SPRINGERBRIEFS IN SPEECH TECHNOLOGY
STUDIES IN SPEECH SIGNAL PROCESSING, NATURAL LANGUAGE UNDERSTANDING, AND MACHINE LEARNING

Manjunath K.E.

# Multilingual Phone Recognition in Indian Languages

Springer

# SpringerBriefs in Speech Technology

Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning

**Series Editor**

Amy Neustein, Fort Lee, NJ, USA

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the-art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, standardized manuscript preparation and formatting guidelines, and expedited production schedules.

The goal of the **SpringerBriefs in Speech Technology** series is to serve as an important reference guide for speech developers, system designers, speech engineers and other professionals in academia, government and the private sector. To accomplish this task, the series will showcase the latest findings in speech technology, ranging from a comparative analysis of contemporary methods of speech parameterization to recent advances in commercial deployment of spoken dialog systems.

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

Manjunath K. E.

# Multilingual Phone Recognition in Indian Languages

Springer

Manjunath K. E.
U R Rao Satellite Centre
Indian Space Research Organisation
Old Airport Road, Bengaluru
Karnataka, India

ISSN 2191-737X ISSN 2191-7388 (electronic)
SpringerBriefs in Speech Technology
ISBN 978-3-030-80740-5 ISBN 978-3-030-80741-2 (eBook)
https://doi.org/10.1007/978-3-030-80741-2



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

India is a land of many languages, among them 122 languages are spoken by at least 10,000 people each, with 22 of them constitutionally recognised. Several of the Indian languages do not have sufficient labelled data to develop a separate phone recogniser for themselves. This necessitates an investigation into alternative ways of performing phone recognition, such as multilingual phone recognition. A Multilingual Phone Recognition System (Multi-PRS) is a language-independent, universal Phone Recognition System (PRS) that can recognise the phonetic units present in a speech utterance independent of the language of the speech utterance.

In this book, various aspects of multilingual phone recognition such as development, analysis, performance improvement, and applications of Multi-PRSs are studied for $six$ Indian languages – Kannada (KN), Telugu (TE), Bengali (BN), Odia (OD), Urdu (UR), and Assamese (AS). Among the $six$ Indian languages considered, Chaps. 3, 4, and 5 use only $four$ languages (KN, TE, BN, OD), while the Chap. 6 uses all the $six$ languages. The International Phonetic Alphabets (IPA) based transcription is used for deriving a *common multilingual phone-set* by grouping the acoustically similar phonetic units from multiple languages. Both *Gaussian Mixture Model (GMM)-Hidden Markov Models (HMM)* and *Deep Neural Network (DNN)-HMMs* are explored for training the Multi-PRSs using Mel-frequency Cepstral Coefficients (MFCCs) as features under *context independent* and *context dependent* settings. The behaviour of Multi-PRSs across $two$ language families namely – Dravidian and Indo-Aryan – is studied and analysed by developing separate Multi-PRSs for Dravidian and Indo-Aryan language families. The performance of Multi-PRSs is analysed and compared with that of the Monolingual Phone Recognition Systems (Mono-PRS).

Articulatory Features (AFs) are explored to improve the performance of Multi-PRSs. The AFs for five AF groups – place, manner, roundness, frontness, and height – are predicted from the MFCCs using DNNs. The oracle AFs, which are derived from the ground truth IPA transcriptions, are used to set the best performance realizable by the predicted AFs. The performance of predicted and oracle AFs are compared. In addition to the AFs, the phone posteriors are explored to further boost the performance of Multi-PRS. Multitask Learning (MTL) is explored to improve the prediction accuracy of AFs and thereby reducing the Phone Error Rate (PER) of Multi-PRS. Fusion of AFs is done using two approaches: (i) lattice rescoring approach and (ii) AFs as tandem features. It is found that the oracle AFs by feature fusion with MFCCs offer a remarkably low target PER of 10.4%, which is 24.7% absolute reduction compared to baseline Multi-PRS with MFCCs alone. The fusion of phone posteriors and the AFs derived from the MTL yields the best performance. The best performing system using predicted AFs has shown reduction of 3.2% in absolute PER (i.e. 9.1% reduction in relative PER) compared to baseline Multi-PRS.

Applications of multilingual phone recognition in code-switched and non-code-switched scenarios are discussed. Two different approaches for multilingual phone recognition using code-switched and non-code-switched test sets are compared and evaluated. First approach is a front-end Language Identification (LID) system followed by monolingual phone recognisers (LID-Mono) trained individually on each of the languages present in multilingual dataset, while the second approach uses a *common multilingual phone-set* without requiring a front-end LID based switching. Bilingual code-switching experiments are conducted using the code-switched test sets of Kannada and Urdu languages. The state-of-the-art i-vectors are used to perform LID in first approach. It is found that the performance of *common multilingual phone-set* based approach is superior compared to more conventional LID-Mono approach in both non-code-switched and code-switched scenarios. The performance of LID-Mono approach heavily depends on the accuracy of the LID system, and the LID errors cannot be recovered. However, the *common multilingual phone-set* based approach by virtue of not having to do a front-end LID switching and designed based on the common multilingual phone-set derived from several languages is not constrained by the accuracy of the LID system, and hence performs effectively on non-code-switched and code-switched speech, offering low PERs than the LID-Mono system.

This book is mainly intended for researchers working in the area of multilingual speech recognition. This book will be useful for the young researchers who want to pursue research in speech processing with an emphasis on multilingual speech recognition. Hence, this may be recommended as a text or reference book for the postgraduate-level advanced speech processing course. The book has been organised as follows:

Chapter 1 introduces basic concepts of multilingual speech recognition. The multilingual AFs and code-switched speech recognition are briefly introduced. Chapter 2 describes the prior work on multilingual speech recognition systems with primary focus on multilingual AFs and code-switched speech recognition. Chapter 3 describes the development, evaluation, and analysis of Multi-PRS for four Indian languages. Chapter 4 discusses the proposed approaches to derive the multilingual AFs from spectral features. Chapter 5 discusses the use of predicted multilingual AFs to improve the performance of Multi-PRS. Chapter 6 describes the applications of multilingual phone recognition in code-switched and non-code-switched scenarios. Two approaches for multilingual phone recognition are compared using code-switched and non-code-switched test sets. Chapter 7 provides a brief summary and conclusion of the book with a glimpse towards the scope for possible future work.

I am grateful to my PhD supervisors Prof. V. Ramasubramanian and Prof. Dinesh Babu Jayagopi at the International Institute of Information Technology, Bangalore (IIITB), for their constant support, guidance, and encouragement to carry out this work. This book is based on my doctoral thesis work. I am also grateful to my MS supervisor Prof. K. Sreenivasa Rao at IIT Kharagpur for providing the speech corpora to carry out this work. I thank all the professors and research scholars of IIITB who have helped me to carry out this work. Special thanks to ISRO management and to my colleagues at URSC, ISRO for their cooperation and encouragement during the course of editing and publishing this book. Last but not the least, I am grateful to my parents, my in-laws, my wife, and my daughter for their constant support and encouragement. Finally, I thank all my friends and well-wishers.

Bengaluru, India Dr. Manjunath K. E.

# Contents

# Acronyms

| | |
|---|---|
| AF | Articulatory Feature |
| AF-PER | AF-Prediction Error Rate |
| AF-Predictor | Articulatory Feature Predictor |
| AF-Tandem | Combination of AFs as Tandem features |
| AS | Assamese |
| ASR | Automatic Speech Recognition |
| BN | Bengali |
| CD | Context-Dependent |
| CI | Context-Independent |
| DNN | Deep Neural Network |
| DP | Dynamic Programming |
| FFNN | Feed-Forward Neural Network |
| GMM | Gaussian Mixture Model |
| GMM-UBM | Gaussian Mixture Universal Background Model |
| HL | Hidden Layer |
| HMM | Hidden Markov Model |
| Hz | Hertz |
| IPA | International Phonetic Alphabet |
| ISA | Intrinsic Spectral Analysis |
| KN | Kannada |
| LFV | Language Feature Vector |
| LID | Language Identification |
| LID-Mono | LID-switched Monolingual Approach |
| LRA | Lattice Rescoring Approach |
| LVCSR | Large Vocabulary Continuous Speech Recognition |
| MFCC | Mel-frequency Cepstral Coefficient |
| MLP | Multi-layer Perceptron |
| Mono-PRS | Monolingual Phone Recognition System |
| ms | Millisecond |
| MSE | Mean Squared Error |
| MTL | Multitask Learning |