th order LP analysis to remove the second relations among the samples. Frames of 40 samples of the LP residual are used as input to the AANN. Successive blocks are formed with one sample shift. Each block of 40 samples is normalized to unit magnitude before applying as input to the network. The structure of the AANN model used for source features is *40L 48N 12N 48N 40L*. One model per language is created by training each network for 200 epochs using backpropagation algorithm. For testing, block of 40 samples normalized to unit magnitude is given as input to each AANN model. The confidence value is calculated as in the previous case.

The LP residual obtained after removal of the spectral features is shown in Fig. 4.3(b). It contains information about the strength and sequence of excitation. The model trained using the LP residual will be dominated by the strength of excitation. This strength of excitation is removed in the residual phase as shown in Fig. 4.3(d), and hence the model built using residual phase as input can capture the sequence information present in samples. Blocks of 40 samples of $cos\theta(n)$ values are given as input to train the AANN model. One model is trained for each language. Results of the experimental study given in Table 4.4 shows better performance for phase of the LP residual compared to LP residual. This is may be due to the absence of more speaker-specific strength excitation in phase of the LP residual.

Table 4.4: Performance of language identification system for four languages. The entries from columns 2 to 5 represent the percentage of identification accuracy for the spectral, source, phase and combined systems, respectively.

	Identification accuracy (in %)			
Language	**Spectral**	**Source**	**Phase**	**Combined**
Hindi	100	65	90	100
Kannada	80	52.5	52.5	87.5
Tamil	77.5	77.5	75	80
Telugu	95	72.5	100	100

4.5 COMBINING EVIDENCE FROM SPECTRAL AND RESIDUAL FEATURES

To make use of the complementary information present in spectral and residual features, the evidence may combined. In the training phase, three AANN models per language are created, based on spectral (WLPCC), source (LP residual) and phase (phase of LP residual) features. While testing, spectral, source and phase features extracted from the test utterance are given to the corresponding AANN models as shown in Fig. 4.9. The average confidence value is computed for the given test utterance. The scores of the spectral, source and phase models of each language are added to get the combined evidence. The language of the model which gives the highest evidence is hypothesized as the language of the test utterance. The performance of different features and their combination for 10 sec test utterances is given in Table 4.4. The performance is better when the scores of all the three models are combined by addition.

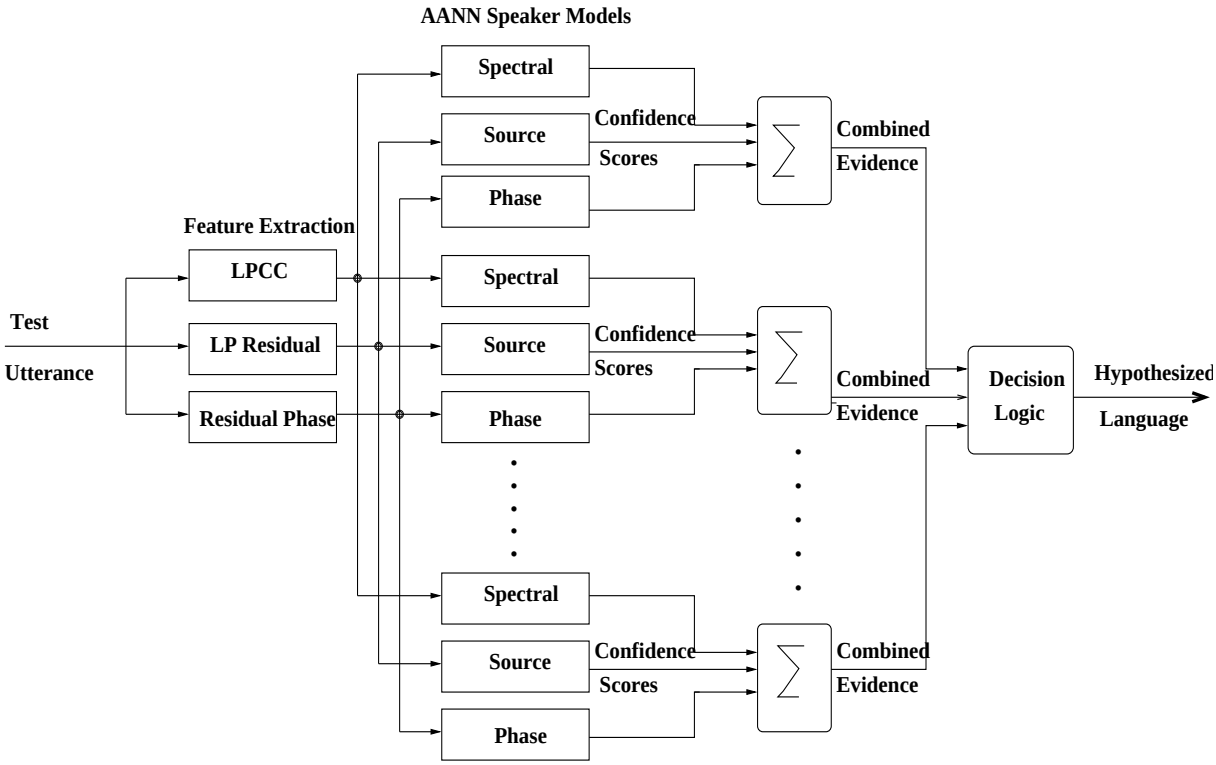


Fig. 4.9: Block diagram of LID system based on spectral, source and phase features.

4.6 SUMMARY AND CONCLUSIONS

In this chapter, we have explored various features derived from short frames of the speech signal for the purpose of language identification. To model the language, distribution of spectral features represented by WLPCCs were captured using AANN models. In order to capture the speaker and channel variability, use of multiple models for single language is suggested. Additional features explored in this study are excitation source characteristics represented by the LP residual and the phase of LP residual.

A detailed review of approaches to LID reveals that improved performance of majority of the LID systems are due to their use of higher level linguistic information. This is achieved by using large corpora of transcribed training speech which may not

be available in all languages required by a specific application. In this chapter, we have shown that it is possible to implement an LID system with reasonable performance, which requires only speech samples. By examining the various frame level features for LID, the following observations are made:

(a) Difference among languages are not obvious at the frame level. But difference in phoneme/syllable realizations and vocabulary are reflected in the frame level features.

(b) Frame level features are derived from short frames ($\leq$ 20msec) of speech, and therefore duration of speech required for training as well as testing is less.

(d) When speaker variability is more, one model per language is not enough to capture all the variability in a language. Multiple speaker models and N-best scoring help to reduce the effect of speaker variability.

Features represented at the frame level do not reflect the sound variations directly. The acoustic features represented at the level of phonemes/syllables, may help to capture the variation in realizations in a better way. In the next chapter, we study the syllable level spectral features, derived directly from the speech signal for language identification.

CHAPTER 5

SYLLABIC FEATURES FOR LANGUAGE IDENTIFICATION

5.1 INTRODUCTION

In the previous chapter, frame level features were explored for language identification. Features at frame level corresponds to smaller span of speech, and hence do not directly represent variations of sound units among languages. To represent variations of sound units, features need to be derived from larger spans of speech signal corresponding to these units. In this chapter, an approach is proposed for using variations in the realization of syllables among languages, for the purpose of language identification (LID). To extract and model the syllable level features, the following issues need to be addressed:

(a) Identifying the regions corresponding to syllables automatically from continuous speech

(b) Deriving fixed dimensional representation from varying length pattern of syllables

(c) Speaker variability within each language

(d) Limited syllable samples for training and testing

In the remaining part of the chapter, we will discuss these issues, and propose methods for addressing them.

This chapter is organized as follows: Section 5.2 discusses the choice of consonant-vowel (CV) units for syllable-based language identification studies. Section 5.3 describes how the CV units can be located in continuous speech using the knowledge of

vowel onset points (VOP). This section also describes the representation of the CV units. Section 5.4 discusses the need for reducing the effect of speaker and channel variability, and also outlines the framework of the proposed LID system. The results of language identification experiments using CV-based features are also discussed in this section. Section 5.5 gives a summary of this study.

5.2 CHOICE OF CONSONANT-VOWEL AS BASIC UNIT

Each language has a finite set of phonemes and syllables. Since the vocal apparatus used in the production of languages is universal, there is a significant overlap in these sets. There may be differences in the realization of the same unit in different languages. We hypothesize that the acoustic realization of the same syllable may differ among languages due to the difference in pronunciation and vocabulary. Since each sound corresponds to a unique articulatory configuration of the vocal tract, it is possible to represent these acoustic-phonetic variations with the help of spectral features. Syllables are chosen as the basic unit for representing this variations as they are context-dependent units. Syllables typically include multiple phonemes, and therefore may capture some significant co-articulation effects, that may be language-specific.

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. In this study, we use an alternative approach for LID, using the variations at the syllable level.

Syllables in general can be of the form C^mVC^n, $m, n \geq 0$, 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 [71] 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 syllable in the world languages [72]. The articulatory movements for the vowel in CV starts at the same time as the movements for the initial consonant [73]. The characteristics of the consonant part is influenced by the succeeding vowel, and therefore CV units capture significant coarticulation effects.

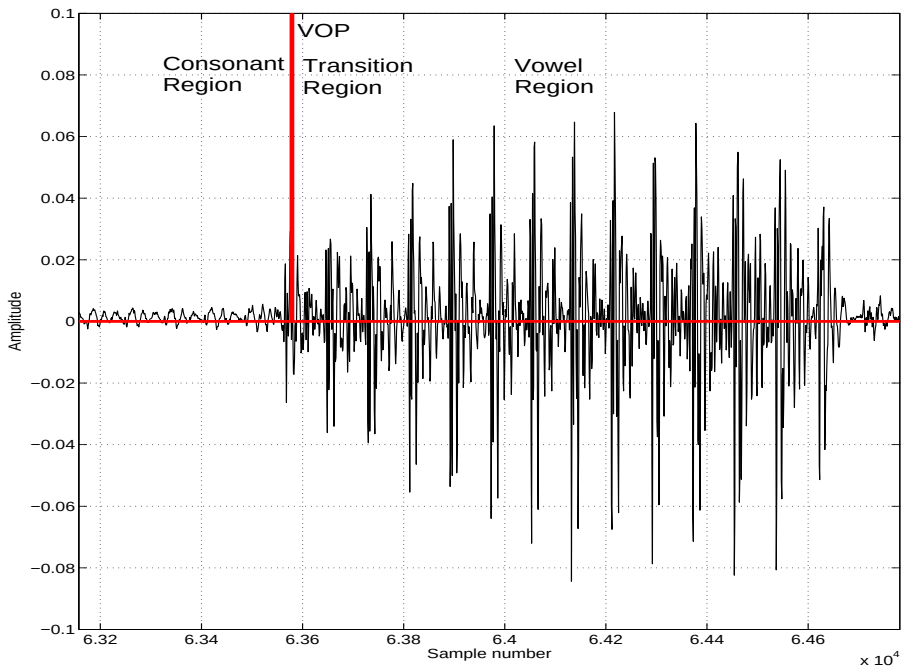


Fig. 5.1: VOP as an anchor point for automatic spotting of CV units in continuous speech.

In CV-based language identification, the first step is to identify CV regions directly from speech signal. The patterns of 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 Fig. 5.1. 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) [74–76]. Now we discuss a method for identifying the regions

corresponding to CV units from continuous speech.

5.3 EXTRACTING CV UNITS FROM CONTINUOUS SPEECH

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.

5.3.1 Acoustic Cue for Detection of Vowel Onset Point

Vowel onset points are the instants at which the consonant part ends and the vowel part starts in a CV type of syllable. It is an important event in speech production, which may be described in terms of changes in the vocal tract system and excitation source characteristics [75]. Fig. 5.2 shows one such acoustic cue, which enables the detection of VOP from continuous speech. It shows the speech waveform corresponding to a syllable, the LP residual and Hilbert envelope of the LP residual. As shown in Fig. 5.2(c), Hilbert envelope of the LP residual represents the strength of excitation. The strength of excitation shows a significant change at the transition from consonant to vowel, and hence can be used as an acoustic cue for detecting the VOP event. The strength of excitation for voiced sound is generally higher compared to that of unvoiced sound. In particular, the strength of excitation for vowels is higher compared to the strength of voiced consonants. Therefore the places with significant change in strength of excitation gives the evidence for the detection of VOPs.

5.3.2 Residual Based Approach for Locating VOP

In this study, a technique based on the excitation source information is used for automatically locating VOPs from continuous speech [77]. This VOP evidence is obtained

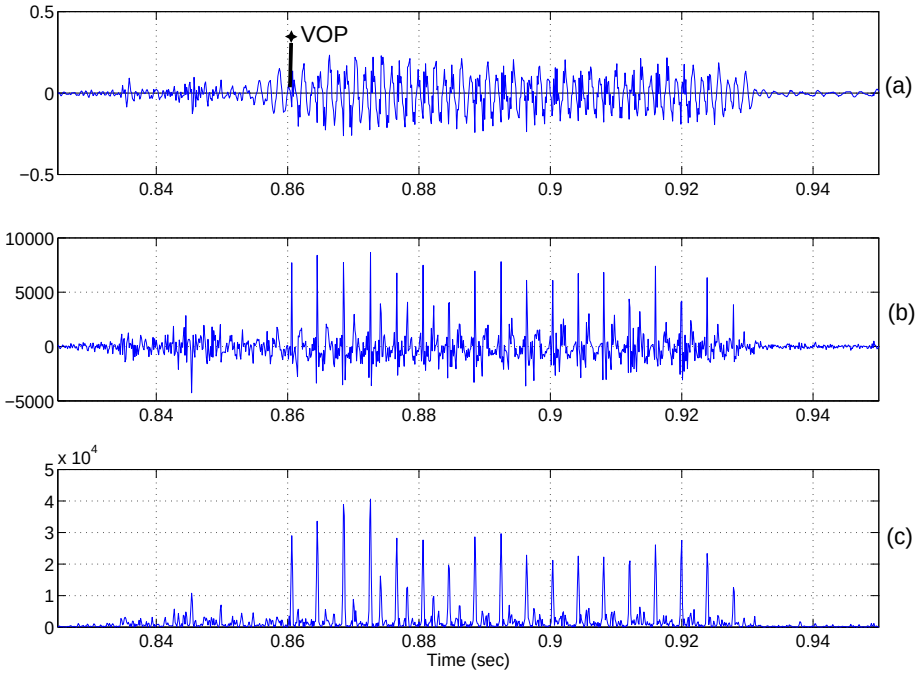


Fig. 5.2: (a) Speech waveform for a syllable with manual marked VOP, (b) LP residual and (c) Hilbert envelope of LP residual.

from the Hilbert envelope of the LP residual, by convolving it with the impulse response of a Gabor filter. A Gabor filter with spatial spread $\sigma = 100$, angular frequency of the sinusoidal component $\omega = 0.0114$ and filter length $n = 800$ is used. The Gabor filter is shown in Fig. 5.3. The peaks in the VOP evidence plot are located using a peak picking algorithm. A few spurious peaks can be eliminated as shown in Fig. 5.4, using the fact that there exist a negative region in VOP evidence plot between two true successive VOPs due to the presence of vowel.

The steps for the detection of VOPs in speech signal is summarized in Table 5.1. Once the VOPs are detected automatically from continuous speech, the next step is to extract features for representing the CV units.

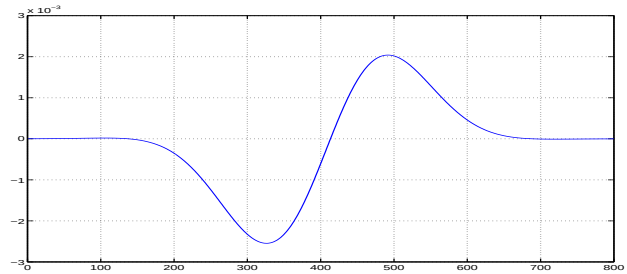


Fig. 5.3: A Gabor filter with $\sigma = 100$, $\omega = 0.0114$ and $n = 800$.

Table 5.1: Steps for detection of VOPs.

1.	Preemphasize input speech.
2.	Compute LP residual using 10th order LP analysis.
3.	Compute Hilbert envelope of the LP residual.
4.	Obtain the VOP evidence plot from the Hilbert envelope by convolving it with the Gabor filter.
5.	Identify peaks in the VOP evidence plot.
6.	Eliminate the spurious peaks based on the following conditions. For two consecutive peaks, eliminate the first peak if there is - no negative region in the VOP evidence plot - distance of separation less than 50 msec
7.	Hypothesize the remaining peaks as the VOPs.

5.3.3 Representation of CV Units

The important characteristics of a CV utterance lie in the region around the VOP. It was shown that a region of fixed duration around VOP, typically 25 msec to the left and 40 msec to the right of VOP, is sufficient to represent the characteristics of a CV unit [76]. For feature extraction, 10 overlapping frames around the VOP, with 20 msec frame size and 5 msec frame shift is considered. Feature vector corresponding to each frame consists of 13 Mel-frequency cepstral coefficients (MFCC), 13 delta and 13 acceleration coefficients leading to a dimension of 39. Such feature vectors

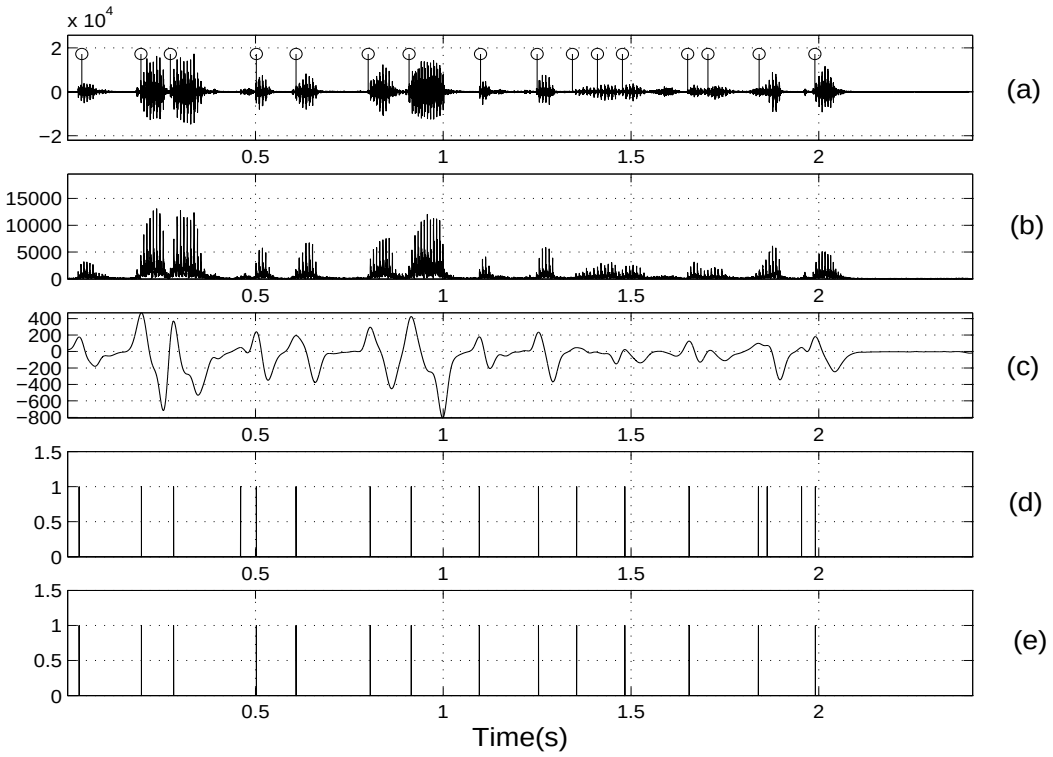


Fig. 5.4: (a) Speech waveform with manual marked VOPs, (b) Hilbert envelope of LP residual, (c) VOP evidence plot, (d) Output of peak picking algorithm and (e) Hypothesized VOP after eliminating few spurious peaks.

corresponding to 10 consecutive frames constitute 390 dimensional feature vector to represent a CV utterance. The algorithm used for deriving MFCCs is described in Appendix C. In this approach, all syllables are treated as CV type, and if they are of other type, only the CV part is processed for feature extraction. This enables us to have a fixed pattern representation irrespective of varying duration. In the next section, we discuss the modeling of CV-based features to capture the characteristics of a language.

5.4 MODELING FEATURES OF CV UNITS FOR LANGUAGE IDENTIFICATION

The spectral features derived from a CV unit contain characteristics of the unit, speaker, language, environment and channel. To enhance the language-specific part in the spectral features, the effect due to other factors should be reduced. One way is to train separate models with features derived from several occurrences of similar CV units corresponding to different speakers and channels. But this require several hours of speech for training. Unfortunately most of the time, as in the case of OGI database, there may not be sufficient number of occurrences of the same unit in a limited training data to capture the variability. To deal with the issues of limited data and speaker variability, we propose multiple speaker-specific AANN models as shown in Fig. 5.5, which is similar to the framework used for frame level spectral features. While testing, suitable scoring techniques are used to reduce the influence of speaker and channel characteristics, thereby enhancing the evidence of the language.

The CV based features derived from each training speech (corresponding to a unique speaker) is used to train a separate AANN model. These multiple speaker-specific models in each language help to capture the speaker variabilities within a language. During identification, each language is scored using the framework shown in Fig. 5.5. The CV based features derived from test utterance are tested against all the models which are grouped in terms of languages. Due to limited duration of speech available for training, models get trained only for those CV units present in the training speech. Therefore, instead of considering the average score of a model for the entire test utterance as in case of frame level features, each CV unit in the test utterance is scored separately. The output of each model for each feature vector is computed, and the squared error is calculated. The error is transformed to a confidence value, which indicate the similarity in terms of identity of unit, speaker, language and channel characteristics. Considering the scores of all the models in a language for calculating effective score may not be a good idea, as the models are trained only for those units

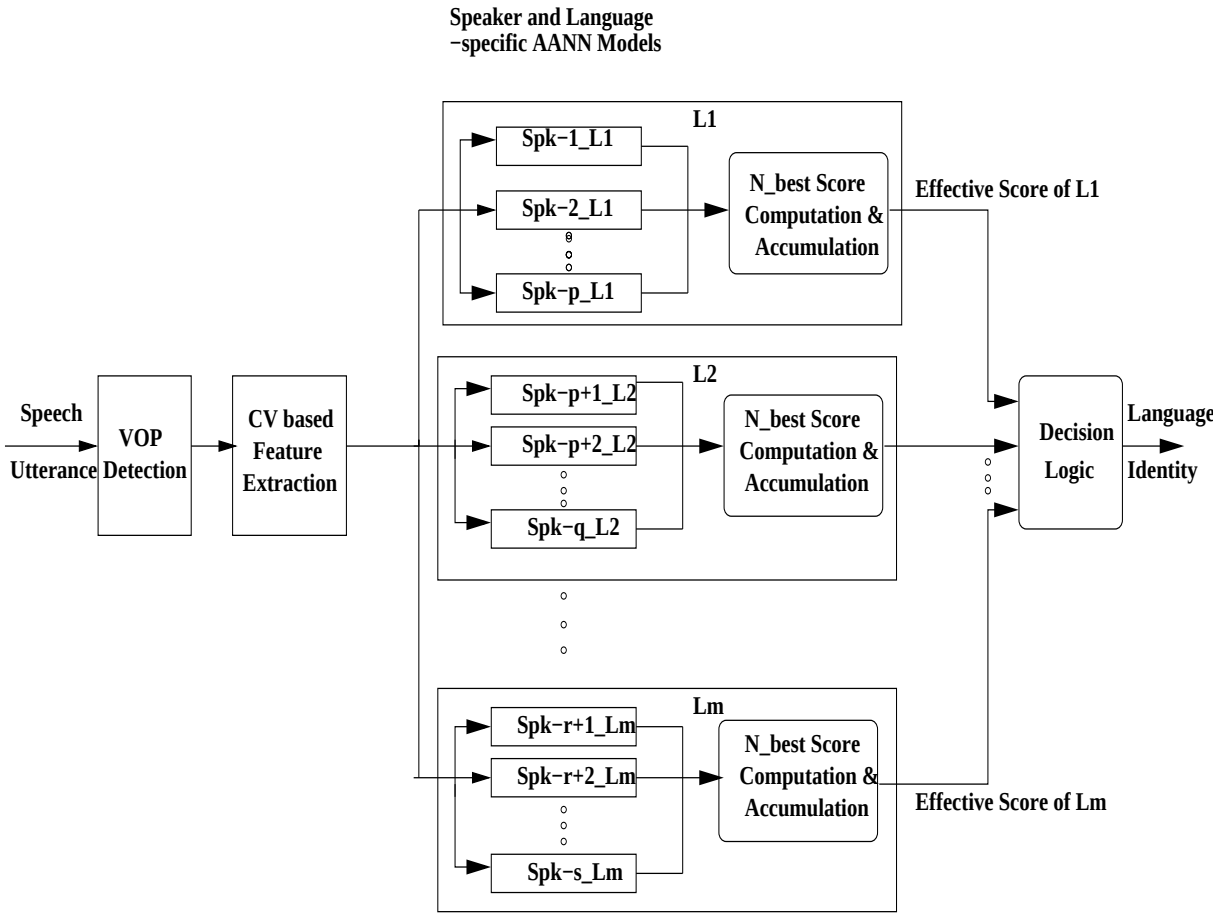


Fig. 5.5: Framework of the proposed CV based LID system.

present in training speech. Considering more than one model score may help to reduce the effect of speaker variability. Therefore for each CV unit in the test utterance, the N-best model scores are used for computing the effective score. The effective scores for all the CV units in the test utterance are added resulting in a high score for the genuine language.

5.4.1 Performance Evaluation on OGI Database

As described in Section 5.3, VOPs are detected first, and the MFCC features are extracted from the CV regions anchored around the VOPs in speech. To capture the

distribution of 390-dimensional CV based feature vectors, AANN models with structure *390L 580N 40N 580N 390L* are used, where L and N represent the neuron with linear and nonlinear activation functions, respectively. Separate models are trained for 40 different speakers in each language. The top five scores among the 40 model scores in each language group is considered for each CV unit in the test utterance. Scores for all the CV units present in the test utterance are accumulated and the language with the highest score is hypothesized as the language of the test utterance.

Table 5.2: Performance of CV based LID system for eleven languages in OGI database.The entries from columns 2 to 4 represent the identification accuracy (in %) considering cases where model of genuine language secured (i) first rank (k=1) (ii) first or second rank (k=2) (iii) first, second or third rank (k=3), respectively.

Language	**k-best performance**		
	k=1	**k=2**	**k=3**
En	90	90	95
Fa	65	85	90
Fr	70	80	90
Ge	50	75	92
Hi	47	65	82
Ja	82	94	94
Ko	75	81	88
Ma	24	47	59
Sp	47	88	88
Ta	60	80	80
Vi	38	63	69

The performance of the CV based LID system considering all the eleven languages in OGI database, namely, English (En), Farsi (Fa), French (Fr), German (Ge), Hindi (Hi), Japanese (Ja), Korean (Ko), Mandarin (Ma), Spanish (Sp), Tamil (Ta) and Vietnamese (Vi), are given in Table 5.2. Identification accuracy is good in the case of English, French, Japanese and Korean. Poor performance for Mandarin and Vietnamese may be due to the lesser syllable examples and variability in channel char-

acteristics. The overall identification accuracy for 1-best, 2-best and 3-best cases are 59%, 77% and 84%, respectively. The improvement in identification accuracy while considering 2-best and 3-best cases indicate scope for further enhancement. The use of complementary features may be helpful to improve the performance.

5.5 SUMMARY AND CONCLUSIONS

The effectiveness of CV based features for LID was demonstrated in this study. The LID system described here utilizes the language-specific variations of most frequently occurring CV type syllables. Regions corresponding to CV units were automatically identified with the help of VOPs, eliminating the need for any transcribed corpora. Since CVs are the most frequently occurring units in most of the languages, the system is expected to work well for short durations of training as well as test speech. The system addresses the issue of speaker variability by building speaker and language-specific AANN models. By considering top N scores while testing, the influence of speaker variability is reduced. The proposed system does not use any prior language-specific knowledge, therefore it is easy to introduce a new language into the system. By examining the CV based features for language identification, the following observations are made:

(a) Spectral features corresponding to CV units are useful for representing their variations, and it is effective for LID.

(b) Syllables being context-dependent units, syllable level features are more effective for language identification compared to the frame level features explored in the previous chapter.

(c) Multiple speaker dependent models and suitable scoring technique help to address the issues of speaker variability and limited amount of training data.

(d) The increase in performance while considering the 2-best, and 3-best cases suggest the use of other features, for further resolving the top ranked languages.

The prosodic features, being an additional source of information, may provide some complementary evidence to spectral-based LID system. In the next chapter, we discuss the use of prosodic and phonotactic features represented at multisyllabic level, for language identification.

CHAPTER 6

MULTISYLLABIC FEATURES FOR LANGUAGE IDENTIFICATION

6.1 INTRODUCTION

The studies in Chapters 4 and 5 have been on the use of features represented at frame level and syllable level for identifying language. In Chapter 4, spectral and residual features were explored to find the possible presence of language-specific characteristics. In Chapter 5, the acoustic variations in the realization of syllables, represented using spectral features, were used for language identification. In this chapter, we examine the usefulness of phonotactic and prosodic features represented at multisyllabic level for language identification (LID).

Human beings apply some constraints on the sequence of sound units while producing speech. These are characteristics 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 [78]. These properties are also capable of conveying many more lexical and nonlexical 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 [4, 15]. But their performance degrade due to noise and channel mismatch. Prosodic features derived from pitch contour, amplitude contour and duration are relatively less affected by channel variations and noise. Though the systems based on spectral features outperform the prosody-based LID systems, their combined performance may provide the needed robustness.

This chapter is organized as follows: In Section 6.2, details of some perceptual experiments are described to illustrate the importance of prosody and phonotactics. Section 6.3 and Section 6.4 describe the phonotactic and the prosodic differences among languages, respectively. In Section 6.5, a study on three Indian languages regarding the possible use of phonotactics and prosody for LID is discussed. In Section 6.6, we propose an approach for deriving prosodic features directly from the speech signal. Section 6.7 describes the language-prosodic features derived using the proposed approach. The experimental study conducted on OGI multi-language telephone-based speech (MLTS) corpus to demonstrate the effectiveness of the derived prosodic features are discussed in Section 6.8. Final section gives a summary of the study on multisyllabic features for language identification.

6.2 HUMAN LANGUAGE IDENTIFICATION - DETAILS OF PERCEPTION EXPERIMENTS

Prosody has a great deal to offer for effective human language identification. Many human perception studies have demonstrated that language identification is possible even when segmental/spectral information is reduced or degenerated. The results of experiments on human language identification confirm that prosodic information are used for language identification when the intelligibility of sound units are less. Perception studies were conducted using signal obtained after spectral envelope removal (SER)

[8, 79]. In SER signal, the spectral envelope is removed using inverse filtering. Human beings could still identify languages fairly successfully from SER signals, indicating the capability of humans to identify languages using nonspectral information.

Ramus and Mehler [7] examined the ability of native French speakers to discriminate between English and Japanese-like stimuli, where the stimuli were constructed by resynthesizing original utterances with all consonants and vowels replaced by archetypes. The resulting stimuli were designed to preserve (1) intonation (all segments replaced by /a/, original pitch contour preserved), (2) rhythm (all consonants by /s/ and all vowels by /a/, flat intonation contour), (3) both rhythm and intonation (all consonants by /s/ and all vowels by /a/, original pitch contour preserved) and (4) rhythm, intonation and broad phonotactics (by replacing all fricatives with /s/, all stop consonants with /t/, all liquids with /l/, all nasals with /n/, all glides with /j/, and all vowels with /a/). They found good discrimination in all cases except (1), which has an interesting implication that suprasegmental cues are sufficient to discriminate these languages. The differences among languages in terms of phonotactics and prosody are discussed in the next two sections.

6.3 PHONOTACTIC DIFFERENCES AMONG LANGUAGES

Phonotactics deals with restrictions in a language on the permissible combinations of phonemes. It defines the permissible syllable structure, consonant clusters and vowel sequences by means of phonotactic constraints. Each language has its own rules on forming sequences, based on the permissible sound sequences in the spoken language. Certain phoneme/syllable clusters which are very common in a particular language may not be legal in some other language [80]. For example, in Japanese, consonant clusters like /st/ are not allowed, but they are common in English. In Japanese, a liquid (/r/) can never follow a stop consonant (/p/, /b/, /k/), unlike in English or French. It is safe to say that no two languages in the world have exactly the same rules about this. Many languages, such as Japanese, never put more than one consonant

before a vowel. On the other end, there are many languages which allow very elaborate consonant clusters. Consonant clusters can be reliable pointers for identifying language of a particular text [1].

Among the language-specific features, the phonotactic constraints are shown to be the most powerful feature for LID [2,15]. LID researchers make use of this property by recognizing the phonemes using language independent/dependent front-end phoneme recognizer. These phoneme labels are then used for building separate statistical models for each of the language to model the phonotactic constraints. Explicit LID systems such as PRLM, PPLM, and PPR described in Chapter 2 make use of the phonotactics for identifying language.

6.4 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 [78, 81, 82]. But in spite 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 nontonal, 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 [83]. 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.

6.4.1 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 two languages, namely Farsi and Mandarin in Fig. 6.1. It can be observed that in general Mandarin has large variations in F_0 values compared to Farsi, in spite of the variations in speaker characteristics.

In this study, our goal is to represent these pitch contour with suitable features to bring 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 [84].

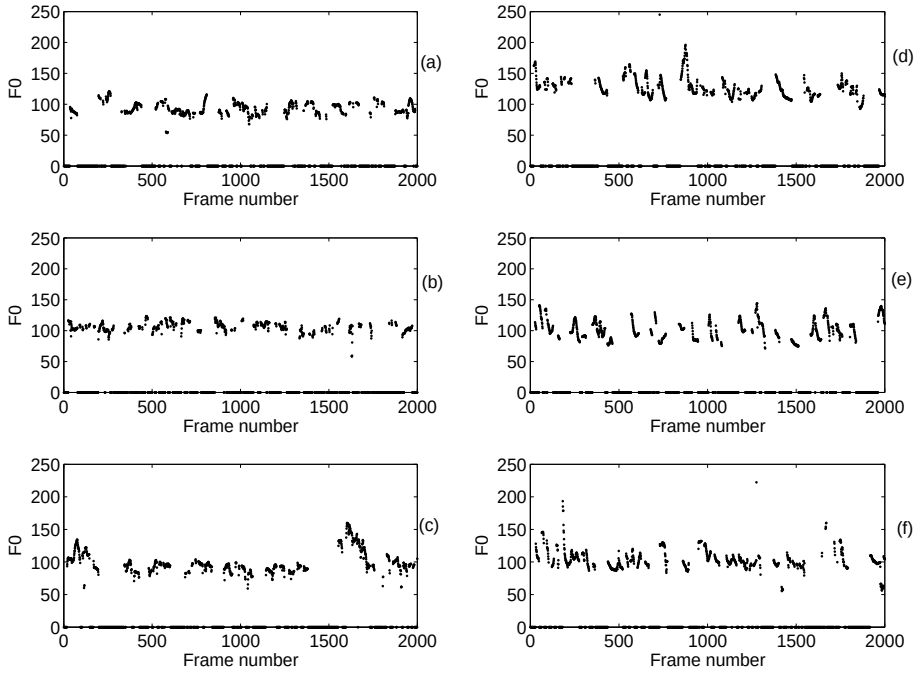


Fig. 6.1: Variation in dynamics of F_0 contour for utterances in Farsi and Mandarin, spoken by three male speakers each. (a), (b) and (c) correspond to Farsi (d), (e) and (f) correspond to Mandarin utterances (taken from OGI MLTS database).

6.4.2 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 [7]. According to the frame/content theory of speech production [85], 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 [86], 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 [87]. Both these measures will be directly influenced by segmental inventory and the phonotactic regularities of a specific language.

6.4.3 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:

(a) 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.

(b) 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.

(c) 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 [67].

6.5 PHONOTACTIC AND PROSODIC FEATURES FOR LID - A STUDY ON THREE INDIAN LANGUAGES

As preliminary step toward understanding the significance of prosody and phonotactics for language identification, a study was conducted on a database of three Indian languages. This database is chosen due to the availability of manual segmentation and syllable level transcription. Features are derived using this segmentation and transcription information. This study is conducted for understanding and verifying the effectiveness of phonotactic and prosodic features for LID.

6.5.1 Description of Continuous Speech Corpus

Speech corpus consisting of recording of television broadcast news bulletins for three Indian languages namely, Tamil, Telugu and Hindi are used in this study. It contains 19 Hindi, 20 Telugu and 33 Tamil broadcast news bulletins, where each bulletin (session) contains 10 to 15 minutes of speech from a single (male or female) speaker. The bulletins in the corpus are segmented into phrases, each approximately of 3 sec duration, and labeled manually. This is further parsed into syllables by human experts using Indian language transliteration (ITRANS) code [88].

6.5.2 Feature Representation

The characteristics explored in this study are the following:

(a) Phonotactics

(b) Broad phonotactics

(c) Prosody along with phonotactics

(d) Prosody alone

The role of phonotactics for LID is studied using the syllable transcription available in the database. The identity of each syllable is represented using a unique code. For syllable coding, each syllable is assumed to have four constituents (consonants and vowels), and each constituent is given a unique numerical code (ranging from 10 to 69), so that each syllable is represented by a distinct four digit code [89]. Word boundary is represented by a unique code which indicates the absence of syllable. In this study, the feature vectors are formed by concatenating the identity code of three consecutive syllables, for representing the phonotactics.

Studies have shown that phonotactics represented in terms of broad category of phonemes (such as vowels, nasals, fricatives, semivowels, stop consonants etc.) is effective for representing the phonotactic differences among languages [6, 7]. Syllable constituents are labeled in terms of eight broad phonetic categories namely vowels, nasals, semivowels, fricatives, unvoiced unaspirated stop, unvoiced aspirated stop, voiced unaspirated stop and voiced aspirated stop, using the transcription information available in the database. The stop consonants are classified as aspirated and unaspirated, based on the manner of articulation. The syllable identity is represented by this broad phonetic labels of the constituents, and the feature vector is obtained by concatenating the features of three consecutive syllables.

Later phonotactic features are clubbed with some prosodic features to see improvement due to the addition of prosodic knowledge. Rhythmic features of languages are represented by the structure of syllable and its duration. Syllable structure is represented by the number of constituents N_t, number of constituents before and after vowel N_{c1}, N_{c2}, respectively [89]. Since it is difficult to compute the durations of syllable constituents (consonants and vowels) separately, duration of the syllable is used

instead. As rhythm is perceived due to succession of syllables, a syllable in isolation can not represent the rhythmic characteristics. In this study, it is approximated to a sequence of three consecutive syllables, and represented by concatenating the features derived from them. In natural speech, the prosody of a syllable also depends on its position with respect to the word and phrase [90]. The positional details of the syllable with reference to the particular phrase and word [89] as used in this study are the following:

Syllable position in the phrase:

(a) Position from starting of the phrase

(b) Position from ending of the phrase

(c) Total number of syllables in the phrase

Syllable position in the word:

(a) Position from starting of the word

(b) Position from ending of the word

(c) Total number of syllables in the word

The intonation characteristics are represented by the following:

(a) The change ΔF_0 of pitch within the syllable

(b) The distance of F_0 peak with respect to the onset of syllable

(c) The mean pitch F_{0_μ} of the syllable

In order to account for the local variations of pitch, the above measures of the preceding and following syllable are concatenated for representing intonation. Table 6.1 summarizes various features used in this experimental study.

Table 6.1: Summary of phonotactic and prosodic features used in the experimental study on three Indian languages.

Characteristics	**Variable**	**# Parameters**	**Representation**	**Dimension**
Phonotactics	Syllable identity	4	trisyllabic	12
Broad phonotactics	Category of syllable	4	trisyllabic	12
Phonotactics and prosody	Syllable identity	4	trisyllabic	39
	Syllable structure	3	trisyllabic	
	Intonation	3	trisyllabic	
	Duration	1	trisyllabic	
	Positional	6	monosyllabic	
Prosody	Syllable structure	3	trisyllabic	27
	Intonation	3	trisyllabic	
	Duration	1	trisyllabic	
	Positional	6	monosyllabic	

6.5.3 Neural Network Classifiers for Language Identification

Human beings acquire the prosodic and phonotactic knowledge of a language over a period of time. The process by which this happens can not be explained or formulated in terms of rules. In this study, we propose the use of multilayer feedforward neural network (MLFFNN) classifiers for LID, using the phonotactic and prosodic features. The MLFFNN classifier as shown in Fig. 6.2, is trained using phonotactic and/or prosodic features derived from training speech of all the three languages. When features derived from the test utterance are applied at the input of the MLFFNN classifier, its output indicate evidence of different languages.

6.5.4 Phonotactics

The identity of three consecutive syllables is used to capture the phonotactic regularities of languages. The syllable identity obtained from the transcription information

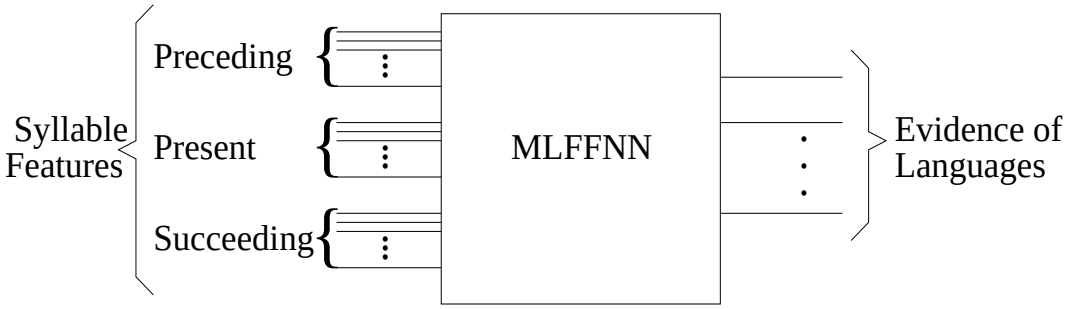


Fig. 6.2: Neural network classifier for language identification using phonotactic and prosodic features.

Table 6.2: Performance of LID system based on phonotactic features. The entries from columns 2 to 5 represent the percentage of the identification accuracy (in %).

	Test duration=20 syllables		**Test duration=50 syllables**	
Language	**Rank based**	**Accumulation**	**Rank based**	**Accumulation**
Tamil	99	98.5	100	100
Telugu	72.4	85.8	83.8	95.6
Hindi	96.2	98.5	100	100

is uniquely represented using four codes [89]. These feature vectors (normalized to -1 to +1) are used for training the MLFFNN classifier as shown in Fig. 6.2. The output corresponding to the language of the training speech is set to +1 and other outputs to -1. The network is trained using backpropagation algorithm. For each language, approximately 40,000 examples of syllables are used for training the classifier. An average of 600 test utterances (each having length of 20 syllables), and 250 test utterances (each having length of 50 syllables) are used for performance evaluation.

The performance of the neural network based classifier using phonotactic features is given in Table 6.2. When a feature vector (obtained from three consecutive syllables) derived from the test utterance is applied to the classifier, output of the classifier indicate evidence of different languages. Methods such as accumulation of evidence and

majority voting (rank-based) have been used for computing the evidence of languages. In rank-based method, number of first ranks obtained for each language is counted for all the feature vectors of the test utterance. The language having maximum first ranks is decided as the winner. In the second method, evidence (classifier output) are accumulated for all the feature vectors of the test utterance. The language having highest accumulated score is hypothesized as the language of the test utterance. In this study, identification accuracy is better for evidence accumulation.

6.5.5 Broad Phonotactics

The syllable constituents are represented using their broad phonetic category instead of their exact identity. The categories used in this study are vowels, nasals, semivowels, fricatives, unvoiced unaspirated stop, unvoiced aspirated stop, voiced unaspirated stop and voiced aspirated stop. The broad phonotactic features are obtained by coding the syllables in terms of its broad classification, and it is modeled using the MLFFNN classifier. The performance of the LID system based on broad phonotactic features is given in Table 6.3. The good identification accuracy obtained from the broad phonotactic features suggests that labeling of syllables in terms of broad categories is sufficient to represent the phonotactic constraints of languages. This will eliminate the need for proper transcribed speech for training as well as for testing of the neural network. The broad phoneme categories used for labeling the syllable constituents should be optimized for languages included in the identification task. The stop consonants are categorized in terms of manner of articulation, namely aspirated and unaspirated. As given in Table 6.3, due to the absence of aspirated sounds in Tamil, identification accuracy is better for Tamil compared to other two languages.

6.5.6 Phonotactics and Prosody

Prosodic features along with phonotactic features are used to train the MLFFNN based classifier as shown in 6.3. Rhythm is represented by the structure of syllable and its

Table 6.3: Performance of LID system on broad phonotactic features. The entries from columns 2 to 5 represent the identification accuracy (in %).

	Test duration=20 syllables		**Test duration=50 syllables**	
Language	**Rank based**	**Accumulation**	**Rank based**	**Accumulation**
Tamil	99.8	99.25	100	100
Telugu	47.65	82.7	49.86	92.6
Hindi	29.6	88.3	81.71	96.34

duration. Intonation is represented by the average pitch of syllable, range of pitch and location of maximum pitch. Since the features corresponding to a single syllable alone are not sufficient for representing the prosodic pattern, sequence of three syllables is taken as the basic unit. The feature vector is obtained by including features from the preceding and succeeding syllables along with the features of the present syllable.

For each language, features derived from approximately 25,000 syllables are used for training the classifier. The results in Table 6.4 show that by including the prosodic features along with the phonotactic features, the classifier shows an improvement in performance, even though the training is done with fewer syllable examples. It can be observed from Table 6.4 that the identification accuracy of Hindi has improved after the inclusion of prosodic features. Hindi belongs to the Aryan group of languages where as Tamil and Telugu are Dravidian languages. Therefore prosodic features can be significant while discriminating languages belong to different classes.

6.5.7 Prosody

In this case, prosodic features alone are used to train a MLFFNN classifier. The evaluation results in Table 6.5 indicate the potential of prosodic features for language identification. But in implicit LID systems, it is important to extract features directly from the speech signal without using manual segmentation and transcription. Therefore as a next step, a subset of prosodic features obtained without any transcription

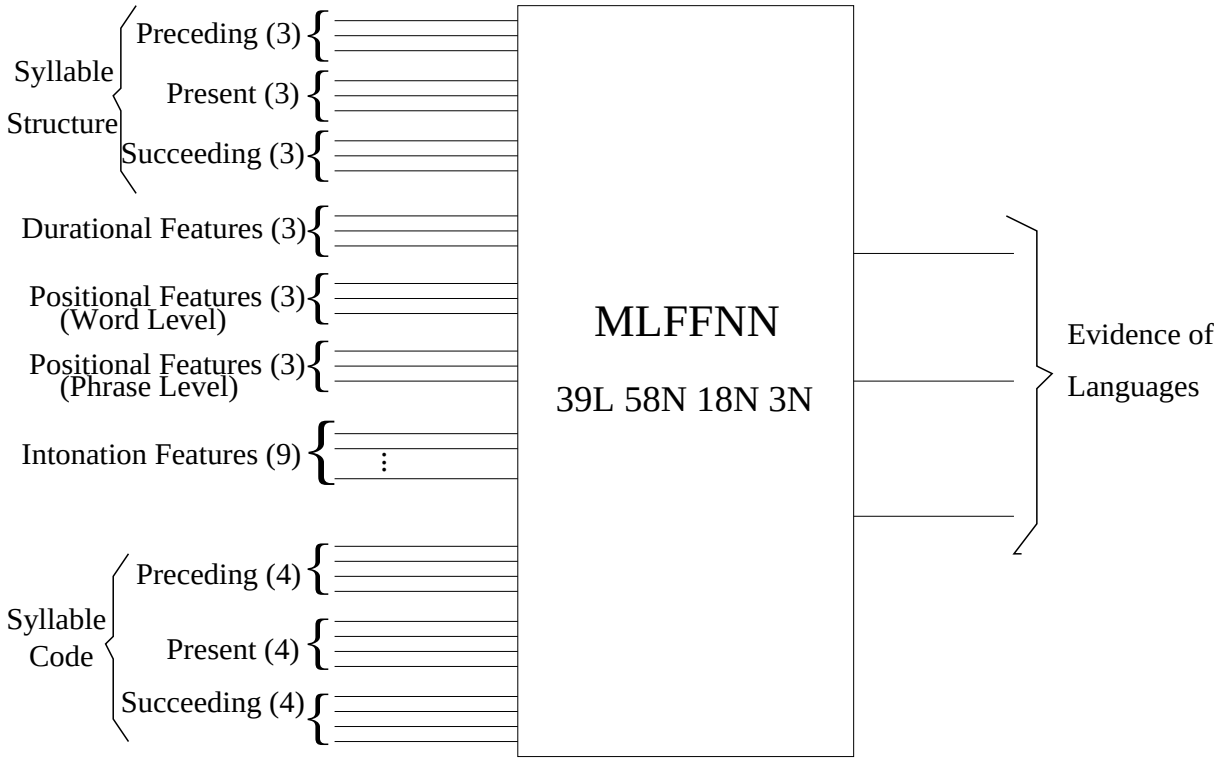


Fig. 6.3: The phonotactic and prosodic features modeled using MLFFNN classifier for language identification.

information are used for LID, and the results are given in Table 6.6. The encouraging results of this system motivated us to device an approach for deriving prosodic features automatically from the speech signal.

6.6 PROSODIC FEATURE EXTRACTION FROM SPEECH SIGNAL

Approaches for extraction of prosodic features can be broadly categorized based on the use of automatic speech recognizer (ASR) as (1) ASR-based approach and (2) ASR-free approach. The ASR-based approach uses segment boundaries obtained from ASR, for extracting the prosodic features [54]. But for applications like language and speaker recognition, the use of ASR may not be needed. In the second approach, inflection points and start or end of voicing of pitch are used for segmentation [22]. The pitch contour dynamics is then represented using parameters derived from linear

Table 6.4: Performance of LID system based on phonotactic and prosody features. The entries from columns 2 to 5 represent the identification accuracy (in %).

	Test duration=20 syllables		**Test duration=50 syllables**	
Language	**Rank based**	**Accumulation**	**Rank based**	**Accumulation**
Tamil	95.15	96.46	99.53	100
Telugu	67	82	90	100
Hindi	99.40	100	100	100

Table 6.5: Performance of LID system based on prosodic features alone. The entries from columns 2 to 5 represent the identification accuracy (in %).

	Test duration=20 syllables		**Test duration=50 syllables**	
Language	**Rank based**	**Accumulation**	**Rank based**	**Accumulation**
Tamil	91	97.4	99.5	99.5
Telugu	59.8	68.6	70	82
Hindi	91.6	93	98	98.5

Table 6.6: Performance of LID system based on prosodic features derived without using transcription information available in the database. The entries from columns 2 to 5 represent the identification accuracy (in %).

	Test duration=20 syllables		**Test duration=50 syllables**	
Language	**Rank based**	**Accumulation**	**Rank based**	**Accumulation**
Tamil	86.6	93	97.5	97
Telugu	41.2	60.6	49.5	74
Hindi	87.8	89.8	95	95

stylized pitch segments [49, 51–53]. This approach has the advantage that prosodic features can be derived directly from the speech signal. In both the approaches, the segmented trajectories are quantized and represented using a small set of labels that describe the dynamics of pitch and energy. The n-grams of these labels are formed to model the characteristics of a speaker or a language.

In this work, we propose a new technique for extraction and representation of prosodic features. The proposed method utilizes the location of vowel onset points (VOP) for identifying the syllable-like regions in continuous speech. This method combines the salient features of the existing approaches mentioned above, namely, the association with the syllabic pattern as in the first approach, and the extraction of features without using ASR as in the second approach.

6.6.1 Choice of Syllable as the Basic Unit

Studies have shown that prosodic characteristics such as tone and stress are strongly linked to the underlying syllable structure [84]. All spoken utterances can be considered as sequence of syllables which constitute a continual rhythmic alternation due to the opening and closing of mouth while speaking [85]. Syllable of CV type provides an articulatory pattern beginning with a tight restriction, and ending with an open vocal tract, resulting in a kind of rhythm that is especially suited both to the production and perception mechanisms [86]. Because syllable has a basic timing structure for speech sounds, a tone can be aligned only relative to the syllable rather than to any individual segments in the syllable [84]. It was demonstrated that the tonal events are aligned to the segmental events such as the onset and/or offset of a syllable [91]. Since syllable by definition invariably consists of a vowel, it will have associated pitch contour and high energy corresponding to the syllable nucleus. Prosody is linked to the underlying syllable sequence [84], and it is meaningful to associate the prosodic pattern to the syllabic sequence. Therefore, syllable appears to be a natural choice for the basic unit for representing prosody.

6.6.2 Association of Prosody with Syllable Sequence

For representing syllable-based rhythm, intonation and stress, the speech signal should be segmented into syllables. Segmenting speech into syllables is typically a language-specific mechanism, and thus it is difficult to develop a language independent algorithm for this. In this work, it is accomplished with the knowledge of VOPs as illustrated in Fig. 6.4, where the VOP refers to the instant at which the onset of vowel takes place in a syllable. The availability of pitch values helps in further reduction of spurious VOPs. For example, the absence of voicing between two VOPs numbered as '10' and '11' shown in Fig. 6.4(b), helps to eliminate the spurious peak '10'.

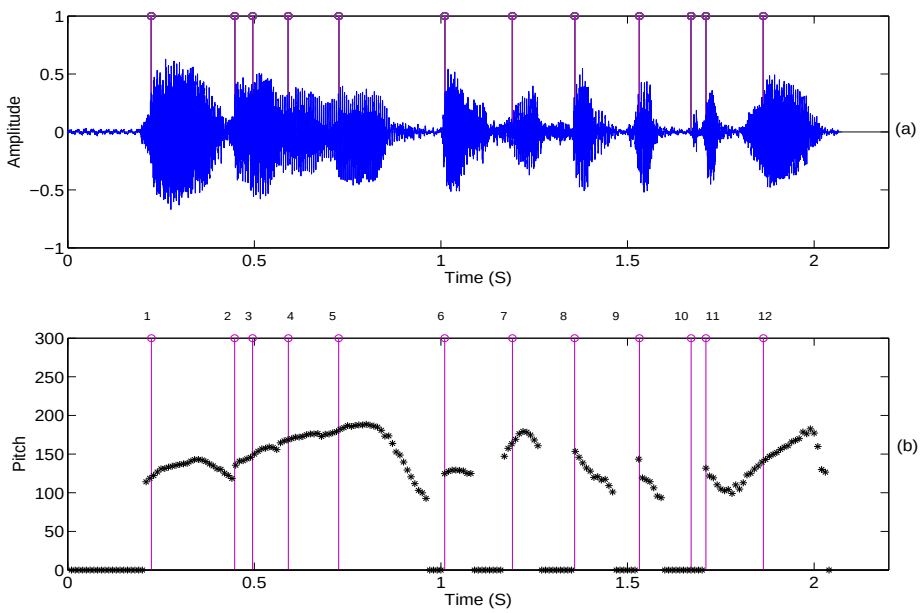


Fig. 6.4: (a) Segmentation of speech into syllable-like units using automatically detected VOPs. (b) F_0 contour associated with VOPs.

To represent the dynamics of F_0 contour, it should be segmented in a linguistically meaningful manner. This is done by segmenting the speech into syllable-like regions. The association of syllable sequence with F_0 contour, as used in this study is shown in Fig. 6.5. After hypothesizing the locations of VOP, these locations are associated

with pitch contour for feature extraction. The continuous portion of the F_0 contour with nonzero values, located within the region of two consecutive VOPs, is treated as one segment of F_0 contour. A set of parameters derived from F_0 contour, intensity and duration are used for representing each segment.

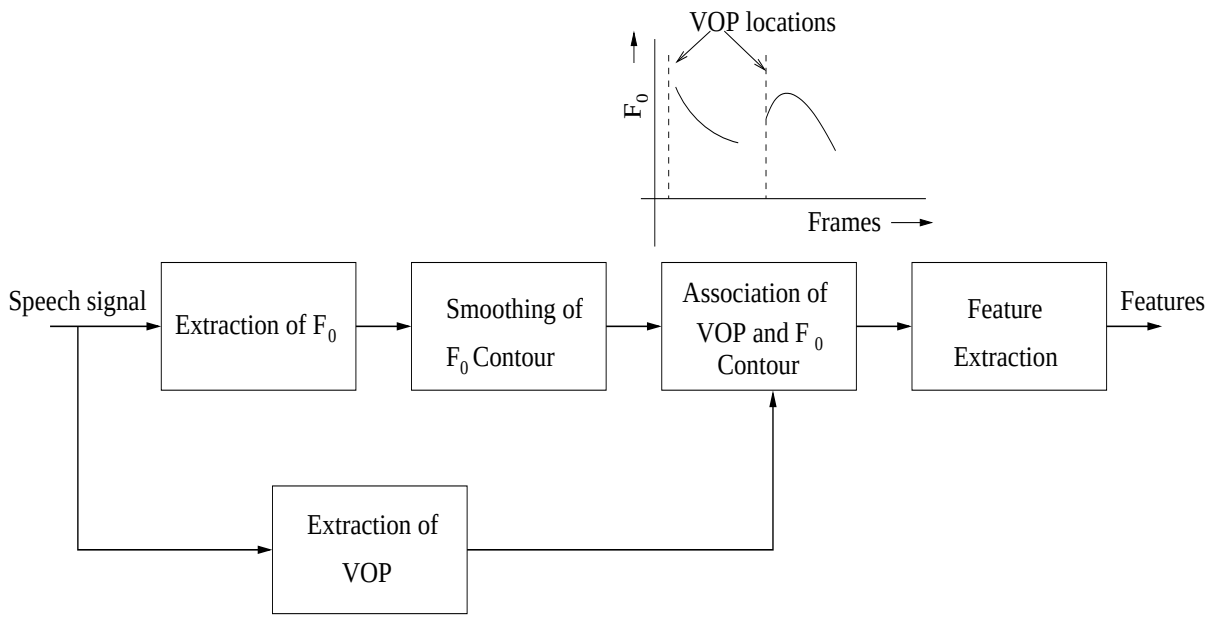


Fig. 6.5: Association of locations of VOP with F_0 contour for prosodic feature extraction.

6.7 PROSODIC FEATURES FOR LANGUAGE IDENTIFICATION

The term prosody refers to certain properties of speech signal such as audible changes in pitch, loudness and syllable length. Prosodic events appear to be time-aligned with syllables or group of syllables [92]. 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 [92]. 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. Features corresponding to the syllable-like regions are derived to represent syllable-based intonation, rhythm, and stress.

6.7.1 Representation of Intonation

Pitch analysis generates F_0 curves with finer variations. But the finer variations referred as microprosody cannot be perceived and have no function in intonation. Therefore F_0 contour obtained using pitch extraction algorithm 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 Fig. 6.6.

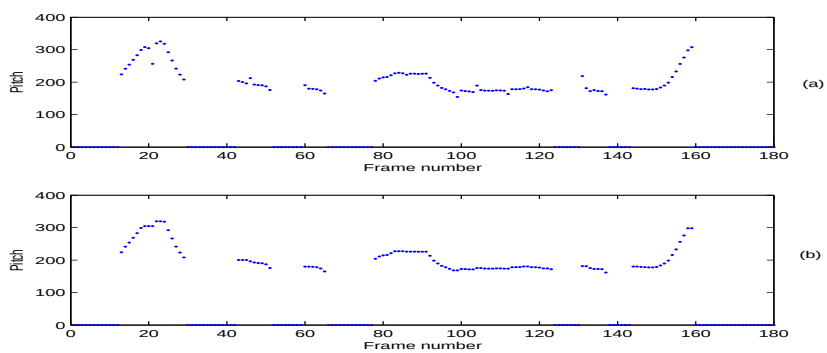


Fig. 6.6: Median filtering for smoothing the F_0 contour. (a) Raw F_0 contour. (b) Smoothed F_0 contour obtained using 7-point median filtering.

As illustrated in Fig. 6.1, 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 Fig. 6.7) 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,

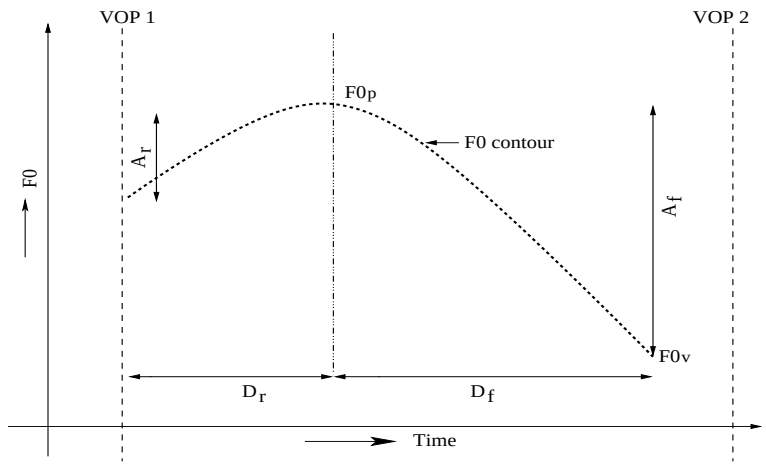


Fig. 6.7: A segment of F_0 contour. Tilt parameters A_t and D_t defined in terms of A_r, A_f, D_r, and D_f represent the dynamics of a segment of F_0 contour.

some parameters are derived.

With reference to Fig. 6.7, tilt parameters [93], namely amplitude tilt (A_t) and duration tilt (D_t) for a segment of F_0 contour are defined as follows:

$$A_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|}, \tag{6.1}$$

and

$$D_t = \frac{|D_r| - |D_f|}{|D_r| + |D_f|}, \tag{6.2}$$

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_{0_p}. Similarly D_r and D_f represent the duration taken for rise and fall respectively. It can be observed from Figs. 6.8 (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 listener's judgment of prominence reflect the role of F_0 variation in relation to variation in prominence [94]. To express the height of the F_0 peak, the difference between peak and valley fundamental frequency ($\Delta F_0 = F_{0_p} - F_{0_v}$) 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 [94]. In this study, this is represented using the distance of F_0 peak location with respect to VOP (D_p).

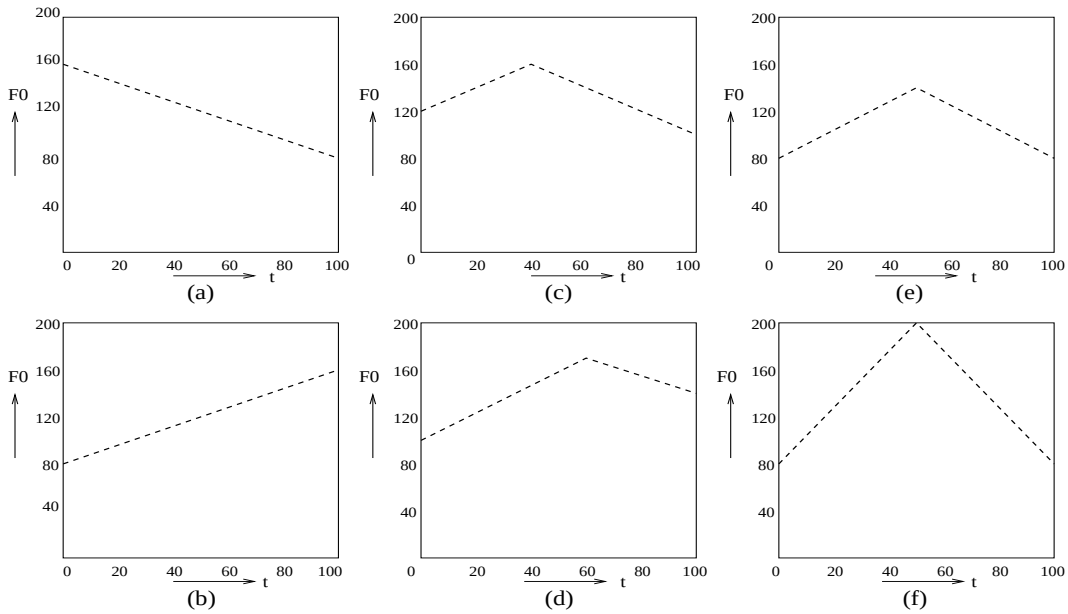


Fig. 6.8: Illustration of F_0 contours with various tilt parameters. (a) A_t =-1, D_t=-1; (b) A_t=1, D_t=1; (c) A_t=-0.2, D_t=-0.2; (d) A_t=0.4, D_t=0.2; (e) A_t=0, D_t=0; (f) A_t=0, D_t=0.

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

(a) Change in F_0 (ΔF_0)

(b) Distance of F_0 peak with respect to VOP (D_p)

(c) Amplitude tilt (A_t)

(d) 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.

6.7.2 Representation of Rhythm

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.

As the voicing information obtained from the pitch extraction algorithm is highly smoothed, we use a technique based on excitation information for voice activity detection [95–97]. As shown in Fig. 6.9, whenever there is a significant excitation of the vocal tract system, it is indicated by a large amplitude in the Hilbert envelope of the LP residual. This 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:

(a) Syllable duration (D_s)

(b) Duration of voiced region (D_v) within each syllable

6.7.3 Representation of Stress

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 [92]. Therefore along with F_0 and duration features mentioned above, we use change in log energy (ΔE) within voiced region to represent the stress.

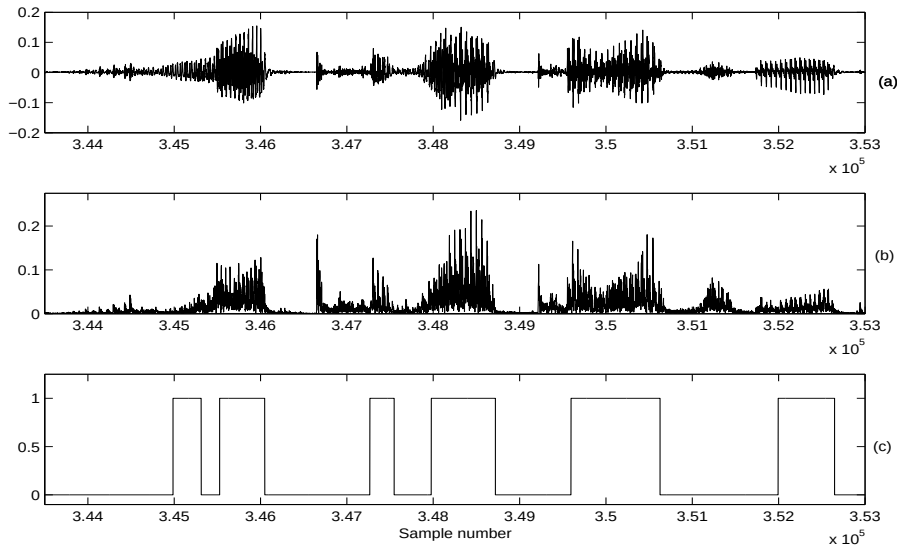


Fig. 6.9: Detection of voiced regions in speech using the strength and periodicity of excitation at the instants of glottal closure showing (a) Speech signal, (b) Hilbert envelope, and (c) Binary waveform having unity amplitude corresponding to the voiced regions.

6.7.4 Modeling of Language-specific Prosody

It has been observed that tones of adjacent syllables influence both the shape and height of the F_0 contour of a particular syllable [98], and prominence of a syllable is estimated based on the pitch characteristics of the contour around it [94]. 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. Features corresponding to such regions are not included in the set of training/testing examples. Since the specific interaction between pitch

movements, intensity and duration play an important role in determining the prosody, these parameters together are used to form a feature vector for modeling, as shown in Fig. 6.10.

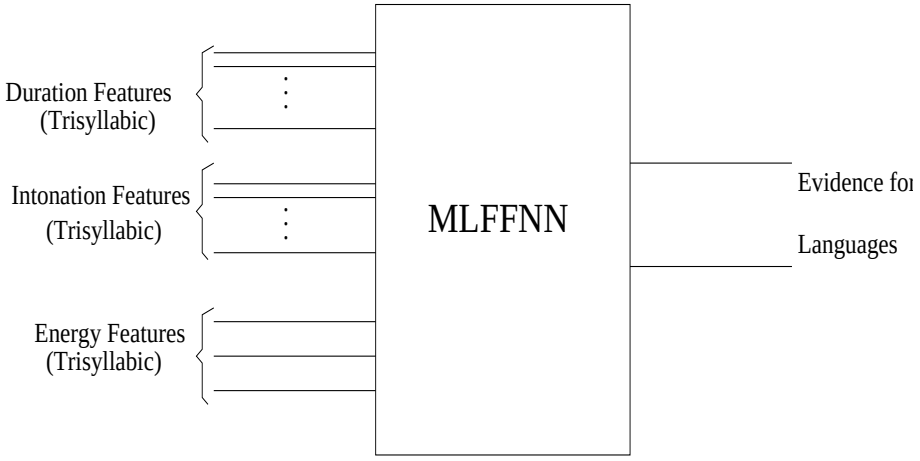


Fig. 6.10: Prosody-based neural network classifier for language identification.

6.8 PERFORMANCE EVALUATION ON OGI DATABASE

To demonstrate the effectiveness of prosodic features mentioned above, a study was conducted using the OGI database. For all the languages, 40 speech (unrestricted spontaneous) files (each with an average duration 45 sec) corresponding to 40 different speakers are used for training. An average of 20 speech files from different speakers are used for evaluating the proposed LID system. The training and testing speaker sets are different. Separate MLFFNN models are trained for each language pair in the OGI database. For example, to build a model for discriminating English from Mandarin, an MLFFNN classifier is trained with examples from English and Mandarin, with output set to $\{+1, -1\}$, and $\{-1, +1\}$, respectively as shown in Fig. 6.10. For evaluating the performance of this model, features derived from three consecutive syllables in the test utterance are applied to the classifier, and the evidence from output are accumulated to obtain the evidence of languages.

Evaluation is done with different combinations of prosodic features and representation to determine the suitable feature set. The results for four pairs are given in Tables 6.7 and 6.8.

Table 6.7: Effectiveness of various prosodic features at trisyllabic level for language discrimination in case of four different language pairs. Entries from columns 3 to 6 represent the percentage of utterance identified correctly.

Prosodic Features	**Feature Dimension**	**Model**			
		En-Ge	**Fr-Ma**	**Ge-Sp**	**Ja-Sp**
ΔF_0, A_t, D_t, D_p (4)	12	67	84	72	76
ΔF_0, A_t, D_t, D_p, D_s, D_v, ΔE (7)	21	69	84	79	85

Table 6.8: Monosyllabic vs trisyllabic features for prosody based language discrimination. Entries from columns 3 to 6 represent the percentage of utterance identified correctly.

Feature Representation	**Feature Dimension**	**Model**			
		En-Ge	**Fr-Ma**	**Ge-Sp**	**Ja-Sp**
Monosyllabic	7	50	84	45	53
Trisyllabic	21	69	84	79	85

The feature vector of dimension 21 corresponding to three consecutive syllables, consisting of 7 parameters from each syllable gave the best performance as shown in Table 6.7. These are syllable duration (D_s), duration of voiced region (D_v), change in F_0 (ΔF_0), distance of F_0 peak with reference to VOP (D_p), amplitude tilt (A_t), duration tilt (D_t) and change in log energy (ΔE). To illustrate the effectiveness of trisyllabic feature representation, a study was conducted using the same features at monosyllabic vs trisyllabic level. The results given in Table 6.8 show better performance for the case of trisyllabic representation in most of the cases.

The results of pair-wise language discrimination task on OGI database are given in Table 6.9. The results of two recent studies are given in square bracket for comparison

Table 6.9: Performance of pair-wise language discrimination task on OGI database. The entries from column 2 to 11 denote the percentage of test utterances identified correctly, for a model corresponding to the languages in first column and first row. For comparison, results of Rouas's and Lin's work are given in square brackets.

Lang.	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
En	63 [76][62]	85 [52][54]	69 [60][56]	73 [-][-]	70 [68][84]	78 [79][75]	78 [75][76]	57 [68][53]	90 [77][64]	70 [68][80]
Fa	-	67 [69][87]	78 [72][73]	58 [-][-]	76 [67][85]	67 [75][70]	81 [76][82]	60 [67][73]	77 [70][71]	61 [67][69]
Fr	-	-	60 [56][42]	85 [-][-]	90 [56][65]	86 [55][54]	84 [61][69]	84 [64][57]	90 [60][44]	78 [58][76]
Ge	-	-	-	88 [-][-]	86 [66][77]	86 [71][65]	72 [62][84]	79 [59][49]	90 [70][59]	71 [66][69]
Hi	-	-	-	-	89 [-][-]	67 [-][-]	92 [-][-]	60 [-][-]	77 [-][-]	78 [-][-]
Ja	-	-	-	-	-	76 [66][75]	62 [50][78]	85 [63][81]	85 [59][79]	88 [69][89]
Ko	-	-	-	-	-	-	91 [74][80]	70 [76][59]	81 [62][58]	81 [56][73]
Ma	-	-	-	-	-	-	-	82 [81][71]	89 [74][69]	85 [50][79]
Sp	-	-	-	-	-	-	-	-	85 [65][48]	85 [62][61]
Ta	-	-	-	-	-	-	-	-	-	88 [71][77]

[21, 23]. It is observed that prosodic features are more effective for discriminating languages that fall into different categories based on rhythm or tonal characteristics. For example, Japanese and Mandarin are discriminated very well from other languages, whereas, discrimination between them is somewhat poor. Due to the limited size of the speech data available in the OGI database, it was difficult to extend this study to multi-class LID problem, as noted by the other researchers [19, 21, 23].

6.9 SUMMARY AND CONCLUSIONS

The effectiveness of prosodic and phonotactic features for LID was demonstrated in the preliminary study on three Indian languages. The MLFFNN classifier, trained using the phonotactic features and/or prosodic features was shown to be effective for LID. The phonotactic features in terms of broad phonetic categories were also examined for LID.

A new technique is proposed for deriving the prosodic features from the speech signal. The effectiveness of the derived prosodic features was demonstrated. The results on the OGI database shows that prosodic features are effective for discriminating languages. The performance is better for discriminating languages that fall into different categories based on rhythm/tonal characteristics. But it was difficult to extend the pair-wise language discrimination to a multi-class LID problem, due to the limited size of the speech data available in the OGI database. A summary of the LID studies carried out in this chapter are given in Table 6.10.

The inclusion of phonotactic features to this framework along with prosodic features may be helpful for extending the results for more languages. Phonotactic features may be in the form of pre-linguistic form such as cepstral coefficients or code book indices. Adding information of syllable structure to the prosodic features may provide better representation of the rhythmic characteristics.

In the next chapter, we address some issues in speaker verification. In particular, we examine the use prosodic features for modeling speakers.

Table 6.10: Summary of the studies on language identification using multisyllabic features.

1.	A preliminary study was conducted on three Indian languages using features derived from manual segmentation and transcription information. The role of following characteristics for LID were examined in this study: (a) Phonotactics (b) Broad phonotactics (c) Prosody along with phonotactics (d) Prosody
2.	A VOP based method is proposed for the extraction of prosodic features from the speech signal.
3.	An experimental study was conducted on selected pairs of languages in OGI database to optimize the set prosodic features and their representation (monosyllabic vs. multisyllabic) to highlight the language-specific characteristics.
4.	A pair-wise language discrimination study was conducted on OGI database to demonstrate the effectiveness of the prosodic features derived using the proposed approach.

CHAPTER 7

PROSODIC FEATURES FOR SPEAKER VERIFICATION

7.1 INTRODUCTION

Speech signal contains information about the message, speaker characteristics and language characteristics, besides emotional state of the speaker and the environment in which the signal is collected. Short-time (10-30 msec) spectrum analysis is performed to extract the time-varying spectral envelope characteristics, attributing them to the shape of the vocal tract system. Generally the residual of the speech signal, obtained after removing the spectral envelope information, is considered not useful for many speech applications. But the residual signal contains information, both at the subsegmental (less than a pitch period) level and at the suprasegmental (>100 msec containing several pitch periods in voiced segments) level. The information at the subsegmental level corresponds to the excitation, mainly due to glottal vibration. The information at the suprasegmental level is mostly contributed by the prosodic characteristics. Since it is difficult to extract and represent the information at the suprasegmental levels for speech applications, the information present at this level is generally ignored. In this chapter we show that it is indeed possible to extract the speaker-specific information present at the suprasegmental level, and represent it in a manner useful for speaker verification. The main focus is on the use of prosodic features for speaker verification task.

In Section 7.2 we discuss features relevant for speaker verification task. Speaker-specific features manifested at various levels of speech signal are also discussed in this section. In Section 7.3, speaker-specific aspect of prosody is discussed. Section 7.4 describes the extraction and representation of prosodic features for speaker verification.

In Section 7.5, the results of prosody-based speaker verification studies on National Institute of Standards and Technology (NIST) 2003 extended data are discussed. In Section 7.6, we study the use of subsegmental, segmental and suprasegmental features for speaker verification. We demonstrate that combining evidence from spectral features and prosodic features, represented at segmental and suprasegmental levels of speech signal, respectively, improve the speaker verification performance in case of NIST extended data task. Section 7.7 summarizes the speaker verification studies.

7.2 FEATURES FOR SPEAKER VERIFICATION

The objective of speaker verification is to accept/reject a person's claim on his/her identity, using features derived from his/her speech. Speaker characteristics are manifested in speech signal as a result of anatomical differences inherent in the speech production organs and differences in the learned speaking habits of individuals [32]. Speaker characteristics differ in:

(a) Vocal tract size and shape

(b) Excitation

(c) Prosody

(d) Idiolect

The magnitude of the frequency spectrum encodes information about the speaker's vocal tract shape via resonances (formants) [33, 35, 37, 39, 99]. Most current speaker verification systems rely on spectral features which exploits the spectral difference. Apart from the characteristics of the vocal tract, information from the excitation source and learning habits of the speaker are known to be exploited by the humans for identifying speakers. The main source of excitation for production of speech is the glottal vibration. In each glottal cycle, the instant of glottal closure is the instant at which significant excitation of vocal tract takes place. Hence a small region (1-5 msec) around the instant of glottal closure contains significant information about

the speaker, which may be exploited for developing speaker verification systems. It is known that human beings use certain higher level features such as prosody and idiolect to identify a familiar speaker. These are the habitual attributes of the speaker rather than physiological characteristics of the speech production system. Speaking habits are reflected in characteristics such as the usage of certain words and phrases, and to the features such as intonation, stress and timing. The idiolect-based system, models the speaker idiosyncrasies by identifying speaker-specific phrases/words obtained from the output of an automatic speech recognition system. We consider only implicit features, and not the information that can be derived from text transcriptions. Thus idiolect-based features are precluded in this study.

The differences among speakers in terms of characteristics of the excitation source, vocal tract and prosody can be represented by features at three different levels, namely, subsegmental, segmental and suprasegmental levels, respectively.

1. Subsegmental Features

   In order to represent the excitation source characteristics of a speaker, a window of size which is less than one pitch period (1-5 msec) is considered. The information corresponding to the vocal tract system and the excitation source may be separated approximately from the speech signal using LP analysis [37]. Since in the LP residual the vocal tract features are removed, it contains mostly the information about the excitation source [43, 44]. Hence the LP residual is used in this study for representing the excitation source characteristics for speaker verification.

2. Segmental Features

   Vocal tract characteristics are obtained assuming quasi-periodic assumption for a segment that contains a few pitch periods (10-30 msec). Features derived from such a segment is referred as segmental features. In this study, the weighted linear prediction cepstral coefficients (WLPCC) at segmental level are used for representing the spectral characteristics.

3. Suprasegmental Features

   Features corresponding to a larger span (> 100 msec) of speech which go beyond segments are referred to as suprasegmental features. To model speaker-specific aspects of prosody, features are represented at the suprasegmental level.

7.3 SPEAKER-SPECIFIC ASPECTS OF PROSODY

It is not just the physiological aspects of speech production organs of a speaker that influence the way an utterance is spoken. It is also influenced by the habitual aspect of a particular speaker. The acquired speaking habits are characteristics learned over a period of time, mostly influenced by the social environment, and also by the characteristics of the first/native language in the 'critical period' (lasting roughly from infancy until puberty) of learning. The prosodic characteristics as manifested in speech give important information regarding the speaking habits of a person.

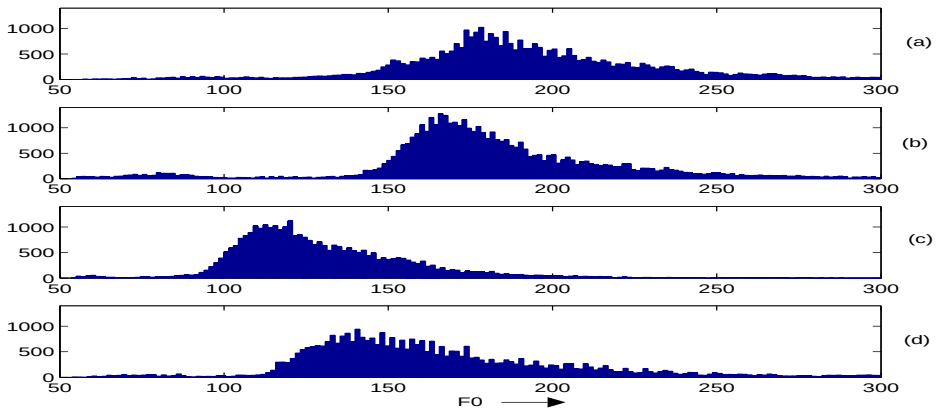


Fig. 7.1: Variation in histogram of F_0 for (a), (b) Two female and (c), (d) Two male speakers.

The physical correlate of pitch is the fundamental frequency (F_0) of vibration of vocal folds. Fundamental frequency reflects speaker-specific characteristics due to the difference in physical structure of vocal folds among speakers. The F_0 distribution is specific to a speaker [100] as illustrated in Fig. 7.1. The global statistics of F_0 values

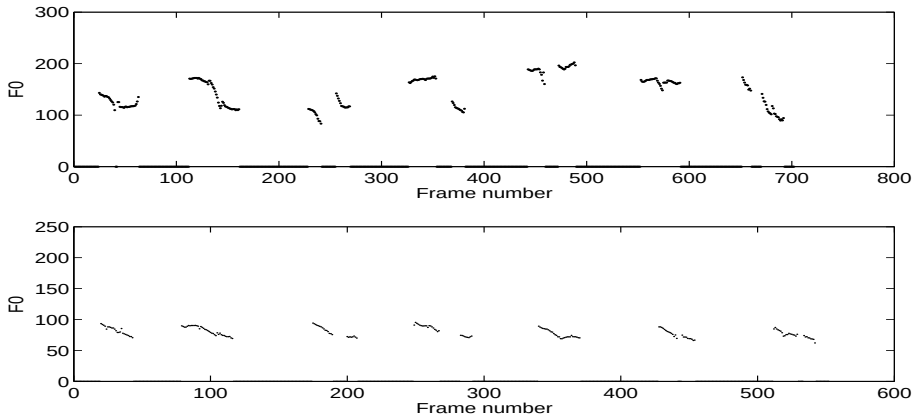


Fig. 7.2: Variation in dynamics of F_0 contour of two different male speakers while uttering *Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday.*

of a speaker can be captured using appropriate distributions for speaker verification task [48].

The variation of F_0 as a function of time (intonation) reflects certain speaking habits of a person. The local dynamics of the F_0 contour can be different among speakers due to different speaking style and accent. It has been shown that the dynamics of F_0 contour can contribute to speaker verification task [49, 52]. Dynamics of F_0 contour corresponding to a sound unit is influenced by several factors such as the identity of sound unit spoken, its position with respect to phrase/word, context (the units that precede and succeed), speaking style of a particular speaker, intonation rules of the language, type of sentence (interrogative or declarative), etc. The dynamics of F_0 contour are different for two speakers, even when they utter the same text as illustrated in Fig. 7.2. But when the same text is repeated by the same speaker, F_0 contour characteristics are consistent as in Fig. 7.3. The speaker-specific information present in F_0 contour can be used for characterizing a speaker. This property is used in text-dependent speaker verification, by comparing F_0 contours using dynamic time warping (DTW).

Dynamics of F_0 contour is useful for characterizing speakers even when the text

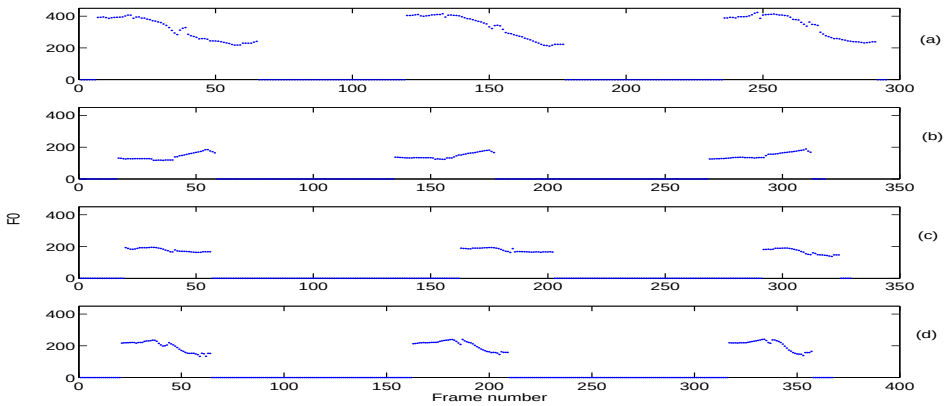


Fig. 7.3: Variation in dynamics of F_0 contour (a) A child (b), (c) Two different males and (d) A female speaker while repeating the same text *Sunday, Sunday, Sunday.*

spoken are different (text-independent) as illustrated in Fig. 7.4. But to capture these characteristics for speaker modeling, more speech data may be required for text-independent tasks.

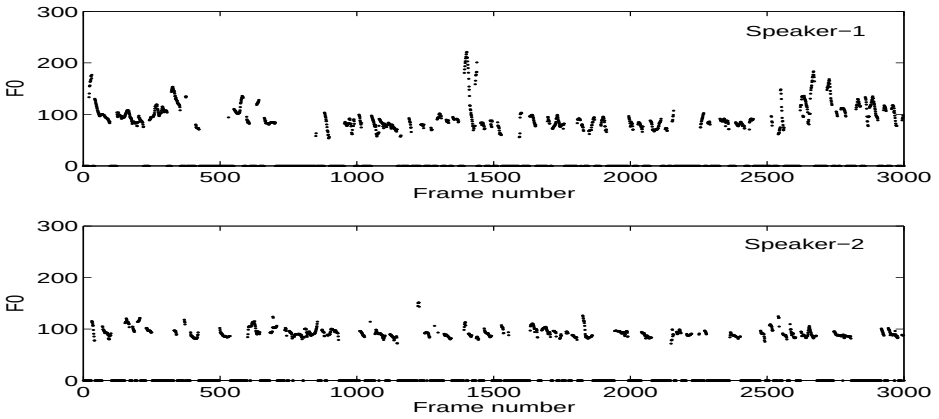


Fig. 7.4: Variation in dynamics of F_0 contour of two different male speakers for different texts.

7.4 PROSODIC FEATURES FOR SPEAKER VERIFICATION

Human beings use several levels of perceptual cues for speaker recognition ranging from high-level cues such as semantics, pronunciation, idiosyncrasies and prosody, to low-level cues such as acoustic cues of speech [9]. Prosodic cues such as pitch gestures, accents and stress reflect the physiological as well as habitual aspect of a speaker. Current text-independent speaker verification systems rely mostly on spectral features derived through short-time spectral analysis. This approach does not attempt to model the long-term speaker-specific characteristics present in the speech signal. The long-term features are relatively less affected by channel mismatch and noise. In order to incorporate long-term features, system generally require significantly more data for training. Hence in 2001, NIST introduced the extended data task which provides multiple conversation sides for speaker training [50]. This helps in the study of long-term features for speaker recognition. A workshop was conducted at the John Hopkins University (JHU) to explore a wide range of features for speaker identification using NIST 2001 extended data task as its testbed [51, 101].

7.4.1 Robustness of Prosodic Features

The F_0 contour characteristics are attractive due to its robustness to channel variations. The effect of channel variations on F_0 contour and spectral feature vectors are illustrated in Figs. 7.5 and 7.6. The same sentence *Don't carry an oily rag like that* spoken by the same speaker, collected over different channels available in Texas Instruments and Massachusetts Institute of Technology (TIMIT) database, is used for comparing the effect channel variations. Channels correspond to TIMIT, NTIMIT and CTIMIT represent speech collected over close-speaking microphone, noisy channel, and cellular environment, respectively. Fig. 7.5 shows the shift of LPCC features in the feature space due to variability in the channel characteristics, whereas Fig. 7.6 illustrates the robustness of F_0 contour characteristics against channel variations. In Fig. 7.6, the F_0 contours remain same for all the cases except some change in duration

of voiced region in (b) and (c) compared to (a).

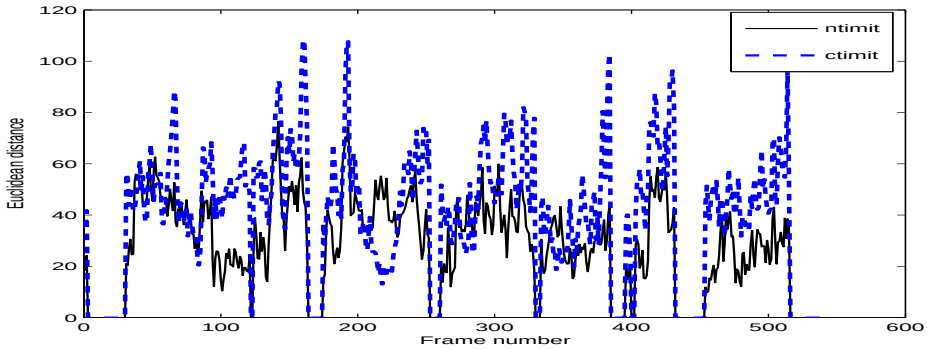


Fig. 7.5: Euclidean distance of LPCC feature vectors on a frame to frame basis for the same speaker and text *Don't carry an oily rag like that.* The solid line corresponds to the distance of NTIMIT data and dashed line corresponds to CTIMIT data with respect to TIMIT data.

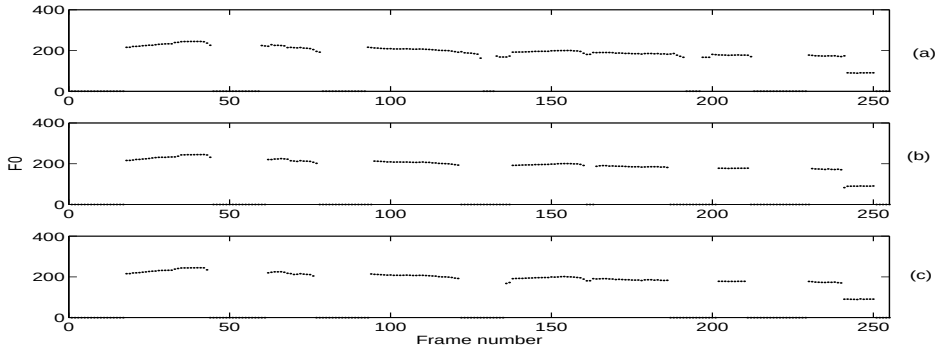


Fig. 7.6: F_0 contours of (a) TIMIT (b) NTIMIT, and (c) CTIMIT sentence of the same speaker for the same sentence *Don't carry an oily rag like that* showing its robustness against channel variations.

7.4.2 Extraction and Representation of Speaker-specific Prosody

Segmentation into syllable-like regions is accomplished with the knowledge of vowel onset points (VOP) as illustrated in Fig. 7.7(a). The locations of VOP are then associated with F_0 contour as in Fig. 7.7(b) for feature extraction. The continuous portion

of the F_0 contour with nonzero values, located within the region of two consecutive VOPs, is treated as one segment of F_0 contour for feature extraction.

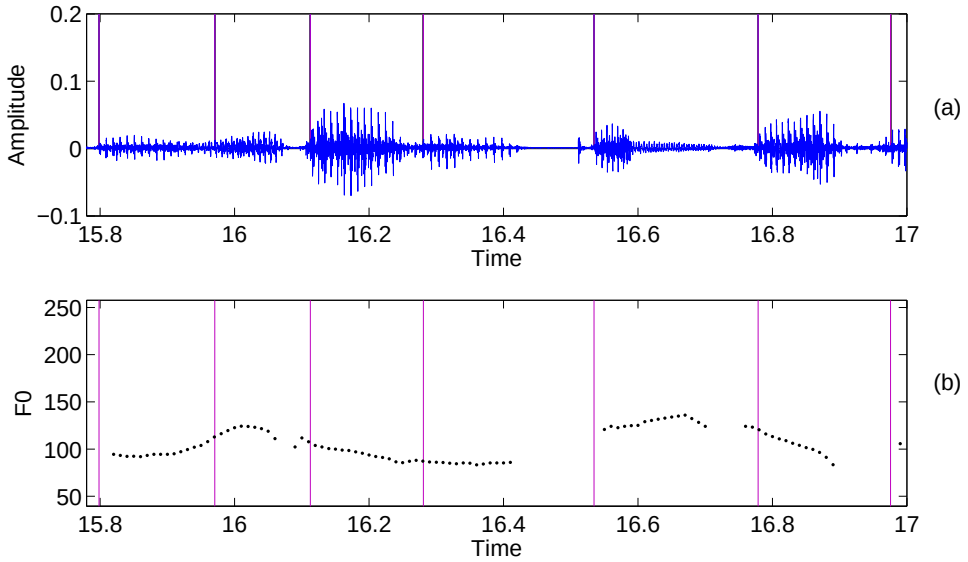


Fig. 7.7: (a) Segmentation of speech into syllable-like units using automatically detected VOPs. (b) F_0 contour associated with VOPs.

As demonstrated in Figs. 7.2, 7.3 and 7.4, the dynamics of the F_0 contour reflects certain learned habits of a speaker. Therefore it is important to represent it using suitable parameters. We use tilt parameters [93] for representing the dynamics of the F_0 contour. An investigation on the role of articulatory constraints in shaping the F_0 contour has revealed that the maximum speed of F_0 change limits how fast the F_0 movements can be produced [102]. The coordination of laryngeal and supralaryngeal movements controls the alignment of syllables and tone [102]. Studies have also indicated that listeners are more sensitive to variations in F_0 peak (F_{0_p}) than F_0 valley (F_{0_v}) [94]. Hence change in F_0 (ΔF_0), distance of F_0 peak (D_p) and the peak value of pitch (F_{0_p}) for each F_0 segment may be useful for speaker recognition. An increase in pitch may be obtained by increasing the vocal fold tension, by increasing the subglottal pressure, or a combination of them. Therefore F_0 peak (F_{0_p}) and F_0 mean (F_{0_μ}) of each segment of F_0 contour may reflect some physiological as well as

habitual aspects of a speaker. The change in log energy (ΔE) along with ΔF_0 gives a quantitative measure of stress characteristics, therefore may be specific to a particular speaker. The F_0 and energy related parameters used in this study for characterizing the speaker-specific aspect of prosody are the following:

(a) Mean value of F_0 (F_{0_μ})

(b) Maximum value of F_0 (F_{0_p})

(c) Change in F_0 (ΔF_0)

(d) Distance of F_0 peak with respect to VOP (D_p)

(e) Amplitude tilt (A_t)

(f) Duration tilt (D_t)

(g) Change in log energy ΔE

Thus each segment corresponding to the voiced region of a syllable is represented using a 7-dimensional feature vector.

We hypothesize that the distribution of prosodic feature vectors form a unique cluster in the feature space. Fig. 7.8 shows the nonlinearly compressed prosodic feature vectors, derived from speech corresponding to two male speakers in NIST 2003 database. The nonlinear compression of the 7-dimensional prosodic feature vector has been obtained using AANN model with a structure *7L 14N 3N 14N 7L* , where L represents linear activation function, N represent nonlinear activation function, and the numerals represent the number of units in the layers. The compressed feature vector is obtained from the dimension-compression hidden (middle) layer having three units. Next section discusses the prosody-based speaker verification studies.

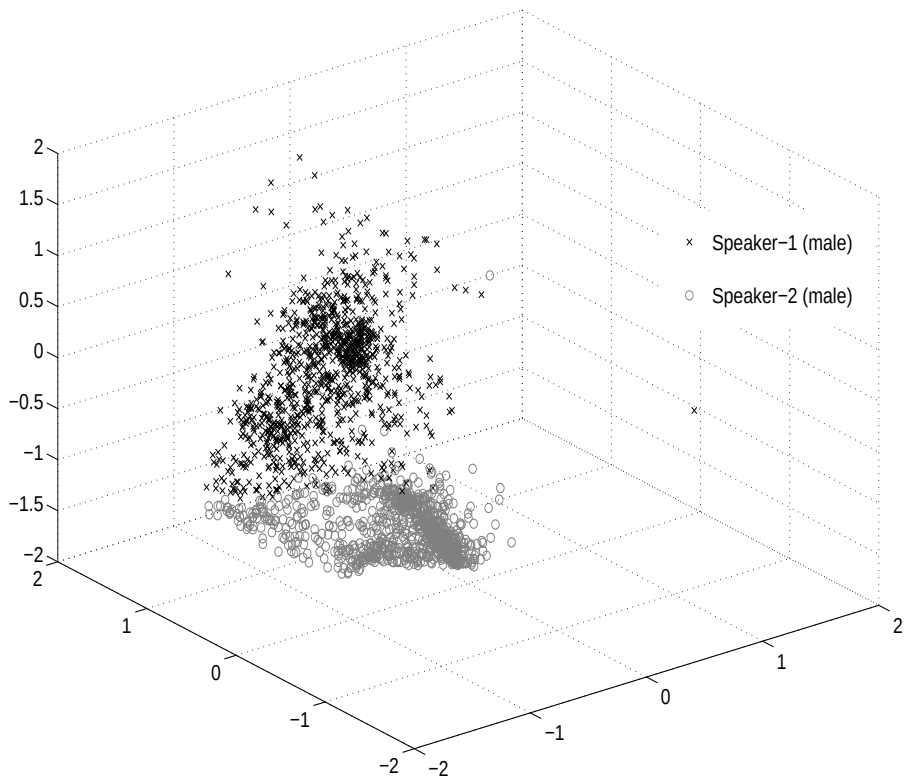


Fig. 7.8: Compressed prosodic feature vectors for two male speakers (taken from NIST 2003 extended data).

7.5 EXPERIMENTAL STUDIES ON PROSODY-BASED SPEAKER VERIFICATION

7.5.1 Database - NIST 2003 Extended data

To demonstrate the significance of prosodic features, we use the first subset of NIST 2003 extended data task [103]. Unlike the traditional speaker recognition tasks, the extended data task provides larger amount of speech data for training the models. This task consists of 16-side, 8-side, and 4-side cases providing 16, 8 and 4 conversation

sides, respectively, for training the target speaker model. Duration of a conversation is about 5 minutes, and one side of it will be approximately 2.5 minutes. Each target model is tested with a set of 1 side test utterance where the task is to find out whether the particular test utterance belongs to the target speaker or not. The subset chosen for our study consists of 137, 54 and 74 speaker models for the 16-side, 8-side and 4-side cases, respectively. The models are evaluated using 1076, 1238 and 1258 test utterances for the 16-side, 8-side, and 4-side cases, respectively. We use equal error rate (EER) and detection error tradeoff (DET) curves as measures for evaluating the performance of our system [104].

7.5.2 Performance Evaluation on NIST 2003 Database

The block diagram of the proposed speaker verification system is shown in Fig. 7.9. To capture the speaker-specific distribution of the prosodic feature vectors, AANN models are used. For each target speaker, an AANN model is developed to capture the distribution of the prosodic features. A set of background (BG) models built from a known set of impostor speakers helps to fix a global threshold for verification, to decide whether the test utterance belongs to the target speaker or not [105–107]. The background models consists of a set of male and female models.

The structure of the AANN model used for capturing the distribution of the speaker-specific prosodic features is *7L 28N 2N 28N 7L*, where L represents linear activation function, N represent nonlinear activation function and the numerals represent the number of units in the layers. The background models consists of 15 female speaker models and 15 male speaker models, trained using a different subset in the database. During testing, each prosodic feature vector derived from the test utterance, is applied to the target speaker model as well as background models. The error between the output and the input of AANN models is converted into a confidence value C_i. The average confidence for each model is computed as $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where C_i is the confidence value for the ith syllable, and N is the number of prosodic feature

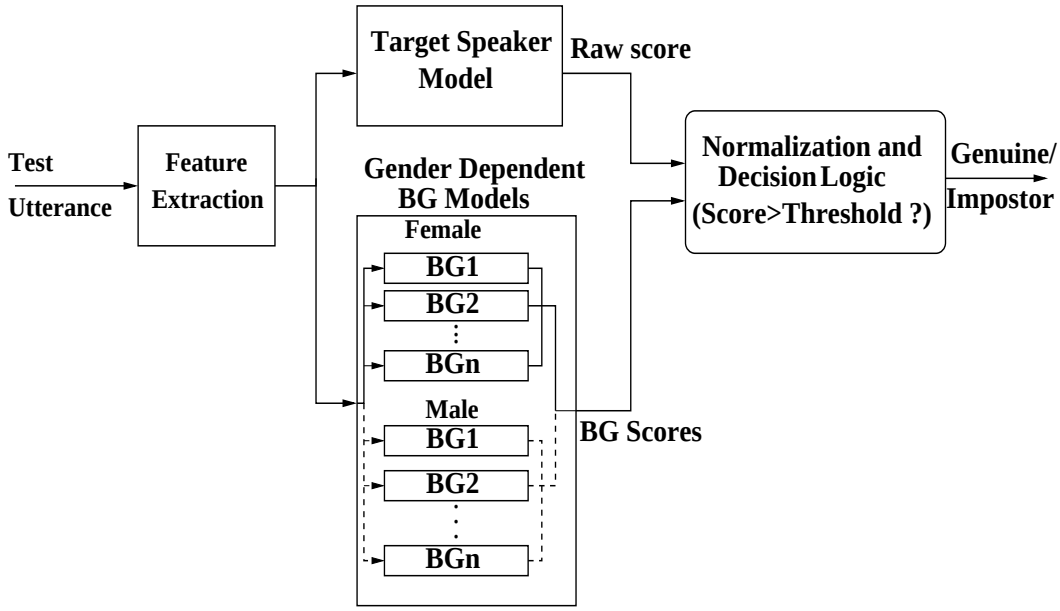


Fig. 7.9: Block diagram of prosody-based speaker verification system, showing the testing of an unknown utterance against target speaker model and a set of background models.

vectors in the test utterance.

Score normalization is used for scaling the likelihood scores, which helps to find a global speaker independent threshold for the decision making process. Feature vectors obtained from the test utterance is presented to the target speaker model as well as to a set of background models as shown in Fig. 7.9. For each test utterance, the decision on the gender is made based on the average score of male/female background model set. The raw score obtained from target speaker model is test normalized [106] using scores of the background models. The normalized score C_n is computed as:

$$C_n = C - \mu_g/\sigma_g \tag{7.1}$$

where μ_g and σ_g represent mean and standard deviation of BG scores corresponding to the hypothesized gender of test utterance.

Pitch related features seems to carry more speaker-specific information. Addition of ΔE slightly improves the EER from 15.8 to 12.4 as given in Table 7.1. Performance

Table 7.1: Performance of prosodic features, for cases where 16 conversation sides are available for training the model of target speaker.

Features	**EER**
Pitch (6 dimension)	15.8
Pitch+Energy (7 dimension)	12.4

of 16-side conversational cases is shown using DET curves in Fig. 7.10. Prosody based system resulted in an EER of 12.4, 15, 26 for 16-side, 8-side, and 4-side conversational cases of the particular data set, respectively.

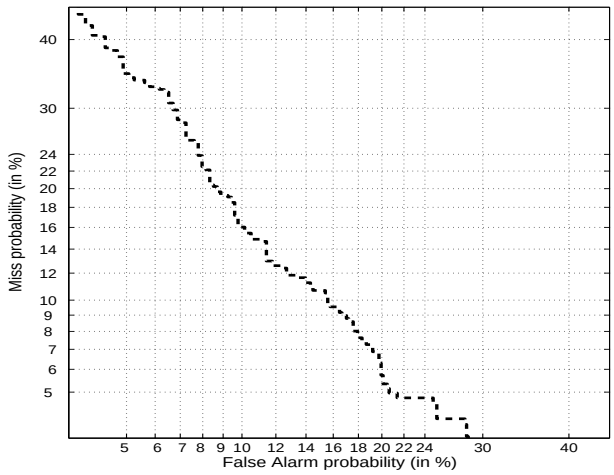


Fig. 7.10: DET curve showing the performance of prosody-based speaker verification system for 16-side conversational case.

7.6 MULTILEVEL FEATURES FOR SPEAKER VERIFICATION

7.6.1 Subsegmental and Segmental Features for Speaker Verification

For the LP residual, the AANN model is used as a nonlinear model to capture the implicit speaker-specific characteristics. The structure of the AANN model used is

40L 48N 12N 48N 40L. Blocks of 40 samples from the voiced regions are used as input to the AANN. Successive blocks are formed with a shift of one sample. Each block is normalized to the range -1 to 1. One AANN model is trained per speaker using 200 epochs. Since the block size is less than a pitch period, only the characteristics of the excitation source within each glottal pulse are captured. For testing, blocks of 40 samples of the LP residual from the voiced regions of speech are given as input to the AANN models. The output of each model is compared with its input to compute the squared error for each block. LP residual-based system resulted in an EER of 33, 33.5, 34.5 for 16-side, 8-side, and 4-side conversational cases of the particular data set, respectively.

AANN model trained using WLPCCs are expected to capture the distribution of spectral feature vectors which is unique for a language or a speaker. The structure of the AANN model used for capturing the distribution of the speaker-specific spectral features is *19L 38N 4N 38N 19L*. One AANN model is trained for each speaker, using the features from all the conversation sides available. Using single AANN model (trained using all the conversation sides) per target speaker gives an EER of 12.6 for the 16-side cases.

The EER is further improved by reducing the effects of channel mismatch between training and testing speech. A similar technique used for spectral-based LID is used for reducing the channel variability. In case of 16-side conversation, all the 16-sides available for the target speaker may not be collected over similar channels. Therefore instead of using all the 16 conversation sides to train a single model, sixteen separate AANN models are trained using single conversation side of 2.5 minutes duration. During verification, test utterance is tested against all the 16 target speaker models as well as background models. If the test utterance belongs the target speaker, majority of the models may give high scores, but a few models may still give low scores due to the mismatch in channel characteristics. Therefore, out of 16 target speaker scores, only N-best scores are considered for computing effective score of target speaker. We

have taken average of best eight confidence scores as the score of the target speaker. This score is normalized as shown in Fig. 7.9. The spectral-based system employing multiple models resulted in an EER of 9.5 for 16-side conversation case for the same data set giving a reduction of 3.1 in EER compared to the single model per target speaker approach. For 8-side and 4-side cases, this approach resulted in an EER of 11.6 and 14, respectively, for the particular data set.

7.6.2 Combining Evidence from Multilevel Features

As spectral features are affected by channel mismatch and noise, the use of prosodic features can play important role in improving the robustness of the speaker verification system. The evidence of the speaker from different features may be combined in several ways to achieve better performance. One simple approach is the addition of evidence from different systems. Prosody-based evidence provide complementary information, while combining with the spectral-based evidence. Combining by addition (giving equal weight to spectral and prosodic features) results in an EER of 6.8 for 16-side showing the effect of complementary information in these features as illustrated in Fig. 7.11.

Combining evidence from spectral and prosodic features by addition result in an EER of 9.3 and 11 for 8-side and 4-side, respectively. Even though the performance of prosodic features is inferior to spectral features in 4-side case, the combination of evidence significantly improve the performance as shown in Fig. 7.12, illustrating their complementary nature for speaker verification task.

Speaker verification system based on spectral features still appears to be performing better than other systems. Since spectral features are affected by channel mismatch and noise, use of prosodic features become important in speaker verification task. The contribution of prosodic features becomes significant when amount of speech data available for training is sufficient to capture the speaker-specific prosodic characteristics, as for the 16-side cases. The residual features represent the excitation

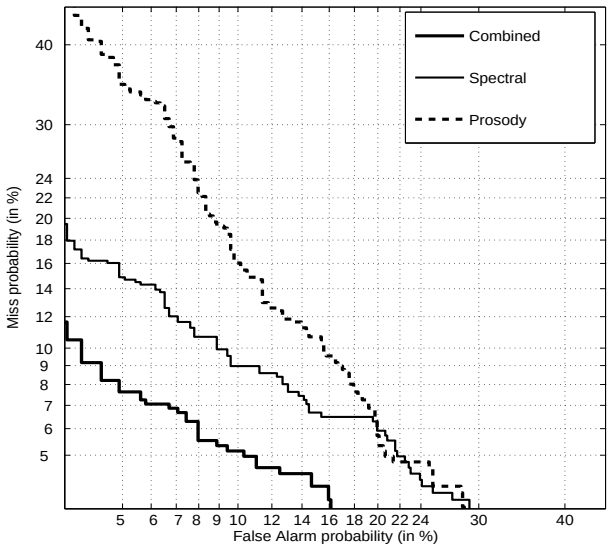


Fig. 7.11: DET curve showing the performance of spectral-based system, prosody-based system and combined system for 16-side conversational case.

source characteristics, and it is helpful for modeling the speaker from limited size of speech data.

7.7 SUMMARY AND CONCLUSIONS

The goal of this study was to explore the usefulness of prosodic features extracted from the suprasegmental levels of the speech signal for speaker verification task. Prosodic features derived from F_0 contour and energy variation reflect both physiological and learned aspect of a speaker. We demonstrate that the distribution of prosodic feature vectors is speaker-specific, and is useful for speaker verification. We have demonstrated using experimental studies that, apart from the spectral features and excitation features, prosodic features also contain significant speaker-specific information. It was observed that excitation characteristics represented at the subsegmental level contribute to the speaker verification, especially when the size of speech data available for training is less. A few seconds (10-20 sec) of speech is all that is required

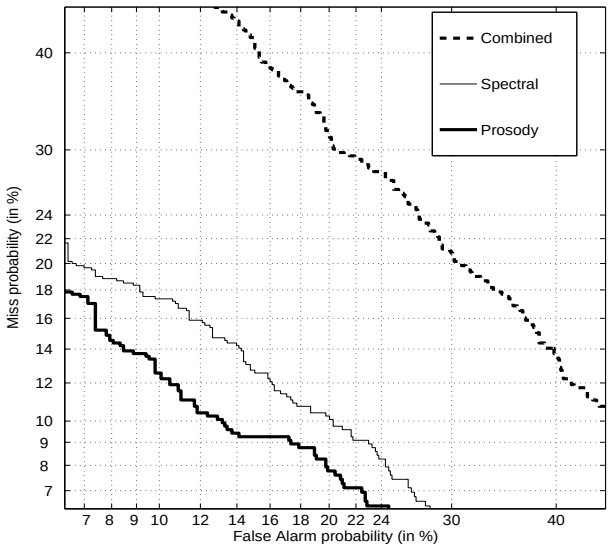


Fig. 7.12: DET curve showing the performance of spectral-based system, prosody-based system and combined system for 4-side conversational case.

for modeling excitation characteristics of a speaker. Prosodic features represented at suprasegmental level can play an important role in speaker verification when large amount of speech data is available for training, and also when there is likely to be channel/handset mismatch between training and testing. In such situations, combining evidence from spectral-based and prosody-based systems improve the robustness of speaker verification system. A summary of the speaker verification studies carried out in this chapter is given in Table 7.2.

Table 7.2: Summary of the studies on speaker verification.

1.	An approach is proposed for the extraction and representation of speaker-specific prosodic features from speech signal.
2.	Prosodic features derived using the proposed method is modeled by capturing their distribution.
3.	The role of excitation, vocal tract and prosodic characteristics represented at subsegmental, segmental and suprasegemntal levels, respectively, are examined for the task of speaker verification for different size of training data.
4.	Demonstrated that combining evidence from prosodic features and spectral features improve the overall performance of the speaker verification system.

CHAPTER 8

SUMMARY AND CONCLUSIONS

8.1 SUMMARY OF THE WORK

In this thesis, we have focused on the extraction and representation of language and speaker-specific features at multiple levels of speech signal, for the purpose of language/speaker recognition. We hypothesize that features derived from the speech signal reflect both language as well as speaker characteristics, and it is difficult to separate the language and speaker parts in these features. For recognizing a language or a speaker, different cues are available in speech. Languages differ in acoustic-phonetics, prosody, phonotactics and vocabulary. Speaker characteristics vary due to difference in the excitation source, vocal tract dimensions, prosody and idiolect. 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/speaker recognition 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 artificial neural network. The evidence obtained from different levels was later combined to arrive at a decision.

For automatic language identification, language-specific features at three different levels of speech signal, namely, frame, syllabic and multisyllabic levels are explored. At the frame level, spectral features and source features derived from lower span of speech signal ($\leq$ 20 msec) were used for language identification. Features such as weighted

linear prediction cepstral coefficients (WLPCC), linear prediction (LP) residual and phase of LP residual are examined. The presence of language-specific information in these features was demonstrated using four Indian languages. It was observed that the frame level features are dominated by the physiological characteristics of the speaker. Multiple speaker models and the N-best scoring are proposed to reduce the effect of speaker variability on language identification. Effectiveness of the proposed approach is demonstrated using OGI database.

To represent variations in realization of syllables among languages, features are represented at the syllable level. Spectral features corresponding to CV type of syllables has been used in this study. Regions corresponding to CV type of syllables are automatically identified with the help of vowel onset points, eliminating the need for any explicit segmentation and transcription. It was observed that more language-specific evidences are obtained while representing spectral features at the syllabic level compared to the frame level representation, as demonstrated for eleven languages in OGI database.

As a first step towards using prosodic features for language identification, segment boundaries from manually labeled corpora were used. The study on three Indian languages revealed the usefulness of prosodic features for language identification. To extract prosodic features automatically from the speech signal, a VOP based approach is proposed. In this approach, syllable has been used as the basic unit for extracting prosodic features. This approach do not require automatic speech recognizer for the extraction of prosodic features, but still gives association of prosodic features with the corresponding syllable sequence. This is done with the knowledge of VOPs, detected automatically from the Hilbert envelope of the LP residual of the speech signal. The region between two successive VOPs is considered as a syllabic region, and parameters are derived to represent duration, dynamics of F_0 contour and energy variations corresponding to each region.

The prosodic characteristics of languages differ in terms of rhythm, intonation

and stress. These characteristics are manifested in acoustic speech signal in terms of duration, F_0 contour and energy. It is hypothesized that the notion of rhythm and stress are generated from the succession of syllables. Therefore prosodic features should be represented at multisyllabic level to highlight the language-specific part. This was approximated to a level of three consecutive syllables, and the prosodic feature vectors were obtained by concatenating the features from them. The effectiveness of prosodic features for language discrimination is demonstrated using OGI database.

Features such as excitation, vocal tract and prosody, represented at multiple levels of speech signal, namely, subsegmental, segmental and suprasegmental, were studied in the context of speaker verification. The vocal tract system and excitation source characteristics reflects the physiological differences among speakers. Prosodic differences are attributed by both physiological and learned aspects of speakers. In spite of the difference in the spoken message and language, speaker characteristics are present in the dynamics of F_0 contour and energy. It was hypothesized that the distribution of prosodic feature vectors are specific to a speaker. The study on speaker verification using NIST 2003 extended data demonstrated the potential of prosodic features for modeling the speaker characteristics. Performance of the prosody based speaker verification system is significant, especially for cases having large training data. Also the prosodic features derived from F_0 contour and energy, are less affected by the variations in channel characteristics and noise. We have demonstrated that, the performance of speaker verification system can be improved by combining the evidence from spectral and prosodic features.

8.2 KEY IDEAS PRESENTED IN THE THESIS

The following are the key ideas presented in this thesis:

1. Differences among languages and speakers are manifested at multiple levels of speech. It is represented using features derived from appropriate spans of acoustic speech signal.

2. An approach based on multiple speaker-specific models and N-best scoring is proposed for minimizing the effect of speaker variability in spectral-based LID.
3. A new method is proposed for extracting prosodic features directly from the speech signal, for language and speaker recognition. In this method, syllable has been treated as the basic unit for deriving prosodic features. Instead of using automatic speech recognizer for obtaining the segment boundaries, the locations of vowel onset points are used for segmenting continuous speech into syllable-like regions.
4. Prosodic features are effective for both language identification and speaker verification. But the level of representation is crucial for highlighting the language part and the speaker part. Language-specific prosody is better represented at a larger span of speech compared to that of speaker-specific prosody.
5. Prosody-based speaker evidence can enhance the robustness of the speaker verification system in combination with spectral-based evidence.

8.3 SCOPE FOR FUTURE WORK

The following are some directions for future work:

Multilevel implicit features used in this study have shown to carry both language and speaker-specific characteristics. Therefore it is possible to combine language and speaker recognition tasks in a unified framework. This was not explored in the present study due to the nonavailability of suitable database (in which same speaker speaking multiple languages). The framework proposed for syllable-based language identification may be useful for this. In this framework, the speaker-specific model giving the highest likelihood will indicate the speaker identity, whereas the language having highest N-best score will give evidence for the language identity. Further studies may be carried out to enhance these evidence.

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 [92]. 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. This possibility may be studied using a database having large size of speech data in multiple languages.

The present study uses rhythm represented in terms of durational features to discriminate languages. In natural conversation, duration characteristics also convey some speaker-specific evidence. The usefulness of duration features such as syllable duration and pause characteristics may be explored along with other prosodic features for speaker verification task.

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.

APPENDIX A

AUTOASSOCIATIVE NEURAL NETWORK FOR PATTERN RECOGNITION TASKS

Autoassociative neural network (AANN) is a special class of FFNN architecture, having some interesting properties which can be exploited for some pattern recognition tasks. Studies on AANN models have been reported extensively in the literature [108]. In an AANN model the input and output layers have the same number of units, and all these units are linear. It was shown that a three layer network with a middle compression layer (# units < # units in the input/output), do not have any specific advantage over normal linear PCA analysis. Even nonlinear units in the compression layer do not provide any advantage for obtaining a better estimation of PCA components [109]. In fact, the nonlinearity may result in suboptimal solutions due to local minima problems in the optimization of network weights using the backpropagation learning. Even using more than three layers also do not provide any significant advantage, if the compression layer is the last but one layer in the network [108].

However, if the AANN has five or more layers, with nonlinear compression layer in the middle with at least two layers after the compression layer, then such a network has interesting properties which can be exploited for many pattern recognition tasks [66]. We focus on two such properties, namely, distribution capturing of the feature vectors, and capturing the second and higher order relations among the samples in the input data. If the input data is speech, say, like a frame of speech, and the network is trained with frames shifted by one sample, then the network captures primarily the significant correlation among the samples, usually the second order correlations. When we give as input and output a signal with significant correlation removed, like in the linear prediction (LP) residual, then the network can be used to extract features in

the higher order relations among the samples, and thus the network can be used as a feature extractor. If feature vectors belonging to a single class are used for training, then the network can be used to capture the distribution of the feature vectors in the feature space. The distribution capturing ability of the AANN model, is demonstrated through an experimental study. The details of the study given here is taken from [110].

Let us consider the five layer AANN model shown in Fig. A.1, which has three hidden layers. The processing units in the first and third hidden layer are nonlinear, and the units in the second compression/hidden layer can be linear or nonlinear. As the error between the actual and the desired output vectors is minimized, the cluster of points in the input space determines the shape of the hypersurface obtained by the projection onto the lower dimensional space. Fig. A.2 shows the space spanned by the 1-dimensional compression layer for the 2-dimensional data shown in Fig. A.2(a) for two different network structures. The structures of the two networks are *2L 4N 1N 4N 2L* and *2L 10N 1N 10N 2L*, where L denotes a linear unit, and N denotes a nonlinear unit. The numbers denote the number of units for each layer. The nonlinear output function for each unit is $\tanh(\lambda x)$, where λ is arbitrarily chosen to be equal to 0.66. The networks were trained using backpropagation algorithm [61, 62]. The solid lines shown in Figs. A.2(b) and A.2(c) indicate mapping of the given input points due to the 1-dimensional compression layer. The second network having more hidden units seems to represent the data through the 1-dimensional compression better, as shown in Fig. A.2(c), compared to Fig. A.2(b) for the network with fewer units in the hidden layers. Thus, one can say that the AANN captures the distribution of the data points depending on the constraints imposed by the structure of the network, just as the number of mixtures and Gaussian functions do in the case of GMMs.

In order to visualize the distribution better, one can plot the error for each input data point in the form of some probability surface as shown in Fig. A.3 for the cases in Figs. A.2(b) and A.2(c) [110]. The error E_i for the data point i in the input space is plotted as $f_i = e^{(-E_i/\alpha)}$, where α is a constant. Note that f_i is not strictly a probability

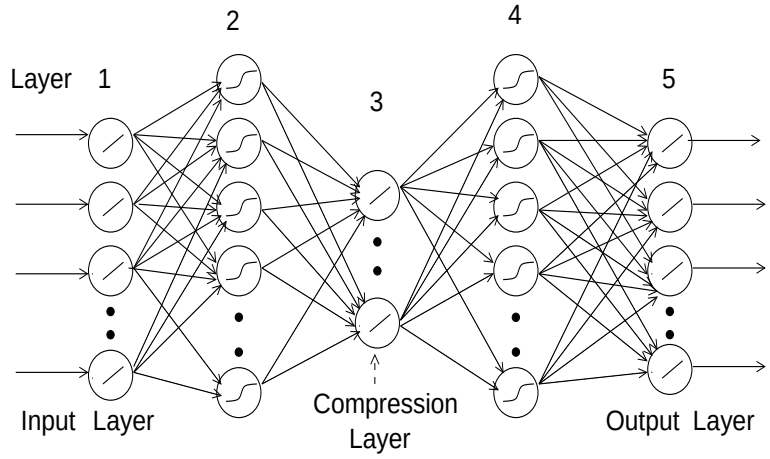
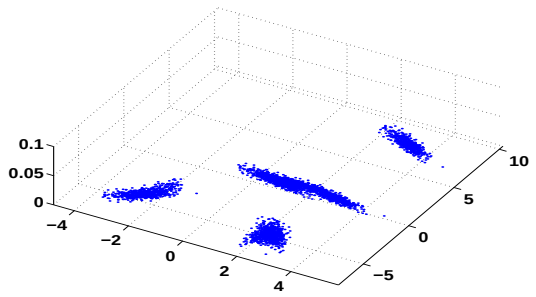
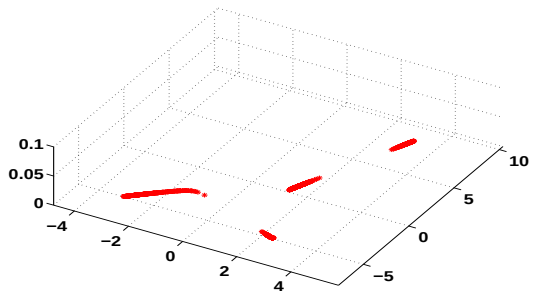


Fig. A.1: A five layer AANN model.

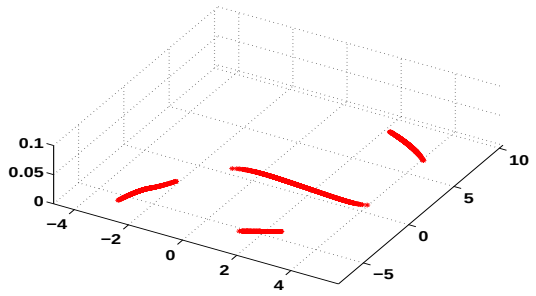
density function, but we call the resulting surface as *probability surface*. The plot of the probability surface shows a large amplitude for smaller error E_i, indicating better match of the network for that data point. The constraints imposed by the network can be seen by the shape the error surface takes in both the cases. One can use the probability surface to study the characteristics of the distribution of the input data captured by the network.



(a)

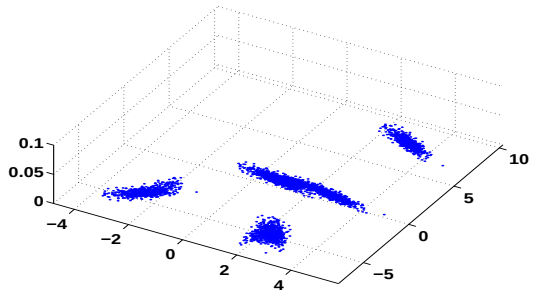


(b)

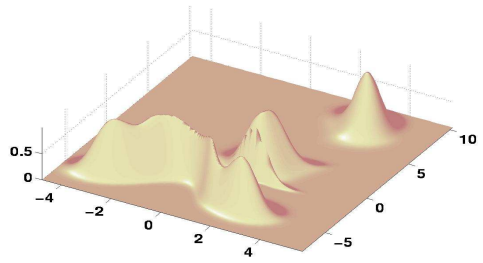


(c)

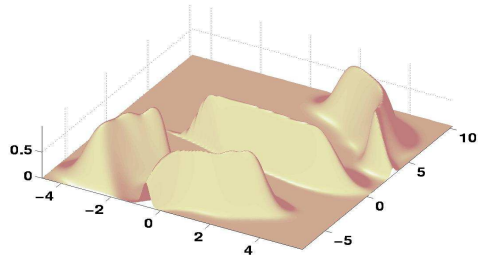
Fig. A.2: (a) Artificial 2-dimensional data. (b) 2-dimensional output of AANN model with the structure *2L 4N 1N 4N 2L* (c) 2-dimensional output of AANN model with the structure *2L 10N 1N 10N 2L*.



(a)



(b)



(c)

Fig. A.3: Probability surfaces realized by two different network structures (b) *2L 4N 1N 4N 2L* (c) *2L 10N 1N 10N 2L*, for the 2-dimensional data shown in (a).

APPENDIX B

WEIGHTED LINEAR PREDICTION CEPSTRAL COEFFICIENTS ANALYSIS

In this appendix, we present an algorithm used for extraction of Weighted Linear Prediction Cepstral Coefficient (WLPCC) representation from speech signal. This algorithm is taken from [42].

The digitized speech signal s(n), is preemphasized by implementing the following difference equation:

$$\tilde{s}(n) = s(n) - 0.95 * s(n-1) \tag{B.1}$$

Let N be the frame size and M be the separation between adjacent frames specified in number of speech signal samples. Then the lth frame of speech signal is denoted by:

$$x_l(n) = \tilde{s}(Ml + n), \quad n = 0, 1, \ldots, N-1, l = 0, 1, \ldots, L-1 \tag{B.2}$$

where L is the number of frames in the entire speech signal. Each frame is windowed using a Hamming window as given below:

$$\tilde{x}_l(n) = x_l(n)w(n), \quad 0 \leq n \leq N-1 \tag{B.3}$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{B.4}$$

Each frame of windowed signal is autocorrelated as follows:

$$\tilde{r}_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), \quad m = 0, 1, \ldots, p \tag{B.5}$$

where p is the order of the linear prediction analysis. Linear prediction coefficients are derived from autocorrelation coefficients using Durbin's method given below. The subscript l is omitted for convenience.

$$E^{(0)} = r(0) \tag{B.6}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|)}{E^{(i-1)}}, \quad 1 \leq i \leq p \tag{B.7}$$

$$\alpha_i^{(i)} = k_i \tag{B.8}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \tag{B.9}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{B.10}$$

The above equations are solved recursively for $i = 1, 2, \ldots, p$, and the final solution gives the linear prediction coefficients a_m, as follows:

$$a_m = \alpha_m^{(p)}, \quad 1 \leq m \leq p \tag{B.11}$$

The cepstral coefficients, c_m, are derived from linear prediction coefficients by recursion of the following equation:

$$c_0 = \ln \sigma^2 \tag{B.12}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p \tag{B.13}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad p \leq m \leq Q \tag{B.14}$$

where σ^2 is the gain term in linear prediction analysis and Q is the number of cepstral coefficients. The cepstral coefficients are weighted using a bandpass filter in the cepstral domain as given below to obtain weighted cepstral coefficients $\hat{c}_m$.

$$\hat{c}_m = w_m c_m \tag{B.15}$$

where

$$w_m = 1 + \frac{Q}{2} \sin\left(\frac{m\pi}{Q}\right), \quad 1 \leq m \leq Q \tag{B.16}$$

APPENDIX C

MEL-FREQUENCY CEPSTRAL COEFFICIENTS ANALYSIS

In this appendix, we present an algorithm used for extraction of Mel-Frequency Cepstral Coefficient (MFCC) representation from speech signal. This algorithm is taken from [111].

The human ear resolves frequencies non-linearity across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner may improve recognition performance. A popular alternative to linear prediction based analysis is therefore filter bank analysis since this provides a much more straightforward route for obtaining the desired non-linear frequency resolution. Mel-Frequency Cepstral Coefficients (MFCCs) are calculated using the following steps:

The output speech sampling rate of the analog-to-digital converter is assumed to be 8 kHz. The speech samples are divided into overlapping frames. The frame length is 20 msec (160 samples) and the frame rate is 5 msec (40 samples).

Each frame is windowed using the Hamming window function:

$$S_w(n) = \left\{ 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \right\} * s(n), \quad 0 \leq n \leq N-1. \tag{C.17}$$

Here N is the frame length and $s(n)$ and $s_w(n)$ are the input and output of the windowing block, respectively.

A Fast Fourier Transform (FFT) of block length L (L=256) is used to compute the magnitude spectrum for each windowed frame. The first 129 bins from the magnitude spectrum are retained for further processing.

$$b_k = \left| \sum_{n=0}^{L-1} S_w(n) e^{jnk\frac{2\pi}{L}} \right|, \quad k = 0, 1, \ldots, L-1. \tag{C.18}$$

Here $S_w(n)$ is the input to the FFT block, L is the block length (256), and b_k is the absolute value of the resulting complex vector.

The power spectrum is computed by taking the square of the magnitude spectrum. Complete frequency range (0-4000 Hz) is used for computing the Mel-warped spectrum. This band is divided into 23 channels equidistant in Mel-frequency scale.

$$m(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{C.19}$$

$$f_{c_i} = m^{-1}\left(i * m\left(\frac{f_s/2}{23+1}\right)\right), \quad i = 1, 2, \ldots, 23 \tag{C.20}$$

$$c_i = floor\left(\frac{f_{c_i}}{f_s} * L\right) \tag{C.21}$$

where $floor(.)$ stands for rounding downwards the nearest integer.

The output of the Mel-filter is the weighted sum of the FFT power spectrum values (b_i) in each band. Triangular, half-overlapped windowing is used as follows:

$$M_k = \sum_{j=c_{k-1}}^{c_k} \frac{j - c_{k-1}}{c_k - c_{k-1}} b_i + \sum_{c_i}^{c_{i+1}} \frac{c_{k-1} - j}{c_{k+1} - c_k} \tag{C.22}$$

where $k = 1, 2, \ldots, 23$ and M_k is the kth MFCC coefficient.

c_0 and c_{24} denote the FFT bin indices corresponding to the starting frequency and half of the sampling frequency, respectively,

$$c_0 = 0 \tag{C.23}$$

$$c_{24} = floor\left(\frac{f_s/2}{f_s} * L\right) = L/2 \tag{C.24}$$

REFERENCES

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

[2] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE signal Processing Magazine*, vol. 11, pp. 33–41, Oct. 1994.

[3] H. J. M. Steeneken, "Editorial - Multilingual interoperability in speech technology," *Speech Communication*, vol. 35, pp. 1–3, 2001.

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

[5] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, pp. 1297–1313, Aug. 2000.

[6] 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.

[7] 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.

[8] 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.

[9] L. P. Heck, *Integrating high-level information for robust speaker recognition.* http: www.cslp.jhu.edu/ws2002/groups/supersid: In John Hopkins University workshop on SuperSID, Baltimore, Maryland, Jul. 2002.

[10] G. Doddington, "Speaker recognition based on idiolectic differences between speakers," in *Proc. EUROSPEECH*, (Aalborg, Denmark), pp. 2521–2524, Sep. 2001.

[11] M. Sigmund, *Speaker Recognition - Identifying People by Their Voices.* Master's thesis, Institute of radio electronics, Brno University of Technology, Czech Republic 2000.

[12] T. J. Hazen, V. W. Zue, "Segment-based automatic language identification system," *J. Acoust. Soc. Am.*, vol. 101, pp. 2323–2331, Apr. 1997.

[13] M. Sugiyama, "Automatic language recognition using acoustic features," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, (Toronto, Canada), pp. 813–816, May 1991.

[14] S. Nakagawa, Y. Ueda and T. Seino, "Speaker-independent text-independent language identification by HMM," in *Int. Conf. Spoken Language Processing*, vol. 2, (Alberta, Canada), pp. 1011–1014, Oct. 1992.

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

[16] P. A. Torres-Carassquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Int. Conf. Spoken Language Processing*, (Orlando, Florida), pp. 89–92, Sep. 2002.

[17] K. P. Li, "Automatic language identification using syllabic spectral features," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Adelaide, Australia), pp. 297–300, Apr. 1994.

[18] A. E. Thyme-Gobbel, and S. E. Hutchins, "On using prosodic cues in automatic language identification," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, (Philadelphia, PA, USA), pp. 1768–1772, Oct. 1996.

[19] F. Cummins, F. Gers, and J. Schmidhuber, "Language identification from prosody without using explicit features," in *Proc. EUROSPEECH*, (Budapest, Hungary), pp. 371–374, Sep. 1999.

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

[21] J. L. Rouas, J. Farinas, F. Pellegrina, and R. A. Obrech, "Modeling prosody for language identification on read and spontaneous speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Hongkong), pp. 40–43, May 2003.

[22] G. Adami and H. Hermansky, "Segmantation of speech for speaker and language recognition," in *Proc. EUROSPEECH*, (Geneva), pp. 841–844, Sep. 2003.

[23] C. Lin and H. Wang, "Language identification using pitch contour information," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Philadelphia, USA), pp. 601–604, Apr. 2005.

[24] L. F. Lamel and J. L. Gauvain, "Language identification using phone-based acoustic liklihoods," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Adelaide, Australia), pp. 40–43, May 1994.

[25] T. J. Hazen and V. W. Zue, "Automatic language identification using a segment-based approach," in *Proc. EUROSPEECH*, vol. 2, (Berlin, Germany), pp. 1303–1306, Sep. 1993.

[26] K. M. Berkling, T. Arai, and E. Barnard, "Analysis of phoneme-based features for language identification," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Seattle, Washington), pp. 289–292, Apr. 1998.

[27] A. K. Sai Jayaram, V. Ramasubramanian, and T. V Sreenivas, "Language identification using parallel subword recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Hongkong), pp. 32–35, Apr. 2003.

[28] T. Nagarajan and H. A. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Montreal, Canada), pp. 401–404, May 2004.

[29] S. Kadambe and J. Hieronymus, "Language identification with phonological and lexical models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 5, (Detroit, USA), pp. 3507–3511, May 1995.

[30] T. Schultz, I. Rogina, and A. Waibel, "LVCSR-based language identification," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, (Atlanta, GA, USA), pp. 781–784, May 1996.

[31] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newman, "Automatic language identification using large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, (Atlanta, GA, USA), pp. 785–788, May 1996.

[32] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437–1462, Sep. 1997.

[33] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18–32, Oct. 1994.

[34] J. M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp. 42–48, Jan. 1990.

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

[36] S. Guruprasad, *Exploring Features and Scoring Methods for Speaker Verification.* M. S thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Oct. 2004.

[37] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[38] B. S. Atal, "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification ," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, Jun. 1974.

[39] S. Furui, "Cepstral analysis technique for automatic speaker verification ," *IEEE Trans. on Speech and Audio Processing*, vol. 29, pp. 254–272, Apr. 1981.

[40] S. P. Kishore, *Speaker Verification using Autoassociative Neural Network Models.* MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Dec. 2000.

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

[42] L. R. Rabiner and B. -H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs New Jersey: PTR Prentice-Hall, 1993.

[43] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, 1995.

[44] M. Faundez-Zanuy and D. Rodriguez-Porcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, (Australia), pp. 121–124, Nov-Dec. 1998.

[45] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Salt Lake City, Utah, USA), pp. 409–412, May 2001.

[46] G. Fant, "Glottal flow: models and interaction," *Journal of Phonetics*, vol. 14, pp. 393–399, 1986.

[47] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 569–586, Sep. 1999.

[48] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Proc. EUROSPEECH*, vol. 3, (Rhodes, Greece), pp. 1391–1394, Sep. 1997.

[49] M. K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker variation," in *Proc. Int. Conf. Spoken Language Processing*, vol. 7, (Sydney, Australia), pp. 3189–3192, Dec. 1998.

[50] NIST 2001 speaker recognition evaluation website: *http://www.nist.gov/speech/tests/spk/2001/index.htm.*

[51] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 4, (Hong kong, China), pp. 784–787, Apr. 2003.

[52] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 4, (Hong kong, China), pp. 788–791, Apr. 2003.

[53] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 4, (Hong kong, China), pp. 792–795, Apr. 2003.

[54] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosody for speaker recognition," *Speech Communication*, vol. 46, pp. 455–472, 2005.

[55] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proc. EUROSPEECH*, vol. 1, (Geneva), pp. 2017–2020, Sep. 2003.

[56] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and V. R. R. Gadde, "Speaker recognition using prosodic and lexical features," in *Proceedings of the IEEE speech recognition and understanding workshop* , (St. Thomas, U. S. Virgin Islands), pp. 19–24, Sept. 2003.

[57] W. Andrews, M. Kohler, J. Campell, J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, (Orlando, Florida), pp. 149–152, May 2002.

[58] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetc speaker recognition using maximum likelihood binary decision tree models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 4, (Hong kong, China), pp. 796–799, Apr. 2003.

[59] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell, "Conditional pronounciation modeling in speaker detection," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 4, (Hong kong, China), pp. 804–807, Apr. 2003.

[60] C. M. Bishop, *Neural Networks for Pattern Recognition.* New York: Oxford University Press Inc., 1995.

[61] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New Jersey: Prentice-Hall International, 1999.

[62] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice-Hall of India, 1999.

[63] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, 2002.

[64] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, (Albuquerque, NM, USA), pp. 1361–1364, Apr. 1990.

[65] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1972.

[66] B. Yegnanarayana, S. V. Gangashetty, and S. Palanivel, *Autoassociative Neural Network Models for Pattern Recognition Tasks in Speech and Image.* Soft Computing Approach to Pattern Recognition and Image Processing: World Scientific Publishing Co.: in A. Gosh and S. K. pal (Eds.), 2002.

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

[68] K. Sri. Rama Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, pp. 52–55, Jan. 2006.

[69] L. Mary, K. S. Murty, S. R. Mahadeva Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. Odyssey-2004*, (Toledo, Spain), pp. 156–159, Jun. 2004.

[70] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, (Banf, Alberta, Canada), pp. 895–898, Oct. 1992.

[71] A. Nayeemulla 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.

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

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

[74] 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.

[75] S. R. Mahadeva 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.

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

[77] S. R. Mahadeva Prasanna, *Event-based Analysis of Speech.* Ph. D thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Mar. 2004.

[78] 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.

[79] M. Kometsu, K. Mori, T. Arai, and Y. Murahara, "Human language identification with reduced segmental information: Comparison between Monolinguals and Bilinguals," in *Proc. EUROSPEECH*, vol. 1, (Scandinavia), pp. 149–152, Sep. 2001.

[80] K. Nagamma Reddy, *Speech Technology: Issues and Implications in Indian Languages.* Thiruvananthapuram, India: Dravidian Linguistics Association, International School of Dravidian Linguistics, 2000.

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

[82] 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.

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

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

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

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

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

[88] A. Chopde, "Itrans Indian language transliteration package version 5.3 source," *http://www.aczoom.com/itrans/.*

[89] K. Sreenivasa Rao, *Acquisition and Incorporation of Prosody Knowledge for Speech Systems in Indian Languages.* Ph. D thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, May 2005.

[90] E. Klabbers, *Segmental and prosodic improvements to speech generation.* Ph. D thesis, Eindhoven University of Technology, Netherlands, Mar. 2000.

[91] M. Atterer and D. R. Ladd, "On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German," *Journal of Phonetics*, vol. 32, pp. 177–197, 2004.

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

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

[94] 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.

[95] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-23, pp. 562–570, Dec. 1975.

[96] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[97] M. Chaitanya, *Single Channel Speech Enhancement.* M. S thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Apr. 2005.

[98] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, pp. 55–105, 1999.

[99] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[100] B. Z. Keller, "F0 and intensity distribution of Marsec speakers: Types of prosody," in *Proc. Int. Conf. Non-linear Speech Processing*, (Barcelona, Spain), pp. 19–22, Apr. 2005.

[101] SuperSID Project Website, "http://www.clsp.jhu.edu/ws2002/groups/supersid," 2002.

[102] Y. Xu and S. Xuejing, "Maximum speed of pitch change and how it may relate to speech," *J. Acoust. Soc. Am.*, vol. 111, pp. 1399–1413, Mar. 2002.

[103] NIST 2003 speaker recognition evaluation website: *http://www.nist.gov/speech/tests/spk/2003/.*

[104] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, vol. 4, (Rhodes, Greece), pp. 1895–1898, Sep. 1997.

[105] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. EUROSPEECH*, vol. 4, (Rhodes, Greece), pp. 963–966, Sep. 1997.

[106] M. C. R. Auckenthaler and H. L. Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[107] J. Mariethoz and S. Bengio, "A unified framework for score normalization techniques applied to text-independent speaker verification," *IEEE Signal Processing Letters*, vol. 12, pp. 532–535, Jul. 2005.

[108] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks.* New York: John Wiley & Sons, Inc., 1996.

[109] Bourlard and Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, pp. 291–294, 1988.

[110] B. Yegnanarayana and S. P. Kishore, "AANN-An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[111] Steve Young et al., *The HTK Book (for HTK Version 2.2).* Cambridge: Entropic Ltd., 1999.

List of Publications

In Refereed Journals

1. Leena Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," revised and submitted to *Speech Communication.*

In International Conferences

1. Leena Mary and B. Yegnanarayana, "Prosodic features for speaker verification," in *Proc. Interspeech - Int. Conf. Spoken Language Processing (Pittsburgh PA, USA)*, pp. 917–920, Sep. 2006.
2. Leena Mary and B. Yegnanarayana, "Consonant-vowel based features for language identification," in *Proc. Int. Conf. Natural Language Processing (IIT Kanpur, India)*, pp. 103–106, Dec. 2005.
3. Leena Mary, K. Srinivasa Rao, and B. Yegnanarayana, "Neural network classifiers for language identification using phonotactic and prosodic features," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing (Chennai, India)*, pp. 404–408, Jan. 2005.
4. Leena Mary, K. Srirama Murty, S. R. Mahadeva Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in *Proc. Odyssey-2004 (Toledo, Spain)*, Jun. 2004.
5. Leena Mary, K. Srinivasa Rao, S. V. Gangashetty, and B. Yegnanarayana, "Neural network models for capturing duration and intonation knowledge for speaker and language identification," in *Proc. Int. Conf. Cognitive and Neural Systems (Boston, USA)*, pp. 34, May 2004.
6. Leena Mary and B. Yegnanarayana, "Autoassociative neural network models for language identification," in *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing (Chennai, India)*, pp. 317–320, Jan. 2004.

CURRICULUM VITAE

1. **NAME**: Leena Mary
2. **DATE OF BIRTH**: 02 March 1967
3. **EDUCATIONAL QUALIFICATIONS**:
   - 1988 Bachelor of Engineering (B.E.)
   - 1990 Master of Technology (M.Tech.)
   - 2006 Doctor of Philosophy (PhD)
4. **PERMANENT ADDRESS**:

   Puthuchira

   Kumplampoika P. O

   Pathanamthitta (Dist.)

   Kerala, India - 689661

   Ph: +91-0473-5252208

DOCTORAL COMMITTEE

1. **CHAIRPERSON**: Prof. T. A. Gonsalves
2. **GUIDE**: Prof. B. Yegnanarayana
3. **MEMBERS**:
   - Prof. S. Mohan
   - Dr. K. S. Swaroop
   - Dr. Deepak Khemani
   - Dr. C. Chandra Sekhar