SPRINGER BRIEFS IN ELECTRICAL AND COMPUTER ENGINEERING · SPEECH TECHNOLOGY

K. Sreenivasa Rao
Dipanjan Nandi

Language Identification Using Excitation Source Features

Springer

SpringerBriefs in Electrical and Computer Engineering

Speech Technology

Series editor

Amy Neustein, Fort Lee, NJ, USA

Editor's Note

The authors of this series have been hand-selected. They comprise some of the most outstanding scientists—drawn from academia and private industry—whose research is marked by its novelty, applicability, and practicality in providing broad based speech solutions. The SpringerBriefs in Speech Technology series provides the latest findings in speech technology gleaned from comprehensive literature reviews and *empirical investigations* that are performed in both laboratory and *real life* settings. Some of the topics covered in this series include the presentation of real life commercial deployment of spoken dialog systems, contemporary methods of speech parameterization, developments in information security for automated speech, forensic speaker recognition, use of sophisticated speech analytics in call centers, and an exploration of new methods of soft computing for improving human-computer interaction. Those in academia, the private sector, the self service industry, law enforcement, and government intelligence, are among the principal audience for this series, which is designed to serve as an important and essential reference guide for speech developers, system designers, speech engineers, linguists and others. In particular, a major audience of readers will consist of researchers and technical experts in the automated call center industry where speech processing is a key component to the functioning of customer care contact centers.

Amy Neustein, Ph.D., serves as Editor-in-Chief of the International Journal of Speech Technology (Springer). She edited the recently published book "Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics" (Springer 2010), and serves as quest columnist on speech processing for Womensenews. Dr. Neustein is Founder and CEO of Linguistic Technology Systems, a NJ-based think tank for intelligent design of advanced natural language based emotion-detection software to improve human response in monitoring recorded conversations of terror suspects and helpline calls. Dr. Neustein's work appears in the peer review literature and in industry and mass media publications. Her academic books, which cover a range of political, social and legal topics, have been cited in the Chronicles of Higher Education, and have won her a pro Humanitate Literary Award. She serves on the visiting faculty of the National Judicial College and as a plenary speaker at conferences in artificial intelligence and computing. Dr. Neustein is a member of MIR (machine intelligence research) Labs, which does advanced work in computer technology to assist underdeveloped countries in improving their ability to cope with famine, disease/illness, and political and social affliction. She is a founding member of the New York City Speech Processing Consortium, a newly formed group of NY-based companies, publishing houses, and researchers dedicated to advancing speech technology research and development.

More information about this series at http://www.springer.com/series/10043

K. Sreenivasa Rao · Dipanjan Nandi

Language Identification Using Excitation Source Features

Springer

K. Sreenivasa Rao
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal
India

Dipanjan Nandi
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal
India

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISSN 2191-737X ISSN 2191-7388 (electronic)
SpringerBriefs in Speech Technology
ISBN 978-3-319-17724-3 ISBN 978-3-319-17725-0 (eBook)
DOI 10.1007/978-3-319-17725-0

Library of Congress Control Number: 2015936375

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Language identification (LID) is a process of determining the language from the uttered speech. Most LID studies have been carried out using vocal tract and prosodic features. However, the characteristics of excitation source have not been explored for LID study. In this book, implicit and explicit features of excitation source have been explored for language discrimination task. Linear prediction (LP) residual signal is used for representing excitation source signal. The implicit relations among the raw LP residual samples, its magnitude, and phase components are explored to capture the language-specific excitation source information. The proposed implicit features consist of raw LP residual samples, its magnitude and phase components at three different levels: (i) sub-segmental level (within a glottal cycle or pitch cycle), (ii) segmental level (within 2–3 successive glottal cycles), and (iii) supra-segmental level (across 50 glottal cycles). These features capture the implicit language-specific phonotactic constraints embedded in excitation source signal. Evidences obtained from each level are combined to derive complete implicit features of excitation source for LID task.

In addition to implicit features of excitation source, LP residual signal has also been parameterized at sub-segmental, segmental, and supra-segmental levels to capture the language-specific phonotactic information. At sub-segmental level, the characteristics of a single glottal pulse have been modeled using glottal flow derivative (GFD) parameters of LP residual signal. Residual mel frequency cepstral coefficients (RMFCC) and mel power difference of spectrum in sub-band (MPDSS) features are explored to derive language-specific excitation source information at segmental level. At supra-segmental level, temporal variations of pitch, epoch strength, and epoch sharpness are explored for capturing language-specific supra-segmental level source information. Evidences from each level are combined to capture complete parametric representation of excitation source. Further, evidences obtained from implicit and parametric features are combined to acquire overall language-specific excitation source information.

The nonoverlapping language-specific information present in excitation source and vocal tract features has also been investigated in this book. The robustness of proposed excitation source features has been examined by varying (i) amount of

training data, (ii) length of test samples, and (iii) background noise characteristics. From experimental studies, it has been observed that the excitation source is more robust, compared to vocal tract features for LID task.

This book is mainly intended for researchers working on language identification area. The book is also useful for young researchers who want to pursue research in speech processing with an emphasis on excitation source features. Hence, this may be recommended as a text or reference book for the postgraduate level advanced speech processing course. The book has been organized as follows:

Chapter 1 introduces the basic concept of language identification (LID) and the various features used in LID. The application of LID system has been demonstrated. Chapter 2 provides a review of the methods reported in prior works of LID. Chapter 3 discusses the implicit features of excitation source for language recognition task. Chapter 4 explores the parametric features of excitation source for language discrimination study. Chapter 5 investigates the robustness of excitation source features in the context of language identification. Chapter 6 provides a brief summary and conclusion of the book with a glimpse toward the scope for possible future work.

We would especially like to thank all professors of School of Information and Technology, IIT Kharagpur for their moral encouragement and technical discussions during the course of editing and organization of the book. Special thanks to our colleagues at Indian Institute of Technology, Kharagpur, India for their cooperation to carry out the work. We are grateful to our parents and family members for their constant support and encouragement. Finally, we thank all our friends and well wishers.

K. Sreenivasa Rao
Dipanjan Nandi

Contents

Acronyms

AIR	All India Radio
ANN	Artificial Neural Networks
BP	Block Processing
CMS	Cepstral Mean Subtraction
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
ESC	Epoch Strength Contour
ESNC	Epoch Sharpness Contour
GCI	Glottal Closing Instants
GFD	Glottal Flow Derivative
GMM	Gaussian Mixture Models
GOI	Glottal Opening Instants
GVV	Glottal Volume Velocity
HE	Hilbert Envelope
HMM	Hidden Markov Model
Hz.	Hertz
IFT	Inverse Fourier Transform
IITKGP-MLILSC	Indian Institute of Technology KharaGPur-Multi-Lingual Indian Language Speech Corpus
kHz.	Kilo Hertz
LF	Liljencrants-Fant
LID	Language Identification
LP	Linear Prediction
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
LPR	Linear Prediction Residual
MFCC	Mel Frequency Cepstral Coefficients
MPDSS	Mel Power Difference of Spectrum in Sub-bands
ms	Milli Seconds
OGI-MLTS	Oregon Graduate Institute database Multi-Language Telephone-based Speech

PC	Pitch Contour
RMFCC	Residual Mel Frequency Cepstral Coefficient
RP	Residual Phase
seg	Segmental
SNR	Signal-to-Noise Ratio
sub	Sub-segmental
supra	Supra-segmental
SVM	Support Vector Machines
VQ	Vector Quantization
ZFF	Zero Frequency Filter
ZFFS	Zero Frequency Filtered Signal

Chapter 1
Introduction

Abstract This chapter introduces the basic goal of language identification (LID) and its impacts on real-life applications. A brief overview of the basic features used for developing LID systems has been given and different categories of LID systems are also discussed here. Eventually, the primary issues in developing LID systems and the major contributions of this book towards solving those issues have been highlighted.

Keywords Language identification · Identification of Indian languages · Explicit language identification · Implicit language identification · Issues in language identification · Excitation source features for language identification · Speech features for language identification

1.1 Introduction

Speech is mainly intended to convey message among human beings. Speech signal along with the message information, also carries significant information about the speaker, language and the emotion associated to message as well as speaker. The primary objective of an automatic language identification (LID) system is to determine the language identity from the uttered speech. Due to several real-life applications of automatic LID systems such as, speech to speech translation systems, information retrieval from multilingual audio databases and multilingual speech recognition systems, it has become an active research problem in India and abroad. Indian languages belong to several language groups and sub-groups. The major two language groups are the Indo-Aryan languages spoken by 76.86 % of Indian citizens and the Dravidian languages spoken by 20.82 % Indians [1]. The Indo-Aryan languages are highly influenced by Sanskrit, whereas, the Dravidian languages has a history, independent of Sanskrit [1]. However, Dravidian languages Telugu and Malayalam are influenced by Sanskrit. Most of the languages in India share common set of phonemes and also follow similar grammatical structure. Therefore, deriving the language discriminative information in Indian context is really a challenging task.

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_1

Important applications of automatic LID system are as follows—(i) development of multilingual speech recognition system: Speech recognition is the task of identifying the sequence of sound units based on the information available in speech signal. Multilingual speech recognition system should have the phonetic information of all the languages. During evaluation if language identity of input speech utterance is unknown, then speech recognizers of all the languages have to run simultaneously, and hence the computational complexity increases by multi-fold. Therefore, it is necessary to know the language identity of input speech utterance. Language identification system is used at front-end to recognize the language of the input speech utterance. Thus, only one speech recognizer corresponding to a particular language will be activated. (ii) Development of multilingual audio search engine: Basic goal of audio retrieval is to find the occurrences of a given query in a large audio database. In case of information retrieval from multilingual audio database, whole database has to be searched if the language identity of input query is unknown. Hence, the searching time and computation complexity involved are very high. Language identification system can be used at front-end to determine the language identity of the input audio query. Thus, audio database corresponding to a particular language need to be searched, which reduces the computation complexity and searching time. (iii) Development of speech-to-speech translation system: In a multilingual country like India, it is difficult to convey messages through verbal communication between two persons with different language backgrounds. Hence, the message in source language is to be converted to the desired target language. For this task, it is necessary to recognize the source language first. Therefore, the LID system is used at the front-end of a speech-to-speech translation system to determine the language identity of input speech utterance.

1.2 Types of Language Identification Systems

The language identification systems can be of two types, namely, *explicit* and *implicit* LID systems. Research on *explicit* LID systems started a few decades back. However, the development of *implicit* LID systems is still an active area of research. A succinct description of these two LID systems are given below.

1.2.1 Explicit Language Identification System

In the *explicit* LID system, phonetic information is first extracted from the speech signal using a speech or phoneme recognizer, which gives a sequence of phoneme labels. The development of phoneme recognizer is most important stage in *explicit* LID system. The language models are used following the phoneme recognizer, to estimate the probability of occurrence of particular phoneme sequences within each of the target languages. To develop speech or phoneme recognizer with high accuracy

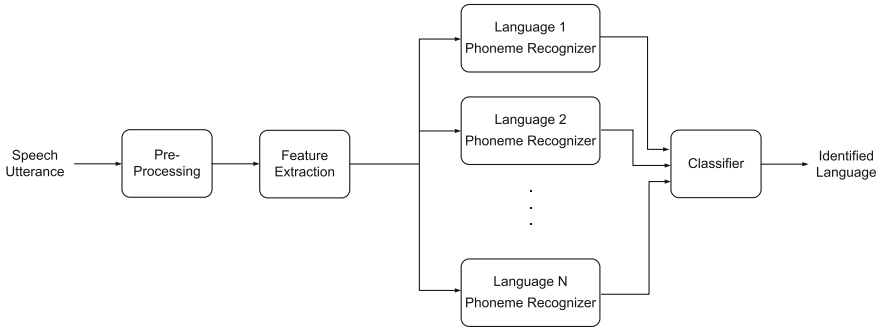

Fig. 1.1 Block diagram of *explicit* LID system

requires a large amount of transcribed or labelled speech data, which is the limitation of *explicit* LID system. Adding a new language into the *explicit* LID system is not a trivial task, because it requires labelled speech data for that new language. Therefore, developing an *explicit* LID system is a cumbersome task. However, the *explicit* LID system provides high LID accuracy. So, there is a trade-off between the accuracy and computational complexity to develop an *explicit* LID system. The block diagram of an *explicit* LID system is shown in Fig. 1.1.

1.2.2 Implicit Language Identification System

In case of *implicit* LID systems, labelled speech data and speech or phoneme recognizer are not required. The language-specific information is extracted directly from raw speech data by applying signal processing techniques. Non-linear modelling techniques such as, Gaussian mixture models (GMMs) and Neural networks (NNs) are used to build the language models. Adding a new language to the existing system is not a cumbersome task. The complexity of *implicit* LID systems is less compared to *explicit* systems. However, the performance of *explicit* LID systems is superior to *implicit* systems. So, the trade-off between performance and simplicity has become inevitable, if the number of languages under consideration is large. Our research focuses on the *implicit* LID systems. A generalized block diagram of an *implicit* LID system is shown in Fig. 1.2.

1.3 Features Used for Developing Speech Systems

Speech production system has two major components: vocal tract system and source of excitation. The characteristics of both vocal tract system and excitation source are

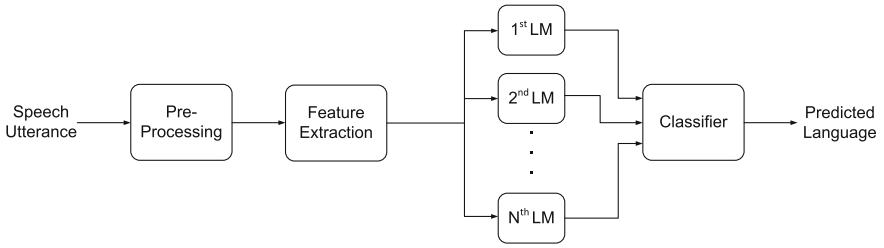

Fig. 1.2 Block diagram of *implicit* LID system

embedded in speech signal. In general, features used for most of the speech tasks are given below.

Spectral Features

During speech production, vocal tract system behaves like a time varying resonator or filter, and characterizes the variation of the vocal tract shapes in the form of resonances and antiresonances. The shape of the vocal tract resonator system is captured through spectral envelope of speech signal. Several parameterization techniques such as, linear prediction cepstral coefficients (LPCCs) and mel-frequency cepstral coefficients (MFCCs) are available for modeling vocal tract information. Spectral features are used for developing several speech based systems such as, speech recognition, speaker recognition, language recognition and emotion recognition systems. In speech recognition, spectral features are used to extract the acoustic information corresponding to each of the sound units [2–5]. Spectral envelope of the phoneme captures the unique set of dominant frequencies. The length, shape and dynamics of the vocal tract system vary from one speaker to another. Speaker-specific information is captured by modeling the spectral information extracted from spoken utterances of a speaker [6–8]. The characteristics of same sound unit may vary from one language to another due to co-articulation effects. Such variations can be represented by spectral features [9–16]. Spectral features are also used for recognizing different emotions from speech signal [17, 18].

Prosodic Features

Human beings impose some constraints on the sequence of sound units while producing speech [19], which incorporates naturalness to speech. The variation of pitch with respect to time incorporates some special characteristics to speech, which is known as *intonation*. *Intonation* is an important feature which is used to recognize the attitude and emotion of a speaker, grammatical structure and the types of a spoken utterance (i.e., statement or question). The duration of sound units or syllables vary with respect to time and forms a special pattern which incorporates *rhythm* to human speech. *Stress* is defined as the emphasis given to certain syllables or words in a sentence. The acoustic correlates correspond to *intonation*, *rhythm* and *stress* are pitch, duration and energy, respectively. These three properties (*intonation*, *rhythm*

and *stress*) are collectively known as *prosody*. Prosodic features have been used for developing speaker recognition [20, 21], emotion recognition [13–15, 22, 23], emotional speech synthesis [24] and language identification systems [12, 25].

Excitation Source Features

The constriction of expiration of air acts as excitation source during the production of speech. The quasi-periodic air pulses generated by the vocal folds vibration acts as source of excitation for voiced speech production. During the production of unvoiced speech, the expiration of air constraints at different places in the vocal tract. This information can be captured by passing the speech signal through the inverse filter [26]. To capture the excitation source information, linear prediction (LP) residual signal can be analyzed at three different levels: (i) sub-segmental level (within a glottal cycle or pitch cycle), (ii) segmental level (within 2–3 successive glottal cycles) and (iii) supra-segmental level (across 50 glottal cycles). Excitation source features have been explored for many speech tasks such as, speech enhancement [27], speaker recognition [28–31], audio clip classification [32], emotion recognition [33] and classification of infant cries [34, 35].

1.4 Issues in Developing Language Identification Systems

Recognizing a language without having an explicit information about the phonotactic rules, syntax and morphology is really a challenging task. In developing automatic language identification system it is assumed that, speakers in test and training sets do not overlap. The matching between the test utterance and the reference sample is always from unconstrained utterances of two different speakers. Hence, between the two utterances there exist differences such as, text, speaker, emotion of the speaker, dialect, environment and language. Therefore, developing an automatic language identification system with high accuracy, needs to derive the language discriminative characteristics apart from other information. In this section some issues about language identification system have been discussed.

- Variation in speaker characteristics: Different speakers have different speaking styles which provides a large variability of the speakers within a language. Thus, it is essential to reduce the speaker variability while building language models.
- Variation in accents: Accent is a distinctive way of pronouncing a language, especially one associated with a particular country, area, or social class. From the accent one can easily distinguish whether the person is native speaker of a particular language or not.
- Variation in background environment and channel characteristics: Speech signal has strong influence on the environmental conditions where the data is collected and on channel through which it is transmitted. These factors can have significant effect on the features derived from the short time spectral analysis. Therefore, it

is necessary to have the features which are robust enough to achieve a significant accuracy in language identification.
- Variation in dialects: Dialect is a special form of a language, which is unique for a specific region or social group. It is distinguished by pronunciation, grammar, or vocabulary, especially a variety of speech differing from the standard literary language or speech pattern of the culture in which it exists.
- Similarities in language: Most of the Indian languages are originated from the Sanskrit language. Hence, there is a lot of similarity in Indian languages. Most of the Indian languages have common set of phonemes and also follow similar grammatical structure. To develop a language identification system it is necessary to derive non-overlapping language-specific information for each language. Therefore, building a robust automatic language identification system in Indian context is really a challenging task.

1.5 Objective and Scope of the Work

The primary goal of this book is to present robust and efficient techniques to develop LID system in context of Indian languages. This work focuses on modeling the excitation source information for language discrimination task. Excitation source information can be captured by processing the linear prediction (LP) residual signal [26]. Linear prediction residual signal mostly contains higher order relations [36] and it is difficult to capture these relations using parametric techniques. We hypothesize that, language-specific information may present in the higher-order relations of the LP residual samples. In this work, we have proposed both implicit and explicit approaches to analyze the higher order relations of LP residual samples for modeling language discriminative information efficiently. The complementary information present between the vocal tract and excitation source features has been investigated by combining the evidences obtained from these features. Robustness of proposed excitation source features, compared to vocal tract features is also examined by varying amount of training data, length of test utterances and background noise characteristics.

1.6 Contributions of the Book

The primary contributions of this book are summarized as follows:

- Implicit features from LP residual, its magnitude and phase components are explored for language discrimination task.
- Explicit parametric features from LP residual signal are explored for capturing language-specific information.

- Evidences obtained from proposed implicit and parametric features of excitation source are combined to enhance the LID performance.
- Combination of the evidences obtained from overall excitation source and vocal tract features has been explored to investigate the existence of complementary language-specific information present in these two features.
- The robustness of excitation source information has been examined for language identification task by varying (i) background noise, (ii) amount of training data and (iii) duration of test samples.

1.7 Organization of the Book

- This chapter provides brief introduction about language identification and its applications. Different types of LID systems are described briefly. General features used for various speech tasks are described. Challenging issues in developing LID systems in Indian context are mentioned. The objectives and scope of the present work have been discussed. The major contributions of the book and chapter-wise organization are provided at the end of this chapter.
- Chapter 2 provides compendious reviews about both the *explicit* and *implicit* LID systems present in the literature. Existing works related to language identification in Indian context are briefly discussed. The related works about the excitation source features are also presented here. Various speech features and models proposed in the context of language identification are briefly reviewed in this chapter. The motivation for the present work from the existing literatures is briefly discussed.
- Chapter 3 discusses about the proposed approaches to model the implicit features of excitation source information for language identification. In this chapter, the raw LP residual samples, its magnitude and phase components are processed at three different levels: sub-segmental, segmental and supra-segmental levels to capture different aspects of excitation source information for LID task.
- Chapter 4 describes the proposed methods to extract parametric features at sub-segmental, segmental and supra-segmental levels to capture the language-specific excitation source information.
- Chapter 5 explains the combination of *implicit* and parametric features of excitation source to enhance the LID accuracy. Further, complementary nature of excitation source and vocal tract features is exploited for improving the LID accuracy. The robustness of proposed language-specific excitation source features is investigated at the end of this chapter.
- Chapter 6 summarizes the book contributions and provides future directions.

References

1. V.M. Vanishree, Provision for Linguistic Diversity and Linguistic Minorities in India. Master's thesis. Applied Linguistics, St. Mary's University College, Strawberry Hill, London, February 2011
2. F. Runstein, F. Violaro, An isolated-word speech recognition system using neural networks. Circuits Syst. **1**, 550–553 (1995)
3. A. Kocsor, L. Toth, Application of Kernel-based feature space transformations and learning methods to phoneme classification. Appl. Intell. **21**, 129–142 (2004)
4. R. Halavati, S.B. Shouraki, S.H. Zadeh, Recognition of human speech phonemes using a novel fuzzy approach. Appl. Soft Comput. **7**, 828–839 (2007)
5. T. Hao, M. Chao-Hong, L. Lin-Shan, An initial attempt for phoneme recognition using Structured Support Vector Machine (SVM), in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 4926–4929 (2010)
6. S. Furui, Cepstral analysis techniques for automatic speaker verification. IEEE Trans. Audio Speech Lang. Process. **29**(2), 254–272 (1981)
7. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Audio, Speech Lang. Process. **3**(1), 4–17 (1995)
8. D.A. Reynolds, Speaker identification and verification using gaussian mixture speaker models. Speech Commun. **17**, 91–108 (1995)
9. M. Sugiyama, Automatic language recognition using acoustic features, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 813–816, May 1991
10. K.S. Rao, S. Maity, V.R. Reddy, Pitch synchronous and glottal closure based speech analysis for language recognition. Int. J. Speech Technol. (Springer) **16**(4), 413–430 (2013)
11. J. Balleda, H.A. Murthy, T. Nagarajan, Language identification from short segments of speech. in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1033–1036, October 2000
12. V.R. Reddy, S. Maity, K.S. Rao, Recognition of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. (Springer) **16**(4), 489–510 (2013)
13. S.G. Koolagudi, K. Sreenivasa Rao, Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. Int. J. Speech Technol. (Springer) **15**(3), 495–511 (2012)
14. K. Sreenivasa Rao, S.G. Koolagudi, *Emotion Recognition using Speech Features*. (Springer, 2012). ISBN 978-1-4614-5142-6
15. K. Sreenivasa Rao, S.G. Koolagudi, Robust Emotion Recognition Using Spectral And Prosodic Features. (Springer, 2012). ISBN 978-1-4614-6359-7
16. S.G. Koolagudi, D. Rastogi, K. Sreenivasa Rao, Spoken language identification using spectral features. *Communications in Computer and Information Science (CCIS): Contemporary Computing*, vol. 306, (Springer, 2012), pp. 496–497
17. D. Neiberg, K. Elenius, K. Laskowski, Emotion recognition in spontaneous speech using GMMs, in *Internation Speech Communication and Association (INTERSPEECH)*, September 2006
18. D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition. Speech Commun. **52**(7), 613–625 (2009)
19. K.S. Rao, B. Yegnanarayana, Modeling durations of syllables using neural networks. Comput. Speech Lang. **21**, 282–295 (2007)
20. A.G. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey, Modeling prosodic dynamics for speaker recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, April 2003
21. L. Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition. Speech Commun. **50**(10), 782–796 (2008)
22. K.S. Rao, S.G. Koolagudi, R.R. Vempada, Emotion recognition from speech using global and local prosodic features. Int. J. Speech Technol. **16**(2), 143–160 (2013)
23. K. Sreenivasa Rao, S.G. Koolagudi, Identification of hindi dialects and emotions using spectral and prosodic features of speech. J. Syst. Cybern. Inform. **9**(4), 24–33 (2011)

24. J. Yadav, K. Sreenivasa Rao, Emotional-speech synthesis from neutral-speech using prosody imposition, in *International Conference on Recent Trends in Computer Science and Engineering (ICRTCSE-2014)*, Central University of Bihar, Patna, India, 8–9, February 2014
25. D. Martinez, L. Burget, L. Ferrer, N. Scheffer, i-vector based prosodic system for language identification, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4861–4864, March 2012
26. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
27. B. Yegnanarayana, T.K. Raja, Performance of linear prediction analysis on speech with additive noise, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1977)
28. C.S. Gupta, S.R.M. Prasanna, B. Yegnanarayana, Autoassociative neural network models for online speaker verification using source features from vowels, in *IEEE International Joint Conference Neural Networks* May 2002
29. D. Pati, S.R.M. Prasanna, Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. Int. J. Speech Technol. (Springer) **14**(1), 49–63 (2011)
30. D. Pati, D. Nandi, K. Sreenivasa Rao, Robustness of excitation source information for language independent speaker recognition, in *16th International Oriental COCOSDA Conference*, Gurgoan, India, November 2013
31. D. Pati, S.R.M. Prasanna, A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information. Sadhana (Springer) **38**(4), 591–620 (2013)
32. A. Bajpai, B. Yegnanarayana, Exploring features for audio clip classification using LP residual and AANN models, in *International Conference on Intelligent Sensing and Information Processing*, pp. 305–310, January 2004
33. K.S. Rao, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. (Springer) **16**, 181–201 (2013)
34. A.V. Singh, J. Mukhopadhyay, K. Sreenivasa Rao, K. Viswanath, Classification of infant cries using dynamics of epoch features. J. Intell. Syst. **22**(3), 253–267 (2013)
35. A.V. Singh, J. Mukhopadyay, S.B.S. Kumar, K. Sreenivasa Rao, Infant cry recognition using excitation source features, in *IEEE INDICON*, Mumbai, India, December 2013
36. S.R.M. Prasanna, C.S. Gupta, B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**, 1243–1261 (2006)

Chapter 2
Language Identification—A Brief Review

Abstract This chapter provides compendious reviews about both the explicit and implicit LID systems present in the literature. Existing works related to language identification in Indian context are briefly discussed. The related works about the excitation source features are also presented here. Various speech features and models proposed in the context of language identification are briefly reviewed in this chapter. The motivation for the present work from the existing literature is briefly discussed.

Keywords Prior works on explicit language identification · Prior works on implicit language identification · Prior works on excitation source features · Motivation for using source features for language identification

2.1 Prior Works on Explicit Language Identification System

In 1974, Dodington and Leonard [1], Leonard [2] have explored frequency of occurrences of certain reference sound units in different languages. The average LID accuracy of 64 % and 80 % have been achieved using five and seven languages, respectively.

In 1977, House and Neuberg [3] conducted LID studies on manually phonetic transcribed data. The language related information has been extracted from a broad phonetic transcription instead of using acoustic features extracted from speech signal. Speech signal has been considered as a sequence of symbols chosen from a set. The elements of the set are defined as follows: stop consonant, fricative consonant, vowel and silence. Language identification experiment has been carried out on eight languages. In this work, Hidden Markov Model (HMM) has been trained using broad phonetic labelled data derived from phonetic transcription. This work had shown perfect discrimination of eight languages and demonstrated that excellent language identification can be achieved by exploiting phonotactic information.

In 1980, Li and Edwars [4] developed automatic LID system based on automatic acoustic-phonetic segmentation of speech. By using six different acoustic-phonetic classes, automatic LID system has been developed using five languages. These six acoustic-phonetic classes are (i) syllable nuclei, (ii) non-vowel sonorants, (iii) vocal

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_2

murmur, (iv) voiced frication, (v) voiceless frication and (vi) silence and low energy segments. Hidden Markov Models (HMM) have been used for developing language models. Recognition accuracy of 80 % has been achieved with this approach.

In 1993 and 1994, Lamel and Gauvain [5, 6] conducted cross-lingual experiments by exploring phone recognition for French and English languages. A set of 35 phones were used to represent the French language corpus and a set of 46 phones were used to represent the English language data. Three-state left-to-right continuous density HMM with Gaussian mixture model (GMM) observation density has been used to build the phone models. It has been observed that, the French language is easier to recognize at the phone level but, harder to recognize at the lexical level due to the larger number of homophones.

In 1994, Muthusamy et al. [7] have proposed a perceptual benchmark for language identification task. Perceptual studies with listeners from different language backgrounds have been conducted. The experiments have been conducted on ten languages from OGI-MLTS database. The results obtained from the subjects reported as the benchmark for evaluating the LID performances obtained from automatic LID systems. The experimental analysis showed that, the duration of utterances, familiarity of languages and the number of known languages are the important factors to recognize a language. The comparison between the subjective analysis and machine performance concluded that, increased exposure to each language and longer training sessions contribute to improved classification performance. Therefore, to develop the speech recognizer for any language, the primary requirement is large amount of segmented and labelled speech corpus.

In 1994, Berkling et al. [8] have analyzed phoneme based features for language recognition. They have performed the LID study on three languages: English, Japanese and German from OGI-MLTS speech corpus. A superset of phonemes for the three languages has been considered. The phonemes which can provide the best discrimination between language pairs have used to build the superset. The experimental analysis drawn the conclusion that, to develop a LID system with large number of languages, it may be useful to reduce the number of features despite a small loss in LID accuracy.

In 1994, Tucker et al. [9] have conducted LID experiments with the languages belong to same language family. Sub-word models for English, Dutch and Norwegian languages have been developed for carrying out the LID study. Two types of language models: language independent and language-specific models have been developed in this study. Three techniques namely, (i) the acoustic differences between the phonemes of each language, (ii) the relative frequencies of phonemes of each language and (iii) the combination of previous two sources have been explored for classifying the languages. The third technique provides average LID accuracy of 90 % for three languages.

In 1994, Zissman and Singer [10] have carried out a comparative study using four approaches: (i) Gaussian mixture model based classification, (ii) phoneme recognition followed by language modeling (PRLM), (iii) parallel PRLM (PRLM-P) and (iv) language-dependent parallel phoneme recognition (PPR). The OGI-MLTS corpus has been used to evaluate the performances of the four LID approaches. The LID

study showed that, best performance is obtained with PRLM-P system, which does not require labelled speech corpus for developing language models.

In 1995, Kadambe and Hieronymus [11] have developed LID systems using phonological and lexical models to distinguish the languages. The LID study has been carried out on four languages: English, German, Mandarin and Spanish from OGI-MLTS speech corpus. Identification accuracy of 88 % has been achieved with four languages. It has been observed that, English and Spanish languages are distinguishable by their lexical information. This study concludes that, the language-specific information can also be captured by analyzing the higher level linguistic knowledge.

In 1995, Yan and Bernard [12] have developed language-dependent phone recognition systems for language discrimination task. Six languages (English, German, Hindi, Japanese, Mandarin and Spanish) from OGI-MLTS corpus have been used for LID study. Continuous HMMs are used to build the language-dependent phone recognizers. Acoustic and duration models are exploited for developing LID system. Forward and backward bigram based language models are proposed. A neural network based approach has been proposed for combining the evidences obtained from the above mentioned acoustic, language and duration models.

In 1997, Navratil and Zhulke [13] have proposed two approaches to build language models: (i) modified bigrams with a context mapping matrix and (ii) language models based on binary decision trees. To build the binary decision tree two approaches are proposed. These two approaches are, (i) building the whole tree for each class and (ii) adapting from a universal background model (UBM). Both the models are incorporated in a phonetic language identifier with a double bigram decoding architecture. The LID study has been carried out on NIST'95 language database.

In 1997, Hazen and Zue [14] have developed automatic LID system utilizing the phonotactic, acoustic-phonetic and prosodic information within a unified probabilistic framework. The evidences obtained from three different sources are combined to improve the LID accuracy. Experimental results showed that, the phonotactic information present in the speech utterances is the most useful information for language discrimination task. It has been observed that, acoustic-phonetic and prosodic information can also be useful for increasing the system's accuracy, especially when the short duration utterances are used for evaluation.

In 2001, K. Kirchhoff and S. Parandekar [15] have developed LID systems based on n-gram models of parallel streams of phonetic features and sparse statistical dependencies between these streams. The LID study has been conducted on OGI-MLTS database. It has been shown that, the proposed feature-based approach outperforms phone-based system. They have also reported that, proposed approach shows significantly better identification accuracy using test utterances of very short duration ($\leq$3 s). In future, data-driven measures for predicting optimal cross-stream dependencies, as well as different schemes for score integration can be explored.

In 2001, Gleason and Zissman [16] have demonstrated two methods to enhance the accuracy of parallel PRLM (PPRLM) system. They have explored Composite background (CBG) modeling technique, which allows us to identify target language in an environment where labelled training data is unavailable or limited.

In 2003, V. Ramasubramanian et al. [17] have shown the theoretical equivalence of parallel sub-word recognition (PSWR) and Ergodic-HMM (E-HMM) based LID. In this work, the sub-word recognizer (SWR) at the front-end represents the states and the language model (LM) of each language at the back-end represents the state-transition of E-HMM in that language. The proposed equivalence unifies two distinct approaches of langauge identification: parallel phone (sub-word) recognition and E-HMM based approaches. This LID study has been carried out on 6 languages from OGI-MLTS database. The performance of E-HMM based system is superior compared to GMM, which indicates the effectiveness of the E-HMM based approaches.

In 2004, J. Gauvain et al. [18], Shen et al. [19] proposed a novel method using phone lattices for developing automatic LID system. The use of phone lattices both in training and testing significantly improves the accuracy of a LID system based on phonotactics. Decoding is done by maximizing the expectation of the phonotactic likelihood for each language. Neural network has been used to combine the scores of multiple phone recognizers for improving the recognition accuracy. NIST 2003 corpus is used for carrying out the study.

In 2007, Li et al. [20] have proposed a novel approach for spoken language identification task based on vector space modeling (VSM). The hypothesis is that, the overall characteristics of all languages can be covered by a universal set of acoustic units, which can be characterized by the acoustic segment models (ASMs). The ASM framework further extended to language independent phone models for LID task by introducing an unsupervised learning procedure to circumvent the need for phonetic transcription. The spoken utterance has been converted to a feature vector with its attributes representing the co-occurrence statistics of the acoustic units. Then a vector space classifier has been built for language identification. The proposed framework has been evaluated on NIST 1996 and 2003 LRE databases.

In 2008, Sim and Li [21] have proposed a new approach for building a parallel phone recognition followed by language model (PPRLM) system. A PPRLM system comprises multiple parallel sub-systems, where each sub-system employs a phone recognizer with a different phone set for a particular language. This method aims at improving the acoustic diversification among its parallel sub-systems by using multiple acoustic models. The acoustic models are trained on the same speech data with the same phone set but using different model structures and training paradigms. They have examined the use of various structured precision (inverse covariance) matrix modeling techniques as well as the maximum likelihood and maximum mutual information training paradigms to produce complementary acoustic models. The results show that, acoustic diversification, which requires only one set of phonetically transcribed speech data, yields similar performance improvements compared to phonetic diversification. In addition, further improvements were obtained by combining both diversification factors. The proposed approach has been evaluated on NIST 2003 and 2005 LRE databases.

In 2008, Tong et al. [22] have proposed a target-oriented phone tokenizers (TOPT), each having a subset of phones that have high discriminative ability for a target language. Two phone selection methods are proposed to derive such phone subsets from a phone recognizer. It has been shown that, the TOPTs derived from a universal

phone recognizer (UPR) outperform those derived from language specific phone recognizers. The TOPT front-end derived from a UPR also consistently outperforms the UPR front-end without involving additional acoustic modeling. The proposed method has been evaluated on NIST 1996, 2003 and 2007 LRE databases.

In 2012, Botha and Barnard [23] used n-gram statistics as features for LID study. A comparative study has been carried out using different classifiers such as, support vector machines (SVMs), naive Bayesian and difference-in-frequency classifiers. The work has been carried out by varying the values of n. Experimental results conclude that, the SVM classifier outperforms other classifiers.

In 2012, Barroso et al. [24] have proposed hybrid approaches to build LID system based on the selection of system elements by several classifiers (Support Vector Machines (SVMs), Multilayer Perceptron classifiers and Discriminant analysis). The LID study has been carried out on three languages: Basque, Spanish and French. The proposed approach improves the system performance.

In 2013, Siniscalchi et al. [25] proposed a novel universal acoustic characterization approach for language recognition. Universal set of fundamental units has been explored, which can be defined across all the languages. This LID study has exploited some speech attributes like manner and place of articulations of sound units to define the universal set of language-specific fundamental units. Summary of the prior works related to explicit LID studies mentioned above is provided in Table 2.1.

Table 2.1 Summary of prior works on *explicit* language identification studies

Sl. no.	Features	Models/ Classification techniques	Number of languages and databases	Remarks	Reference
1.	Broad phonetic transcription (i.e., stop consonant, fricative consonant, vowel and silence)	HMM	8 languages	Phonotactic information is language-specific	[3]
2.	Acoustic-phonetic information	HMM	5 languages	Recognition accuracy of 80 % has been achieved	[4]
3.	PLP coefficients with 56 dimensions	ANN	3 languages from OGI-MLTS database	To develop LID system with large number of languages, it may be useful to reduce the number of features despite a small loss in LID accuracy	[8]

(continued)

Table 2.1 (continued)

Sl. no.	Features	Models/ Classification techniques	Number of languages and databases	Remarks	Reference
4.	Acoustic differences between the phonemes, relative frequency of phonemes and combination of previous two sources of information	Sub-word models were built using HMM	8 languages from EUROM 1 database	90 % LID accuracy is achieved	[9]
5.	MFCC	GMM, PRLM, PRLM-P, PPR	10 languages from OGI-MLTS database	PRLM-P provides best accuracy of 79.2 %	[10]
6.	Phoneme inventory, phonemotactics, syllable structure, lexical and prosodic differences	HMM	4 languages from OGI-MLTS database	88 % accuracy is achieved. Language-specific information can be captured using higher order linguistic knowledge	[11]
7.	Acoustic and duration models	HMM for phoneme recognizer and forward ANN for combining the scores	6 languages from OGI-MLTS database and backward bigram based language model	91.06 % accuracy is achieved for test sample length of 45 s	[12]
8.	Information from a wider phonetic context	Modified bigrams with a context mapping matrix and language models based on binary decision trees	9 languages NIST'95 LRE database	Error rate of 9.4 % is achieved with 45 s test sample duration	[13]
9.	Phonotactic, acoustic-phonetic and prosodic information	Interpolated trigram model and GMM	OGI-MLTS database	Phonotactic information is most useful information for LID task	[14]

(continued)

Table 2.1 (continued)

Sl. no.	Features	Models/ Classification techniques	Number of languages and databases	Remarks	Reference
10.	Phonetic features like, voicing, consonantal place of articulation, manner of articulation, nasality and lip rounding	HMM for phone recognition and *n*-gram language model	OGI-MLTS database	Proposed feature-based approach outperforms phone-based system	[15]
11.	MFCC	Parallel sub-word recognition and Ergodic HMM based LID	6 languages of OGI-MLTS database	The performance of E-HMM based system is superior compared to GMM	[17]
12.	Lexical constraints and phonotactic patterns	PPRLM	NIST 1996, 2003 and 2007 LRE databases	TOPTs derived from UPRs outperform those from language-specific phone recognizers	[22]
13.	*n*-gram statistics as features used for text based LID	SVM, naive Bayesian and difference-in-frequency classifiers	11 South African languages	The SVM classifier outperforms other classifiers and 99.4 % accuracy is achieved	[23]
14.	Morphological features	Hybrid system using SVM, Multilayer Perceptron classifiers and Discriminant analysis	3 languages in Basque context	Hybrid approach works well for under-resourced languages	[24]
15.	Manner and place of articulations of sound units	SVM and maximal figure-of-merit (MFoM)	NIST 2003	Universal set of language-specific fundamental units is proposed	[25]

2.2 Prior Works on Implicit Language Identification System

In 1986, Foil [26] has explored two different approaches to carry out LID study in noisy background. In first approach, language-specific prosodic features are captured by processing pitch and energy contours for LID task. Even though the languages with very similar phoneme sets, the frequency distribution of phonemes often

vary between the languages. In second method, formant vectors are computed only from the voiced segments for each language which is used to discriminate the same phonemes present in different languages. K-means clustering algorithm has been used for formant classification. The conclusion has been drawn from this LID study is that, formant features are better than the prosody features for LID task.

In 1989, Goodman et al. [27] have improved the LID accuracy obtained by Foil [26] in 1986. An important modification has been suggested to training algorithm. The training data has been split into "clean" and "noisy" vectors. K-means clustering algorithm has been used in this experiment. Experiments were also carried out to determine whether pitch information is useful in performing language identification in such noisy conditions or not. The use of syllabic rate as a language discriminative feature has also been investigated.

In 1991, Muthusamy et al. [7] have proposed a phonetic segment-based approach for developing automatic spoken language identification system. The idea was that, the acoustic structure of languages can be estimated by segmenting speech into broad phonetic categories. The language-specific phonetic and prosodic information has been extracted to develop automatic LID system. The LID study has been carried out on American English, Japanese, Mandarin Chinese and Tamil languages. Identification accuracy of 82.3 % has been achieved.

In 1991, Sugiyama [28] has explored linear prediction coefficients (LPCs) and cepstral coefficients (LPCCs) for language recognition. Vector quantization (VQ) of different code book sizes has been proposed for language recognition task. Different distortion measurement techniques like cepstral distance and weighted likelihood ratio have been explored in this LID study. In [9], VQ histogram algorithm has also been proposed for language prediction. Morgan et al. [29] and Zissman [30] have proposed the Gaussian mixture models (GMMs) [31] for language identification study.

In 1994, Itahashi et al. [32] and Shuichi and Liang [33] have developed LID systems based on fundamental frequency and energy contours with the modeling technique based on a piecewise-linear function.

In 1994, K. Li [34] explored spectral features at syllable level to capture the language discriminative information. The syllable nuclei (vowels) are detected automatically. The spectral feature vectors are then computed from the regions near the syllable nuclei instead of computing feature vectors from the whole training data.

In 1999, F. Pellegrino and R. Andre-Obrecht [35] have designed a unsupervised approach based on vowel system modeling. In this work, the language models are developed only using the features extracted from the vowels of each language. Since this detection is unsupervised and language independent, no labelled data is required. GMMs are initialized using an efficient data-driven variant of the LBG algorithm: the LBG-Rissanen algorithm. This LID study are carried out on 5 languages from OGI-MLTS database which provides 79 % recognition accuracy.

In 2005, Rouas et al. [36] have proposed an approach for language identification study based on *rhythmic* modelling. Like phonetics and phonotactics, *rhythm* is also an important feature which can be used for capturing language-specific information. In [36] an algorithm has been proposed to extract the *rhythm* for LID task. They

have used a vowel detection algorithm to segment the *rhythmic* units related to syllables. Several parameters are extracted (consonantal and vowel duration, cluster complexity) and modelled with a Gaussian Mixture. This LID study has been carried out on read speech collected from seven languages.

In 2007, Rouas [37] have developed a LID system based on modelling the prosodic variations. n-gram models were used to model the short-term and long-term language-dependent sequences of labels. The performance of the system is demonstrated by experiments on read speech and evaluated by experiments on spontaneous speech. An experimental study has also been carried out to discriminate the Arabic dialects. It has been shown that the proposed system was able to clearly identify the dialectal areas, leading to the hypothesis that, Arabic dialects have prosodic differences.

In 2010, Sangwan et al. [38] have proposed a language analysis and identification system based on the speech production knowledge. The proposed method automatically extracts key production traits or "hot-spots" which have significant language discriminative capability. At first, the speech utterances were parsed into consonant and vowel clusters. Subsequently, the production traits for each cluster is represented by the corresponding temporal evolution of speech articulatory states. It was hypothesized that, a selection of these production traits are strongly tied to the underlying language, and can be exploited for identifying languages. The LID study has been carried out on 5 closely related languages spoken in India namely, Kannada, Tamil, Telugu, Malayalam, and Marathi. The LID accuracy of 65 % is achieved with this approach. Furthermore, the proposed scheme was also able to identify automatically the key production traits of each language (e.g., dominant vowels, stop-consonants, fricatives etc.).

In 2012, Martnez et al. [39] have proposed an i-vector based prosodic system for language identification system. They have built an automatic language recognition system using the prosody information (*rhythm*, *stress*, and *intonation*) from speech and makes decisions about the language with a generative classifier based on i-Vectors.

In Indian context, J. Balleda et al. [40], have first attempted to identify Indian languages. VQ and 17 dimensional mel-frequency cepstral coefficients (MFCCs) have been explored for language recognition task. Nagarajan [41], have explored different code book methods for LID study. Automated segmentation of speech into syllable like units and parallel syllable like unit recognition have been explored to build *implicit* LID system. Sai Jayaram et al. [42], have proposed trained sub-word unit models without any labelled or segmented data, which are clustered using K-means clustering algorithm. Hidden Markov models (HMM) are used for predicting the language. In 2004, Leena Mary and B. Yegnanarayana have explored the autoassociative neural networks (AANN) for capturing language-specific features for developing LID system [43]. They have also explored prosodic features for capturing the language-specific information [44]. In K.S. Rao et al. [45], have explored spectral features using block processing (20 ms block size), pitch synchronous and glottal closure region (GCR) based approaches for discriminating 27 Indian languages. The language-specific prosodic features have also been explored by V. R. Reddy

et al. [46]. In this work, prosodic features are extracted from syllable, word and sentence levels to capture language-specific information. Jothilakshmi et al. [47], have explored a hierarchical approach for identifying the Indian languages. This method first identifies the language group of a given test utterance and then identifies the particular language inside that group. They have carried out the LID task by using different acoustic features such as, MFCC, MFCC with velocity and acceleration coefficients, and shifted delta cepstrum (SDC) features. In 2013, Bhaskar et al. [48] have carried out LID study using gender independent, gender dependent and hierarchical grouping approaches on 27 Indian languages. Vocal tract features are used to capture the language-specific information. Summary of the prior works related to implicit LID studies mentioned above is provided in Table 2.2.

Table 2.2 Summary of prior works on *implicit* language identification studies

Sl. no.	Features	Models/ Classification techniques	Number of languages and databases	Remarks	Reference
1.	Prosodic and formant features	K-means clustering	Recorded noisy radio signals as database	Formant features are better than the prosody features for LID task	[26]
2.	LPCs and LPCCs	Vector Quantization	20 languages	Accuracy of 65 % is achieved	[9]
3.	Spectral features at syllable level	ANN	5 languages from OGI-MLTS database	Syllabic spectral feature is useful for LID. 95 % accuracy is achieved	[34]
4.	MFCC	GMM	5 languages from OGI-MLTS database	79 % accuracy is achieved	[35]
5.	*Rhythm* at syllable level	GMM	7 languages from MULTEXT corpus	88 % accuracy is achieved	[36]
6.	Production knowledge of vowels and consonants	HMM	5 languages from South Indian Language (SInL) corpus	65 % accuracy is achieved	[38]
7.	Prosody information (rhythm, stress, and intonation)	i-vector based classification	NIST LRE 2009	Prosodic features contain language-specific knowledge	[25]
8.	MFCC	VQ	5 languages	Presence of some CV units is crucial for LID	[11]

(continued)

Table 2.2 (continued)

Sl. no.	Features	Models/ Classification techniques	Number of languages and databases	Remarks	Reference
9.	Weighted linear prediction cepstral coefficients (WLPCC)	AANN	4 languages	93.75 % accuracy is achieved	[44]
10.	MFCC with delta and delta-delta and shifted delta spectrum (SDC) features	Hierarchial based LID system using GMM, HMM and ANN	9 languages	80.56 % accuracy is achieved	[47]
11.	MFCC using block processing, pitch synchronous and glottal closure based approaches	GMM	27 languages from IITKGP-MLILSC database	Glottal closure based approach performs better than other methods	[45]
12.	Prosodic features extracted from syllable, word and phrase levels	GMM	27 languages from IITKGP-MLILSC database	Word level features provide better LID accuracy	[46]

2.3 Prior Works on Excitation Source Features

The LP residual signal has been processed for several speech related tasks such as, speech enhancement, speaker recognition, audio clip classification and emotion recognition. Few works related to the excitation source features are described as below. B. Yegnanarayana and T. K. Raja [49] have analyzed the LP residual signal while the speech signal has been corrupted with additive white noise. It has been observed that, the features obtained from LP residual signal perform well even though the signal to noise ratio (SNR) is low. The excitation source information has also been exploited for robust speaker recognition task. In B. Yegnanarayana et al. [50], have developed a text-dependent speaker verification system using source, supra-segmental and spectral features. The supra-segmental features such as, pitch and duration are explored. Excitation source features extracted from LP residual signal is modeled by auto associative neural network (AANN). Although the supra-segmental and source features individually does not provide good performance. However, combining the evidences from these features improve the performance of the speaker verification system significantly. In this study, Neural network models are used to combinethe evidences from multiple sources of information. In [51], AANN

is proposed for capturing speaker-specific source information present in LP residual signal. Speaker models are built for each vowel to study the speaker information present in each vowel. Using this knowledge an online speaker verification system has been developed. This study shows that, excitation source features also contain significant speaker-specific information. In [52], LP residual signal, its magnitude and phase components are implicitly processed at sub-segmental, segmental and supra-segmental levels to capture speaker-specific information. The speaker identification and verification studies performed using NIST-99 and NIST-03 databases. This study demonstrates that, the segmental level features provide best performance followed by sub-segmental features. The supra-segmental features provide least performance. In [53], segmental level excitation source features are used for language independent speaker recognition study. In [54], LP residual signal has been explored for capturing the audio-specific information. Autoassociative neural network models have been used to capture the audio-specific information extracted from LP residual signal. In [55], the excitation source component of speech has been explored for characterizing and recognizing the emotions from speech signal. In this work, excitation source information is extracted from both LP residual and glottal volume velocity (GVV) signals. In this study, sequence of LP residual samples and their phase information, parameters of epochs and their dynamics at syllable and utterance levels have been used for characterizing emotions. Further, samples of GVV signal and its parameters also explored for emotion recognition task. In [56], a method has been proposed for duration modification using glottal closure instants (GCIs) and vowel onset points (VOPs). The VOPs are computed using the Hilbert envelope of LP residual signal. Manipulation of duration is achieved by modifying the duration of the LP residual with the help of instants of significant excitation as pitch markers. The modified residual is used to excite the time-varying filter. Perceptual quality of the synthesized speech is found to be natural. In [57], GCIs are computed from LP residual signal by using the property of average group-delay of minimum phase signals. The modification of pitch and duration was achieved by manipulating the LP residual with the help of the knowledge of the instants of significant excitation. The modified residual signal was used as excitation signal to the vocal tract resonator. The proposed method is evaluated using waveforms, spectrograms, and listening tests and it is found that, the perceptual quality of synthesized speech has been improved and there were no significant distortion. In K.S. Rao et al. [58], have proposed a time-effective method for determining the instants of significant excitation (GCIs) in speech signals. The proposed methods consist of two phases: (i) at first phase approximate epoch locations using the Hilbert envelope of LP residual signal and (ii) at second phase, accurate locations of the instants of significant excitation is determined by computing the group delay around the approximate epoch locations derived from the first phase. In [59], pitch contours are modified by using the significant instant of excitation and this technique can be used in voice conversion, expressive speech synthesis applications. In [60], excitation source features have been used for voice conversion tasks. The basic goal of the voice conversion

system is to modify the speaker-specific characteristics, keeping the message and the environmental information contained in the speech signal intact. In [60], a neural network models for developing mapping functions at each level has been proposed. The features used for developing the mapping functions are extracted using pitch synchronous analysis. In this work, the instants of significant excitation are used as pitch markers to perform the pitch synchronous analysis. Instants of significant excitation are computed from LP residual signal by using the property of average group-delay of minimum phase signals. In [61], a method has been proposed which is capable of jointly converting prosodic features, spectral envelope and excitation signal maintaining the correlation between them and this method has been used in voice conversion application.

2.4 Motivation for the Present Work

From the prior works related to LID studies mentioned in Sects. 2.1 and 2.2, it is observed that the existing LID systems are mostly developed using spectral features representing the vocal tract system characteristics and prosodic features representing the supra-segmental characteristics of the languages. The excitation source component of speech has still not been explored for LID task. From the literature, it has been observed that excitation source information represented by LP residual signal has been explored for several speech tasks. But, it has not been investigated for language discrimination task. Therefore, in this book, we want to explore excitation source features for language discrimination task. The human speech production system consists of time varying vocal tract resonator and the source for provoking the resonator. Speech sounds are produced as a consequence of acoustical excitation of the human vocal tract resonator. During the production of voiced sounds, the vocal tract is excited by a series of nearly periodic air pulses generated by the vocal cords vibration. State-of-the-art LID systems mostly approximate the dynamics of vocal tract shape and use this vocal tract information for discriminating the languages. However, the demeanor of the vocal folds vibration also changes from one sound unit to another. Although there is a significant overlap in the set of sound units in different languages, but the same sound unit may differ across different languages due to the co-articulation effects and dialects. Hence, we conjecture that, the characteristics of excitation source may contain some language-specific information. In present work, we have explored the excitation source features for capturing language-specific phonotactic information. A theoretical study has been carried out in Sect. 2.4 to support our hypothesis.

Correlation Among the Languages from Excitation Source Point of View

In this section, the significance of the excitation source information for language identification task is shown by their respective correlation coefficients for within and between languages. Correlation determines the degree of similarity between two

Table 2.3 Correlation coefficients across the languages derived from excitation source features

Languages	Correlation coefficients																										
Arunachali	**1.8**	0.83	0.52	0.8	1.02	0.97	0.7	0.49	0.38	0.87	0.93	0.5	0.66	0.42	0.98	0.32	0.78	1.21	0.48	1.21	1.98	0.63	1.1	0.64	0.5	1.4	1.59
Assamese	0.83	**3.26**	0.59	1.18	1.32	0.69	1.68	0.94	0.69	1.09	1.32	0.58	0.91	1.06	1.7	0.78	1.07	1.46	0.41	1.31	1.3	1.4	1.72	1.23	0.83	1.53	1.88
Bengali	0.52	0.59	**2.09**	0.88	0.61	0.57	0.56	0.82	0.82	0.71	0.57	0.63	0.66	0.51	0.54	0.39	0.36	0.74	0.54	0.67	0.87	0.67	0.57	0.36	0.74	0.8	0.68
Bhojpuri	0.8	1.18	0.88	**2.8**	0.72	0.72	0.67	0.95	0.82	1.2	1.26	0.91	1.07	0.79	1.29	0.59	1.19	1.07	0.58	1.21	1.53	1.16	1.53	0.66	1.17	1.39	0.92
Chhattisgarhi	1.02	1.32	0.61	0.72	**3.42**	1.21	1.2	0.74	0.63	0.88	1.53	0.63	0.91	0.88	1.47	0.73	0.91	2.73	0.72	2.18	2.71	0.93	2.23	0.84	0.92	2.74	2.46
Dogri	0.97	0.69	0.57	0.72	1.21	**2.06**	0.62	0.38	0.54	0.71	1.11	0.67	0.69	0.57	1.2	0.53	0.71	1.57	0.59	1.19	1.66	0.55	1.47	0.72	0.64	1.53	1.91
Gojri	0.7	1.68	0.56	0.67	1.2	0.62	**4**	0.76	1.11	0.87	1.37	0.49	0.76	1.19	1.41	1.38	1.04	2.8	0.65	0.91	0.04	0.77	0.72	0.94	0.5	1.63	1.29
Gujarati	0.49	0.94	0.82	0.95	0.74	0.38	0.76	**1.91**	0.64	0.69	0.8	0.55	0.57	0.72	0.75	0.42	0.55	0.87	0.54	0.78	1.63	0.86	0.94	0.64	0.72	1.07	0.92
Hindi	0.38	0.69	0.82	0.82	0.63	0.54	1.11	0.64	**1.67**	0.69	1.12	0.46	0.91	0.93	0.59	0.54	0.47	0.83	0.57	0.75	0.86	0.8	0.93	0.38	0.95	1	0.79
Indian English	0.87	1.09	0.71	1.2	0.88	0.71	0.87	0.69	0.69	**2.53**	0.98	0.55	1.09	0.9	0.75	0.74	0.73	1.28	0.5	1.21	0.87	0.83	1.15	0.93	0.74	0.92	1.05
Kannada	0.93	1.32	0.57	1.26	1.53	1.11	1.37	0.8	1.12	0.98	**2.41**	0.53	1.12	1.05	1.18	0.91	1.02	1.96	0.56	1.13	1.14	0.7	1.63	0.83	0.59	1.62	2.33
Kashmiri	0.5	0.58	0.63	0.91	0.63	0.67	0.49	0.55	0.46	0.55	0.53	**1.28**	0.69	0.46	0.85	0.46	0.65	1.23	0.47	0.92	1.2	0.73	0.71	0.44	0.53	1.01	1.02
Konkani	0.66	0.91	0.66	1.07	0.91	0.69	0.76	0.57	0.91	1.09	1.12	0.69	**3.1**	1.15	1.08	0.57	0.6	0.48	0.8	0.76	2.2	0.67	1.32	0.56	0.94	1.35	0.93
Malayalam	0.42	1.06	0.51	0.79	0.88	0.57	1.19	0.72	0.93	0.9	1.05	0.46	1.15	**2.02**	0.96	1.02	0.6	0.76	0.56	0.84	1.09	1.05	1.5	0.67	0.97	1.13	1
Manipuri	0.98	1.7	0.54	1.29	1.47	1.2	1.41	0.75	0.59	0.75	1.18	0.85	1.08	0.96	**2.9**	0.56	1.48	1.84	0.42	1.54	1.59	1.02	2.25	1.06	1.23	1.69	2.42
Marathi	0.32	0.78	0.39	0.59	0.73	0.53	1.38	0.42	0.54	0.74	0.91	0.46	0.57	1.02	0.56	**1.34**	0.52	1.59	0.41	0.61	0.15	0.54	0.36	0.55	0.5	1.06	0.9
Mizo	0.78	1.07	0.36	1.19	0.91	0.71	1.04	0.55	0.47	0.73	1.02	0.65	0.6	0.6	1.48	0.52	**1.67**	1.35	0.48	1.21	1.24	0.63	1.1	0.51	0.68	1.24	1.59
Nagamese	1.21	1.46	0.74	1.07	2.73	1.57	2.8	0.87	0.83	1.28	1.96	1.23	0.48	0.76	1.84	1.59	1.35	**9.89**	0.41	3.02	0.14	0.99	1.12	1.22	1.09	3.12	3.62
Nepali	0.48	0.41	0.54	0.58	0.72	0.59	0.65	0.54	0.57	0.5	0.56	0.47	0.8	0.56	0.42	0.41	0.48	0.41	**1.27**	0.48	0.88	0.6	0.74	0.43	0.61	1.23	0.53

(continued)

Table 2.3 (continued)

Languages	Correlation coefficients																										
Oriya	1.21	1.31	0.67	1.21	2.18	1.19	0.91	0.78	0.75	1.21	1.13	0.92	0.76	0.84	1.54	0.61	1.21	3.02	0.48	**3.98**	1.84	0.98	1.97	0.92	1.22	2.29	2.23
Punjabi	1.98	1.3	0.87	1.53	2.71	1.66	0.04	1.63	0.86	0.87	1.14	1.2	2.2	1.09	1.59	0.15	1.24	0.14	0.88	1.84	**13.79**	1.79	4.49	0.84	1.57	3.45	3.03
Rajasthani	0.63	1.4	0.67	1.16	0.93	0.55	0.77	0.86	0.8	0.83	0.7	0.73	0.67	1.05	1.02	0.54	0.63	0.99	0.6	0.98	1.79	**1.98**	1.44	0.66	0.62	1.23	1.17
Sanskrit	1.1	1.72	0.57	1.53	2.23	1.47	0.72	0.94	0.93	1.15	1.63	0.71	1.32	1.5	2.25	0.36	1.1	1.12	0.74	1.97	4.49	1.44	**4.82**	0.88	1.18	3.07	2.62
Sindhi	0.64	1.23	0.36	0.66	0.84	0.72	0.94	0.64	0.38	0.93	0.83	0.44	0.56	0.67	1.06	0.55	0.51	1.22	0.43	0.92	0.84	0.66	0.88	**1.67**	0.4	1.19	1.44
Tamil	0.5	0.83	0.74	1.17	0.92	0.64	0.5	0.72	0.95	0.74	0.59	0.53	0.94	0.97	1.23	0.5	0.68	1.09	0.61	1.22	1.57	0.62	1.18	0.4	**2.17**	1.28	0.92
Telugu	1.4	1.53	0.8	1.39	2.74	1.53	1.63	1.07	1	0.92	1.62	1.01	1.35	1.13	1.69	1.06	1.24	3.12	1.23	2.29	3.45	1.23	3.07	1.19	1.28	**6.01**	3.82
Urdu	1.59	1.88	0.68	0.92	2.46	1.91	1.29	0.92	0.79	1.05	2.33	1.02	0.93	1	2.42	0.9	1.59	3.62	0.53	2.23	3.03	1.17	2.62	1.44	0.92	3.82	**5.26**

signals. Suppose that we have two real signal sequences $x(n)$ and $y(n)$ each of which has finite energy. The *cross-correlation* of $x(n)$ and $y(n)$ is a sequence $r_{xy}(l)$, which is defined as follows:

$$r_{xy}(l) = \sum_{n=1}^{p} x(n)y(n-l),\ l = 0, \pm 1, \pm 2, ... \tag{2.1}$$

where, l is the time shift parameter. The x and y are the two signals being correlated. If the signals are identical, then the correlation coefficient is maximum and if they are orthogonal then the correlation coefficient is minimum. When $x(n) = y(n)$, the procedure is known as *autocorrelation* of $x(n)$. From each language database, one male speaker's data of 5 min duration is considered and the LP residual has been extracted. The LP residual is then decimated by factor 4 to suppress the sub-segmental level information and then the LP residual samples are processed in block size of 20 ms with a shift of 2.5 ms which provides segmental level information. To normalize the speaker variability between the languages, the mean subtraction is imposed to all the feature vectors across all languages. Then the *seg* level feature vectors are modeled with GMM for each language. The average *mean* vectors are considered as the signal for a particular language to compute the correlation coefficients. To portray the significance of *seg* level LP residual feature in language discrimination task these correlation coefficients are used. The correlation coefficients between two signals is a sequence of length $(2l - 1)$. The average of the $(2l - 1)$ correlation coefficient values is considered in our work which is shown in Table 2.3. The values of first row of the Table 2.3 indicates the correlation coefficients of first language with respect to itself and other 26 languages. The correlation coefficient within a language has been computed from two different speech utterances spoken by a speaker. The first element of first row indicates the auto-correlation coefficient of first language calculated from the average *mean* vectors of two utterances within one language. The other 26 values of first row represents the cross-correlation coefficients between the first language and other 26 languages. Lower the cross-correlation coefficient value between two languages indicate more dissimilarity between them. We have taken the average of the 26 cross-correlation coefficients from the 2nd column to 27th column of the 1st row which represents average cross-correlation coefficient of first language with respect to other 26 languages. This average cross-correlation coefficient value (0.85) is less than the auto-correlation coefficient value (1.8) which resides in the 1st column of the 1st row. This explains that the *seg* level LP residual feature has significant language discriminative capability. If we analyze the other rows of the Table 2.3, similar characteristics can be observed. This theoretical discussion elicits the significance of the excitation source features in language identification task which is the motivation of the present work.

2.5 Summary

In this chapter, the existing works related to both the *explicit* and *implicit* LID systems have been described. Prior works based on excitation source features are also discussed. It has been observed that, the excitation source component of speech has not been explored for language discrimination task, which is the motivation of present work. Hence, in this work, excitation source information has been explored to capture language-specific phonotactic information for LID task.

References

1. R. Leonard, G. Doddington, Automatic language identification. Technical Report RADC-TR-74-200 (Air Force Rome Air Development Center, Technical Report) August 1974
2. R. Leonard, Language Recognition Test and Evaluation. Technical Report RADCTR-80-83 (Air Force Rome Air Development Center, Technical Report). March 1980
3. A.S. House, E.P. Neuberg, Toward automatic identification of the languages of an utterance. J. Acoust. Soc. Am. **62**(3), 708–713 (1977)
4. K.P. Li, T.J. Edwards, Statistical models for automatic language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 884–887, April 1980
5. L.F. Lamel, J.L. Gauvain, Cross lingual experiments with phone recognition. in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 507–510, April 1993
6. L.F. Lamel, J.L. Gauvain, Language identification using phonebased acoustic likelihoods, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, , pp. I/293–I/296, April 1994
7. Y. Muthusamy, R. Cole, M. Gopalakrishnan, A segment-based approach to automatic language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 353–356, April 1991
8. K.M. Berkling, T. Arai, E. Bernard, Analysis of phoneme based features for language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I/289–I/292, April 1994
9. R.C.F. Tucker, M. Carey, E. Parris, Automatic language identification using sub-word models, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I/301–I/30, April 1994
10. M.A. Zissman, E. Singer, Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1 pp. I/305–I/308, (1994)
11. S. Kadambe, J. Hieronymus, Language identification with phonological and lexical models, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3507–351, May 1995
12. Y. Yan, E. Barnard, An approach to automatic language identification based on language-dependent phone recognition, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3511–3514, May 1995
13. J. Navratil, W. Zuhlke, Phonetic-context mapping in language identification. Eur. Speech Commun. Assoc. (EUROSPEECH) **1**, 71–74 (1997)
14. T.J. Hazen, V.W. Zue, Segment-based automatic language identification. J. Acoust. Soc. Am. **101**, 2323–2331 (1997)
15. K. Kirchhoff, S. Parandekar, Multi-stream statistical n-gram modeling with application to automatic language identification, in *European Speech Communication Association (EUROSPEECH)*, pp. 803–806, (2001)

16. T. Gleason, M. Zissman, Composite background models and score standardization for language identification systems, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 529–532 (2001)
17. V. Ramasubramanian, A.K.V.S. Jayram, T.V. Sreenivas, Language identification using parallel sub-word recognition - an ergodic HMM equivalence, *European Speech Communication Association (EUROSPEECH)* (Geneva, Switzerland), September 2003
18. J. Gauvain, A. Messaoudi, H. Schwenk, Language recognition using phone latices, in *International Speech Communication Association (INTERSPEECH)*, pp. 25–28 (2004)
19. W. Shen, W. Campbell, T. Gleason, D. Reynolds, E. Singer, Experiments with lattice-based PPRLM language identification, in *Speaker and Language Recognition Workshop*, pp. 1–6 (2006)
20. H. Li, B. Ma, C.H. Lee, A vector space modeling approach to spoken language identification. IEEE Trans. Audio Speech Lang. Process. **15**(1), 271–284 (2007)
21. K.C. Sim, H. Li, On acoustic diversification front-end for spoken language identification. IEEE Trans. Audio Speech Lang. Process. **16**(5), 1029–1037 (2008)
22. R. Tong, B. Ma, H. Li, E.S. Chng, A target-oriented phonotactic front-end for spoken language recognition. IEEE Trans. Audio Speech Lang. Process. **17**(7), 1335–1347 (2009)
23. G.R. Botha, E. Barnard, Factors that affect the accuracy of text-based language identification. Comput. Speech Lang. **26**(5), 307–320 (2012)
24. N. Barroso, K. Lopez de Ipina, C. Hernandez, A. Ezeiza, M. Grana, Semantic speech recognition in the Basque context Part II: language identification for under-resourced languages. Int. J. Speech Technol. **15**(1), 41–47 (2012)
25. S.M. Siniscalchi, J. Reed, T. Svendsen, C.-H. Lee, Universal attribute characterization of spoken languages for automatic spoken language recognition. Comput. Speech Lang. **27**(1), 209–227 (2013)
26. J.T. Foil, Language identification using noisy speech, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 861–864, (1986)
27. F. Goodman, A. Martin, R. Wohlford, Improved automatic language identification in noisy speech, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 528–531, May 1989
28. M. Sugiyama, Automatic language recognition using acoustic features, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 813–816, May 1991
29. D. Morgan, L. Riek, W. Mistretta, C. Scofield, P. Grouin, F. Hull, Experiments in language identification with neural networks. Int. Joint Conf. Neural Netw. **2**, 320–325 (1992)
30. M. Zissman, Automatic language identification using gaussian mixture and hidden markov models, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 399–402, April 1993
31. D.A. Reynolds, R.C. Rose, Robust text -independent speaker identification using gaussian mixture speaker models. IEEE Trans. Audio Speech Lang. Process. **3**(1), 72–83 (1995)
32. S. Itahashi, J. Zhou, K. Tanaka, Spoken language discrimination using speech fundamental frequency, in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1899–1902, (1994)
33. I. Shuichi, D. Liang, Language identification based on speech fundamental frequency, in *European Speech Communication Association (EUROSPEECH)*, pp. 1359–1362 (1995)
34. K.P. Li, Automatic language identification using syllabic spectral features, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I/297–I/300, April 1994
35. F. Pellegrino, R. Andre-Obrecht, An unsupervised approach to language identification, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 833–836, Mar 1999
36. J.L. Rouas, J. Farinas, F. Pellegrino, R. Andr-Obrecht, Rhythmic unit extraction and modelling for automatic language identification. Speech Commun. **47**, 436–456 (2005)
37. J.L. Rouas, Automatic prosodic variations modeling for language and dialect discrimination. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1904–1911 (2007)

38. A. Sangwan, M. Mehrabani, J. Hansen, Automatic language analysis and identification based on speech production knowledge, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5006–5009, March 2010
39. D. Martinez, L. Burget, L. Ferrer, N. Scheffer, i-vector based prosodic system for language identification, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4861–4864, March 2012
40. J. Balleda, H.A. Murthy, T. Nagarajan, Language Identification from Short Segments of Speech, in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1033–1036, October 2000
41. T. Nagarajan, Implicit system for spoken language identification, Ph.D. dissertation, Indian Institute of Technology Madras, India (2004)
42. A.K.V.S. Jayaram, V. Ramasubramanian, T.V. Sreenivas, Language identification using parallel sub-word recognition, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 32–35, April 2003
43. L. Mary, B. Yegnanarayana, Autoassociative neural network models for language identification. in *Internatioanl Conference on Intelligent Sensing and Information Processing*, pp. 317–320 (2004)
44. L. Mary, Multilevel implicit features for language and speaker recognition, Ph.D. dissertation, Indian Institute of Technology Madras, India (2006)
45. K.S. Rao, S. Maity, V.R. Reddy, Pitch synchronous and glottal closure based speech analysis for language recognition. Int. J. Speech Technol. (Springer) **16**(4), 413–430 (2013)
46. V.R. Reddy, S. Maity, K.S. Rao, Identification of indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. (Springer) **16**(4), 489–511 (2013)
47. S. Jothilakshmi, V. Ramalingam, S. Palanivel, A hierarchical language identification system for Indian languages. Digital Signal Process. (Elsevier) **22**(3), 544–553 (2012)
48. B. Bhaskar, D. Nandi, K.S. Rao, Analysis of language identification performance based on gender and hierarchial grouping approaches, in *International Conference on Natural Language Processing*, December 2013
49. B. Yegnanarayana, T.K. Raja, Perfoemance of linear prediction analysis on speech with additive noise, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1977)
50. B. Yegnanarayana, S.R.M. Prasanna, J. Zachariah, C. Gupta, Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. Audio Speech Lang. Process. **13**(4), 575–582 (2005)
51. C.S. Gupta, S.R.M. Prasanna, B. Yegnanarayana, Autoassociative neural network models for online speaker verification using source features from vowels, in *IEEE International Joint Conference Neural Networks*, May 2002
52. D. Pati, S.R.M. Prasanna, Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. Int. J. Speech Technol. (Springer) **14**(1), 49–63 (2011)
53. D. Pati, D. Nandi, K. Sreenivasa Rao, Robustness of excitation source information for language independent speaker recognition, in *16th International Oriental COCOSDA Conference*, Gurgoan, November 2013
54. A. Bajpai, B. Yegnanarayana, Exploring features for audio clip classification using LP residual and AANN models, in *International Conference on Intelligent Sensing and Information Processing*, pp. 305–310, January 2004
55. K.S. Rao, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. (Springer) **16**, 181–201 (2013)
56. K.S. Rao, B. Yegnanarayana, Duration modification using glottal closure instants and vowel onset points. Speech Commun. **51**(12), 1263–1269 (2009)
57. K.S. Rao, B. Yegnanarayana, Prosody modification using instants of significant excitation. IEEE Trans. Audio Speech Lang. Process. **14**(3), 972–980 (2006)
58. K.S. Rao, S.R.M. Prasanna, B. Yegnanarayana, Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. IEEE Signal Process. Lett. **14**(10), 762–765 (2007)

59. K.S. Rao, Unconstrained pitch contour modification using instants of significant excitation. Circuits Syst. Signal Process. (Springer) **31**(6), 2133–2152 (2012)
60. K.S. Rao, Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput. Speech Lang. **24**(3), 474–494 (2010)
61. R. Hussain Laskar, K. Banerjee, F. Ahmed Talukdar, K. Sreenivasa Rao, A pitch synchronous approach to design voice conversion system using source-filter correlation. Int. J. Speech Technol. (Springer) **15**(3), 419–431 (2012)

Chapter 3
Implicit Excitation Source Features for Language Identification

Abstract This chapter discusses about the proposed approaches to model the implicit features of excitation source information for language identification. Excitation source features such as raw LP residual samples, its magnitude and phase components are processed at three different levels: sub-segmental, segmental and supra-segmental levels to capture different aspects of excitation source information for LID task. Further, LID systems are developed by combining the evidences obtained from LID systems built using individual features.

Keywords Implicit excitation source features · Linear prediction residual · Magnitude and phase components of LP residual · Hilbert envelope · Linear prediction residual phase · Indian language speech database · IITKGP-MLILSC · Sub-segmental features · Segmental features · Suprasegmental features

3.1 Introduction

The constriction during expiration of air acts as excitation source during the production of speech. The quasi-periodic air pulses generated by the vocal folds vibration acts as source of excitation for voiced speech production. During the production of unvoiced speech, the expiration of air constraints at different places in the vocal tract. In speech production, the majority of excitation occurs due to the production of voiced speech, which reflects the dynamic nature of the vocal folds vibration. The excitation source signal can be derived by passing the speech signal through the inverse filter [1]. In this work, 10th order LP analysis followed by inverse filter is used for estimating LP residual signal. Linear prediction analysis represents the second order statistical features in terms of the autocorrelation coefficients. So, the LP residual signal does not represent any significant second order relations corresponding to the vocal tract resonator and it contains only the higher order relations [2]. It is difficult to capture the higher order relations present in the LP residual signal using parametric techniques. We surmise that, the language-specific information may present in the higher-order relations of the LP residual samples. In this chapter, the implicit relations among the LP residual samples are analyzed to model the language-specific excitation source

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_3

information at different levels. The intuition is that, the implicit relations among the raw LP residual samples which reflect the excitation source information can be captured by processing the raw LP residual samples at three levels: (i) sub-segmental level (within a glottal cycle or pitch cycle), (ii) segmental level (within 2–3 successive glottal cycles) and (iii) supra-segmental level (across 50 glottal cycles). These levels are also termed as *sub*, *seg* and *supra* correspond to sub-segmental, segmental and supra-segmental levels, respectively. To capture the amplitude and phase information of excitation source it is necessary to separate these two components from LP residual signal and process them independently. Analytic signal representation of LP residual has been explored for separating magnitude and phase components of LP residual signal [3]. In this chapter, the magnitude and phase components of LP residual signal are also processed implicitly at three different levels: *sub*, *seg* and *supra* and eventually combined to capture the language-specific excitation source information.

This chapter is organized as follows: Sect. 3.2 describes the speech corpus used in this work. Section 3.3.1 explains the analytic representation of LP residual signal. Section 3.3.2 explains the proposed *implicit* processing of LP residual signal. Section 3.3.3 describes the *implicit* processing of magnitude and phase components of LP residual signal. In Sect. 3.4, development of LID systems using the proposed excitation source features is described. In Sects. 3.5 and 3.6, performance evaluation of LID systems is discussed. Section 3.7 summarizes the contents of this chapter.

3.2 Speech Corpus

In this work, we have evaluated the proposed excitation source features on Indian Institute of Technology Kharagpur Multi-Lingual Indian Language Speech Corpus (IITKGP-MLILSC) [4] and Oregon Graduate Institute Multi-Language Telephone-based Speech (OGI-MLTS) [5]. The detailed description of these two speech corpus is given in the following subsections.

3.2.1 Indian Institute of Technology Kharagpur Multi-Lingual Indian Language Speech Corpus (IITKGP-MLILSC)

In this work, LID study has been carried out on Indian Institute of Technology Kharagpur—Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) [4]. This database contains 27 Indian regional languages. Sixteen languages are collected from news bulletins of broadcasted radio channels and the remaining are recorded from broadcasted TV talk shows, live shows, interviews and news bulletins. An average of 73 min of speech data per language spoken by male and female speakers is present in the database. The broadcasted television channels are accessed using

Table 3.1 Description of the IITKGP-MLILSC language database

Languages	Region	Speaking population @ 2001 census (Mil)	No. of speakers		Duration (min.)
			F	M	
Arunachali	Arunachal Pradesh	0.41	6	6	175.84
Assamese	Assam	13.17	6	7	43.41
Bengali	West Bengal	83.37	13	9	68.63
Bhojpuri	Bihar	38.55	2	6	26.75
Chhattisgarhi	Chhattisgarh	11.50	6	6	54.58
Dogri	Jammu and Kashmir	2.28	5	4	101.81
Gojri	Jammu and Kashmir	20.00	2	6	70.03
Gujarati	Gujarat	46.09	4	7	76.72
Hindi	Uttar Pradesh	422.05	6	12	85.56
Indian English	All over India	125.23	6	6	63.01
Kannada	Karnataka	37.92	1	5	29.43
Kashmiri	Jammu and Kashmir	5.53	1	14	47.67
Konkani	Goa and Karnataka	2.49	5	4	66.92
Malayalam	Kerala	33.07	7	7	60.49
Manipuri	Manipur	1.47	7	7	107.54
Marathi	Maharashtra	71.94	5	6	41.34
Mizo	Mizoram	0.67	6	6	110.13
Nagamese	Nagaland	0.03	5	6	104.31
Nepali	West Bengal	2.87	6	6	54.19
Oriya	Orissa	33.02	7	4	53.86
Punjabi	Punjab	29.10	0	8	64.01
Rajasthani	Rajasthan	50.00	5	6	69.43
Sanskrit	Uttar Pradesh (UP)	0.014	0	12	58.2
Sindhi	Gujarat and Maharashtra	2.54	6	6	109.83
Tamil	Tamilnadu	60.79	8	10	75.38
Telugu	Andhra Pradesh (AP)	74.00	0	9	99.08
Urdu	UP and AP	51.54	1	8	54.03

VentiTV software and the Pixelview TV tuner card. Audacity software is used for recording the speech data from TV channels. The language data of broadcasted Radio channels are collected from the archives of Prasar Bharati, All India Radio (AIR) website [6]. The detail description of the database is given in Table 3.1.

3.2.2 Oregon Graduate Institute Multi-Language Telephone-Based Speech (OGI-MLTS) Database

Oregon Graduate Institute (OGI) Multi-Language Telephone-based Speech (MLTS) database consists of 11 languages. Muthusamy et al. [5] have collected the following 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese. Later, Hindi language data has been added. During collection of this database, each speaker was asked a series of questions designed to elicit: (i) Fixed vocabulary speech, (ii) Domain-specific vocabulary speech, (iii) Unrestricted vocabulary speech. The "unrestricted vocabulary speech" was obtained by asking the callers to speak on any topic on their choice. The "unrestricted vocabulary speech" of each speaker consists of two separate utterances. The duration of one utterance is 50 s and duration of another utterance is 10 s. In our work, we have used only the utterances of 50 s from each speaker for LID study. From each language we have considered calls both from male and female speakers.

3.3 Extraction of Implicit Excitation Source Information from Linear Prediction Residual

Speech sounds are generated due to the excitation of the human vocal tract resonator. The quasi-periodic air pulses generated by the vocal folds vibration, are used as the source of excitation to the vocal tract resonator during the voiced sound production. The excitation source information can be captured by processing the LP residual signal [1]. Therefore, to capture the knowledge about the excitation source, the raw LP residual samples can be used as features directly. However, the magnitude component of LP residual signal may prevails over phase component of LP residual signal during the processing of LP residual signal directly. The phase component of LP residual signal may also contain some information about the excitation source of human speech production system. In this work, we have processed the raw LP residual samples, the magnitude and phase components of LP residual signal independently to capture different aspects of excitation source information for language identification. The magnitude and phase components of LP residual signal can be separated by deriving the analytic signal of LP residual [3]. The analytic signal representation of LP residual is discussed in the following subsection.

3.3.1 Analytic Signal Representation of Linear Prediction Residual

Speech can be viewed as convolution of excitation source component of vocal tract system response, which is given by

$$s(n) = e(n) * h(n) \tag{3.1}$$

where, $s(n)$ is speech signal, $e(n)$ is excitation source signal and $h(n)$ is vocal tract system response. It is necessary to separate the two components for processing them independently. Linear prediction (LP) analysis has been proposed to separate the excitation source and the vocal tract response [1]. In the LP analysis, each speech sample is predicted as a linear combination of past p samples, where p is the order of prediction [1]. Each speech sample $s(n)$ can be predicted as follows

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{3.2}$$

where, $\hat{s}(n)$ is the predicted speech sample and a_k represent LP coefficients (LPCs). The error between original and predicted signal is called as prediction error or LP residual [1] which is denoted by $r(n)$ and is given by

$$r\,(n) = s\,(n) - \hat{s}(n) = s\,(n) + \sum_{k=1}^{p} a_k s\,(n-k) \tag{3.3}$$

The LP residual signal $r(n)$ is obtained by passing the speech signal through an inverse filter $A(z)$ [1] given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.4}$$

The predicted samples $\hat{s}(n)$ approximately model the vocal tract system in terms of linear prediction coefficients (LPCs). Suppression of vocal tract system response from the original speech signal $s(n)$ gives the LP residual signal $r(n)$. So, LP residual signal mostly represents the excitation source information. Proper representation of excitation source information by the LP residual depends upon the order of prediction (p). In [2], it has been shown that for a speech signal sampled at 8 kHz, the LP residual extracted using LP order in the range of 8–14 represents the excitation source information accurately. We use LP order of 10 followed by inverse filtering the speech signal (sampled at 8 kHz) for estimating the LP residual signal.

The analytic signal of LP residual denoted as $r_a(n)$ is given by

$$r_a(n) = r(n) + jr_h(n) \tag{3.5}$$

where, $r(n)$ is the corresponding LP residual signal and $r_h(n)$ is the Hilbert transform of the $r(n)$. $r_h(n)$ is obtained by performing Inverse Fourier Transform (IFT) on $R_h(\omega)$

$$r_h(n) = IFT[R_h(\omega)] \tag{3.6}$$

where

$$R_h(\omega) = \begin{cases} -jR(\omega), & 0 \leq \omega < \pi \\ jR(\omega), & 0 > \omega \geq -\pi \end{cases} \tag{3.7}$$

$R(\omega)$ is the Fourier transform of $r(n)$. The magnitude of the analytic signal is known as Hilbert envelope (HE) of the LP residual signal [7] which is written as

$$|r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{3.8}$$

The cosine of the phase is called as residual phase (RP) [7] which is written as

$$\cos(\theta(n)) = \frac{Re(r_a(n))}{|r_a(n)|} = \frac{r(n)}{|r_a(n)|} \tag{3.9}$$

In Fig. 3.1, a segment of voiced speech, corresponding LP residual signal, HE and RP of LP residual have been shown. The unipolar nature of HE can be observed from Fig. 3.1c, which only reflects the amplitude variation of LP residual signal. Whereas, the RP shown in Fig. 3.1d reflects only the phase information of LP residual signal. The analytic signal of LP residual represents the magnitude and phase information independently. The proposed implicit processing of LP residual signal, HE and RP are described in Sects. 3.3.2 and 3.3.3.

3.3.2 Implicit Processing of Linear Prediction Residual Signal

Linear prediction analysis represents the second order statistical features in terms of the autocorrelation coefficients. So, the LP residual signal does not represent any significant second order relations corresponding to the vocal tract resonator. Hence the autocorrelation function of the LP residual has low correlation values for nonzero time lags [2]. We surmise that, the language-specific information may present in the higher-order relations of the LP residual samples. However, it is difficult to capture accurately the higher order relations present in the LP residual signal using any parameteric approaches. In this work, we have explored raw LP residual samples for deriving language-specific information present in higher order relations. LP residual samples, magnitude of LP residual represented by HE and phase information of LP residual represented by RP have been proposed to characterize the language-specific excitation source information at different levels. The LP residual samples,

magnitude and phase information of LP residual can be analyzed individually at three levels: (i) sub-segmental level (within a glottal cycle or pitch cycle), (ii) segmental level (within 2-3 successive glottal cycles) and (iii) supra-segmental level (across 50 glottal cycles).

At sub-segmental level, the LP residual signal, is processed at 5 ms block size with a shift of 2.5 ms and corresponding LP residual samples are used as sub-segmental level features to capture the subtle variations within individual glottal cycles such as, glottal closing instants (GCI), glottal opening instants (GOI), glottal cycle peaks and its locations. At the segmental level, first the LP residual signal is decimated by a factor 4 to suppress the sub-segmental level information and then processed the LP residual signal, in block size of 20 ms with a shift of 2.5 ms and corresponding LP residual samples are used as segmental level features. The instantaneous pitch and *epoch* strength of a segment of speech are captured by segmental level processing of LP residual signal. At supra-segmental processing, the LP residual is first decimated by a factor 50 to eliminate the sub-segmental and segmental level information and then LP residual signal is processed in blocks of 250 ms with shift of 6.25 ms and corresponding LP residual samples are used as supra-segmental level features. At supra-segmental level the slow varying pitch and energy contour information are captured by processing LP residual signal. The decimation is performed on LP residual signal at segmental and supra-segmental levels to suppress the previous level's information and to reduce the dimensionality. In Fig. 3.2, the LP residual signal and its corresponding HE and RP have been shown at sub-segmental, segmental and

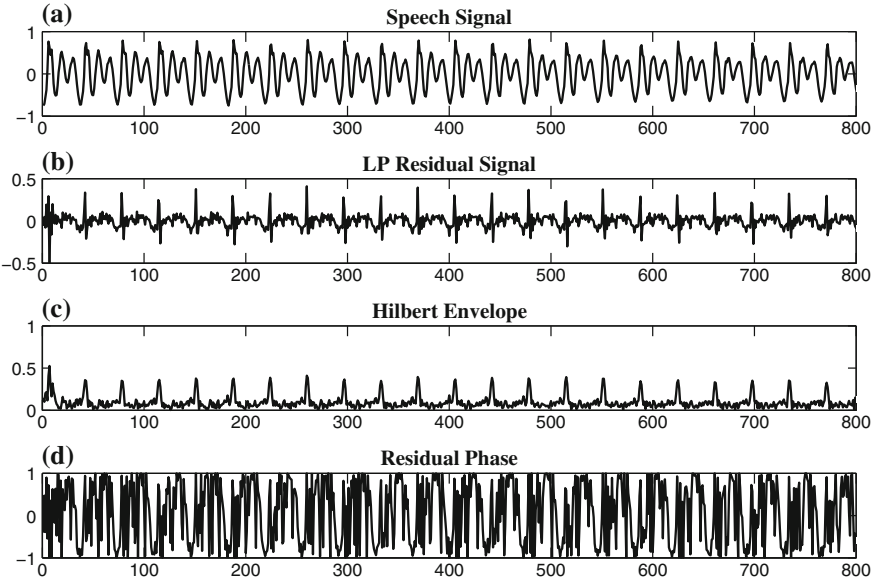

Fig. 3.1 **a** A segment of voiced speech and its **b** LP residual signal. **c** and **d** are Hilbert envelope and phase of corresponding LP residual signal, respectively

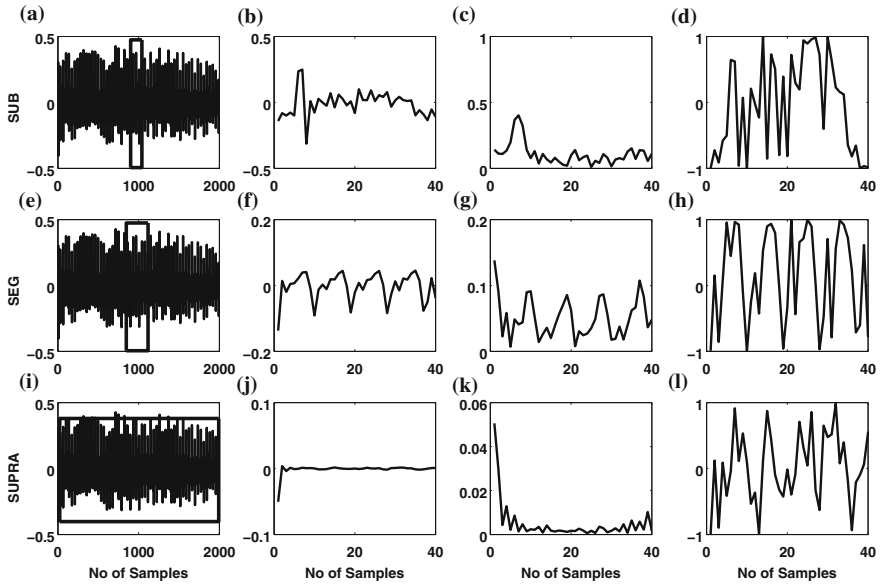

Fig. 3.2 **a**, **e** and **i** are LP residual signals marked by bounding boxes indicating sub-segmental, segmental and supra-segmental frames, respectively. **b**, **c** and **d** are sub-segmental level LP residual, Hilbert envelope and residual phase frames, respectively. **f**, **g** and **h** are segmental level (decimated by factor 4) LP residual, Hilbert envelope and residual phase frames, respectively. **j**, **k** and **l** are supra-segmental level (decimated by factor 50) LP residual, Hilbert envelope and residual phase frames, respectively

supra-segmental levels. Figure 3.2a shows the LP residual signal with 5 ms bounding box to show sub-segmental level frame. Figure 3.2b portrays the 5 ms windowed segment of LP residual signal. Figure 3.2e shows the LP residual signal with 20 ms bounding box to show segmental level frame. The 20 ms frame has been decimated by factor 4, which is shown in Fig. 3.2f. Figure 3.2i shows the LP residual signal with 250 ms bounding box to show supra-segmental level frame. Figure 3.2j depicts 250 ms frame after decimated by 50. From the Fig. 3.2b, f and j, it can be observed that the characteristics of LP residual samples are distinct at three different levels. Figure 3.2b delineates the minute variation of the LP residual samples within one glottal cycle. Periodic nature is observed from Fig. 3.2f, which represents the pitch and energy information of 2–3 consecutive glottal cycles. Figure 3.2j shows the pitch and energy contour information of several glottal cycles (50 cycles) which represents the supra-segmental level information.

3.3.3 *Implicit Processing of Magnitude and Phase Components of Linear Prediction Residual*

The magnitude component of linear prediction (LP) residual signal may predominates over phase component of LP residual signal during direct processing of raw LP residual samples. The amplitude component is not reliable, and it varies with respect to loudness or intensity of speech. Whereas, phase may be robust to above variations. The phase component of LP residual signal may also contain some information about the excitation source of speech production system. Hence, we have separated these two components by analytic signal representation of LP residual signal, which is discussed in Sect. 3.3.1. The magnitude and phase components of LP residual are represented by Hilbert envelope (HE) and residual phase (RP) of LP residual signal, respectively. We have proposed the implicit processing approach for HE and RP independently to capture the language-specific excitation source information.

The HE and RP of LP residual signal are also precessed implicitly at three different levels: sub-segmental, segmental and supra-segmental levels. Detailed description of implicit processing approach has been given in Sect. 3.3.2. Figure 3.2c, d shows the HE and RP estimated from 5 ms LP residual block shown in Fig. 3.2a, which represents the magnitude and phase information of LP residual signal at sub-segmental level, respectively. In Fig. 3.2e, LP residual signal has been shown and bounding box represents the 20 ms segmental level frame. Figure 3.2g and h portray the HE and RP estimated from 20 ms block after decimating by factor 4, which represents the magnitude and phase information of LP residual signal at segmental level, respectively. Figure 3.2i shows LP residual signal with a bounding box of 250 ms supra-segmental level frame. Figure 3.2k and l portray the HE and RP estimated from the 250 ms block after decimating by factor 50, which represents the magnitude and phase information of LP residual signal at supra-segmental level, respectively. From the Fig. 3.2c, g and k, it can be observed that the characteristics of HE of LP residual is distinct at three different levels. Similar observation can be made for RP of LP residual from Fig. 3.2d, h and l. Hence, HE and RP of LP residual signal may provide language-discriminative information at sub-segmental, segmental and supra-segmental levels. Therefore, in this work, we have processed the HE and RP separately to capture different demeanors of excitation source of speech production system for discriminating the languages.

3.4 Development of Language Identification Systems Using Implicit Excitation Source Features

In this work, LID systems using proposed implicit excitation source features are developed at three phases. The description of LID systems are given below. The block diagram of LID systems developed in various phases is shown in Fig. 3.3.

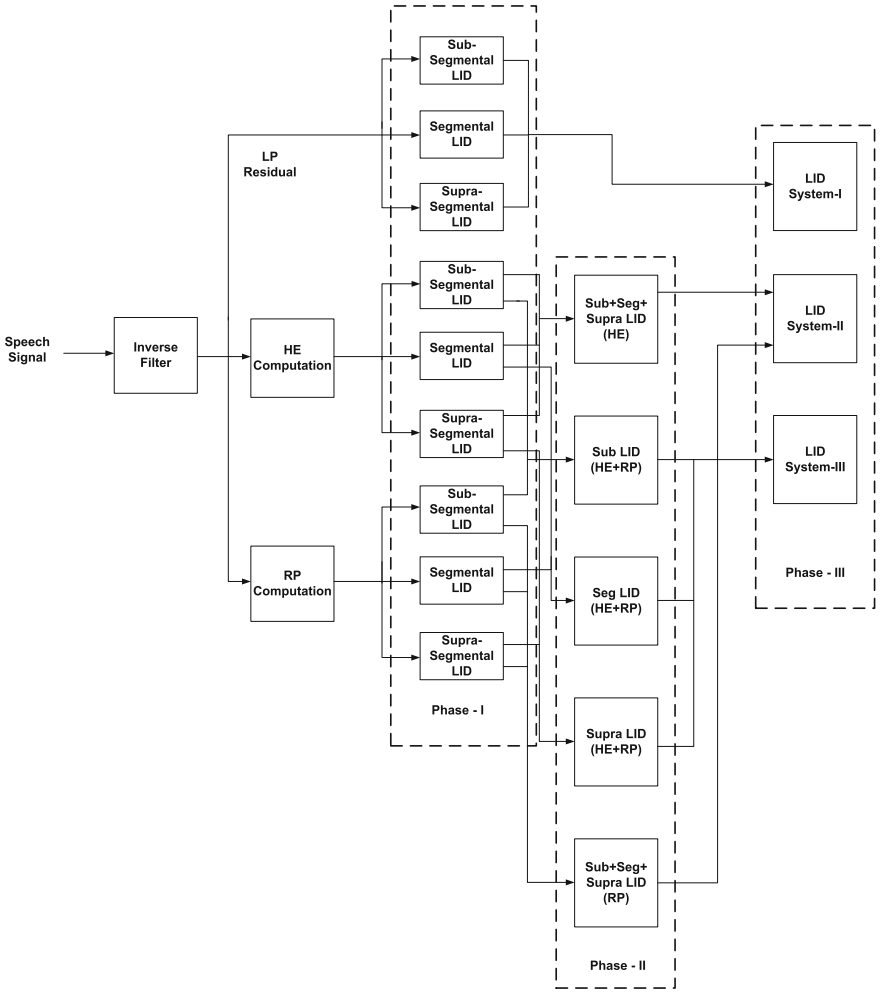

Fig. 3.3 Language identification systems using *implicit* excitation source features

Phase-I

At the phase-I, 9 different LID systems are developed:

- Three LID systems are developed by using the raw LP residual samples directly, extracted from sub-segmental (*sub*), segmental (*seg*) and supra-segmental (*supra*) levels.
- Three LID systems are developed by using the raw samples of HE of LP residual signal extracted from sub-segmental, segmental and supra-segmental levels.
- Three LID systems are developed by using the raw samples of RP of LP residual signal extracted from *sub*, *seg* and *supra* levels.

All nine different LID systems are developed using GMMs [8]. The LID systems are developed by using optimal number of Gaussian mixtures with respect to identification accuracy. The features at *sub*, *seg* and *supra* level contain partial information about the excitation source. Therefore, to achieve complete excitation source information for discriminating languages, we have combined the evidences from the LID systems developed by partial features. The block diagram shown in Fig. 3.3 represents different LID systems developed by proposed LP residual, HE and RP features and their combinations. In this study, we have performed eight different combinations which are shown in Fig. 3.3 at phase-II and phase-III.

Phase-II

In phase-II, we have performed the following five combinations:

- Evidences are combined from the LID systems developed using samples of HE extracted from *sub*, *seg* and *supra* levels.
- Evidences are combined from the LID systems developed using samples of RP extracted from *sub*, *seg* and *supra* levels.
- Since, HE and RP represents the magnitude and phase components of LP residual, the language-specific knowledge from these two components may be different at *sub*, *seg* and *supra* levels. Hence, we have explored the combinations of HE and RP at *sub*, *seg* and *supra* levels separately which is denoted as HE + RP in Fig. 3.3. The above mentioned combinations are as follows:
  - The evidences of HE and RP features are combined at *sub* level.
  - The evidences of HE and RP features are combined at *seg* level.
  - The evidences of HE and RP features are combined at *supra* level.

Phase-III

In phase-III, we have performed the following three combinations:

- Evidences are combined from the LID systems developed using LP residual samples extracted from *sub*, *seg* and *supra* levels.
- The first and fifth combinations shown in phase-II of Fig. 3.3 refer to LID systems with overall HE and RP features, respectively. These two features (HE and RP) represent magnitude and phase components of LP residual. Hence, the evidences obtained from these two systems are further combined to achieve the language-specific information completely from the excitation source view point. The LID system-II at phase-III of Fig. 3.3 is developed by the above mentioned combination.
- However, the combination of HE and RP at each level (HE + RP) represents partial information about excitation source. Hence, to acquire the complete language-specific excitation source information we have combined the evidences of (HE+RP) features from *sub*, *seg* and *supra* levels. The LID system-III at phase-III of Fig. 3.3 indicates the corresponding combination.

3.5 Performance Evaluation of LID Systems Developed Using Implicit Excitation Source Features

In this work, we have carried out the evaluation of the language models using leave-two-speaker-out approach. In this approach, $(n - 2)$ speakers are used from each language to develop the language models and the other two speakers of each language who have not participated during training phase are considered for evaluation. In each iteration, two speakers will be kept for evaluation and the other speaker's data is used for developing the language models. Average performances obtained from all iterations are shown here. 20 test utterances from each language, each of 10 s duration have been considered for evaluation of the language models. The Gaussian mixture models (GMMs) [8] are used to build the language models. The detailed description of training GMMs has been given in Appendix A. Different Gaussian mixtures (32, 64, 128 and 256) have been explored for capturing the language-specific excitation source information.

Phase-I

At phase-I, 9 different LID systems are developed by deriving the source information from *sub*, *seg* and *supra* levels of LP residual, HE and RP samples. Table 3.2 portrays the identification performances of individual languages as-well-as the average performances of LID systems. The first column represents the 27 Indian languages used in this LID study. The next three columns represent the LID performances obtained by processing the HE samples at *sub*, *seg* and *supra* levels. The average performances of HE feature at *sub*, *seg* and *supra* levels are 22.77, 32.03 and 25.18 %, respectively. The processing of HE samples of LP residual at segmental level represents the instantaneous pitch and energy information present in the LP residual envelopes within 2–3 consecutive glottal cycles, which yields better accuracy compared to other two levels. The different nature of raw HE samples at *sub*, *seg* and *supra* levels, which has been portrayed in the Fig. 3.2c, g and k can also be observed in the LID performances at the corresponding levels. For example, the LID performances of both Nagamese and Urdu languages are 50 % at *seg* level whereas, the *sub* and *supra* level implicit HE features are not able to identify these two languages.

The 5, 6 and 7th columns represent the LID performances obtained by processing the RP samples at *sub*, *seg* and *supra* levels, respectively. The average LID performances of RP features at *sub*, *seg* and *supra* levels are 42.03, 50.00 and 43.14 %, respectively. The *seg* level phase information provides better LID accuracy than the phase information present at *sub* and *supra* levels. The phase of LP residual contains more significant language-specific information at each level than the magnitude of LP residual signal, which can be inferred by comparing the average LID performances obtained by processing HE features and RP features. The processing of raw RP samples at *sub*, *seg* and *supra* levels also contain different aspects of excitation source which can be observed by analyzing the individual language performances. Such as, the Gujarati language has not identified by the phase

Table 3.2 LID Performances using Hilbert envelope (HE), residual phase (RP) and linear prediction residual (LPR) features at phase-I (Frame size of *sub*, *seg* and *supra* levels are 5, 20 and 250 ms, respectively)

Languages	Average recognition performances (%)								
	Phase-I								
	HE			RP			LPR		
	sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
Arunachali	45	5	10	50	45	45	0	35	70
Assamese	10	0	0	5	15	10	0	35	0
Bengali	0	5	10	5	5	0	0	5	0
Bhojpuri	0	5	0	10	25	55	50	35	10
Chhattisgarhi	30	50	100	50	100	95	0	100	100
Dogri	70	10	10	30	50	20	35	50	15
Gojri	15	25	100	30	95	100	95	80	95
Gujarati	0	30	40	40	0	80	0	15	45
Hindi	0	15	0	25	20	0	5	15	0
Indian English	75	90	95	100	100	100	100	95	90
Kannada	45	0	0	10	25	40	10	45	35
Kashmiri	0	70	10	85	95	35	100	95	25
Konkani	0	100	75	0	95	100	0	85	100
Malayalam	70	15	0	75	50	0	50	75	0
Manipuri	0	0	5	0	0	65	0	0	15
Marathi	0	0	0	0	0	5	0	0	0
Mizo	65	10	30	30	10	0	0	0	0
Nagamese	0	50	0	55	50	50	70	70	20
Nepali	0	35	20	25	20	25	60	0	10
Oriya	0	40	5	0	40	0	0	40	5
Punjabi	45	100	25	50	100	100	0	100	45
Rajasthani	15	10	70	100	65	100	0	35	45
Sanskrit	55	50	5	85	100	10	0	15	0
Sindhi	30	0	5	50	10	0	70	45	0
Tamil	45	0	10	65	60	15	0	35	20
Telugu	0	100	55	100	100	100	20	100	95
Urdu	0	50	0	60	75	15	0	0	15
Average performance	**22.77**	**32.03**	**25.18**	**42.03**	**50.00**	**43.14**	**24.62**	**44.62**	**31.66**

information at *seg* level, whereas, the phase of LP residual at *sub* and *supra* levels provide 40 and 80 % identification accuracy. The *sub* level RP feature provides 5, 25, 75, 30 and 50 % LID performances for Bengali, Hindi, Malayalam, Mizo and

Sindhi languages, respectively. The *seg* level RP feature provides 5, 20, 50, 10 and 10 % LID performances for the corresponding languages, respectively. However, the *supra* level phase information was unable to identify the corresponding languages.

The 8, 9 and 10th columns represent the LID performances obtained by processing the raw LP residual samples (LPR) directly at three levels. The average LID performances at *sub*, *seg* and *supra* levels are 24.62, 44.62 and 31.66 %, respectively. The LP residual features also contain different aspects of excitation source at three different levels which can be inferred by comparing the individual language performances at different levels. The LID performances of Arunachali, Chhattisgarhi, Gujarati, Punjabi, Rajasthani and Tamil languages obtained from *seg* level LPR feature are 35, 100, 15, 100, 35 and 35 %, respectively. Similarly, the *supra* level LPR feature provides LID performances of 70, 100, 45, 45, 45 and 20 % for the above mentioned languages, respectively. However, the *sub* level LPR feature could not identify the corresponding languages.

The magnitude or envelope information predominates over phase information while direct processing of LP residual signal. Therefore, we have processed separately the magnitude and phase components of LP residual signal. It can be elicited by analyzing the LID performances obtained from HE, RP and LPR features that, the phase of LP residual signal represented by RP contains significant language-specific information. The LID performances provided by the HE and RP features for a particular language are different, which indicates their distinct nature. So, empirically it is observed that, magnitude and phase components of LP residual reflects distinct aspects of excitation source. For example, the LID performances of Assamese, Kannada, Sindhi and Tamil languages provided by RP feature at *seg* level are 15, 25, 10 and 60 %. However, the *seg* level implicit HE feature was unable to identify the corresponding languages.

Phase-II

The magnitude component of LP residual signal represented by HE and phase component of LP residual represented by RP contains partial information about excitation source. Therefore, to derive the complete excitation source information, we have combined the evidences obtained from the LID systems developed by HE and RP. The combinations at score level can be performed in different ways, which have been shown at the phase-II and phase-III in Fig. 3.3. We use adaptive weighted combination scheme [9]. In this scheme, the evidences from different modalities are combined at the score level. The combined score C is given by

$$C = \sum_{i=1}^{k} w_i c_i \tag{3.10}$$

where, w_i and c_i denotes weighting factor and confidence score of ith evidence and k denotes number of modalities considered. The Weighting factor w_i varies from 0 to 1 with a step size of 0.01 and sum up to 1 (i.e., $\sum_{i=1}^{k} w_i = 1$). The evidences

Table 3.3 Performance of LID systems developed using combinations of implicit source features shown in phases-II and III of Fig. 3.3

Languages	Average recognition performances (%)							
	Phase-II					Phase-III		
	HE ($sub + seg + supra$)	RP ($sub + seg + supra$)	HE + RP			*src1*	*src2*	*src3*
			sub	*seg*	*supra*			
Arunachali	45	50	70	45	50	50	50	55
Assamese	0	50	5	25	10	50	45	25
Bengali	25	0	10	10	0	0	15	5
Bhojpuri	15	30	35	25	55	30	25	30
Chhattisgarhi	100	100	60	100	100	100	100	100
Dogri	60	100	65	50	5	100	90	95
Gojri	100	100	50	100	100	100	100	100
Gujarati	40	50	55	0	90	50	50	50
Hindi	20	20	35	15	0	20	20	5
Indian English	100	100	95	100	100	100	100	100
Kannada	10	40	30	20	30	40	45	45
Kashmiri	75	100	70	100	35	100	100	100
Konkani	100	100	0	95	100	100	100	100
Malayalam	55	60	90	55	5	60	65	75
Manipuri	0	55	0	0	70	55	30	0
Marathi	0	0	0	0	5	0	0	0
Mizo	50	5	60	10	5	5	20	0
Nagamese	45	100	0	50	50	100	100	100
Nepali	35	20	45	25	25	20	20	15
Oriya	35	30	0	40	5	30	40	35
Punjabi	100	100	90	100	100	100	100	100
Rajasthani	90	100	100	80	100	100	100	100
Sanskrit	45	100	100	100	10	100	100	30
Sindhi	10	45	50	10	0	45	50	85
Tamil	50	70	80	45	20	70	70	55
Telugu	100	100	55	100	100	100	100	100
Urdu	50	90	40	75	15	90	85	85
Average performance	**50.18**	**63.51**	**47.77**	**50.92**	**43.88**	**63.51**	**63.70**	**58.88**

obtained from LID systems developed by HE feature at *sub*, *seg* and *supra* levels are combined which provides 50.18 % average LID performance shown at 2nd column of Table 3.3. The identification accuracy of 63.51 % is obtained by combining the evidences obtained from RP features at *sub*, *seg* and *supra* levels (see 3rd column of Table 3.3).

However these two features contain different information which can be observed by comparing the individual language performances shown at 2nd and 3rd columns of Table 3.3. The LID performances of Assamese and Manipuri languages are 50 and 55 % obtained from combined RP feature. However, the LID performances of both the languages obtained from combined HE feature are 0 %. Similarly, the LID accuracy of Bengali language obtained from HE feature is 25 %, whereas, the RP feature could not identify the Bengali language. The evidences obtained from HE and RP features are also combined at each level which are shown at 4, 5 and 6th columns of Table 3.3. The combined HE + RP information provides 47.77, 50.92 and 43.88 % identification performances at *sub*, *seg* and *supra* levels, respectively. The *seg* level information gives better accuracy compared to other levels. The combined HE + RP feature at each level contains different information which can be observed by comparing the individual language performances, which are shown at 4, 5 and 6th columns of Table 3.3. For example, the LID performances at *seg* level for Konkani and Nagamese languages are 95 and 50 %. The *supra* level performances for corresponding languages are 100 and 50 %. Whereas, the combined HE + RP feature at *sub* level did not identify the corresponding languages.

Phase-III

Three LID systems are developed at phase-I of Fig. 3.3 by processing the LP residual samples at *sub*, *seg* and *supra* levels. The evidences obtained from these three LID systems provide partial information from excitation source point of view. Therefore, to derive the complete information, we have combined the evidences from the first three LID systems at phase-I to develop the LID system-I at phase-III in Fig. 3.3. Identification accuracy of 58.88 % is obtained from LID system-I at phase-III shown in Fig. 3.3. The combined HE feature at *sub*, *seg* and *supra* levels contain only the magnitude information of LP residual signal which is shown in 2nd column of Table 3.3. The combined RP feature at *sub*, *seg* and *supra* levels contain only the phase information of LP residual signal which is shown in 3rd column of Table 3.3. Therefore, to obtain the complete information about the excitation source we have again combined the evidences obtained from these two features to develop the LID system-II at phase-III. The 7th column of Table 3.3 represents the individual as-well-as the average LID performance obtained from LID system-II at phase-III, which is denoted as *src1*. *src1* provides LID performance of 63.51 % which is equivalent to the LID performance contributed by only RP feature. This empirical observation infers that, the phase component of LP residual signal is more significant for language identification than magnitude component of LP residual. The HE and RP features have also been combined at each level which is shown at 4, 5 and 6th columns of Table 3.3. The combined HE + RP features at *sub*, *seg* and *supra* levels reflect partial information from excitation source point of view. Therefore, we have further combined the evidences obtained from 2nd, 3rd and 4th LID systems of phase-II, to build LID system-III at phase-III in Fig. 3.3. The average LID performance of 63.70 % shown in 8th column of Table 3.3 denoted as *src2* is obtained from the LID system-III at phase-III. In this work, we have developed LID systems by processing language-specific excitation

Table 3.4 Performances of LID systems developed using implicit HE, RP and LPR features and evaluated by male speaker set

Languages	Average recognition performances (%)								
	HE			RP			LPR		
	sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
Arunachali	0	0	0	28.6	0	42.9	0	0	71.5
Assamese	0	0	0	14.3	28.6	0	0	71.5	0
Bengali	0	0	14.3	0	0	0	0	0	0
Bhojpuri	0	14.3	0	14.3	42.9	14.3	100	57.2	0
Chhattisgarhi	0	0	100	0	100	100	0	100	100
Dogri	57.2	0	0	28.6	0	14.3	28.6	0	0
Gojri	14.3	42.9	100	28.6	85.8	100	85.8	100	100
Gujarati	0	57.2	0	42.9	0	71.5	0	0	0
Hindi	0	28.6	0	42.9	42.9	0	0	28.6	0
Indian English	100	85.8	100	100	100	100	100	100	100
Kannada	28.6	0	0	0	28.6	57.2	0	28.6	28.6
Kashmiri	0	71.5	0	71.5	85.8	28.6	100	85.8	0
Konkani	0	100	42.9	0	85.8	100	0	100	100
Malayalam	57.2	28.6	0	57.2	57.2	0	0	57.2	0
Manipuri	0	0	0	0	0	100	0	0	28.6
Marathi	0	0	0	0	0	0	0	0	0
Mizo	0	14.3	28.6	0	14.3	100	0	0	0
Nagamese	0	100	0	71.5	100	0	42.9	100	0
Nepali	0	71.5	28.6	42.9	0	0	100	0	0
Oriya	0	0	0	0	0	0	0	0	14.3
Punjabi	14.3	100	28.6	42.9	100	100	0	100	42.9
Rajasthani	0	28.6	71.5	100	100	100	0	0	85.8
Sanskrit	71.5	57.2	14.3	71.5	100	14.3	0	14.3	0
Sindhi	0	0	0	0	0	0	57.2	71.5	0
Tamil	42.9	0	14.3	71.5	71.5	0	0	57.2	42.9
Telugu	0	100	42.9	100	100	100	28.6	100	85.8
Urdu	0	42.9	0	57.2	85.8	0	0	0	0
Average performance	**14.28**	**34.92**	**21.69**	**36.5**	**49.2**	**42.32**	**23.8**	**43.38**	**29.62**

source information in different ways which are denoted as *src1*, *src2* and *src3* in Table 3.3. The average identification accuracy obtained from *src2* feature is better than the LID performances obtained from *src1* and *src3* systems.

In our study, language models are developed using both male and female speakers present in training dataset. However, we further separated the male and female

Table 3.5 Performances of LID systems developed using implicit HE, RP and LPR features and evaluated by female speaker set

Languages	Average recognition performances (%)								
	HE			RP			LPR		
	sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
Arunachali	80	20	40	60	100	60	0	100	80
Assamese	0	0	0	0	0	0	0	0	0
Bengali	0	0	0	20	20	0	0	20	0
Bhojpuri	0	0	0	0	0	100	0	0	40
Chhattisgarhi	20	100	100	100	100	100	0	100	100
Dogri	100	60	20	40	100	60	60	100	40
Gujarati	0	0	60	40	0	100	0	40	100
Hindi	0	0	0	0	0	0	20	0	0
Indian English	60	100	80	100	100	100	100	100	80
Konkani	0	100	100	0	100	100	0	100	100
Malayalam	80	0	0	100	40	0	100	100	0
Manipuri	0	0	0	0	0	40	0	0	0
Marathi	0	0	0	0	0	20	0	0	0
Mizo	100	0	0	80	0	0	0	0	0
Nagamese	0	0	0	80	0	100	100	40	40
Nepali	0	0	0	0	0	60	0	0	20
Oriya	0	100	0	0	100	0	0	100	0
Rajasthani	20	0	100	100	40	100	0	80	60
Sindhi	40	0	20	100	40	0	100	40	80
Tamil	80	0	0	60	0	0	0	0	0
Average performance	**29**	**24**	**26**	**44**	**37**	**47**	**24**	**46**	**37**

speakers of test dataset to evaluate the LID systems separately. This speaker independent LID study is carried out to examine the effects of gender in language discrimination task. Performances obtained from male and female sets are shown in Tables 3.4 and 3.5, respectively. Some language database does not contain female speakers in test data set such as, Gojri, Kannada, Kashmiri, Punjabi, Sanskrit, Telugu and Urdu. Seven and five speakers per each language have been used during evaluation using male and female speaker sets, respectively. From this study it is observed that, segmental level *implicit* features provide best average LID accuracy followed by supra-segmental and sub-segmental levels, respectively. Similar trend also observed during evaluation without separating male and female speakers (see Table 3.2). However, LID accuracy obtained from HE feature during the evaluation using female speaker set are 29, 24 and 26 % for *sub*, *seg* and *supra* levels, respectively. The LID performances achieved from RP feature using female speaker set are 44, 37 and 47 % for *sub*, *seg* and *supra* levels, respectively.

Table 3.6 LID Performances using *implicit* Hilbert envelope (HE), residual phase (RP), linear prediction residual (LPR) features evaluated on OGI-MLTS database

Languages	Average recognition performances (%)								
	Phase-I								
	HE			RP			LPR		
	sub	*seg*	*supra*	*sub*	*seg*	*supra*	*sub*	*seg*	*supra*
English	0	44.44	100	22.22	88.89	77.78	33.33	44.44	100
Farsi	44.44	44.44	33.33	11.11	77.78	11.11	33.33	100	88.89
French	11.11	100	0	66.67	88.89	0	0	66.67	44.44
German	11.11	100	0	77.78	22.22	11.11	0	11.11	0
Hindi	0	100	11.11	22.22	55.56	66.67	0	55.56	11.11
Japanese	100	33.33	66.67	44.44	0	77.78	0	44.44	11.11
Korean	22.22	11.11	0	44.44	66.67	0	100	55.56	0
Mandarin	0	0	88.89	33.33	11.11	44.44	0	0	0
Spanish	0	0	66.67	77.78	22.22	22.22	44.44	11.11	11.11
Tamil	33.33	88.89	0	0	100	66.67	44.44	77.78	11.11
Vietnamese	0	0	0	44.44	22.22	33.33	88.89	22.22	55.56
Average performance	**20.2**	**47.47**	**33.33**	**40.4**	**50.5**	**37.37**	**31.31**	**44.44**	**30.3**

Table 3.7 Performance of LID systems developed using combined *implicit* source features on OGI-MLTS database

Languages	Average recognition performances (%)							
	Phase-II					Phase-III		
	HE (*sub* + *seg* + *supra*)	RP (*sub* + *seg* + *supra*)	HE + RP			*src1*	*src2*	*src3*
			sub	*seg*	*supra*			
English	55.56	77.78	22.22	55.56	100	77.78	88.89	33.33
Farsi	55.56	55.56	33.33	77.78	11.11	77.78	33.33	100
French	100	55.56	88.89	100	11.11	66.67	55.56	11.11
German	77.78	66.67	77.78	77.78	22.22	66.67	44.44	0
Hindi	88.89	66.67	11.11	88.89	77.78	44.44	100	0
Japanese	100	66.67	77.78	33.33	100	77.78	88.89	0
Korean	22.22	55.56	55.56	77.78	0	55.56	33.33	100
Mandarin	11.11	44.44	33.33	0	55.56	44.44	66.67	0
Spanish	11.11	77.78	77.78	0	44.44	77.78	55.56	11.11
Tamil	33.33	77.78	0	100	33.33	77.78	66.67	100
Vietnamese	22.22	55.56	33.33	11.11	11.11	55.56	33.33	100
Average performance	**52.52**	**63.63**	**46.46**	**56.56**	**42.42**	**65.65**	**60.6**	**41.41**

3.6 Evaluation of LID Systems Developed Using Implicit Excitation Source Features on OGI-MLTS Database

The proposed implicit features of excitation source are also evaluated on OGI-MLTS database. The LID performances obtained from OGI-MLTS database are shown in Tables 3.6 and 3.7. The *seg* level RP phase feature provides better accuracy, compared to others. It can be observed from LID accuracies shown in columns 2 to 7 of Table 3.6 that, magnitude and phase components of LP residual signal provide distinct language-specific information. It can be stated from the LID performances presented in Table 3.6 that, the language-specific knowledge present at *sub*, *seg* and *supra* levels are fundamentally distinct. Hence, the evidences from three different levels are combined to capture complete excitation source features. The best average LID accuracy of 65.65 % is obtained by processing complete excitation source information (see column 7 of Table 3.7).

3.7 Summary

Most of the languages in India share a common set of phonemes but, the characteristics of same sound unit may vary across different languages due to co-articulation effects and dialects. Speech signal carries this phonotactic information which is language-specific. Hence, the characteristics of vocal tract shapes and excitation source also contain the sequential information of phonotactics which is unique for individual languages. In this work, we have focused only the characteristics of excitation source. It is known from the literature that, excitation source information can be captured by processing the LP residual signal. Therefore, we have processed the raw LP residual samples, its magnitude and phase separately for capturing language-specific excitation source information. In this chapter, an unified framework is proposed for capturing the excitation source information by processing LP residual signal, its magnitude and phase components implicitly. Raw LP residual samples, its magnitude and phase components have been processed at three different levels: sub-segmental, segmental and supra-segmental levels for LID task. The minute variations present within one glottal cycle, instantaneous pitch and energy of 2–3 consecutive glottal cycles, and variations of the pitch and energy across 50 glottal cycles are captured at sub-segmental, segmental and supra-segmental levels, respectively. Experimental results illustrate that, the language-specific source information present at sub-segmental, segmental and supra-segmental levels are fundamentally distinct. Therefore, the evidences from each level are combined to capture the complete excitation source information. The information provided by segmental level processing of LP residual is more effective, compared to the other two levels. This indicates the efficacy of instantaneous pitch and energy in language discrimination task. In direct processing of the LP residual, the effect of the magnitude component of LP residual predominates over the phase component of LP residual. To assuage

this, the magnitude and phase information is captured independently using the analytic signal representation of the LP residual. From the empirical observation it can be stated that, combination of magnitude information represented by HE and phase information represented by RP seem to be a better choice than direct processing of LP residual for language discrimination task.

References

1. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
2. S.R.M. Prasanna, C.S. Gupta, B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**, 1243–1261 (2006)
3. L. Cohen, Time frequency distribution: a review. Proc. IEEE **77**, 941–979 (1989)
4. S. Maity, A.K. Vuppala, K.S. Rao, D. Nandi, IITKGP-MLILSC speech database for language identification, in *National Conference on Communication*, February 2012
5. Y.K. Muthusamy, R.A. Cole, B.T. Oshika, The OGI multilanguage telephone speech corpus, in *Spoken Language Processing*, pp. 895–898, October 1992
6. Prasar Bharati, http://www.newsonair.nic.in/. May 2014
7. K.S.R. Murty, B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett. **13**(1), 52–55 (2006)
8. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Audio Speech Lang. Process. **3**(1), 72–83 (1995)
9. V.R. Reddy, S. Maity, K.S. Rao, Identification of indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. (Springer) **16**(4), 489–511 (2013)

Chapter 4
Parametric Excitation Source Features for Language Identification

Abstract This chapter describes the proposed methods to extract parametric features at sub-segmental, segmental and supra-segmental levels to capture the language-specific excitation source information. In this work, glottal pulse, spectral and epoch parameters are used for representing sub-segmental, segmental and supra-segmental information present in excitation source signal. Further, these individual features are combined at score level to enhance the accuracy of LID systems by exploiting the non-overlapping information present among the features.

Keywords Parametric representation of excitation source information · Sub-segmental level features · Segmental level features · Suprasegmental level features · Glottal pulse parameters · Glottal flow derivative · Glottal flow parameters · RMFCC features · MPDSS features · Epoch parameters

4.1 Introduction

Implicit processing of raw LP residual samples in time domain is computationally intensive. There will be information loss due to decimation of residual samples at segmental and supra-segmental levels during the *implicit* processing. Hence, more compact, simple and effective approaches are needed to capture the excitation source information from LP residual signal. However, there is no efficient approach to characterize the excitation source features for capturing language-specific phonotactic information. In this chapter, LP residual signal has been parameterized at sub-segmental (*sub*), segmental (*seg*) and supra-segmental (*supra*) levels to capture different aspects of excitation source information from the language identification point of view. The glottal flow derivative (GFD) parameters are extracted from LP residual signal to capture the *sub* level excitation source information. The frequency domain representation often provides certain cues of a particular signal which may not be captured by temporal processing of signal [1]. The magnitude spectrum of LP residual signal represents the energy and periodicity information relevant to excitation source. Hence, the LP residual signal is processed in spectral and cepstral domains to capture the *seg* level excitation source information. Further, the rate of vocal folds vibration (pitch) and the strength of excitation at glottal closing instants(GCI) also

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_4

varies with time. LP residual signal is processed at *supra* level to derive this temporal variation. The excitation source information obtained from the three levels portray distinct language-specific information. Therefore, the evidences from all three levels are combined to obtain the complete excitation source information for LID task.

Rest of this chapter is organized as follows: The proposed parametric representations of excitation source are described in Sect. 4.2. The description of LID systems developed using the proposed features is delineated in Sect. 4.3. The evaluation of LID systems using proposed excitation source features is given in Sects. 4.4 and 4.5. The Sect. 4.6 summarizes this chapter.

4.2 Parametric Representation of Excitation Source Information

Excitation source signal consists of sequence of quasi-periodic glottal air pulses produced by the vocal folds vibration. This information can be captured by processing the LP residual signal [2] at different levels. In Sects. 4.2.1–4.2.3, proposed parametric approaches have been described to capture the sub-segmental, segmental and supra-segmental level excitation source information, respectively.

4.2.1 Parametric Representation of Sub-segmental Level Excitation Source Information

In this section, language-specific excitation source information has been captured by parameterizing LP residual signal within a glottal cycle. The Liljencrants-Fant (LF) parameters which model the glottal flow derivative (GFD) are used for parameterizing the sub-segmental level source information. Techniques are proposed for estimating the LF model parameters from the LP residual signal.

LF Model of Glottal Flow Derivative

The glottal air pulses are the primary source to provoke the vocal tract resonator during the production of voiced speech. The vibration of the vocal folds during the production of voiced speech generate the glottal flow which is quasi-periodic. A segment of voiced speech and corresponding LP residual signal have been shown in Fig. 4.1a, b, respectively. Figure 4.1c portrays the glottal volume velocity (GVV) waveform corresponding to the speech signal. At sub-segmental level, LP residual signal is analyzed in block size of 5 ms with a shift of 2.5 ms to capture the characteristics of one glottal cycle. A glottal pulse and its glottal flow derivative (GFD) are shown in Fig. 4.2a, b, respectively. Glottal pulse is defined as a single period of the glottal air flow. Each glottal flow cycle is divided into three phases: *closed-phase*, *open-phase* and *return-phase*. During the time of *closed-phase* no air flows through the glottis, because the vocal folds remain closed completely. So, the glottal flow derivative corresponding to *closed-phase* may not contain any significant information about excitation source. The interim during which the vocal

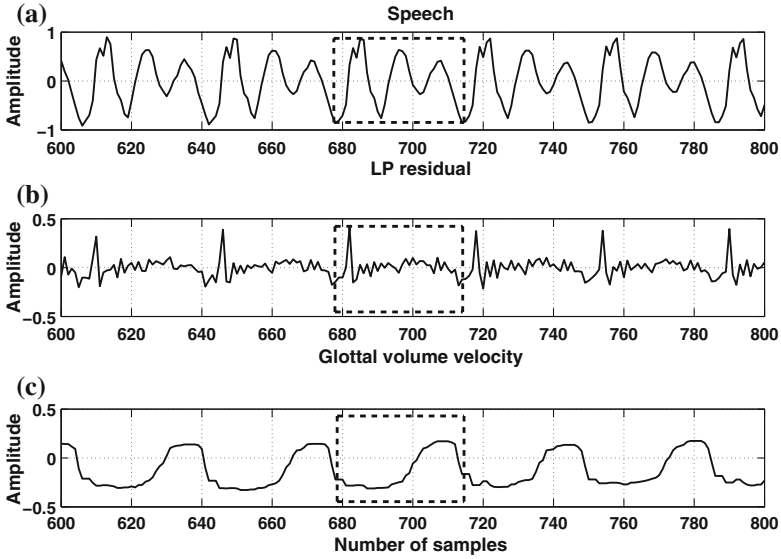

Fig. 4.1 **a** Speech signal, **b** corresponding LP residual signal and **c** glottal volume velocity marked by bounding boxes represent sub-segmental level frame (5 ms)

folds are open and air flows through the glottis is known as *open-phase*. At the beginning of *open-phase*, the interaction between the vocal folds and the vocal tract resonator increases and it continues until the flow becomes constant. The shapes of glottal air pulses depend upon the vocal folds vibration characteristics [3]. The traits which characterize the glottal pulse shape are the rate of increase of the glottal air flow, the magnitude of maximum air flow and the maximum rate of decrease of the air flow and the corresponding time instants. The shape of a GFD cycle corresponding to a glottal pulse is also characterized by the rate of increase or decrease of the glottal air flow, the zero-crossing point represents the instant of the maximum air flow through the glottis and the negative peak of GFD waveform represents the instant of maximum rate of decrease of the air flow. The characteristics of vocal folds vibration is unique for each sound unit and same sound unit may vary across the languages due to phonotactic characteristics. Hence, the shape of the glottal pulse and the corresponding GFD waveform during the *open-phase* interval is also unique across different sound units and may provide language discriminative information. We presume that, the *open-phase* of the GFD cycle may contain significant language-specific knowledge.

Return-phase is defined as the interim between the time instant of maximum rate of air flow decrease and the time instant of zero air flow. *return-phase* for a glottal pulse is shown in Fig. 4.2a. In this phase, vocal folds tend to close. High frequency energy generates due to the sudden close of vocal folds which is also reflected in the *return-phase* characteristics. The duration of this phase and the slope of GFD waveform corresponding to the *return-phase* depends on the rapidness of

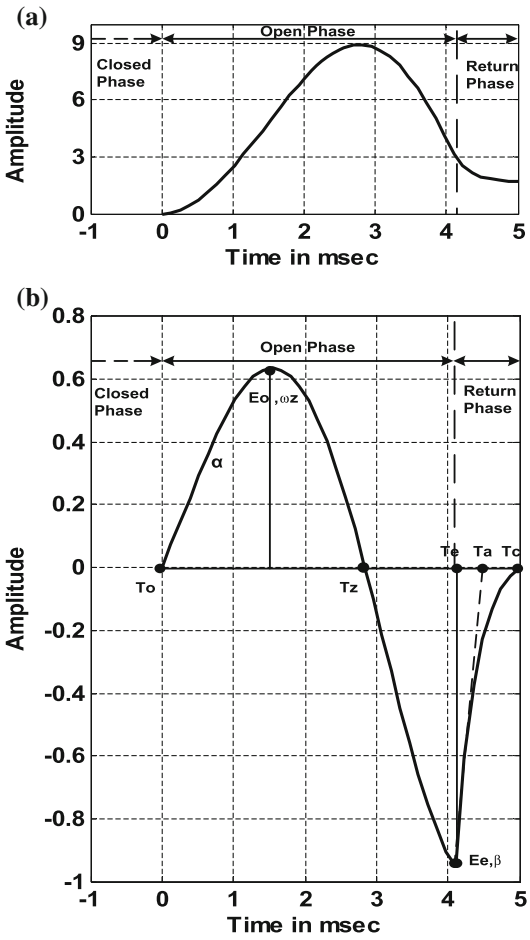

Fig. 4.2 **a** Glottal flow **b** LF model of GFD for one glottal cycle

vocal folds closing. Hence, this phase provides significant information about the excitation source [4]. The rapidness of closing of vocal folds reduces the duration of the *return-phase*, which yields more high frequency energy. These characteristics can be observed from the exponential nature of the GFD during the *return-phase* shown in Fig. 4.2b. The glottal air flow during the *return-phase* generally provides the perceptual importance because it determines the spectral slope [5]. The spectral slope is defined as a measure of how quickly the spectrum of a speech signal tails off towards the high frequencies. The sudden change of glottal pulse in *return-phase* may be context dependent. The rapidness of vocal folds closing for one phoneme may vary from one language to another which reflects the uniqueness of a language. Hence, we conjecture that, the *return-phase* may contain language-specific information.

The LF model [6] of the glottal air flow illustrates the GFD waveform in terms of exponentially growing sinusoid in the *open-phase* and a decaying exponential in the *return-phase* [4]. The LF model parameters of a GFD cycle shown in Fig. 4.2b are listed below.

1. T_c: Time instant of glottal closure (GCI).
2. T_o: Time instant of glottal opening (GOI).
3. E_e: Absolute value of the maximum rate of glottal flow decrease.
4. T_e: The time instant of maximum rate of glottal flow decrease.
5. α: The growth factor defined as the ratio of E_e to maximum rate of glottal flow increase.
6. β: Exponential time constant that determines how quickly glottal flow derivative returns to zero after time T_e.
7. ω_z: Frequency that determines flow derivative curvature to the left of the first GFD zero crossing (T_z), $\omega_z = \frac{\pi}{T_z - T_o}$.

These seven LF model parameters are estimated from each glottal cycle of LP residual signal. The three phases of a single GFD cycle denoted as $e_{LF}(t)$ can be expressed mathematically by the following way [4, 7].

$$\begin{aligned} e_{LF}(t) &= 0, && 0 \leq t < T_o \\ &= E_0 e^{\alpha(t-T_0)} \sin[w_z(t-T_0)], && T_0 \leq t < T_e \\ &= -\frac{E_e}{\beta T_a}[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], && T_e \leq t < T_c \end{aligned} \tag{4.1}$$

where, E_o is a gain constant and T_a is (time constant of the return phase) the time instant where the slope of *return-phase* crosses the time axis. In the LF model analysis it is generally assumed that there is no air flow during the *closed-phase*. Hence, $e_{LF}(t) = 0$ is assumed at *closed-phase*. From the Fig. 4.2b it can be observed that, T_o, T_e, E_e, α, ω_z and E_o characterize the *open-phase* and E_e, β and T_c characterize the *return-phase*. All the seven parameters estimated from each of the glottal cycles of LP residual signal are used as feature vector to capture the sub-segmental level excitation source information for LID task.

Proposed Method for Computation of the LF Model Parameters

The LF model parameters are computed from each glottal cycle (5 ms window) to capture the language-specific information present within one glottal cycle. The proposed method for computing seven LF parameters is discussed below.

1. **Computation of T_c**
   T_c indicates the glottal closing instant (GCI). In our work, we have used *zero-frequency filtering* [8] method to calculate the GCI locations. In this method, the speech signal is passed through the *zero-frequency resonator* twice. The reason of passing the speech signal twice through the *zero-frequency resonator* is to mitigate the effects of high frequency resonances. This yields a filtered output which varies as a polynomial function of time which is known as zero-

frequency filtered signal (ZFFS). The positive zero crossings in the ZFFS indicates the GCI locations [8].

2. **Computation of T_o**
   The steps to determine the GOIs are described below:
   - Compute the pitch period (P_g) of the gth glottal cycle as, $P_g = T_c(g) - T_c(g-1)$, where, $T_c(g)$ and $T_c(g-1)$ are the closing instants of the gth and its just previous glottal cycle.
   - Compute the average pitch period P_{avg}.
   - Compute the opening instant of the $(g+1)$th glottal cycle ($T_{o(g+1)}$) by using Eq. 4.2 [9].

$$T_{o(g+1)} = T_{cg} + 0.3 \times min[T_{c(g+1)} - T_{c(g)}, P_{avg}] \tag{4.2}$$

3. **Computation of E_e**
   E_e indicates the absolute value of the maximum rate of glottal flow decrease. The magnitude of the negative peak of the GFD waveform shown in Fig. 4.2b indicates the maximum rate of glottal flow decrease. This can be calculated by finding the absolute maximum value from LP residual samples within one glottal cycle.

4. **Computation of T_e**
   Once the E_e is computed, the corresponding time instant T_e also can be obtained, which denotes the time instant of maximum rate of glottal flow decrease.

5. **Computation of α and β**
   The amplitude of the GFD cycle is zero during the *closed-phase*. Therefore, there is no significant information present in *closed-phase*. Hence, we can consider that, the effective GFD cycle is from the instant of the glottal opening (T_o) to the instant of the glottal closing (T_c). By assuming $T_o = 0$ (i.e., the glottal opening instant starts from $t = 0$), the Eq. 4.1 will become as follows:

$$e_{LF}(t) = E_o e^{\alpha(t)} \sin[w_z(t)], \qquad 0 \le t < T_e \tag{4.3}$$

$$= -\frac{E_e}{\beta T_a}[e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}], \qquad T_e \le t < T_c \tag{4.4}$$

Now, we can observe from the LF model of GFD cycle shown in Fig. 4.2b that, the value of the negative peak of GFD waveform is $-E_e$ at time instant $t = T_e$ (i.e., $\left[e_{LF}(t)\right]_{t=T_e} = -E_e$). Now, if we put $t = T_e$ in the Eqs. 4.3 and 4.4, we will obtain the

$$\left[e_{LF}(t)\right]_{t=T_e} = E_o e^{\alpha T_e} \sin[w_z(T_e)] \tag{4.5}$$

$$\left[e_{LF}(t)\right]_{t=T_e} = -\frac{E_e}{\beta T_a}[1 - e^{-\beta(T_c-T_e)}] \tag{4.6}$$

So, we can get the following relation:

$$E_o e^{\alpha T_e} \sin(\omega_z T_e) = -E_e$$
$$\Rightarrow \alpha = \frac{1}{T_e} \ln\left[-\frac{E_e}{E_o \sin(\omega_z T_e)} \right] \tag{4.7}$$

Again, by comparing Eq. 4.6 and $\left[e_{LF}(t)\right]_{t=T_e} = -E_e$ we can get the following equation

$$-\frac{E_e}{\beta T_a}[1 - e^{-\beta(T_c - T_e)}] = -E_e$$
$$\Rightarrow 1 - e^{-\beta(T_c - T_e)} = \beta T_a \tag{4.8}$$

In [7], it is assumed that the return flow is relatively faster. So, the assumption $\beta(T_c - T_e) \gg 1$ has been taken. With this assumption we can reduce the Eqs. 4.8–4.10 as follows:

$$\beta(T_c - T_e) \gg 1$$
$$\Rightarrow e^{-\beta(T_c - T_e)} \simeq 0 \tag{4.9}$$

So,

$$\beta T_a = 1 \tag{4.10}$$

A constraint is imposed with the above assumption that, the GFD cycle returns to zero at the end of each cycle. It means, the area under the GFD curve is 0.

$$\int_0^t e_{LF}(t)dt = 0 \tag{4.11}$$

To compute the value of β, we have to solve the Eq. 4.11.

$$\int_0^t e_{LF}(t)dt = 0$$
$$\Rightarrow \int_0^{T_e} e_{LF}(t)dt + \int_{T_e}^{T_c} e_{LF}(t)dt = 0$$
$$\Rightarrow \int_0^{T_e} E_o e^{\alpha(t)} \sin(\omega_z t)dt + \int_{T_e}^{T_c} -\frac{E_e}{\beta T_a}[e^{-\beta(t - T_e)} - e^{-\beta(T_c - T_e)}]dt = 0 \tag{4.12}$$

By solving the above equation we obtain the expression of β as follows:

$$\beta = \frac{E_e(\alpha^2 + \omega_z^2)}{E_o\{e^{\alpha T_e}[\alpha sin(\omega_z T_e) - \omega_o cos(\omega_z T_e)] + \omega_z\}} \tag{4.13}$$

The α and β are modified iteratively until the Eq. 4.10 is satisfied. The number of iterations are fixed based on experimental studies. To terminate the computation procedure it is bounded by 10 iterations which provides best LID performance reported in this work.

6. **Computation of ω_z**
   As the LP residual signal is a noise like signal, it is difficult to find out the T_z and E_o accurately. In our work, we have proposed an iterative procedure to find out these two parameters. The parameters T_z and E_o are related to the *open-phase* of the glottal cycle which is generally larger than the *return-phase*. Initially, we assume that, T_z (the first zero crossing of GFD cycle) is 50 % of the total glottal cycle duration. With this assumption, E_o is measured as the absolute maximum value of the glottal cycle up to T_z. Now, we can easily compute the α and β by Eqs. 4.7 and 4.13, respectively. Thus, to verify the accuracy of the initial estimation of the parameters, the constraint of Eq. 4.10 has been imposed. In every iteration, the T_z value is increased by 5 % of the glottal cycle. The reason for increasing the T_z value is due to the larger duration of the open phase. In our work, we have used 10 iterations which is fixed by experimental observations.

We also have explored the dynamic nature of the LF parameters to capture the fine variations of the glottal activities. The variations in the LF parameters from one glottal cycle to the other may be attributed to the fine variations in the glottal cycles. The dynamic nature can be captured by Delta coefficients. The following equation is used for computing the Delta coefficients.

$$\Delta c[m] = \frac{\sum_{i=1}^{k} i(c[m+i] - c[m-i])}{2\sum_{i=1}^{k} i^2} \tag{4.14}$$

where, $c[m]$ is the mth GFD coefficient. The Delta-Delta (acceleration) coefficients are computed by performing time derivative over the Delta coefficients. To abate the speaker variability in speech corpus, we have carried out speaker normalization using mean subtraction (MS) method. The LID system is developed by imposing the MS on GFD concatenated with dynamic coefficients.

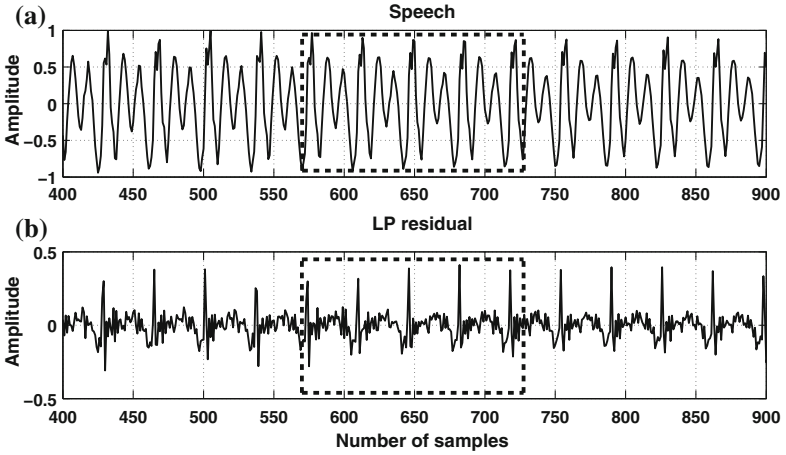

Fig. 4.3 **a** Speech signal and **b** corresponding LP residual signal marked by bounding boxes represent the segmental level frame (20 ms)

4.2.2 Parametric Representation of Segmental Level Excitation Source Information

Parameters extracted from the LP residual signal consisting of (2–3) consecutive glottal cycles may carry segmental level information. At segmental level, the LP residual signal is processed in block size of 20 ms with a shift of 10 ms. The LP residual signal is parameterized at segmental level to capture the language-specific energy and periodicity information present within (2–3) glottal cycles. A segment of voiced speech and corresponding LP residual signal are shown in Fig. 4.3a, b, respectively. The bounding boxes in the Fig. 4.3a, b represent the 20 ms segmental level frame. In this work, the LP residual signal is processed in spectral and cepstral domains to derive the segmental level energy and periodicity information, respectively. The LP residual spectrum mostly represents the energy and harmonic information related to the excitation source. Sections 4.2.2.1 and 4.2.2.2 describes the processing of LP residual signal in cepstral and spectral domains to capture the energy and periodicity information of excitation source, respectively.

4.2.2.1 Language-Specific Energy Information from Cepstral Analysis of LP Residual Spectrum

Cepstral analysis is performed on LP residual signal to capture the energy of excitation source. The filtering operation is performed on LP residual spectrum to obtain the sub-band spectrum. The cepstral analysis is performed on sub-band spectrum of LP residual signal to capture the energy information. The cepstral coefficients

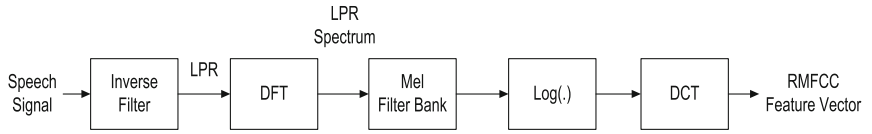

Fig. 4.4 Block diagram for extracting RMFCC features from speech signal

obtained from sub-band spectrum are termed as residual cepstral coefficients (RCC). In this work, the filter-bank which consists of 24 overlapping Mel-filters have been explored. Since, Mel-scale is relevant to human perception, therefore, Mel-filters are exploited in the present work to obtain the sub-band information from residual spectrum. The cepstral coefficients are computed from the Mel filtered spectrum by the following equation.

$$c(i) = \sum_{m=1}^{M} X(m) \cos\left[i \left(m - \frac{1}{2} \right) \frac{\pi}{M} \right] \tag{4.15}$$

where, $i = 1, 2, \ldots, C$ are the cepstral coefficients, (usually, $C < M$) M denotes total number of filters used ($M = 24$), and $X(m) = log_{10}\left(\sum_{k=0}^{N-1} [R_m(k)]^2 \right)$, represents the log energy output of the mth filter. The cepstral coefficients computed from Mel sub-band spectra are termed as Residual Mel Filter Cepstral Coefficients (RMFCC). In this work, only first 13 cepstral coefficients are used as feature vector and are modelled to capture the spectral energy information of excitation source. The sequence of steps involved in computation of RMFCC features is shown in Fig. 4.4.

There is significant overlap in the set of sound units for different languages in India. However, each language has unique phonotactic constraints based on sequence of occurrence of sound units. Therefore, the production characteristics of a particular sound unit may vary from one language to another due to language-specific co-articulation effects. The characteristics of vocal tract system and excitation source also contain unique information due to language-specific phonotactics. In this work, we have focused only on the characteristics of the excitation source. Hence, we have proposed a method to capture the energy of excitation source through RMFCCs extracted from LP residual signal. We hypothesize that, segmental level energy information may contain language-specific phonotactic knowledge.

4.2.2.2 Language-Specific Periodicity Information from Spectral Analysis of LP Residual Signal

Each sound unit corresponds to a unique articulatory configuration of both the vocal tract and excitation source. Hence, the characteristics of vocal tract and excitation source also unique for a particular sound unit. The nature of the residual spectrum varies from one sound unit to another due to the periodic nature of excitation source.

Therefore, the periodic information or the harmonic structure of the excitation source also varies from one sound unit to another. To capture the periodicity information of excitation source, spectral analysis has been performed.

The information about harmonic structure can be obtained by analyzing the power spectrum $p(k) = [R(k)]^2$ of the LP residual signal. This periodicity information can be represented by Power Differences of Spectrum in Sub-bands (PDSS) features [10]. The PDSS feature represents the periodicity nature of the excitation source [11]. PDSS can be represented as spectral flatness (SF) measure of the power spectrum in sub-bands. SF can be measured by the ratio of geometric mean (GM) to arithmetic mean (AM) of the power spectrum. PDSS of residual sub-band spectra is defined as follows [10]:

$$V(i) = 1.0 - \frac{\left[\prod_{k=L_i}^{H_i} p(k)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p(k)} \tag{4.16}$$

where, $N_i = h_i - l_i + 1$ is the sample number of frequency points in the ith filter. Where, l_i and h_i are the lower and upper limits of frequency in ith sub-band, respectively. Since, $0 \leq SF \leq 1$, the values of PDSS also varies from 0 to 1. High SF value indicates that the power of the spectrum is distributed throughout the spectrum and the power spectrum would appear relatively smooth. Low SF value indicates that the spectral power is concentrated only within smaller regions of power spectrum. The spectral flatness can also be measured within a specified sub-band, rather than across the whole spectrum. In present work, the spectral flatness has been measured in sub-bands. If the power spectrum has less dynamic range, for example nearly flat, then GM $\simeq$ AM and the PDSS value will be less than one. Alternatively, if PDSS is low, the spectrum is less periodic. If the spectrum has peaks and dips, for example the dynamic range is more, then GM is less than AM and PDSS value is close to one. In this case the spectrum is more periodic. So, PDSS measure gives information about the periodicity nature of a spectrum. Sub-band spectra are obtained by multiplying the residual power spectrum with a filterbank and PDSS values are computed from each sub-band using Eq. 4.16. In this work, the Mel filter banks are used for computing the PDSS from Mel sub-bands. The dominant information about the excitation source is manifested in the higher range of frequency. So, it is expected that PDSS computed from Mel sub-bands may provide better identification accuracy. Hence, in this work PDSS values have been computed using Mel filter-bank which is termed as Mel PDSS (MPDSS) features. The steps to extract the MPDSS feature are shown in Fig. 4.5.

The pitch of a particular sound unit is context dependent. Since, each language has unique phonotactics, even though sound units may be mostly common across the languages, the co-articulation effect (due to context) will be unique for each of the languages. Due to the pitch difference of two languages the periodicity of voiced sound units may also vary. Therefore, we can presume that, the periodicity information of excitation source captured by MPDSS feature may provide language discriminative information.

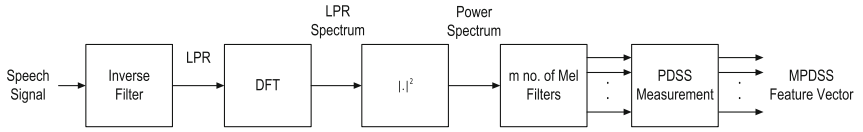

Fig. 4.5 Block diagram for extracting MPDSS features from speech signal

4.2.3 Parametric Representation of Supra-Segmental Level Excitation Source Information

In this section, techniques to parameterize the LP residual signal at supra-segmental level have been discussed. The tension of the vocal folds and the sub-glottal air pressure change continuously during speech production [12, 13]. As a consequence, the average rate of vocal folds vibration or pitch and the strength of excitation at glottal closing instants also vary with time. At supra-segmental level, these temporal variations are captured by processing larger segment of LP residual signal. The peaks of LP residual signal represents the glottal closing instants (GCIs) which are important events in the speech signal. At supra-segmental level, we have explored three different parameters: (i) pitch contour (PC), (ii) epoch strength contour (ESC) and (iii) epoch sharpness contour (ESNC) to capture the language-specific excitation source information. Therefore, in this work, the LP residual signal has been parameterized using 100 ms frame at supra-segmental level to capture the language-specific excitation source information.

Epochs or glottal closure instants (GCIs) are estimated by zero-frequency filtering method [14–16]. The zero-frequency filtering method locates the GCIs by passing the speech signal through a zero-frequency resonator twice. The zero-frequency resonator is a second order infinite impulse response (IIR) filter located at 0 Hz [15]. The purpose of passing the speech signal twice is to reduce the effects of high frequency resonances [15]. Passing the speech signal twice through a zero-frequency resonator is equivalent to four times successive integration. The trend in the filtered signal is removed by subtracting the local mean computed over an interval corresponding to the average pitch period. The resulting mean subtracted signal is called as zero-frequency filtered signal (ZFFS). The negative to positive zero crossings of ZFFS signal are the glottal closing instants (GCIs) [15]. Figure 4.6a–d show a segment of voiced speech, output of two cascaded zero frequency resonators, zero frequency filtered signal after mean subtraction and glottal closing instants. The steps involved to derive the ZFFS are given below.

(i) Difference the speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{4.17}$$

(ii) Pass the difference speech signal $x(n)$ twice through zero-frequency resonator

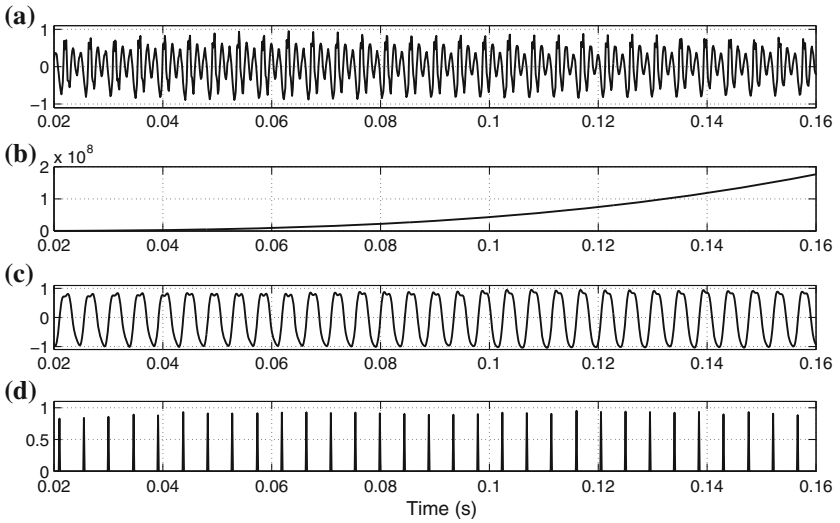

Fig. 4.6 **a** Segment of voiced speech, **b** output of cascade of two 0-Hz resonators, **c** zero frequency filtered signal after mean subtraction and **d** glottal closing instants (GCIs)

$$y_1(n) = -\sum_{k=1}^{2} a_k y_1(n-k) + x(n) \tag{4.18}$$

and,

$$y_2(n) = -\sum_{k=1}^{2} a_k y_2(n-k) + y_1(n) \tag{4.19}$$

where, $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$.

(iii) Compute the average pitch period using the autocorrelation over a 20 ms speech segment

(iv) Remove the trend in $y_2(n)$ by subtracting the mean computed over average pitch period. The resulting signal is ZFFS, which is shown as follows,

$$y(n) = y_2(n) - \frac{1}{2N_a + 1} \sum_{m=-N_a}^{N_a} y_2(n+m) \tag{4.20}$$

where, $2N_a + 1$ corresponds to the number of samples in the window used for mean subtraction. The time interval between two successive GCIs is the instantaneous pitch period T_0. The reciprocal of instantaneous period is instantaneous pitch ($P_0 = \frac{1}{T_0}$) [14].

The slope of the ZFFS around the negative to positive zero crossings corresponding to the locations of GCIs gives the measure of epoch strength (A_0) [16]. In [17], it has

been observed that, the strong peak in the Hilbert envelope of LP residual signal has a corresponding negative to positive zero crossing in the zero-frequency filtered signal and the slope of the ZFFS signal at negative to positive zero crossing is proportional to the amplitude of the corresponding peak in the Hilbert envelope. A window of 0.125 ms duration around each negative to positive zero crossing is assumed to be linear and considered to calculate the slope of positive zero crossings of ZFFS [16]. The strength of excitation of a vowel varies from one language to another language due to the unique phonotactic characteristics of each language. Hence, we presume that, epoch strength contour may contain language-specific information.

The distribution of strength of excitation can be viewed as, epoch sharpness. Epoch sharpness is computed from the Hilbert envelope of LP residual signal. A window of 3 ms is considered around each peak of HE of LP residual signal. The sharpness of each epoch is computed as follows [17]:

$$\eta = \frac{\sigma}{\mu} \tag{4.21}$$

where, the σ is the standard deviation and μ is the mean of HE samples obtained from 3 ms frame around each epoch, respectively. If the value of σ is more, then the sharpness of epoch is more. In this work, we have considered sequence of 10 consecutive epoch sharpness values as a feature vector, and by shifting each epoch a new feature vector can be generated. We hypothesize that, the variation of epoch sharpness across several glottal cycles may contain language-specific source information. The pitch, epoch strength and epoch sharpness contours are processed separately to capture the language-specific information. Finally, the evidences are combined to acquire the complete excitation source information present at supra-segmental level.

4.3 Development of LID Systems Using Parametric Features of Excitation Source

In this chapter, LID systems are developed at three phases by processing the proposed excitation source features. The block diagram of all LID systems developed in various phases is shown in Fig. 4.7.

Phase-I

At the phase-I, 5 different LID systems are developed:

- First LID system is developed by using RMFCC features at *seg* level which is denoted as RMFCC LID in Fig. 4.7.
- Second LID system is developed by using MPDSS features at *seg* level which is denoted as MPDSS LID in Fig. 4.7.
- Third LID system is developed by processing the pitch contour at *supra* level which is denoted as PC LID in Fig. 4.7.

- Fourth LID system is developed by processing the epoch strength contour at *supra* level which is denoted as ESC LID in Fig. 4.7.
- Fifth LID system is developed by processing the epoch sharpness contour at *supra* level which is denoted as ESNC LID in Fig. 4.7.

The above five LID systems are developed using GMMs [18]. The RMFCC and MPDSS features at *seg* level contains energy and periodicity information corresponding to the excitation source. So, to capture the complete language-specific excitation source information at *seg* level, we have combined the evidences from the LID systems (RMFCC LID and MPDSS LID) developed by partial features. Similarly, the evidences obtained from PC, ESC and ESNC LIDs are combined to acquire the complete excitation source information at *supra* level. The block diagram shown in Fig. 4.7 represents different LID systems developed by proposed excitation source features at three different levels and their combinations. In this study, we have performed three different combinations which are shown at phase-II and phase-III in Fig. 4.7.

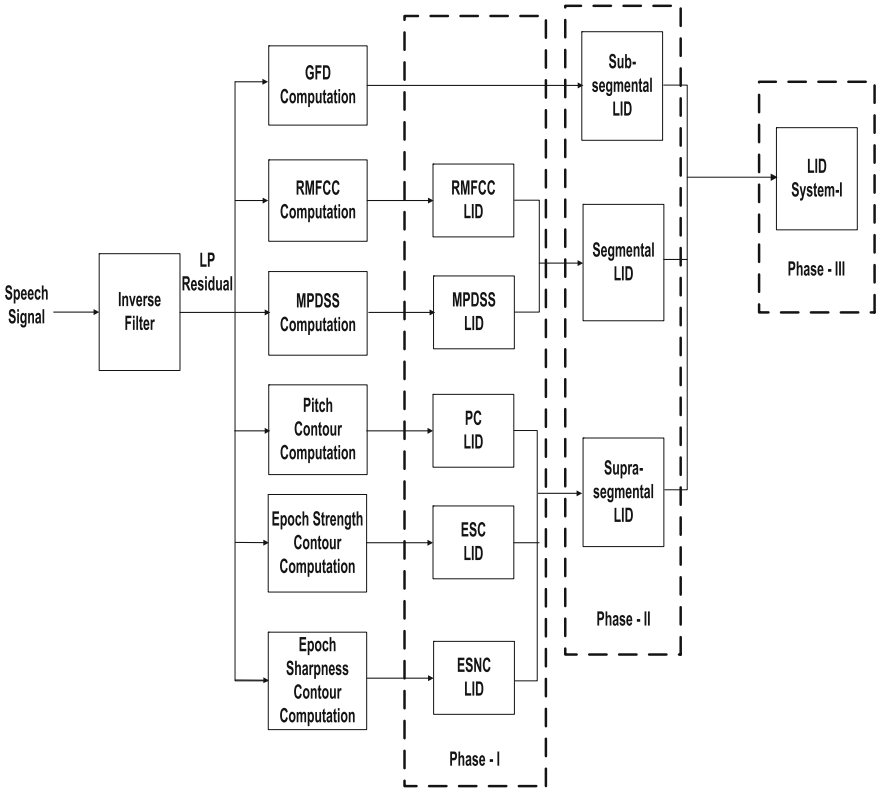

Fig. 4.7 Development of LID systems using parametric features of excitation source

Phase-II

In phase-II, we have developed the following three LID systems:

- The sub-segmental LID system is developed by processing the *sub* level GFD parameters.
- The segmental LID system is developed by combining the evidences obtained from RMFCC and MPDSS LID systems of phase-I to capture the complete excitation source information at *seg* level.
- The supra-segmental LID system is developed by combining the evidences obtained from PC, ESC and ESNC LID systems to capture the complete excitation source information at *supra* level.

Phase-III

In phase-III, we have performed the following combination:

- The language-specific cues obtained from *sub*, *seg* and *supra* level features provide distinct information about the excitation source. Therefore, to capture the complete language-specific excitation source information we have combined the evidences from sub-segmental, segmental and supra-segmental LID systems of phase-II. The LID system-I at phase-III is developed by processing the parametric features of excitation source.

4.4 Performance Evaluation of LID Systems Developed Using Parametric Features of Excitation Source

Evaluation of LID systems developed using parametric features of excitation source has been carried out at three phases. Experimental results are analyzed below.

Phase-I

At phase-I, 5 different LID systems are developed by parameterizing the LP residual signal at two different levels: *seg* and *supra*. The Table 4.1 portrays the identification performances of individual languages as-well-as the average performances of LID systems. The first column represents the 27 Indian languages used in this LID study. 2nd, 3rd, 4th, 5th and 6th columns represent the LID performances obtained by processing the *seg* level RMFCC and MPDSS features and *supra* level pitch, epoch strength and epoch sharpness contours, respectively. The average recognition accuracies obtained from the corresponding LID systems are 61.29, 51.29, 32.77, 18.33 and 30.55 %, respectively. The energy of the excitation source is represented by *seg* level RMFCC feature which provides better LID accuracy compared to other features. Comparison can be made between the individual language performances obtained from RMFCC and MPDSS features. Individual language accuracies portray that, the RMFCC and MPDSS features contribute distinct language-specific knowledge. Hence, at phase-II the evidences obtained

from these two features are combined to capture the complete *seg* level excitation source information. By comparing the 4th, 5th and 6th columns of Table 4.1, it can be observed that, the LID accuracies obtained from pitch, epoch strength and epoch sharpness contours contain distinct information. For example, LID accuracies obtained from ESC and ESNC features are 10 and 35 % for Arunachali language, respectively. Whereas, the PC feature could not identify the Arunachali language. For Bengali language, ESNC feature provides only 5 % accuracy. However, PC and ESC were not able to identify Bengali language. PC and ESC features identified the Gujarati and Kannada languages, whereas, the ESNC feature could not recognize these two languages. From this analysis, distinct nature of these three features is observed. Hence, evidences from PC, ESC and ESNC features are combined at phase-II to capture the complete *supra* level source information.

Phase-II

At phase-II, three different LID systems are developed. GFD parameters are computed by parameterizing the LP residual signal at *sub* level. The 7th column of Table 4.1 shows the LID performance obtained by processing the GFD parameters. The average performance obtained from GFD feature is 40.18 %. The 2nd LID system at phase-II has been developed by combining the evidences from RMFCC and MPDSS features to represent the *seg* level excitation source information. The LID accuracy of 62.22 % is obtained by processing the *seg* level excitation source information, which is shown in 8th column of Table 4.1. The 3rd LID system of phase-II has been developed by combining the evidences from PC, ESC and ESNC features at *supra* level. The LID performance obtained from the 3rd LID system of phase-II is 42.96 % shown in 9th column of Table 4.1. In this work, adaptive weighted combination scheme [19] has been used. In this scheme, the evidences from different modalities are combined at the score level. The combined score C is given by

$$C = \sum_{i=1}^{k} w_i c_i \tag{4.22}$$

where, w_i and c_i denotes weighting factor and confidence score of ith evidence and k denotes number of modalities are considered together. The Weighting factor w_i varies from 0 to 1 with a step size of 0.01 and sum up to 1 (i.e., $\sum_{i=1}^{k} w_i = 1$). The comparison has been carried out among the individual language performances obtained from *sub*, *seg* and *supra* levels shown in 7th, 8th and 9th columns of Table 4.1. The LID performances for Arunachali, Gujarati, Hindi, Manipuri and Marathi languages obtained by processing *seg* level excitation source feature are 50, 50, 10, 100 and 40 %, respectively. However, *sub* level GFD feature was unable to identify the corresponding languages. Similarly, the *supra* level excitation source feature could not identify the Manipuri and Marathi languages. However, the *supra* level information provides 35, 10 and 5 % accuracies for Arunachali, Gujarati and Hindi languages, respectively. For Malayalam language the excitation source information at *sub* and *supra* levels provide 45 and 90 % LID accuracies. However,

Table 4.1 Performances of language identification systems developed by using parametric features of excitation source

Languages	Average recognition performances (%)								
	Phase - I					Phase - II			Phase - III
	RMFCC	MPDSS	Pitch contour	Epoch strength contour	Epoch sharp-ness	GFD para-meters (*sub*)	*seg*	*supra*	*src*
Arunachali	50	50	0	10	35	0	50	35	50
Assamese	20	0	35	5	70	35	10	50	30
Bengali	65	55	0	0	5	25	65	15	65
Bhojpuri	45	40	80	25	50	40	45	50	50
Chhattisgarhi	100	100	45	40	30	85	100	85	100
Dogri	95	50	5	15	20	30	95	25	95
Gojri	100	100	60	40	35	60	100	65	100
Gujarati	35	0	10	5	0	0	50	10	55
Hindi	10	20	0	0	0	0	10	5	10
Indian English	100	100	45	15	90	100	100	80	100
Kannada	40	35	35	5	0	55	40	10	35
Kashmiri	70	70	65	5	15	70	70	60	90
Konkani	20	0	20	35	35	50	20	45	35
Malayalam	0	0	45	15	80	45	0	90	35
Manipuri	100	75	0	0	0	0	100	0	90
Marathi	40	20	0	5	0	0	40	0	35
Mizo	60	40	10	0	10	50	55	30	60
Nagamese	90	50	40	45	0	45	85	60	100
Nepali	95	45	10	5	40	15	95	30	90
Oriya	40	40	35	25	30	20	40	40	40
Punjabi	40	50	95	45	40	90	45	70	50
Rajasthani	100	100	45	70	40	55	100	45	100
Sanskrit	75	95	25	0	30	25	95	65	90
Sindhi	50	35	0	45	5	40	50	10	75
Tamil	50	25	20	10	40	45	45	35	50
Telugu	100	100	100	20	75	100	100	90	100
Urdu	65	90	60	10	50	5	75	60	80
Average performance	**61.29**	**51.29**	**32.77**	**18.33**	**30.55**	**40.18**	**62.22**	**42.96**	**67.03**

Table 4.2 Performances of LID systems developed using parametric excitation source features and evaluated by male speaker set

Languages	Average recognition performances (%)					
	GFD	RMFCC	MPDSS	PC	ESC	ESNC
Arunachali	0	0	0	0	0	0
Assamese	0	0	0	71.5	0	100
Bengali	0	0	0	0	0	0
Bhojpuri	28.6	85.8	71.5	100	14.3	100
Chhattisgarhi	57.2	100	100	28.6	0	28.6
Dogri	0	100	28.6	0	0	0
Gojri	57.2	100	100	57.2	28.6	42.9
Gujarati	0	14.3	0	14.3	14.3	0
Hindi	0	14.3	42.9	0	0	0
Indian English	100	100	100	42.9	0	85.8
Kannada	57.2	28.6	42.9	57.2	0	0
Kashmiri	42.9	71.5	71.5	42.9	0	28.6
Konkani	100	14.3	0	28.6	57.2	71.5
Malayalam	14.3	0	0	71.5	28.6	71.5
Manipuri	0	100	85.8	0	0	0
Marathi	0	71.5	28.6	0	14.3	0
Mizo	0	100	0	28.6	0	14.3
Nagamese	85.8	100	100	85.8	71.5	0
Nepali	0	100	85.8	14.3	0	71.5
Oriya	0	0	0	0	0	0
Punjabi	85.8	28.6	57.2	100	57.2	42.9
Rajasthani	71.5	100	100	14.3	85.8	100
Sanskrit	0	57.2	85.8	14.3	0	57.2
Sindhi	0	0	0	0	42.9	0
Tamil	57.2	14.3	0	0	14.3	57.2
Telugu	100	100	100	100	14.3	71.5
Urdu	0	42.9	85.8	57.2	14.3	28.6
Average performance	**31.74**	**53.43**	**47.61**	**34.39**	**16.93**	**35.97**

the *seg* level excitation source feature could not able to identify the Malayalam language. This empirical analysis delineates the discordant nature of excitation source information present at *sub*, *seg* and *supra* levels from language discrimination aspects.

Phase-III

The parametric features of excitation source contribute distinct language-specific information at three different levels, which has been discussed at phase-II. Hence,

Table 4.3 Performances of LID systems developed using parametric excitation source features and evaluated by female speaker set

Languages	Average recognition performances (%)					
	GFD	RMFCC	MPDSS	PC	ESC	ESNC
Arunachali	0	100	100	0	0	40
Assamese	80	20	0	0	0	40
Bengali	40	100	100	0	0	20
Bhojpuri	20	0	0	80	40	0
Chhattisgarhi	100	100	100	40	40	40
Dogri	100	100	60	0	40	20
Gujarati	60	60	0	0	0	0
Hindi	0	0	0	0	20	0
Indian English	100	100	100	40	20	80
Konkani	0	40	0	40	0	0
Malayalam	60	0	0	20	0	80
Manipuri	0	100	80	0	0	0
Marathi	0	0	0	0	0	0
Mizo	100	40	0	0	0	0
Nagamese	0	100	0	0	40	0
Nepali	60	100	0	0	0	0
Oriya	40	100	100	100	80	80
Rajasthani	20	100	100	60	40	0
Sindhi	80	100	60	0	40	20
Tamil	40	100	100	80	0	0
Average performance	**45**	**68**	**45**	**23**	**18**	**21**

the evidences from these three levels are combined to capture the complete language-specific excitation source information. The LID system-I at phase-III has been developed by using all parametric excitation source features. The LID accuracy obtained by processing complete excitation source information is 67.03 % which is denoted as *src* shown at 10th column of Table 4.1.

Parametric excitation source features also explored for investigating the gender effects in language identification study. The male and female speakers present in test set have been separated to evaluate the LID systems separately. Performances obtained from male and female sets are shown in Tables 4.2 and 4.3, respectively. From this study it is observed that, segmental level parametric features provide best average LID accuracies followed by supra-segmental and sub-segmental levels, respectively. This trend of LID accuracies also observed in LID study without separating the male and female speakers present in test set (see Table 4.1).

Table 4.4 Language identification performances obtained from OGI-MLTS database using parametric features of excitation source

Languages	Average recognition performances (%)								
	Phase - I					Phase - II			Phase - III
	RMFCC	MPDSS	Pitch contour	Epoch strength contour	Epoch sharpness contour	GFD (*sub*)	*seg*	*supra*	*src*
English	66.67	66.67	33.33	88.89	33.33	77.78	77.78	55.55	77.77
Farsi	77.78	100	55.56	100	66.66	66.67	77.78	100	77.77
French	88.89	100	22.22	11.11	22.22	100	88.89	11.11	88.88
German	66.67	77.78	55.56	11.11	0	55.56	77.78	11.11	66.66
Hindi	77.78	100	100	44.44	0	22.22	88.89	77.77	88.88
Japanese	33.33	88.89	44.44	55.56	88.88	22.22	33.33	66.66	55.55
Korean	88.89	22.22	66.67	11.11	55.55	11.11	88.89	44.44	88.88
Mandarin	66.67	33.33	33.33	100	22.22	55.56	66.67	66.66	100
Spanish	44.44	44.44	77.78	55.56	77.77	77.78	33.33	77.77	44.44
Tamil	100	66.67	100	77.78	77.77	33.33	88.89	100	100
Vietnamese	77.78	22.22	11.11	11.11	88.88	55.56	77.78	55.55	77.77
Average performance	**71.71**	**65.65**	**54.54**	**51.51**	**48.48**	**52.52**	**72.72**	**60.60**	**78.78**

4.5 Performance Evaluation of LID Systems Developed Using Parametric Features of Excitation Source on OGI-MLTS Database

The effectiveness of the proposed excitation source features for language identification task has been analyzed on OGI-MLTS database [20]. 11 languages are used for LID task. From each language around 1 h of data is used for building the language models. 3 speakers from each language who have not participated during training phase are considered during evaluation. From each speaker 3 test utterances each of duration 10 s are considered for evaluation. The LID systems have been developed using the excitation source information at three different levels. The LID accuracies obtained from the empirical analysis portray the similar characteristics of excitation source features as it has been observed for Indian languages. The experimental results obtained from OGI-MLTS database are shown in Table 4.4.

4.6 Summary

In this chapter, parametric features of excitation source have been proposed for language discrimination task. The LP residual signal has been parameterized at sub-segmental, segmental and supra-segmental levels. At sub-segmental level, the GFD parameters are estimated from LP residual signal to capture the minute variations present within a glottal cycle. At segmental level, RMFCC and MPDSS features are proposed to capture the energy and periodicity information related to the excitation source. At supra-segmental level, the pitch, epoch strength and epoch sharpness contours are proposed for capturing the variations of the pitch and energy across 50 glottal cycles. Experimental analysis delineates that, the language-specific excitation source information present at three levels are fundamentally distinct. Therefore, the evidences obtained from different features proposed at three levels are combined to capture the complete parametric excitation source information. The combined RMFCC and MPDSS feature at segmental level provides better LID accuracy compared to other two levels. This indicates the potency of segmental level excitation source information for language discrimination task.

References

1. L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, 1978)
2. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
3. D.G. Childers, A.K. Krishnamurthy, A critical review of electroglottography. Crit. Rev. Biomed. Eng. **12**(2), 131–161 (1985)

4. M.D. Plumpe, T.F. Quatieri, D.A. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Audio Speech Lang. Process. **7**(5), 569–586 (1999)
5. R. Veldhuish, A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. J. Acoust. Soc. Am. **103**(1), 566–571 (1998)
6. T.V. Ananthapadmanabha, G. Fant, Calculation of true glottal flow and its components. Speech Commun. **1**, 167–184 (1982)
7. Y. Qi, N. Bi, A simplified approximation of the four-parameter LF model of voice source. J. Acoust. Soc. Am. **96**(2), 1182–1185 (1994)
8. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
9. P. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio Speech Lang. Process. **15**(1), 34–43 (2007)
10. S. Hayakawa, K. Takeda, F. Itakura, Speaker identification using harmonic structure of LP-residual spectrum, *Biometric Personal Authentification*, vol. 1206, Lecture notes (Springer, Berlin, 1997)
11. A.H. Gray, J.D. Markel, A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Trans. Audio Speech Lang. Process. ASSP-**22**(3), 207–217 (1974)
12. J.J. Wolf, Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Am. **51**(2), 2044–2055 (1972)
13. B.S. Atal, Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Am. **52**(6), 1687–1697 (1972)
14. B. Yegnenarayana, K.S.R. Murthy, Event based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Process. **17**(4), 614–624 (2009)
15. K.S.R. Murthy, B. Yegnanarayana, Epoch extraction from speech signal. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1602–1613 (2008)
16. K.S.R. Murthy, B. Yegnanarayana, Characterization of glottal activity from speech signal. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
17. G. Seshadria, B. Yegnanarayana, Perceived loudness of speech based on the characteristics of glottal excitation source. J. Acoust. Soc. Am. **126**(4), 2061–2071 (2009)
18. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Audio Speech Lang. Process. **3**(1), 72–83 (1995)
19. V.R. Reddy, S. Maity, K.S. Rao, Identification of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. (Springer) **16**(4), 489–511 (2013)
20. Y.K. Muthusamy, R.A. Cole, B.T. Oshika, The OGI multilanguage telephone speech corpus, in *Spoken Language Processing*, pp. 895–898 (1992)

Chapter 5
Complementary and Robust Nature of Excitation Source Features for Language Identification

Abstract This chapter discusses about the combination of implicit and parametric features of excitation source to enhance the LID accuracy. Further, complementary nature of excitation source and vocal tract features is exploited for improving the LID accuracy. The robustness of proposed language-specific excitation source features is investigated on various noisy background environments.

Keywords Language identification using vocal tract features · Language identification using excitation source and vocal tract system features · Robust language identification · Language identification in noisy environments · Robust excitation source features

5.1 Introduction

In Chap. 3, the raw LP residual samples, its magnitude and phase components have been processed to capture the *implicit* language-specific excitation source information. In Chap. 4, LP residual signal has been parameterized at *sub*, *seg* and *supra* levels to capture different aspects of excitation source information for language discrimination task. In this chapter, the evidences obtained from implicit and parametric features are combined to capture the complete excitation source information. Further, the LID systems are also developed by processing the vocal tract information represented by Mel-frequency cepstral coefficients (MFCCs). As we know that, vocal tract and source for exciting the vocal tract are two different components of speech production system. Hence, in this chapter, the evidences obtained from complete excitation source features have been combined with the evidences of MFCC features to investigate the existence of complementary nature of these two features. Eventually, the robustness of proposed excitation source features is also examined in this chapter.

The rest of this chapter is organized as follows: Sect. 5.2 briefly describes the vocal tract features. Section 5.3, discusses about the development of LID systems by combining the evidences of complete excitation source and vocal tract features. Sections 5.4 and 5.5 analyze the accuracies of the integrated LID systems. In Sect. 5.6,

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_5

the robustness of excitation source features is illustrated. Section 5.7 summarizes the contents of this chapter.

5.2 Vocal Tract Features

During speech production, vocal tract system behaves like a time varying resonator or, a filter and it characterizes the variations in the vocal tract shape in the form of resonances and antiresonances that occur in the speech spectrum. The shape of the vocal tract resonator system is captured by spectral analysis of speech signal. Several parameterization techniques like, linear prediction cepstral coefficients (LPCCs) and mel-frequency cepstral coefficients (MFCCs) are available for modeling vocal tract information [1]. Since mel-filters are based on human auditory characteristics, state-of-the-art LID systems mostly use MFCC features. Detailed description of MFCC feature has been given in Appendix B. The processing steps of MFCC are given as follows:

- First pre-emphasis is carried out on given speech signal. This step refers filtering which emphasizes the higher frequency components of speech signal. The purpose of pre-emphasis is to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region.
- Speech is a slow varying quasi-stationary signal. Therefore, speech analysis must be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over the range of 10–30 ms frame size [2, 3]. The blocked frames are Hamming windowed. This helps to reduce the edge effect while taking the discrete Fourier transform (DFT) on the signal.
- Fourier transform is performed on each windowed frame to obtain the magnitude spectrum.
- Mel-spectrum is obtained by passing the magnitude spectrum through the Mel-filter bank. A *mel* is a unit of perceived speech frequency or a unit of tone. The *mel* scale is therefore a mapping between the physical frequency scale (Hz) and the perceived frequency scale (Mels). The mel spectrum values or mel frequency coefficients of the magnitude spectrum $X(k)$ are computed by multiplying the magnitude spectrum by each of the triangular mel weighting filters.

$$S(m) = \sum_{k=0}^{N-1} X(k)^2 H_m(k), \qquad 0 \leq m \leq M-1 \tag{5.1}$$

  where, M is total number of triangular mel weighting filters.
- The log operation is performed on Mel-frequency coefficients. Discrete Cosine Transform (DCT) is then applied on log magnitude spectrum to obtain the cep-

stral coefficients. These cepstral coefficients are known as Mel-frequency cepstral coefficients (MFCCs).

The MFCC feature vector represents only the information present at power spectral envelope of a single frame. However, speech signal also carries information in spectral variations with respect to time i.e., the knowledge present in the trajectories of the MFCC coefficients over time. The temporal variation present in a sequence of MFCC feature vectors is derived by computing the time derivative of MFCC feature vector and is concatenated to the corresponding MFCC feature vector. The first and second derivatives of the feature are usually called velocity or Delta coefficients and acceleration or Delta-Delta coefficients, respectively. Velocity and acceleration coefficients are concatenated with the static coefficients to form a feature vector.

The formula used for calculating the Delta coefficients is given below:

$$\Delta c[m] = \frac{\sum_{i=1}^{k} i\,(c[m+i] - c[m-i])}{2\sum_{i=1}^{k} i^2} \tag{5.2}$$

where, $c[m]$ is the mth MFCC coefficient and $2k + 1$ is the window size to compute the Delta coefficients. The Delta-Delta (acceleration) coefficients are computed by performing time derivative over the Delta coefficients.

5.3 Development of Language Identification Systems Using Excitation Source and Vocal Tract Features

In this work, LID systems are developed at four phases which are described below. The block diagram of all LID systems developed in various phases is shown in Fig. 5.1.

Phase-I

At the phase-I, 9 different LID systems are developed:

- First three LID systems are developed by using the *sub* level features. The *sub 1*, *sub 2* and *sub 3* LID systems are developed by processing the raw LP residual samples, HE + RP feature and GFD parameters at sub-segmental level, respectively.
- Next three LID systems are developed by using *seg* level features. The *seg 1*, *seg 2* and *seg 3* LID systems are developed by processing the raw LP residual samples, HE + RP and combined RMFCC and MPDSS features at segmental level, respectively.
- Bottom three LID systems are developed by using *supra* level features. The *supra 1*, *supra 2* and *supra 3* LID systems are developed by processing the raw

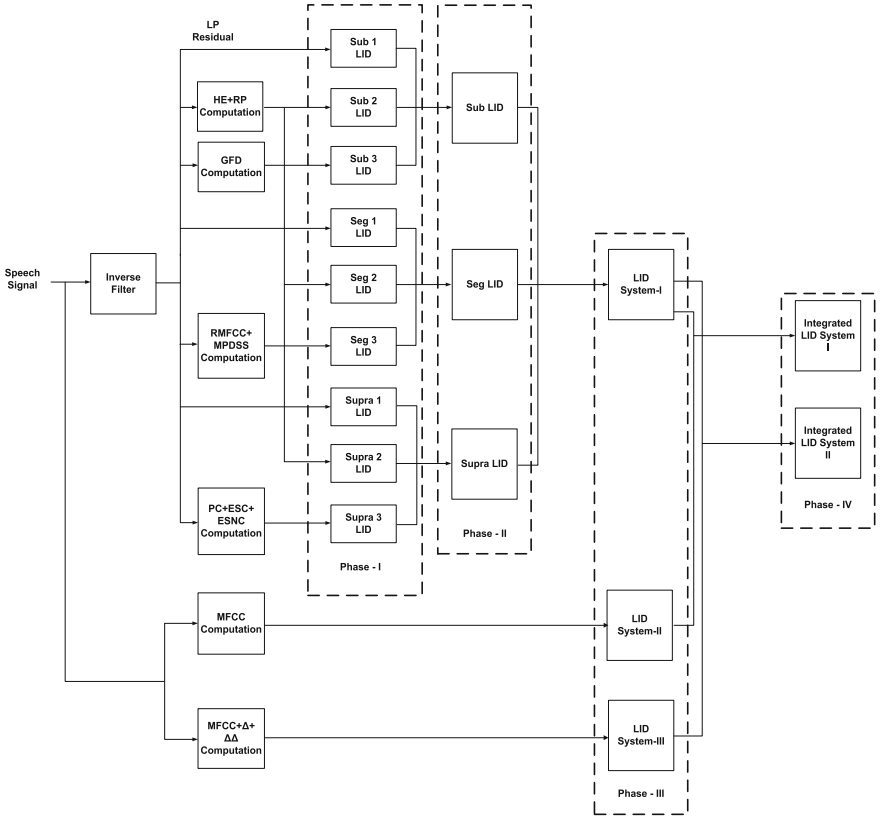

Fig. 5.1 Development of LID systems using excitation source and vocal tract features

LP residual samples, HE + RP feature and combined pitch, epoch strength and epoch sharpness contours at supra-segmental level, respectively.

At each level, language-specific information has been captured using both implicit and parametric features. Hence, evidences are combined to acquire the complete language-specific information at each level. The LID systems developed at phase-II illustrate these combinations.

Phase-II

At phase-I, LID systems are developed by using both implicit and explicit features of excitation source. The language-specific information captured by these features may be different. Therefore, the combination of implicit and explicit features may enhance the performance further. At phase-II, we have performed the following three combinations:

- The first LID system has been built by combining the evidences from all the systems developed using *sub* level features.

- The second LID system has been built by combining the evidences from all the systems developed using *seg* level features.
- The third LID system has been built by combining the evidences from all the systems developed using *supra* level features.

Phase-III

The features at *sub*, *seg* and *supra* level contain partial information about the excitation source. Therefore, to achieve complete excitation source information for discriminating the languages, we have combined the evidences from the LID systems developed by partial features. In phase-III, we have developed LID systems by combining evidences from *sub*, *seg* and *supra* LID systems of phase-II. The vocal tract information and excitation source are two different components of human speech production system. So, these two components contribute different aspects of language-specific information. Therefore, we have developed the LID systems using vocal tract information represented by MFCCs (LID system-II) which is shown at phase-III. We have also developed LID system using the velocity and acceleration coefficients concatenated with MFCCs which has been shown as LID System-III at phase-III in Fig. 5.1.

Phase-IV

We have combined the evidences from the LID systems developed by vocal tract information represented by MFCC feature with excitation source features to investigate the existence of complementary information present in these two features. In phase-IV of Fig. 5.1, the integrated LID systems using vocal tract and excitation source features are illustrated. In phase-IV, we have developed two different integrated LID systems:

- The LID system-I at phase-III is developed by processing the overall excitation source information. The evidences obtained from this system is combined with evidences from the LID system developed by using MFCC feature (LID System-II) at phase-III to develop the integrated LID system-I at phase-IV.
- The LID system-III at phase-III is developed by concatenating the dynamic coefficients with MFCC. The integrated LID system-II at phase-IV is developed by combining the evidences from LID system-I and LID system-III of phase-III,

5.4 Performance Evaluation of Source and System Integrated LID Systems

In this chapter, we have carried out the combination of source and vocal tract system features at different phases. LID accuracies obtained from integrated LID systems are tabulated in Tables 5.1 and 5.2.

Table 5.1 LID performances using implicit and parametric excitation source features

Languages	Average recognition performances (%)											
	Phase - I									Phase - II		
	sub 1	*sub 2*	*sub 3*	*seg 1*	*seg 2*	*seg 3*	*supra 1*	*supra 2*	*supra 3*	*sub*	*seg*	*supra*
Arunachali	0	70	0	35	45	50	70	50	35	35	50	35
Assamese	0	5	35	35	25	10	0	10	50	35	0	50
Bengali	0	10	25	5	10	65	0	0	15	25	50	10
Bhojpuri	50	35	40	35	25	45	10	55	50	55	50	40
Chh	0	60	85	100	100	100	100	100	85	95	100	100
Dogri	35	65	30	50	50	95	15	5	25	50	95	55
Gojri	95	50	60	80	100	100	95	100	65	80	100	100
Gujarati	0	55	0	15	0	50	45	90	10	70	50	65
Hindi	5	35	0	15	15	10	0	0	5	35	30	5
Indian English	100	95	100	95	100	100	90	100	80	100	100	100
Kannada	10	30	55	45	20	40	35	30	10	25	50	20
Kashmiri	100	70	70	95	100	70	25	35	60	85	95	95
Konkani	0	0	50	85	95	20	100	100	45	0	85	95
Malayalam	50	90	45	75	55	0	0	5	90	100	70	90
Manipuri	0	0	0	0	0	100	15	70	0	0	100	45
Marathi	0	0	0	0	0	40	0	5	0	0	0	5
Mizo	0	70	50	0	10	55	0	5	30	50	50	20
Nagamese	70	0	45	70	50	85	20	50	60	90	100	75
Nepali	60	45	15	0	25	95	10	25	30	30	85	30
Oriya	0	0	20	40	40	40	5	5	40	0	40	35
Punjabi	0	90	90	100	100	45	45	100	70	100	100	95
Raj	0	100	55	35	80	100	45	100	45	100	100	100
Sanskrit	0	100	25	15	100	95	0	10	65	95	100	70
Sindhi	70	50	40	45	10	50	0	0	10	50	90	35
Tamil	0	80	45	35	45	45	20	20	35	65	60	45
Telugu	20	55	100	100	100	100	95	100	90	80	100	100
Urdu	0	40	5	0	75	75	15	15	60	25	85	50
Average performance	**24.62**	**47.77**	**40.18**	**44.62**	**50.92**	**62.22**	**31.66**	**43.88**	**42.96**	**54.62**	**71.66**	**57.96**

Table 5.2 LID performances using implicit + parametric (*src*) and vocal tract features

Languages	Average recognition performances (%)				
	Phase - III			Phase - IV	
	src	MFCC	MFCC + $\Delta + \Delta\Delta$	*src* + MFCC	*src* + MFCC + $\Delta + \Delta\Delta$
Arunachali	50	100	100	50	65
Assamese	0	5	0	0	0
Bengali	50	55	65	50	65
Bhojpuri	50	35	50	50	50
Chhattisgarhi	100	100	100	100	100
Dogri	100	65	100	100	100
Gojri	100	100	100	100	100
Gujarati	50	50	50	50	50
Hindi	30	10	20	30	30
Indian English	100	100	100	100	100
Kannada	50	50	0	50	30
Kashmiri	95	75	70	95	95
Konkani	90	60	50	90	85
Malayalam	70	30	35	70	50
Manipuri	100	100	100	100	100
Marathi	0	45	60	0	60
Mizo	50	15	50	50	75
Nagamese	100	100	100	100	100
Nepali	80	90	100	85	100
Oriya	40	40	40	40	40
Punjabi	100	65	50	100	50
Rajasthani	100	100	100	100	100
Sanskrit	100	5	65	100	100
Sindhi	90	95	85	90	95
Tamil	65	25	45	65	75
Telugu	100	100	100	100	100
Urdu	85	70	60	85	90
Average performance	**72.03**	**62.40**	**66.48**	**72.22**	**74.25**

Phase-I

At phase-I, 9 different LID systems are developed. Table 5.1 portrays the identification performances of individual languages as-well-as the average performances of phase-I and phase-II LID systems. The first column represents the 27 Indian languages used in this LID study. LID performances obtained by processing the excitation source features at each level are shown from 2nd to 10th columns of Table 5.1. The average LID accuracies of 9 different LID systems at phase-I of Fig. 5.1 are 24.62, 47.77, 40.18, 44.62, 50.92, 62.22, 31.66, 43.88 and 42.96 %. Individual language performances showed from 2nd to 4th columns are compared to investigate the existence of complementary language-specific information captured by different *sub* level features. For example, *sub 2* feature provides LID performances of 70 and 55 % for Arunchali and Gujarati languages, respectively. However, these two languages are not recognized by *sub 1* and *sub 3* features. Similarly, *sub 2* and *sub 3* features identified Assamese, Bengali, Chhattisgarhi, Mizo, Punjabi, Rajasthani, Sanskrit, Tamil and Urdu languages. However, these languages are not recognized by *sub 1* feature. This shows that, complementary information is present in *sub 1*, *sub 2* and *sub 3* features.
The individual LID performances obtained from *seg 1*, *seg 2* and *seg 3* features are also analyzed. For example, *seg 2* feature provides LID accuracies of 10, 25 and 75 % for Mizo, Nepali and Urdu languages, respectively. LID performances of 55, 95 and 75 % are achieved by processing only the *seg 3* feature for the corresponding languages. However, the *seg 1* feature is unable to recognize the above mentioned languages. LID accuracies of 100 and 40 % are obtained for Manipuri and Marathi languages by processing only *seg 3* feature. These two languages are not recognized by *seg 1* and *seg 2* features. Hence, it can be elicited that, *seg 1*, *seg 2* and *seg 3* features contain distinct language-specific cues.
The individual language performances have shown from 8th to 10th columns are compared to observe the complementary nature of *supra 1*, *supra 2* and *supra 3* features. For Assamese language the *supra 2* and *supra 3* features provide recognition accuracies of 10 and 50 %, respectively. However, the Assamese language has not identified by *supra 1* feature. *supra 3* feature gives 15, 5 and 10 % recognition accuracies for Bengali, Hindi and Sindhi languages, respectively. However, the *supra 1*, *supra 2* features were not able to identify these three language. For Malayalam language, LID accuracies of 5 and 90 % are obtained by processing *supra 2* and *supra 3* features, respectively. However, *supra 1* feature could not identify Malayalam language. Similarly, distinct nature of *supra* level features are observed for other languages also. This experimental analysis shows the complementary nature of all the nine different features used to develop LID systems at phase-I.

Phase-II

Different approaches are proposed to capture language-specific information at each level. The recognition accuracies provided by different features are analyzed

at phase-I. This analysis portrays the complementary nature of the features proposed at each level. Hence, to capture the complete information at each level, we have combined the evidences obtained from different features of each level. In this work, adaptive weighted combination scheme [4] has been used to combine the evidences. The evidences obtained from *sub 1*, *sub 2* and *sub 3* features are combined to capture the complete *sub* level information. The average LID performances of 54.62 % is obtained from complete *sub* level information shown at 11th column of Table 5.1. Similarly, the average recognition accuracies of 71.66 and 57.96 % are achieved by complete *seg* and *supra* level information, respectively. The individual language performances have shown from 11th to 13th columns are also compared to investigate the complementary language information present at each level. For example, LID accuracies of Assamese language using complete *sub* and *supra* level features are 35 and 50 %, respectively. However, the *seg* level feature was unable to identify Assamese language. Similarly, recognition accuracy of 5 % has been achieved for Marathi language by exploring *supra* level features. The *sub* and *seg* level features were unable to identify the Marathi language. For Konkani language, recognition accuracies of 85 and 95 % are achieved by processing *seg* and *supra* level features. However, the *sub* level information was unable to identify Konkani language. This comparative study illustrates the complementary language-specific excitation source information present at *sub*, *seg* and *supra* levels.

Phase-III

The excitation source information contributes distinct language-specific information at three different levels. Hence, the evidences from these three levels are combined to capture the complete language-specific excitation source information. The 1st LID system of phase-III has been developed by processing the complete excitation source information. The LID accuracy obtained by processing complete excitation source information is 72.03 % which is denoted as *src* shown at 2nd column of Table 5.2. The LID system-II at phase-III has been developed by spectral features represented by MFCCs. The identification accuracy obtained from this system is 62.40 % which is shown at 3rd column of Table 5.2. The LID performance obtained by processing the MFCCs concatenated with the velocity and acceleration coefficients is 66.48 % which has been shown at 4th column of Table 5.2. The individual language performances achieved from the proposed excitation source features are compared with the state-of-the-art vocal tract features represented by MFCCs, which shows the disparateness between these two features from language identification point of view. The distinct nature can also be observed by comparing the individual language performances of 2nd column with 4th column shown in Table 5.2.

Phase-IV

By analyzing the recognition accuracies achieved from phase-III LID systems it has been observed that, the vocal tract and excitation source features contribute complementary information for language discrimination task. Hence, to improve

the LID accuracies, we have combined the evidences from these two complementary features at phase-IV. The LID accuracies obtained from the integrated LID system-I and II of phase-IV have been shown in 5th and 6th columns of Table 5.2. The average LID performances achieved from the two integrated LID systems are 72.22 and 74.25 %, respectively, which have been improved compared to individual features.

5.5 Performance Evaluation of Source and System Integrated LID Systems on OGI-MLTS Database

The efficacy of overall excitation source features for language identification task has been analyzed on OGI-MLTS database [5]. The *implicit* and parametric excitation source features at *sub*, *seg* and *supra* levels are combined. Furthermore, the evidences obtained from three levels are also combined to capture overall excitation source features, which is represented by *src*. Language identification systems are also developed using vocal tract features represented by MFCC. Performances obtained from vocal tract and excitation source features are combined to investigate the complementary nature of these two features. The experimental results obtained from OGI-MLTS database are shown in Table 5.3. The LID accuracies obtained from

Table 5.3 Language identification performances obtained from overall excitation source, vocal tract and their combination information evaluated on OGI-MLTS database

Languages	Average recognition performances (%)							
	sub overall	*seg* overall	*supra* overall	*src*	MFCC	MFCC + $\Delta + \Delta\Delta$	*src* + MFCC	*src* + MFCC + $\Delta + \Delta\Delta$
English	55.56	83.33	88.89	100.00	100.00	100.00	100.00	100.00
Farsi	77.78	83.33	66.67	83.33	55.56	94.44	66.67	100.00
French	100.00	83.33	22.22	100.00	100.00	100.00	100.00	100.00
German	88.89	66.67	55.56	66.67	88.89	88.89	100.00	100.00
Hindi	22.22	100.00	66.67	100.00	44.44	77.78	100.00	88.89
Japanese	55.56	33.33	88.89	33.33	55.56	77.78	33.33	88.89
Korean	22.22	100.00	11.11	100.00	100.00	100.00	100.00	100.00
Mandarin	44.44	66.67	55.56	83.33	100.00	100.00	100.00	100.00
Spanish	100.00	50.00	55.56	50.00	88.89	77.78	100.00	88.89
Tamil	11.11	100.00	88.89	100.00	100.00	100.00	100.00	100.00
Vietnamese	55.56	83.33	100.00	83.33	77.78	100.00	66.67	100.00
Average performances	**57.57**	**77.27**	**63.63**	**81.81**	**82.82**	**92.42**	**87.87**	**96.96**

empirical analysis portray similar trend as it has been observed for Indian languages (see Tables 5.1 and 5.2).

5.6 Robustness of Excitation Source Features

In this section, the robustness of excitation source information has been analyzed for language identification task. In this study, the LP residual signal is processed at segmental level to capture the language-specific excitation source information. The spectral features are highly affected by background noise and length of the test utterances during most of the identification tasks. In this work, the raw LP residual samples are used as feature vectors for building the LID systems. The main objective of this study is to demonstrate the robustness of excitation source information extracted from LP residual signal for language identification in view of (i) background noise, (ii) varying amount of training data and (iii) varying length of test samples. In this work, Gaussian mixture models (GMM) are used for building the language models. Finally, the robustness of proposed excitation source features is compared with the well known spectral features using LID performances on IITKGP-MLILSC language database.

5.6.1 Motivation for the Use of Excitation Source Information for Robust Language Identification

Speech production system consists of vocal tract system and a source for exciting the vocal tract resonator. Most of the existing works have been exploited the dynamic behaviour of vocal tract for language identification task. However, the spectral features are susceptible to noisy environment. Hence, there is a need to derive a robust language-specific feature from the speech signal. In this work, we want to investigate whether the excitation source contain any robust language-specific information or not. As we know that, the excitation source information can be captured by processing the LP residual signal [6]. The samples of the LP residual signal are uncorrelated and hence the LP residual samples appear like noise samples. Figure 5.2a portrays LP residual signal corresponding to a vowel segment. The bounding box represents the 20 ms segmental (*seg*) level frame. The *seg* level LP residual frame is decimated by factor 4 and then raw LP residual samples are processed. The reason of performing decimation is to suppress the fine variations of LP residual signal present within a glottal cycle, which can be captured by sub-segmental level processing of LP residual signal. The *seg* level LP residual frame has been depicted in Fig. 5.2b. The log magnitude spectrum corresponding to the *seg* level LP residual frame is shown in Fig. 5.2c. The instantaneous pitch and energy can be observed from *seg* level LP residual frame and its magnitude spectrum shown in Fig. 5.2b, c, respectively. The white noise of

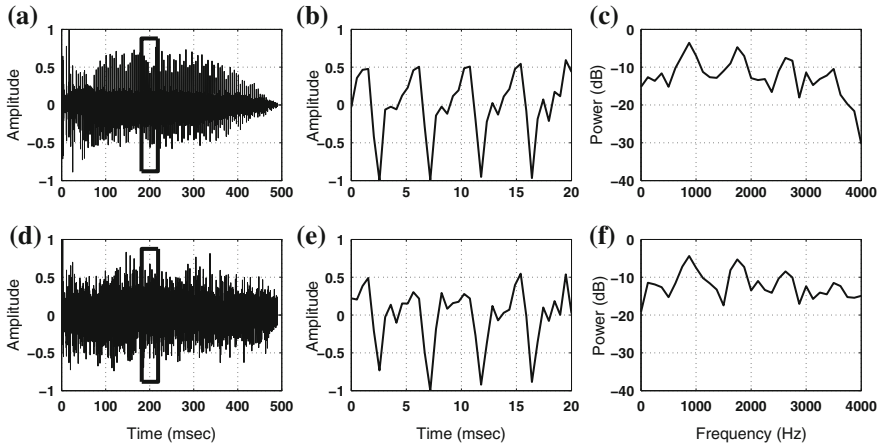

Fig. 5.2 **a** LP residual signal obtained from clean vowel segment and the bounding box represents 20 ms *seg* level frame, **b** LP residual signal obtained from *seg* level frame after decimation by factor 4, **c** log magnitude spectrum corresponding to the *seg* level frame shown in (**b**), **d** LP residual signal after adding 10 dB white noise to vowel segment, **e** noisy *seg* level LP residual frame after decimation by factor 4 and **f** corresponding log magnitude spectrum

10 dB SNR is added with the vowel segment and the LP residual signal is obtained from it. Figure 5.2d shows the LP residual signal corresponding to the noisy vowel segment. In Fig. 5.2e, the 20 ms *seg* level LP residual frame corresponding to the noisy LP residual signal has been portrayed. The Fig. 5.2f delineates the log magnitude spectrum obtained from the *seg* level LP residual frame shown in Fig. 5.2e. The important thing that can be observed is the *seg* level noisy LP residual frame in temporal domain and corresponding magnitude spectrum does not degrade much by 10 dB additive white noise. The instantaneous pitch and energy of vowel segment still preserved after adding 10 dB white noise also. To speculate the robustness of excitation source features, LP residual samples and corresponding log magnitude spectra after adding white noise with different SNR levels using same LP order is portrayed in Fig. 5.3. The notable thing is that, for all SNR levels the characteristics of LP residual samples obtained from *seg* level frame and corresponding magnitude spectra are not degraded greatly. This illustrates the robustness of excitation source information.

The Fig. 5.4 portrays the distribution of 1st coefficients of MFCC feature vectors obtained from the Arunachali language in clean and noisy environments. The solid line indicates the distribution of 1st coefficients of MFCC feature vectors obtained from clean speech. The dashed line represents the distribution of 1st coefficients of MFCC feature vectors obtained from noisy (10 dB additive white noise) speech. As MFCC coefficients are affected by 10 dB additive white noise, the distribution corresponding to noisy data deviates from the original one. It can be conjectured that, the features obtained from LP residual may be robust to noisy environment compared to spectral features represented by MFCCs.

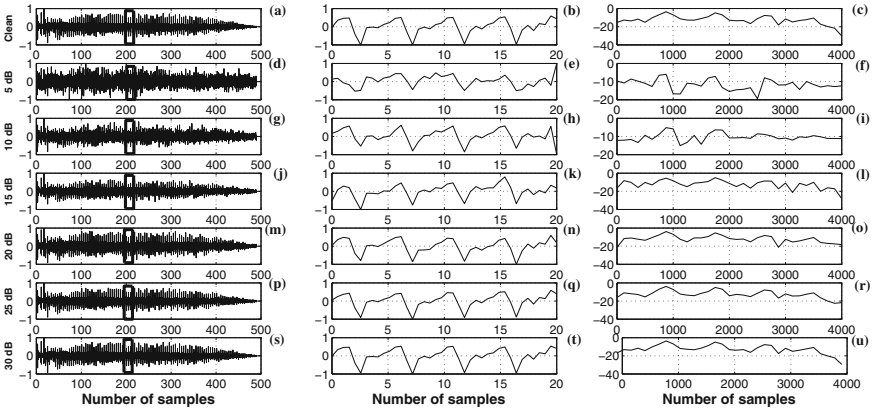

Fig. 5.3 **a** Segment of clean speech and bounding box represents 20 ms segmental level frame. **d**, **g**, **j**, **m**, **p** and **s** are the segments of speech after adding 5, 10, 15, 20, 25, 25 and 30 dB white noise, respectively and the bounding boxes represent 20 ms segmental level frames. **b**, **e**, **h**, **k**, **n**, **q** and **t** are LP residual samples obtained from corresponding 20 ms segmental level frame after decimated by factor 4. **c**, **f**, **i**, **l**, **o**, **r**, **u** are the corresponding log magnitude spectra

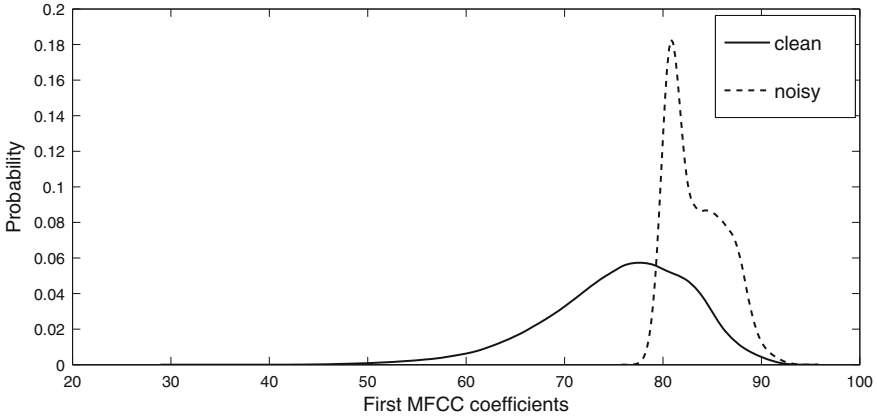

Fig. 5.4 *Solid line* represents the distribution of 1st MFCC coefficients of MFCC feature vectors in clean environment and *dashed line* represents the distribution of 1st MFCC coefficients of MFCC feature vectors after adding 10 dB noise to speech signal

5.6.2 Processing of Robust Excitation Source Features for Language Identification

The characteristics of excitation source can be captured by passing the speech signal through the inverse filter [6]. In this work, the 10th order LP analysis followed by inverse filter is used for estimating LP residual signal. Linear prediction analysis represents the second order statistical features in terms of the autocorrelation

coefficients. So, the LP residual signal does not represent any significant second order relations corresponding to the vocal tract resonator. Without disturbing the magnitude and phase components of LP residual signal, the raw residual samples are used as feature vectors to capture different aspects of excitation source information. We prophesy that, the evidences obtained from excitation source component of speech may contain robust language-specific phonotactic information.

In this study, the LP residual signal is analyzed within 2-3 consecutive glottal cycles known as segmental (*seg*) level information. At *seg* level, first the LP residual signal is decimated by a factor 4 to suppress the finer variations, which occur due to the dynamic nature vocal folds within a glottal cycle. The LP residual signal is then processed, in block size of 20 ms with a shift of 2.5 ms and corresponding raw LP residual samples are used as *seg* level features. The instantaneous pitch and energy of vocal folds vibration are captured by *seg* level processing of LP residual signal. In Fig. 5.2a, the LP residual signal is shown and the bounding box indicates the *seg* level 20 ms frame. Figure 5.2b shows the raw LP residual samples obtained from the *seg* level frame.

5.6.3 Evaluation of Robustness of Excitation Source Features for Language Identification

This LID study has been carried out on 27 Indian languages. In this work, $(n - 2)$ speakers are used from each language to develop the language models and the other two speakers of each language who have not participated during training phase are considered for evaluation. Here, n is the total number of speakers in each language. In this work, the amount of training data and duration of test utterances have been varied to analyze the effects on LID accuracies. We have considered 5, 20 and 45 min of training data per each language to develop the language models. 20 test utterances are collected from both male and female speakers per each language during evaluation. We have considered test samples duration of 2, 5 and 10 s for evaluation purpose. Clean speech data has been used to develop the language models whereas, the language models are evaluated using both clean and noisy speech utterances to examine the robustness of excitation source features. For experimental convenience and due to unavailability of real-life noisy data for each language, the noisy backgrounds were artificially simulated. During evaluation all the test utterances are corrupted with the additive white noise collected from NOISEX-92 database [7]. We explored 32, 64, 128 and 256 Gaussian mixtures for suitable modelling of the excitation source information. The performances of both source and vocal tract features for optimum GMMs are given in Tables 5.4 and 5.5. 1st column of both the tables indicate different amount of training data used in this work. 2nd column represents different lengths of test utterances considered for each of the training condition. 3rd columns of both the tables represent the LID accuracies obtained by processing the LP residual and MFCC features from clean test utterances, respectively. LID systems are evaluated

Table 5.4 LID Performance using LP residual feature in clean and noisy (white noise) backgrounds

Amount of data		LID performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	42.40 (1.30)	41.48 (3.45)	40.74 (5.17)	39.44 (8.19)	38.89 (9.47)	36.11 (15.95)
	5	46.48	47.03 (−1.18)	46.29 (0.41)	45.00 (3.18)	42.03 (9.57)	41.85 (9.96)	39.07 (15.94)
	10	47.96	49.07 (−2.31)	47.96 (0.00)	48.52 (−1.17)	46.29 (3.48)	42.22 (11.97)	42.40 (11.59)
20	2	41.85	41.48 (0.88)	40.37 (3.54)	40.00 (4.42)	37.59 (10.18)	36.85 (11.95)	33.33 (20.36)
	5	46.67	47.22 (−1.18)	46.85 (−0.39)	46.29 (0.81)	44.07 (5.57)	41.85 (10.33)	39.25 (15.90)
	10	47.40	48.33 (−1.96)	47.77 (−0.78)	48.14 (−1.56)	46.29 (2.34)	44.63 (5.84)	40.18 (15.23)
45	2	38.33	37.22 (2.90)	36.67 (4.33)	35.37 (7.72)	35.00 (8.69)	32.59 (14.98)	29.07 (24.16)
	5	44.81	44.07 (1.65)	43.44 (3.06)	43.33 (3.30)	40.56 (9.48)	37.33 (16.69)	33.51 (25.22)
	10	45.56	46.11 (−1.21)	47.22 (−3.64)	47.03 (−3.23)	44.81 (1.65)	42.40 (6.94)	36.85 (19.12)

Table 5.5 LID Performance using MFCC feature in clean and noisy (white noise) backgrounds

Amount of data		LID performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55.00	50.74 (7.75)	45.18 (17.85)	35.74 (35.02)	21.48 (60.95)	8.51 (84.53)	5.18 (90.58)
	5	59.25	58.88 (0.62)	54.25 (8.44)	39.62 (33.13)	25.55 (56.88)	10.00 (83.12)	5.74 (90.31)
	10	62.78	61.66 (1.78)	58.51 (6.80)	44.81 (28.62)	31.29 (50.16)	12.03 (80.84)	7.96 (87.32)
20	2	52.96	46.67 (11.88)	42.03 (20.64)	33.89 (36.01)	17.96 (66.09)	8.33 (84.27)	4.62 (91.28)
	5	59.81	59.25 (0.94)	52.03 (13.01)	38.51 (35.61)	21.67 (63.77)	8.14 (86.39)	5.56 (90.70)
	10	63.70	63.33 (0.58)	58.33 (8.43)	41.67 (34.58)	25.00 (60.75)	9.44 (85.18)	5.92 (90.71)
45	2	53.51	47.22 (11.75)	42.77 (20.07)	32.96 (38.40)	16.29 (69.56)	7.96 (85.12)	4.62 (91.37)
	5	58.14	57.22 (1.58)	49.07 (15.60)	37.40 (35.67)	19.07 (67.20)	9.07 (84.40)	4.62 (92.05)
	10	62.59	62.22 (0.59)	56.85 (9.17)	40.74 (34.91)	21.48 (65.68)	11.29 (81.96)	5.18 (91.72)

by varying the background noise levels. The notable observation can be made when the average recognition accuracies are analyzed in noisy environments. The SNR has been varied from 5 to 30 dB with the steps of 5 dB to study the effects of background noise on LID accuracies for both the features. The LID accuracies obtained by processing LP residual and MFCC features at different SNR levels are tabulated from 4th to 9th columns of Tables 5.4 and 5.5, respectively.

5.6.3.1 Effect of Amount of Training Data on LID Performances

Slight variation in LID accuracies has been observed by varying the amount of training data for both source and spectral features. For example, the LID accuracies obtained from 5, 20 and 45 min of training data are 42.96, 41.85 and 38.33 %, respectively. These accuracies are obtained from LID systems developed using excitation source features, and evaluated by clean test utterances of 2 s duration. Similar observation can also be made for 5 and 10 s test utterances by processing both LP residual and MFCC features at clean and noisy backgrounds. It can be observed that, the LID accuracy is slightly increasing with decreasing the amount of training data. As the duration of training data is less, the speaker bias will be reduced. Hence, slightly better accuracy with less training data has been observed. From the empirical analysis it can be elicited that, both the excitation source and spectral features are not much sensitive to the amount of training data.

5.6.3.2 Effect of Length of Test Utterances on LID Performances

The noteworthy improvement in LID accuracies using spectral features can be observed by varying the length of test samples on a fixed amount of training data. From the 3rd column of Table 5.5, it can be observed that, for 5 min of training data, LID accuracies are 55, 59.25 and 62.78 % using 2, 5 and 10 s clean test utterances, respectively. The maximum LID accuracy is obtained in this training condition is 62.78 %, which is considered as reference performance to compute the percentage decrement of LID accuracies by varying the duration of test utterances. The percentage decrement in LID accuracies for 2 and 5 s test samples are 12.39 and 5.62 %, respectively. The maximum LID accuracy is obtained using LP residual feature for 5 min training data per each language is 47.96 %, which is considered as reference performance to compute the percentage decrement in this case. The percentage decrement of LID accuracies are 10.42 and 3.08 % for 2 and 5 s test samples, respectively. From this observation it can be stated that, the excitation source features represented by LP residual signal is less sensitive to the duration of test utterances compared to the vocal tract features. This empirical analysis delineates the robustness of excitation source features for language identification task.

5.6.3.3 Effect of Background Noise on LID Performances

The conspicuous effects of noise on LP residual and MFCC features can be analyzed by comparing the LID accuracies tabulated in Tables 5.4 and 5.5, respectively. The LID performance decreases by adding white noise with the test utterances during evaluation. The percentage deviation of LID accuracy is computed as,

$$D = \left(\frac{P_c - P_n}{P_c} \right) \times 100\,\% \tag{5.3}$$

where, P_c is the LID performance obtained from clean background and P_n is the LID accuracy achieved at a particular SNR level. The numerical values given inside the parentheses from 4th to 9th columns of Tables 5.4 and 5.5 represent the values of percentage deviation (D) computed by Eq. 5.3. Percentage deviation of LID accuracies by processing both the LP residual and MFCC features are shown in Fig. 5.5. In this case, the amount of training data is considered 5 min per each language and duration of test samples 2, 5 and 10 s are considered. Bottom three curves illustrate the D value obtained from LP residual feature for different lengths of test utterances. Top three curves represent the values of D obtained by processing the MFCC feature for different lengths of test utterances. Crucial observation can be made by comparing the performance deviations obtained from LPR and MFCC features. At high noise levels (i.e., low SNR values), the slopes of the deviation curves obtained from MFCC feature are high, compared to the slopes of the LPR curves. This indicates that, the deviation in performance is increasing drastically for MFCC feature, compared to the LPR feature. It can be observed that, the LPR feature provides better LID accuracy compared to the MFCC feature, while increasing the noise levels. This empirical analysis illustrates that, the language-specific excitation source information captured by processing the LP residual signal is more robust, compared to the vocal tract information represented by MFCCs. The potency of LP residual feature can also be illustrated by comparing the LID performances obtained from LP residual and MFCC features at a particular SNR level. For example, the recognition performance of 36.11 % has been achieved by processing the LP residual feature at 5 dB SNR level. In this case, the amount of training data is 5 min per each language and

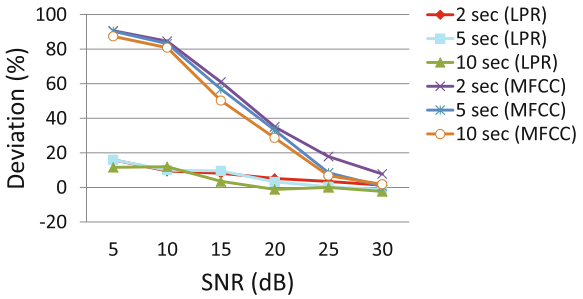

Fig. 5.5 Deviation in performance D for LP residual and MFCC features for white noise at different SNRs

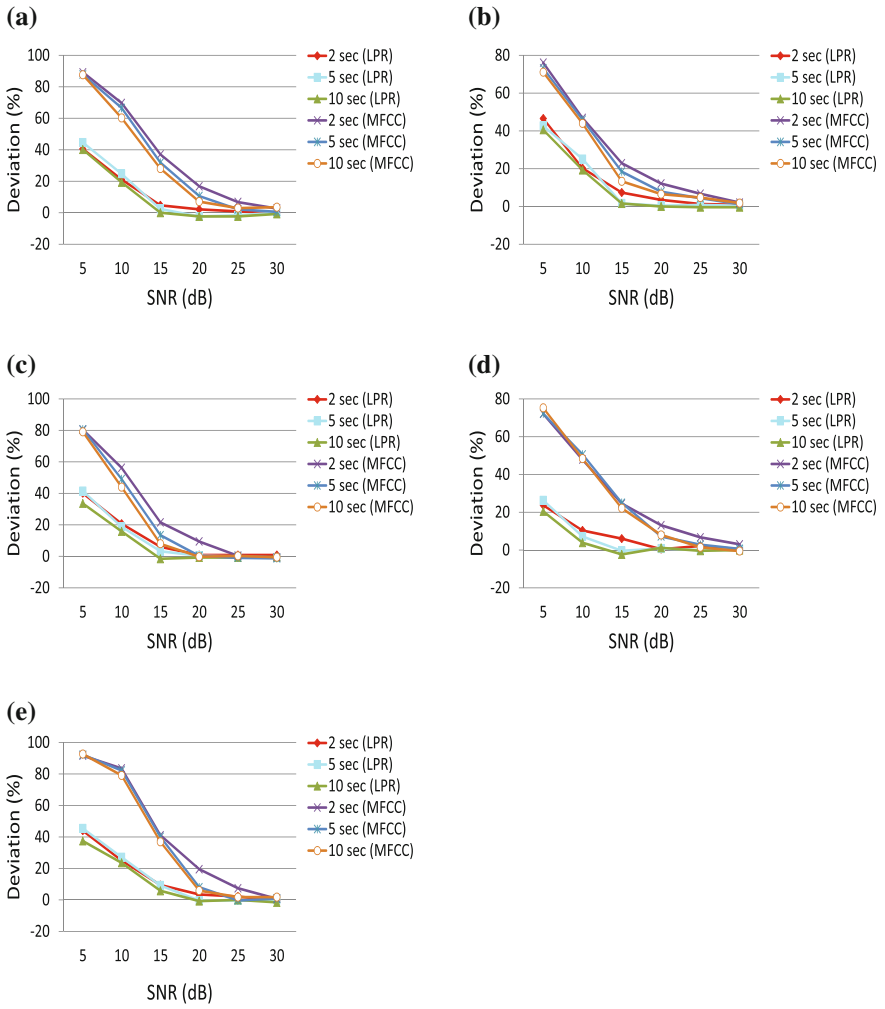

Fig. 5.6 Deviation in performance D of LP residual and MFCC features for fighter jet (Buccanier), destroyer engine, factory, high frequency channel and pink noises at different SNRs. **a** Buccanier noise. **b** Destroyer engine noise. **c** Factory noise. **d** HF channel noise. **e** Pink noise

duration of each test sample is 2 s. The accuracy obtained by processing the MFCC feature at similar conditions is only 5.18 %.

The similar study has been carried out by using various types of noises such as, factory, high frequency (HF) radio channel, pink, fighter jet (Buccaneer) and destroyer engine noises. Experimental results obtained from these noises are given in Appendix C. The percentage deviation with respect to the SNRs is shown in Fig. 5.6. In this case, 5 min of training data per each language and duration of test samples 2, 5 and 10 s are considered. In Fig. 5.6a–e, the bottom three lines represent the deviation

of LID accuracies for LPR feature and top three lines represent performance deviation for MFCC feature. It can be observed from Fig. 5.6a–e that, the deviation of LPR feature is always less, compared to the MFCC feature. Another notable thing can be observed is that, the rate of increasing of slopes of the deviation curves corresponding to MFCC feature is high, compared to the LPR features, while decreasing the SNR values (i.e., increasing the noise levels). This observation delineates the potency of LPR feature in various noisy environments. From this empirical analysis it can be elicited that, the excitation source features contribute robust language-specific information, compared to vocal tract information represented by MFCC feature.

5.7 Summary

In this chapter, the evidences from implicit and parametric source features are combined to capture overall excitation source information for language identification. The overall language-specific excitation source information is compared to the contemporary vocal tract features. The empirical evidences portray the complementary nature of these two features. Hence, we have further combined the evidences obtained from overall excitation source and vocal tract features to develop integrated LID systems. We have also examined the robustness of excitation source features compared to the vocal tract features by evaluating the LID systems in clean and degraded environments with varying the amount of training data and lengths of test utterances. The segmental level raw LP residual samples are processed to capture the robust excitation source information for language identification task. The LID study has been carried out using various noises at different SNR levels to investigate the existence of robust language-specific information. It has been observed that, the spectral features depends on the quality and quantity of data. However, the empirical analysis concludes that, the excitation source information provides robust language-specific information compared to vocal tract information represented by MFCCs.

References

1. K.S. Rao, S. Maity, V.R. Reddy, Pitch synchronous and glottal closure based speech analysis for language recognition. Int. J. Speech Technol. (Springer) **16**(4), 413–430 (2013)
2. J.R. Deller Jr, J.H.L. Hansen, J.G. Proakis, *Discrete-Time Processing of Speech Signal*, 2nd edn. (IEEE Press, New York, 2000)
3. L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, 1978)
4. V.R. Reddy, S. Maity, K.S. Rao, Identification of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. (Springer) **16**(4), 489–511 (2013)
5. Y.K. Muthusamy, R.A. Cole, B.T. Oshika, The OGI multilanguage telephone speech corpus, in Spoken Language Processing, pp. 895–898 (1992)
6. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
7. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**, 247–251 (1993)

Chapter 6
Summary and Conclusion

Abstract This chapter summerizes the overall contents of the book. Major contributions and future scope of work have been highlighted.

Keywords Language identification · Excitation source features for language identification · Implicit excitation source information · Parametric excitation source features · Robust language identification · Robust excitation source features · Language identification using excitation source and vocal tract system features

6.1 Summary of the Book

In this work, language identification (LID) has been carried out by exploring the characteristics of excitation source. The linear prediction (LP) residual signal is considered for representing the excitation source, and it has been analyzed at sub-segmental, segmental and supra-segmental levels to derive different aspects of excitation source for language discrimination task. Methods are proposed to derive implicit features of excitation source. *Implicit* processing is computationally intensive. Hence, LP residual signal is also parameterized at three different levels to capture language-specific information in a simple and effective way. The complementary information present between the vocal tract and excitation source features is also analyzed. Eventually, the robustness of proposed excitation source features is also examined by varying (i) amount of training data, (ii) length of test samples and (iii) background noise levels.

In *implicit* processing approach, raw LP residual samples, its magnitude and phase components are processed at three different levels: sub-segmental, segmental and supra-segmental levels for LID task. The minute variations present within a glottal cycle, instantaneous pitch and energy of 2–3 consecutive glottal cycles, and temporal variations of the pitch and energy across 50 glottal cycles are captured at sub-segmental, segmental and supra-segmental levels, respectively. From the empirical analysis, it has been observed that, the segmental level implicit features provide better accuracy, compared to other two levels. This indicates that, the instantaneous pitch and energy is more effective for language discrimination task. The effect of magnitude component of LP residual signal predominates over the phase component during

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0_6

direct processing of LP residual. Phase does not vary with the amplitude fluctuations. However, the phase of LP residual signal may also contribute some language-specific knowledge. Therefore, we have explored analytic signal representation of LP residual to process the magnitude and phase components of LP residual signal independently. From the empirical observations it can be elicited that, the combination of magnitude and phase information seem to be a better choice than direct processing of raw LP residual samples for language discrimination task [1, 2].

Linear prediction residual signal is also parameterized at sub-segmental, segmental and supra-segmental levels to capture the higher order statistics present in LP residual signal for language discrimination task. At sub-segmental level, glottal flow derivative (GFD) parameters are explored for modeling the glottal pulse characteristics. At, segmental level, LP residual signal is analyzed in cepstral and spectral domains to capture instantaneous energy and periodicity information. Residual mel-frequency cepstral coefficients (RMFCC) and mel power difference of spectrum in sub-bands (MPDSS) have been proposed for modeling language-specific energy and periodicity information, respectively. At supra-segmental level, pitch, epoch strength and epoch sharpness contours are proposed for capturing temporal variation of pitch and energy. Segmental level features provide better LID accuracy compared to other two levels. This indicates the efficacy of segmental level excitation source information for language discrimination task. Experimental observation portrays that, the language-specific excitation source information present at three levels are fundamentally distinct. Therefore, the evidences obtained from different features proposed at three levels are combined to capture the complete parametric excitation source information.

The evidences obtained from *implicit* and parametric features provide different cues for LID task. Therefore, we have combined the evidences from *implicit* and parametric features of excitation source to enhance the LID accuracy. The complementary nature of vocal tract and excitation source features is also investigated. The robustness of proposed excitation source features has been examined by varying (i) amount of training data, (ii) length of test samples and (iii) background noise levels. Experimental results show that, both excitation source and vocal tract features are not much sensitive to the variation of amount of training data. However, the vocal tract information represented by spectral features is more fragile to the length of test utterances than excitation source features. The language models are developed with clean data and evaluated by test samples after adding noise of different SNRs. Performance deviation with respect to the accuracy in clean background has been studied for both the features to observe the effects of noise. The performance deviation increases drastically for vocal tract feature, compared to the excitation source feature, as noise level increases. From the experimental evidences the conclusion can be drawn that, the excitation source feature is more robust to background noise, compared to the vocal tract feature.

6.2 Contributions of the Book

The major contributions of this work can be summarized as follows:

- Methods are proposed to derive implicit features from LP residual, its magnitude and phase components for capturing language-specific information.
- Methods are proposed for deriving explicit parametric features from LP residual signal for language discrimination task.
- Combination of implicit and parametric features of excitation source has been explored to enhance the LID performance.
- Combination of the evidences obtained from overall excitation source and vocal tract features has been explored to investigate the existence of complementary language-specific information present in these two features.
- The robustness of excitation source information has been explored for language identification task in terms of varying (i) background noise, (ii) amount of training data and (iii) duration of test samples.

6.3 Future Scope of Work

- We have analyzed LP residual signal at sub-segmental, segmental and supra-segmental levels by considering the fixed frame size and frame shift. The corresponding LP residual samples are not pitch synchronized. Computational complexity involved in the temporal processing of raw LP residual samples can be reduced by selecting the pitch synchronous blocks.
- In this work, language models are developed by Gaussian mixture model (GMM). Several non-linear modeling techniques like, artificial neural network (ANN) and support vector machines (SVM) can be explored with the proposed excitation source features.
- In future, hybrid and hierarchical LID systems can be explored for improving the LID accuracy.
- In this work, we have considered LP residual signal as excitation signal and present LID study carried out by analyzing LP residual signal at sub-segmental, segmental and supra-segmental levels. In future, the glottal volume velocity (GVV) can be considered as excitation signal and the similar study can be carried out. Further, the Electroglottographic (EGG) signal can also be analyzed for representing excitation source information precisely.
- In this work, we have developed integrated LID system by combining the proposed excitation source features with the vocal tract features. Further, LID system can also be developed by combining excitation source, vocal tract and prosodic features to enhance the LID accuracy.
- In this work, excitation source features are explored for language discrimination task. Explicit phonotactic features are not explored in this work. It may be possible to represent the phonotactic information of a language without using the phone

recognizers. In future, language-specific phonotactic information can be explored by implicit approaches.
- Speech enhancement methods can be incorporated with the proposed excitation source features to enhance the LID accuracy in noisy backgrounds.
- In this book, we have analyzed the robustness of excitation source features using single noise and fixed SNR. However, in real-life applications, the test samples may degrade by various background noises with different SNRs. In future, this work can be extended with varying noise types and noise levels.
- Deep neural networks can be explored to automatically derive the implicit language-specific features.
- In future, the phase component of LP residual signal represented by residual phase (RP) can be used as feature for analyzing the robustness of excitation source features.
- The LID studies based on speaking rate, regional accent, speaking style conditions under clean speech environment (or, high SNR conditions) will be performed in future.

References

1. D. Nandi, D. Pati, K. Sreenivasa Rao, Language identification using Hilbert envelope and phase information of linear prediction residual, in *16th International Oriental COCOSDA Conference*, Gurgoan, India, November 2013
2. D. Nandi, D. Pati, K. Sreenivasa Rao, Sub-segmental, segmental and supra-segmental analysis of linear prediction residual signal for language identification, in *International Conference on Signal Processing and Communications (SPCOM)*, IISc Bangalore, India, July 2014

Appendix A
Gaussian Mixture Model

In speech and speaker recognition, the acoustic events are usually modeled by Gaussian probability density functions (PDFs), described by the mean vector and the covariance matrix. However unimodel PDF with only one mean and covariance are unsuitable to model all variations of a single event in speech signals. Therefore, a mixture of single densities i.e., a Gaussian Mixture Model (GMM) is used to model the complex structure of the density probability. For a D-dimensional feature vector denoted as x_t, the mixture density for speaker Ω is defined as weighted sum of M component Gaussian densities as given by the following [1]

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

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

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

where μ_i is a mean vector and Σ_i covariance matrix for ith component. The mixture weights have to satisfy the constraint [1]

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

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

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

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0

Training the GMMs

To determine the model parameters of GMM of the speaker, the GMM has to be trained. In the training process, the maximum likelihood (ML) procedure is adopted to estimate model parameters. For a sequence of training vectors $X = \{x_1, x_2, .., x_T\}$, the GMM likelihood (assuming independent observations) can be written as [1]

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

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

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

Often, the average log-likelihood value is used by dividing $\log[P(X|\Omega)]$ by T. This is done to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor [2]. The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. The likelihood value is, however, a highly nonlinear function in the model parameters and direct maximization is not possible. Instead, maximization is done through iterative procedures. Of the many techniques developed to maximize the likelihood value, the most popular is the iterative expectation maximization (EM) algorithm [3].

Expectation Maximization (EM) Algorithm

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

1. **Initialization**: In this step an initial estimate of the parameters is obtained. The performance of the EM algorithm depends on this initialization. Generally, LBG [4] or K-means algorithm [5] is used to initialize the GMM parameters.
2. **Likelihood Computation**: In each iteration the posterior probabilities for the ith mixture is computed as [1]:

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

3. **Parameter Update**: Having the posterior probabilities, the model parameters are updated according to the following expressions [1].
   Mixture weight update:

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

   Mean vector update:

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

   Covariance matrix update:

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

In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. It is more common to use diagonal covariance matrices for GMM, since linear combination of diagonal covariance Gaussians has the same model capability with full matrices. Another reason is that speech utterances are usually parameterized with cepstral features. Cepstral features are more compactable, discriminative, and most important, they are nearly uncorrelated, which allows diagonal covariance to be used by the GMMs [1]. The iterative process is normally carried out 10 times, at which point the model is assumed to converge to a local maximum [1].

Testing

In identification phase, mixture densities are calculated for every feature vector for all speakers and speaker with maximum likelihood is selected as identified speaker. For example, if S speaker models $\{\Omega_1, \Omega_{2,\ldots,} \Omega_S\}$ are available after the training, speaker identification can be done based on a new speech data set. First, the sequence of

feature vectors $X = \{x_1, x_2, .., x_T\}$ is calculated. Then the speaker model $\hat{s}$ is determined which maximizes the a posteriori probability $P(\Omega_S|X)$. That is, according to the Bayes rule [1]

$$\hat{s} = \max_{1 \le s \le S} P(\Omega_S|X) = \max_{1 \le s \le S} \frac{P(X|\Omega_S)}{P(X)} P(\Omega_S). \tag{A.11}$$

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

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log P(x_t|\Omega_s) \tag{A.12}$$

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

Decision in verification is obtained by comparing the score computed using the model for the claimed speaker Ω_S given by $P(\Omega_S|X)$ to a predefined threshold θ. The claim is accepted if $P(\Omega_S|X) > \theta$, and rejected otherwise [6].

References

1. D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Audio Speech Lang. Process. **3**(1), 4–17 (1995)
2. F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, D.A. Reynolds, A tutorial on text-independent speaker verification. EURASIP J. Adv. Signal Process. (Special Issue Biometric Signal Process.) **4**(4), 430–451 (2004)
3. A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. **39**(1), 1–38 (1977)
4. Y. Linde, A. Buzo, R. Gray, An algorithm for vector quantizer design. IEEE Trans. Commun. **28**, 84–95 (1980)
5. C.M. Bishop, *Pattern Recognition and Machine Learning*. (Springer, Heidelberg, 2006)
6. D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**(1), 19–41 (2000)

Appendix B
Mel-Frequency Cepstral Coefficient (MFCC) Features

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

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

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

   where the value of b controls the slope of the filter and is usually between 0.4 to 1.0 [1].
2. **Frame blocking and windowing**: The speech signal is a slowly time-varying or quasi-stationary signal. For stable acoustic characteristics, speech needs to be examined over a sufficiently short period of time. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over 20 ms windows, and advanced every 10 ms [1]. Advancing the time window every 10 ms enables the temporal characteristics of individual speech sounds to be tracked and the 20 ms analysis window is usually sufficient to provide good spectral resolution of these sounds, and at the same time short enough to resolve significant temporal characteristics. The purpose of the overlapping

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0

analysis is that each speech sound of the input sequence would be approximately centered at some frame. On each frame a window is applied to taper the signal towards the frame boundaries. Generally, Hanning or Hamming windows are used [1]. This is done to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the Discrete Fourier Transform (DFT) on the signal.

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

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

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

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

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

where f denotes the physical frequency in Hz, and f_{mel} denotes the perceived frequency [1].

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

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

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

where M is total number of triangular mel weighting filters [2]. $H_m(k)$ is the weight given to the kth energy spectrum bin contributing to the mth output band and is expressed as:

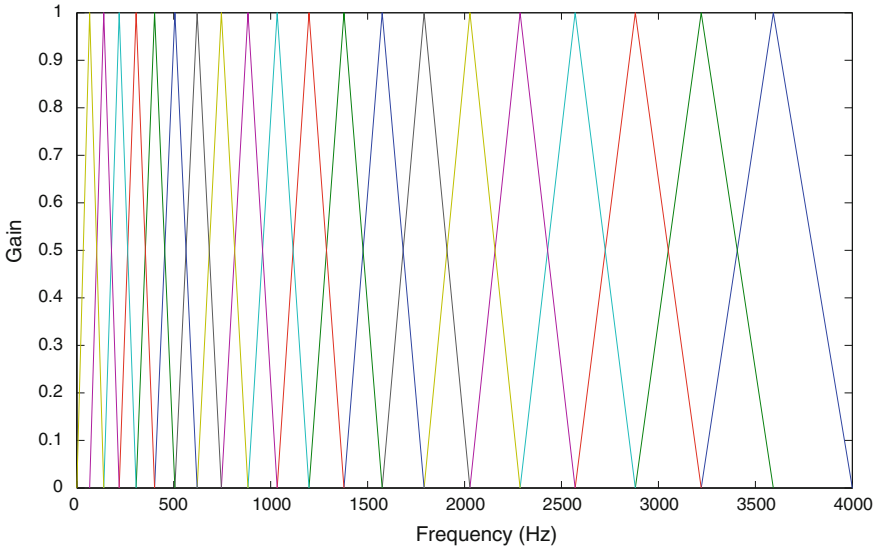

Fig. B.1 Mel filterbank

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

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

5. **Discrete Cosine Transform (DCT)**: Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients produces a set of cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low quefrequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components [1]. Finally, MFCC is calculated as [2]

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

where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. Traditional MFCC systems use only 8–13 cepstral coefficients. The zeroth coefficient is often

excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information.

6. **Dynamic MFCC features**: The cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first and second derivatives of cepstral coefficients [3]. The first order derivative is called delta coefficients, and the second order derivative is called delta-delta coefficients. Delta coefficients tell about the speech rate, and delta-delta coefficients provide information similar to acceleration of speech. The commonly used definition for computing dynamic parameter is

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

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

References

1. L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, 1st edn. (Prentice-Hall, Englewood Cliffs, 1993)
2. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
3. S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**(2), 254–272 (1981)

Appendix C
Evaluation of Excitation Source Features in Different Noisy Conditions

The LID systems developed by using excitation source and vocal tract features are evaluated at different noisy backgrounds. Five different types of noises such as, buccaneer, destroyer engine, factory, high frequency channel and pink noises have been considered for analyzing the robustness of proposed excitation source features. The performances of both excitation source and vocal tract features for above mentioned noises are given in Tables C.1, C.2, C.3, C.4, C.5, C.6, C.7, C.8, C.9 and C.10. For every types of noises it has been observed that, the accuracy obtained from excitation source features is better compared to, vocal tract features at low SNR levels. This empirical analysis portrays the robustness of excitation source features.

K.S. Rao and D. Nandi, *Language Identification Using Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-17725-0

Table C.1 LID performance using LP residual feature in clean and noisy (Buccaneer noise) backgrounds

Amount of data		LID performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	42.963 (−0.006897)	42.593 (0.85523)	42.037 (2.1484)	40.926 (4.7348)	33.889 (21.115)	25.556 (40.513)
	5	46.48	46.481 (−0.0031874)	47.222 (−1.5969)	47.593 (−2.3937)	45.37 (2.3873)	35 (24.699)	25.741 (44.62)
	10	47.96	48.333 (−0.77843)	49.074 (−2.3229)	49.074 (−2.3229)	47.963 (−0.006178)	38.704 (19.3)	28.704 (40.151)
20	2	41.85	42.037 (−0.44692)	41.852 (−0.004425)	40.185 (3.9781)	37.593 (10.173)	31.852 (23.89)	23.889 (42.918)
	5	46.67	46.296 (0.80074)	46.667 (0.0071423)	46.111 (1.1975)	44.63 (4.3719)	34.815 (25.402)	25.556 (45.242)
	10	47.4	48.148 (−1.5784)	48.333 (−1.9691)	48.333 (−1.9691)	48.333 (−1.9691)	36.852 (22.253)	28.333 (40.225)
45	2	38.33	38.333 (−0.0086964)	38.519 (−0.49183)	36.481 (4.8226)	35 (8.6877)	28.704 (25.114)	20.926 (45.406)
	5	44.81	44.259 (1.2291)	44.444 (0.81579)	42.407 (5.3617)	42.037 (6.1883)	32.037 (28.505)	20.926 (53.301)
	10	45.56	46.111 (−1.2096)	47.037 (−3.242)	47.407 (−4.0549)	46.111 (−1.2096)	35 (23.178)	24.815 (45.534)

Table C.2 LID performance using LP residual feature in clean and noisy (destroyer engine noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	42.593 (0.85523)	42.407 (1.2863)	41.481 (3.4416)	39.815 (7.3212)	34.259 (20.253)	22.963 (46.548)
	5	46.48	46.296 (0.39523)	46.111 (0.79365)	46.296 (0.39523)	45.741 (1.5905)	34.815 (25.097)	26.667 (42.628)
	10	47.96	48.148 (−0.3923)	48.148 (−0.3923)	47.963 (−0.006178)	47.222 (1.5383)	38.704 (19.3)	28.519 (40.537)
20	2	41.85	42.222 (−0.88942)	41.296 (1.3231)	40.185 (3.9781)	37.963 (9.288)	32.778 (21.678)	21.481 (48.67)
	5	46.67	46.481 (0.40394)	45.926 (1.5943)	45.926 (1.5943)	44.074 (5.5623)	34.815 (25.402)	22.037 (52.781)
	10	47.4	47.778 (−0.797)	48.148 (−1.5784)	48.519 (−2.3597)	48.519 (−2.3597)	42.037 (11.314)	26.667 (43.741)
45	2	38.33	38.148 (0.47444)	37.222 (2.8901)	35.556 (7.2383)	33.704 (12.07)	28.519 (25.597)	17.593 (54.102)
	5	44.81	44.815 (−0.010745)	44.815 (−0.010745)	43.148 (3.7087)	40.556 (9.4944)	30.926 (30.984)	19.074 (57.433)
	10	45.56	46.111 (−1.2096)	46.481 (−2.0226)	46.667 (−2.429)	45.37 (0.41622)	36.667 (19.52)	24.815 (45.534)

Table C.3 LID performance using LP residual feature in clean and noisy (factory noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	42.593 (0.85523)	42.593 (0.85523)	42.593 (0.85523)	40.37 (6.028)	34.074 (20.684)	25.741 (40.082)
	5	46.48	46.852 (−0.80003)	46.481 (−0.0031874)	46.296 (0.39523)	45 (3.1842)	37.778 (18.723)	27.222 (41.432)
	10	47.96	48.148 (−0.3923)	48.333 (−0.77843)	48.333 (−0.77843)	48.704 (−1.5507)	40.37 (15.825)	31.852 (33.587)
20	2	41.85	42.037 (−0.44692)	40.926 (2.2081)	40.926 (2.2081)	37.778 (9.7305)	32.963 (21.235)	22.593 (46.015)
	5	46.67	46.296 (0.80074)	46.111 (1.1975)	46.111 (1.1975)	43.704 (6.3559)	36.852 (21.037)	27.963 (40.084)
	10	47.4	47.963 (−1.1877)	48.148 (−1.5784)	48.333 (−1.9691)	47.963 (−1.1877)	41.111 (13.268)	30.37 (35.927)
45	2	38.33	38.519 (−0.49183)	37.963 (0.95757)	36.852 (3.8564)	35.185 (8.2046)	28.148 (26.564)	19.074 (50.237)
	5	44.81	45 (−0.42401)	44.259 (1.2291)	42.778 (4.5352)	42.407 (5.3617)	32.593 (27.265)	23.148 (48.342)
	10	45.56	45.37 (0.41622)	46.111 (−1.2096)	46.111 (−1.2096)	46.296 (−1.6161)	37.037 (18.707)	27.593 (39.437)

Table C.4 LID performance using LP residual feature in clean and noisy (high frequency channel noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	42.78 (0.42)	42.04 (2.15)	42.78 (0.42)	40.37 (6.03)	38.52 (10.34)	32.78 (23.7)
	5	46.48	46.3 (0.4)	46.48 (0)	46.11 (0.79)	46.67 (−0.4)	43.15 (7.17)	34.26 (26.29)
	10	47.96	47.96 (−0.01)	48.15 (−0.39)	47.41 (1.15)	49.07 (−2.32)	46.11 (3.86)	38.15 (20.46)
20	2	41.85	41.3 (1.32)	41.67 (0.44)	40.74 (2.65)	39.26 (6.19)	35.74 (14.6)	30.19 (27.87)
	5	46.67	46.11 (1.2)	46.11 (1.2)	46.48 (0.4)	46.3 (0.8)	43.7 (6.36)	34.26 (26.59)
	10	47.4	47.22 (0.38)	47.59 (−0.41)	48.15 (−1.58)	49.44 (−4.31)	47.78 (−0.8)	38.89 (17.96)
45	2	38.33	38.89 (−1.46)	37.78 (1.44)	36.3 (5.31)	35.74 (6.76)	32.41 (15.45)	27.41 (28.5)
	5	44.81	44.63 (0.4)	44.26 (1.23)	43.89 (2.06)	42.04 (6.19)	37.78 (15.69)	30.37 (32.22)
	10	45.56	45.56 (0.01)	46.11 (−1.21)	47.04 (−3.24)	47.41 (−4.05)	43.15 (5.29)	33.52 (26.43)

Table C.5 LID performance using LP residual feature in clean and noisy (pink noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	42.96	43.148 (−0.43796)	42.037 (2.1484)	41.481 (3.4416)	38.889 (9.4765)	32.222 (24.995)	24.074 (43.962)
	5	46.48	46.481 (−0.0031874)	46.667 (−0.40161)	46.667 (−0.40161)	42.222 (9.1605)	33.889 (27.089)	25.37 (45.417)
	10	47.96	48.704 (−1.5507)	47.963 (−0.006178)	48.333 (−0.77843)	45.185 (5.7857)	36.667 (23.547)	30 (37.448)
20	2	41.85	41.296 (1.3231)	41.667 (0.43807)	39.815 (4.863)	37.222 (11.058)	30.37 (27.43)	20.556 (50.883)
	5	46.67	46.481 (0.40394)	46.296 (0.80074)	45.37 (2.7847)	43.704 (6.3559)	33.333 (28.577)	25.926 (44.448)
	10	47.4	48.519 (−2.3597)	47.778 (−0.797)	48.333 (−1.9691)	45.926 (3.1099)	36.111 (23.816)	28.148 (40.616)
45	2	38.33	37.037 (3.3732)	37.778 (1.4407)	36.296 (5.3058)	34.444 (10.137)	26.481 (30.912)	19.074 (50.237)
	5	44.81	44.444 (0.81579)	43.148 (3.7087)	42.778 (4.5352)	38.148 (14.867)	30.37 (32.224)	21.852 (51.234)
	10	45.56	45.741 (−0.39671)	46.111 (−1.2096)	47.037 (−3.242)	43.519 (4.4809)	34.074 (25.211)	27.407 (39.843)

Table C.6 LID performance using MFCC feature in clean and noisy (buccaneer noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55	53.52 (2.69)	51.3 (6.73)	45.74 (16.84)	34.63 (37.04)	16.67 (69.7)	5.93 (89.23)
	5	59.25	58.89 (0.61)	57.96 (2.17)	52.96 (10.61)	40.37 (31.86)	19.82 (66.56)	7.22 (87.81)
	10	62.78	60.56 (3.54)	60.93 (2.95)	58.33 (7.08)	45.19 (28.03)	25 (60.18)	7.78 (87.61)
20	2	52.96	50.56 (4.54)	47.78 (9.79)	42.78 (19.23)	32.96 (37.76)	15.56 (70.63)	6.3 (88.11)
	5	59.81	58.15 (2.78)	57.41 (4.02)	51.85 (13.31)	38.15 (36.22)	20 (66.56)	6.48 (89.16)
	10	63.7	61.67 (3.19)	61.11 (4.06)	58.52 (8.13)	43.7 (31.39)	24.26 (61.92)	8.7 (86.34)
45	2	53.51	49.82 (6.91)	47.59 (11.06)	42.41 (20.75)	31.85 (40.48)	12.41 (76.81)	4.44 (91.69)
	5	58.14	56.67 (2.53)	56.11 (3.49)	48.15 (17.19)	37.96 (34.7)	15.74 (72.93)	5 (91.4)
	10	62.59	60.93 (2.66)	61.11 (2.36)	56.3 (10.06)	41.11 (34.32)	19.44 (68.93)	7.96 (87.28)

Table C.7 LID performance using MFCC feature in clean and noisy (destroyer engine noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55	53.89 (2.02)	51.3 (6.73)	48.33 (12.12)	42.41 (22.9)	29.26 (46.8)	13.15 (76.09)
	5	59.25	58.52 (1.23)	56.67 (4.36)	54.63 (7.8)	48.33 (18.43)	32.04 (45.93)	15.93 (73.12)
	10	62.78	61.67 (1.77)	59.82 (4.72)	58.7 (6.49)	54.44 (13.28)	35.19 (43.96)	18.15 (71.09)
20	2	52.96	50.56 (4.54)	48.89 (7.69)	46.85 (11.53)	41.67 (21.32)	27.04 (48.95)	13.33 (74.82)
	5	59.81	57.59 (3.71)	56.3 (5.87)	53.89 (9.9)	46.48 (22.29)	32.04 (46.44)	15.56 (73.99)
	10	63.7	60.37 (5.23)	59.82 (6.1)	59.07 (7.26)	54.82 (13.95)	37.22 (41.57)	19.44 (69.48)
45	2	53.51	50.74 (5.18)	48.7 (8.98)	45 (15.9)	38.89 (27.32)	23.33 (56.39)	13.52 (74.74)
	5	58.14	57.04 (1.9)	55.19 (5.08)	50.56 (13.05)	44.63 (23.24)	26.48 (54.45)	14.44 (75.16)
	10	62.59	59.82 (4.43)	60.37 (3.55)	59.63 (4.73)	49.44 (21)	30.56 (51.18)	17.22 (72.48)

Table C.8 LID performance using MFCC feature in clean and noisy (factory noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55	55.37 (−0.67)	54.82 (0.34)	49.82 (9.43)	43.15 (21.55)	24.07 (56.23)	10.74 (80.47)
	5	59.25	60 (−1.27)	59.82 (−0.95)	59.07 (0.3)	51.3 (13.42)	30.19 (49.06)	11.48 (80.62)
	10	62.78	63.15 (−0.59)	62.59 (0.3)	62.96 (−0.29)	57.78 (7.97)	35.19 (43.96)	13.15 (79.06)
20	2	52.96	50.74 (4.19)	49.82 (5.94)	46.11 (12.93)	40.56 (23.42)	23.7 (55.24)	9.81 (81.47)
	5	59.81	59.82 (−0.01)	60.19 (−0.63)	57.41 (4.02)	47.22 (21.05)	29.44 (50.77)	10.37 (82.66)
	10	63.7	63.89 (−0.3)	63.33 (0.58)	62.22 (2.32)	55 (13.66)	33.7 (47.09)	12.04 (81.1)
45	2	53.51	51.11 (4.48)	49.82 (6.91)	46.11 (13.83)	38.52 (28.02)	22.04 (58.82)	9.26 (82.7)
	5	58.14	58.15 (−0.01)	56.67 (2.53)	54.07 (6.99)	43.89 (24.51)	25.74 (55.73)	11.85 (79.62)
	10	62.59	61.67 (1.48)	62.22 (0.59)	61.67 (1.48)	50.19 (19.82)	29.63 (52.66)	15.93 (74.56)

Table C.9 LID performance using MFCC feature in clean and noisy (high frequency channel noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55	53.33 (3.03)	51.3 (6.73)	47.78 (13.13)	41.48 (24.58)	28.7 (47.81)	15.37 (72.05)
	5	59.25	58.89 (0.61)	57.59 (2.8)	55 (7.17)	44.44 (24.99)	29.26 (50.62)	16.48 (72.18)
	10	62.78	63.15 (−0.59)	61.85 (1.48)	57.78 (7.97)	48.89 (22.13)	32.41 (48.38)	15.56 (75.22)
20	2	52.96	50.19 (5.24)	47.41 (10.49)	43.52 (17.83)	38.15 (27.97)	29.44 (44.4)	16.67 (68.53)
	5	59.81	57.59 (3.71)	55.74 (6.8)	52.04 (13)	42.96 (28.17)	32.41 (45.82)	16.3 (72.75)
	10	63.7	61.85 (2.9)	60.74 (4.65)	56.3 (11.62)	48.52 (23.83)	33.89 (46.8)	17.96 (71.8)
45	2	53.51	48.89 (8.64)	46.3 (13.48)	43.33 (19.02)	36.85 (31.13)	27.41 (48.78)	13.15 (75.43)
	5	58.14	57.04 (1.9)	55 (5.4)	48.52 (16.55)	39.63 (31.84)	30 (48.4)	13.15 (77.39)
	10	62.59	60.56 (3.25)	59.82 (4.43)	55.37 (11.54)	45.37 (27.51)	32.96 (47.34)	12.41 (80.18)

Table C.10 LID performance using MFCC feature in clean and noisy (pink noise) backgrounds

Amount of Data		LID Performance (in %)						
Train data per language (in mins.)	Test data (in secs.)	Clean environment	Noisy environment					
			30 dB	25 dB	20 dB	15 dB	10 dB	5 dB
5	2	55	54.63 (0.67)	50.93 (7.41)	44.26 (19.53)	32.41 (41.08)	9.07 (83.5)	4.63 (91.58)
	5	59.25	58.7 (0.92)	59.44 (−0.33)	54.44 (8.11)	35.56 (39.99)	10.56 (82.19)	4.63 (92.19)
	10	62.78	61.67 (1.77)	61.67 (1.77)	59.07 (5.9)	39.63 (36.88)	13.15 (79.06)	4.63 (92.63)
20	2	52.96	50.93 (3.84)	47.96 (9.44)	41.67 (21.32)	30.37 (42.65)	7.78 (85.31)	4.63 (91.26)
	5	59.81	58.89 (1.54)	59.63 (0.3)	51.48 (13.93)	35.74 (40.24)	9.44 (84.21)	4.63 (92.26)
	10	63.7	62.04 (2.61)	62.78 (1.45)	57.41 (9.88)	39.82 (37.5)	11.48 (81.98)	4.63 (92.73)
45	2	53.51	50 (6.56)	47.04 (12.1)	41.67 (22.13)	27.96 (47.74)	8.33 (84.43)	5.74 (89.27)
	5	58.14	57.59 (0.94)	57.41 (1.26)	48.15 (17.19)	32.78 (43.62)	9.63 (83.44)	5.93 (89.81)
	10	62.59	62.04 (0.88)	62.59 (0)	55.56 (11.24)	38.7 (38.16)	12.04 (80.77)	7.59 (87.87)