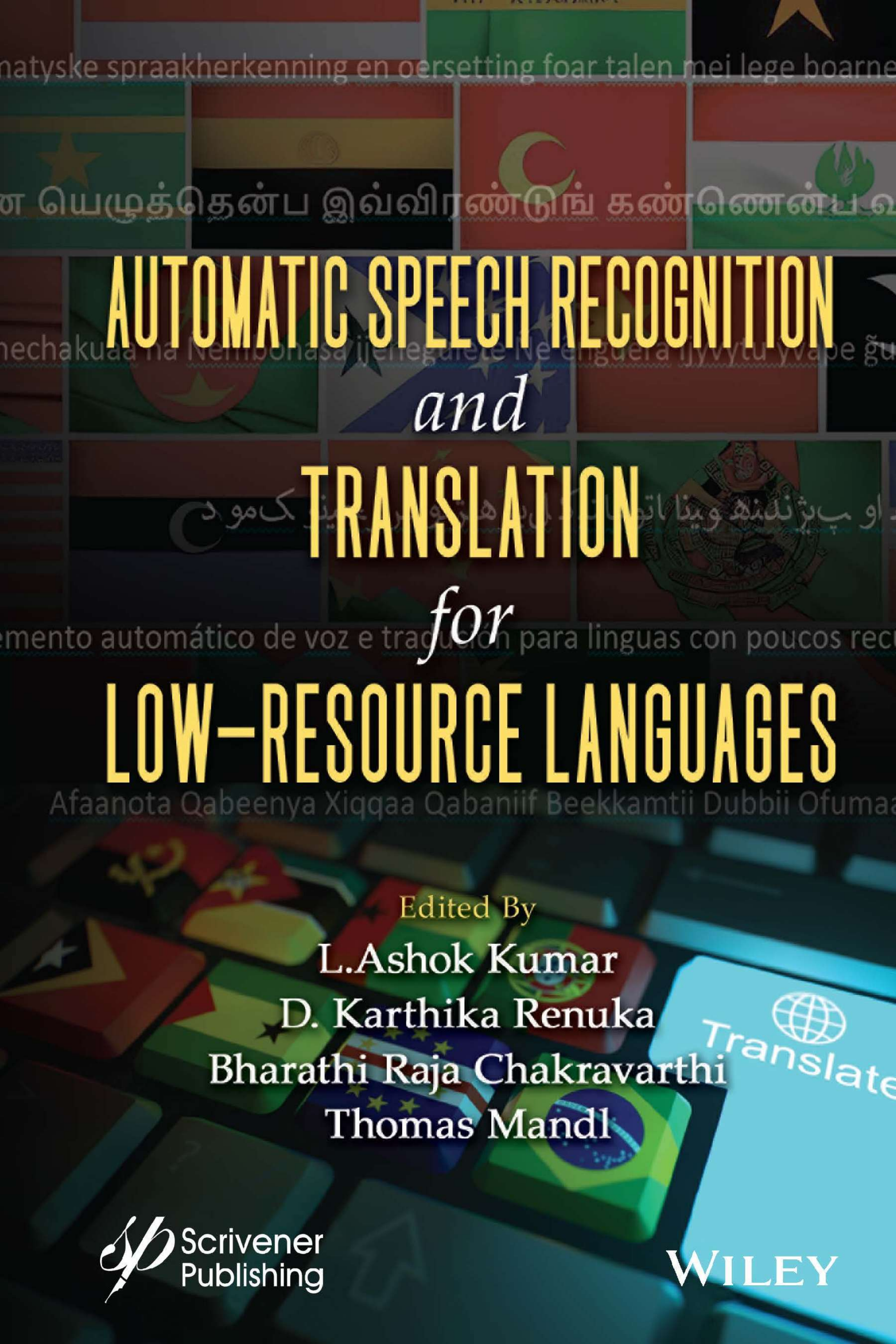
matyske spraakherkenning en oersetting foar talen mei lege boarne
ன யெழுத்தென்ப இவ்விரண்டுங் கண்ணென்ப வ
nechakuaa ha Ñemboñasa ijeheguete Ñe'ẽnguéra ijyvytu yvape gu
emento automático de voz e tradución para linguas con poucos recu
Afaanota Qabeenya Xiqqaa Qabaniif Beekkamtii Dubbii Ofumaa
Translate

AUTOMATIC SPEECH RECOGNITION *and* TRANSLATION *for* LOW-RESOURCE LANGUAGES

Edited By

L.Ashok Kumar
D. Karthika Renuka
Bharathi Raja Chakravarthi
Thomas Mandl

Scrivener Publishing

WILEY

Automatic Speech Recognition and Translation for Low Resource Languages

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Automatic Speech Recognition and Translation for Low Resource Languages

Edited by

L. Ashok Kumar

PSG College of Technology, Coimbatore, India

D. Karthika Renuka

PSG College of Technology, Coimbatore, India

Bharathi Raja Chakravarthi

School of Computer Science, University of Galway, Ireland

and

Thomas Mandl

Institute for Information Science and Language Technology,
University of Hildesheim, Germany

Scrivener
Publishing

WILEY

This edition first published 2024 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-394-21358-0

Cover image: Pixabay.Com
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Dedication

To my wife, Ms. Y. Uma Maheswari, and daughter, A. K. Sangamithra, for their constant support and love.

Dr. L. Ashok Kumar

To my family and friends who have been highly credible and a great source of inspiration and motivation.

Dr. D. Karthika Renuka

Dr. Bharathi Raja Chakravarthi would like to thank his students.

Dr. Bharathi Raja Chakravarthi

Contents

Foreword

Recent advancements in Automatic Speech Recognition (ASR) and Machine Translation (MT) technologies have brought about a new era of hope and possibility for these low-resource languages. The convergence of cutting-edge research, powerful algorithms, and increased computational capacity has paved the way for groundbreaking applications that can revolutionize linguistic accessibility and inclusion.

This book stands as a testament to the transformative potential of ASR and MT technologies for marginalized languages. It brings together a diverse group of experts, researchers, and practitioners who have dedicated their efforts to addressing the unique challenges faced by low-resource languages and finding ways to overcome them with ASR and MT.

The chapters herein explore a wide range of topics related to ASR and MT for low-resource languages. The book delves into the theoretical foundations of ASR and MT, providing readers with a comprehensive understanding of the underlying principles and methodologies. It examines the technical intricacies and practical considerations of developing ASR and MT systems that are specifically tailored to low-resource languages, taking into account the scarcity of data and linguistic resources.

Moreover, this book sheds light on the potential applications of ASR and MT technologies beyond mere transcription and translation. It explores how these technologies can be harnessed to preserve endangered languages, facilitate cross-cultural communication, enhance educational resources, and empower marginalized communities. By offering real-world case studies, success stories, and lessons learned, the contributors provide invaluable insights into the impact of ASR and MT on low-resource languages and the people who speak them.

As you embark on this enlightening journey through the pages of this book, you will discover the tremendous potential of ASR and MT technologies to bridge the digital divide and empower low-resource languages. You will witness the strides made in linguistic accessibility and cultural preservation, and you will gain a deeper appreciation for the profound impact these technologies can have on societies, both large and small.

I extend my heartfelt appreciation to the editors and authors who have contributed their expertise, dedication, and passion to this volume. Their collective wisdom and tireless efforts have given rise to a comprehensive resource that will undoubtedly serve as a guiding light for researchers, practitioners, and policymakers committed to advancing the cause of linguistic diversity and inclusivity.

Together, let us embrace the power of ASR and MT technologies as instruments of empowerment and change. Let us work collaboratively to ensure that no language, no matter how small or remote, is left behind in the digital era. Through our collective endeavors, we can unleash the full potential of low-resource languages, fostering a world where linguistic diversity thrives, cultures flourish, and global understanding is truly within reach.

Sheng-Lung Peng
Dean, College of Innovative Design and Management, National Taipei University of Business, Creative Technologies and Product Design, Taiwan

Preface

In today's interconnected world, effective communication across different languages is vital for fostering understanding, collaboration, and progress. However, language barriers pose significant challenges, particularly for languages that lack extensive linguistic resources and technological advancements. In this context, the field of Automatic Speech Recognition (ASR) and translation assumes paramount importance.

ASR and Translation for Low Resource Languages is a comprehensive exploration into the cutting-edge research, methodologies, and advancements in addressing the unique challenges associated with ASR and translation for low-resource languages. This book sheds light on the innovative approaches and techniques developed by researchers and practitioners to overcome the limitations imposed by scarce linguistic resources and data availability.

To start, the book delves into the fundamental concepts of ASR and translation, providing readers with a solid foundation for understanding the subsequent chapters. Then in explores the intricacies of low-resource languages, analyzing the factors that contribute to their challenges and the significance of developing tailored solutions to overcome them.

The material contained herein encompasses a wide range of topics, ranging from both the theoretical and practical aspects of ASR and translation for low-resource languages. The book discusses data augmentation techniques, transfer learning, and multilingual training approaches that leverage the power of existing linguistic resources to improve accuracy and performance. Additionally, it investigates the possibilities offered by unsupervised and semi-supervised learning, as well as the benefits of active learning and crowdsourcing in enriching the training data.

Throughout the book, emphasis is placed on the importance of considering the cultural and linguistic context of low-resource languages, recognizing the unique nuances and intricacies that influence accurate ASR and translation. Furthermore, we explore the potential impact of these technologies in various domains, such as healthcare, education, and commerce, empowering individuals and communities by breaking down language barriers.

The editors of this book brought together experts, researchers, and enthusiasts from diverse fields to share their knowledge, experiences, and insights in ASR and translation for low-resource languages. We hope that this collaborative effort will contribute to the development of robust and efficient solutions, ultimately fostering inclusive communication and bridging the language divide. We invite readers to embark on this journey of discovery and innovation, gaining a deeper understanding of the challenges, opportunities, and breakthroughs in ASR and translation for low-resource languages. Together, let us pave the way towards a world where language is no longer a barrier, but a bridge that connects individuals, cultures, and ideas.

Dr. L. Ashok Kumar
Professor, PSG College of Technology, India
Dr. D. Karthika Renuka
Professor, PSG College of Technology, India
Dr. Bharathi Raja Chakravarthi
Assistant Professor/Lecturer above-the-Bar School of Computer Science,
University of Galway, Ireland
Dr. Thomas Mandl
Professor, Institute for Information Science and Language Technology,
University of Hildesheim, Germany

Acknowledgement

We bow our head before "The God Almighty" who blessed us with health and confidence to undertake and complete the book successfully. We express our sincere thanks to the Principal and Management of PSG College of Technology, University of Galway, Ireland, University of Hildesheim, Germany for their constant encouragement and support.

We thank our family and friends who always stood beside us and encouraged us to complete the book.

Dr. L. Ashok Kumar is thankful to his wife, Y. Uma Maheswari, for her constant support during writing. He is also grateful to his daughter, A. K. Sangamithra, for her support; it helped him a lot in completing this work.

Dr. D. Karthika Renuka would like to express gratitude to her parents, for their constant support. Her heartfelt thanks to her husband, Mr. R. Sathish Kumar, and her dear daughter, Ms. P. S. Preethi, for their unconditional love which made her capable of achieving all her goals.

Dr. Bharathi Raja Chakravarthi would like to thank his students.

Dr. Thomas Mandl would like to thank his parents and family as well as all his academic colleagues for their inspiration during cooperation.

We would like to acknowledge the help of all the people involved in this project. First, we would like to thank each one of the authors for their contributions. Our sincere gratitude goes to the chapter's authors who contributed their time and expertise to this book. We thank all the authors of the chapters for their commitment to this endeavor and their timely response to our incessant requests for revisions.

Second, the editors wish to acknowledge the valuable contributions of the reviewers regarding the improvement of quality, coherence, and content presentation of chapters. Next, the editors would like to recognize the contributions of editorial board in shaping the nature of the chapters in this book. In addition, we wish to thank the editorial staff at Wiley-Scrivener book for their professional assistance and patience. Sincere thanks to each one of them.

Dr. L. Ashok Kumar
Professor, PSG College of Technology, India
Dr. D. Karthika Renuka
Professor, PSG College of Technology, India
Dr. Bharathi Raja Chakravarthi
Assistant Professor/Lecturer above-The-Bar School of Computer Science, University of Galway, Ireland
Dr. Thomas Mandl
Professor, Institute for Information Science and Language Technology, University of Hildesheim, Germany

1

A Hybrid Deep Learning Model for Emotion Conversion in Tamil Language

Satrughan Kumar Singh[1]*, Muniyan Sundararajan[2] and Jainath Yadav[1]

[1]Department of Computer Science, Central University of South Bihar, Gaya, Bihar, India

[2]Department of Mathematics and Computer Science, Mizoram University, Aizawl, Mizoram, India

Abstract

In speech signal processing, emotion recognition is a challenging task in classifying speech into different emotions. In this chapter, we propose a hybrid model based on FFNN (feed forward neural network) and SVM (support vector machine) for automated emotion conversion in the Tamil language. The use of voice command indeed contributes to a better integrated human-machine interface integration where one can give voice command, which intelligent machine understands and obeys. The Tamil language is mostly syllabic for the synthetical analysis of speech signal recognition. The changes in speech signal processing are mainly observed in several acoustic parameters such as root mean square energy, short-time energy, mel-frequency cepstral coefficient, and zero crossing rate, which are subsequently used for discrimination of the generation of a new set of the feature vector. In this proposed model, firstly, the FFNN model is complemented on the training and test datasets. Thereafter, SVM is used to perform the classification task. In the proposed emotion transformation, emotions such as angry, happy, sad, calm, surprised, fearful, neutral, and disgust are considered as target emotions with the multi-layered signal processing framework. This framework is required for spectral mapping to convert neutral utterance into target emotional utterance that is evaluated by subjective tests. Finally, both subjective and objective tests reveal a high and increased accuracy with the proposed model for spectral mapping and also show that the proposed model is better than Gaussian mixture model (GMM),

**Corresponding author*: satrughanksingh@yahoo.com

L. Ashok Kumar, D. Karthika Renuka, Bharathi Raja Chakravarthi and Thomas Mandl (eds.) Automatic Speech Recognition and Translation for Low Resource Languages, (1–14) © 2024 Scrivener Publishing LLC

FFNN and some pre-trained convolutional neural network (CNN) architectural models.

Keywords**:** Spectral mapping, emotion conversion, GMM, FFNN, pre-trained CNN, FFNN+SVM, objective measure

1.1 Introduction

Speech signal processing for emotion conversion has been a recent emerging domain in the human-machine interface. Presently, people are constantly trying to make computers intelligent so that they can do almost all the work easily like humans [1]. The communication between human and computer occurs in both directions [2]. This communication should have two important features of speech technology, speech recognition and speech synthesis. It is known that humans use emotions frequently to convey the intended message. Therefore, it is expected that the machine should be able to understand and generate desired emotions [3, 26]. Most of the existing speech systems can generate only neutral style speech. In this situation, the transformation of emotion is applied to convert the neutral style speech to desired expressive style speech. The modules of emotion transformation are used for making speaking instruments for disabled people and telling the stories in an automatic way [5, 24]. Generation of emotional speech is a challenging research problem. Some research works have attempted to generate expressive speech using text-to-speech synthesis (TTS) technique. Researchers have used the following methods for expressive speech synthesis: (i) formant synthesis or rule-based synthesis, (ii) di-phone concatenation synthesis, (iii) unit selection synthesis, and (iv) Hidden Markov Model (HMM)-based parametric speech synthesis. Emotion transformation approach differs from expressive speech synthesizers because it takes input as neutral speech, while the input of expressive speech synthesizer is text. It can be used with any speech synthesizers to convert their neutral speech output to the desired emotional speech. It generates emotional speech by creating emotional parameters into neutral speech [6, 7]. A formant vocoder is used to synthesize the speech transformation, showing the contour mapping of the target emotion through neural network [25]. For synthesizing emotional speech, the most important issue is to identify features which carry the emotion-specific information. Among various speech features, the widely used features for discrimination of emotions are prosodic and spectral features. The existing emotion transformation techniques transform neutral to emotional speech using prosody manipulation [8–10].

In this chapter, we have generated emotional speech by mapping of spectral features from neutral to target emotions in Tamil language.

1.2 Dataset Collection and Database Preparation

Spectral feature mapping framework needs parallel utterances of source and target emotions to perform emotion transformation process. Around 30 to 50 parallel utterances are sufficient to build emotion-specific mapping functions [11, 12]. In this work, we selected 100 parallel utterances from the emotional speech database collected from one male and one female speaker in Tamil language. These utterances were recorded in eight emotions such as angry, happy, sad, calm, surprised, fearful, neutral, and disgust. Training facilitates a learning system for creating an acoustic training dataset [4]. For training and testing purposes, we used 70 and 30 parallel utterances, respectively.

1.3 Pre-Trained CNN Architectural Models

1.3.1 VGG16

VGG16 is a convolutional neural network (CNN) model which basically focuses on depth. VGG takes 224 x 224 pixel RGB image. It uses a small receptive field (3 x 3 with stride of 1) followed by a ReLu unit. VGG16 has three fully connected layers; the first two have 4096 channels and the third has 1000 channels, one for each class. All of VGG16's hidden layers use ReLu. VGG has many variants, among which is VGG16, which is famous as its name is derived from its architecture using 16 layers in total among 13 convolution layers, two fully connected layers, and one output layer.

1.3.2 ResNet50

ResNet50 is known as residual network. ResNet works on skip connection. As it is known, deep networks always suffer from vanishing gradients without adjustments. Tiny gradients make learning intractable. To overcome this problem, Microsoft introduced a deep residual learning framework. The skip connection provides the learning network to identity function for passing the input through the block without passing through the other weight layers and allowing the network to traverse through its layers without gaps.

1.4 Proposed Method for Emotion Transformation

In this chapter, feed forward neural network (FFNN) was explored for emotion transformation. In the literature, Gaussian mixture model (GMM) was used for mapping features from one domain to others. However, the weakness of GMM is that it uses the assumption that the shape of mapping function is Gaussian. In addition to it, GMM requires to fix the number of mixtures before the mapping process. These weaknesses motivated us to explore FFNN to develop an emotion transformation system. Normally, it contains two hidden layers for capturing global and local information between input and output parameters [13–15]. Any continuous valued function can be simulated by considering two or more hidden layers in the neural network [16]. Hence, two hidden layers are sufficient for developing mapping functions. We considered three hidden layers in place of two hidden layers to take the additional benefit of symmetric structure. The symmetric structure is useful to map input parameter to output parameter [16–20]. The FFNN is depicted in Figure 1.1. The third hidden layer of FFNN compresses the dimension of input parameters. It captures global information while other hidden layers capture local information required for developing mapping functions. The accurate mapping functions are developed by selecting an appropriate structure of FFNN. The mapping function F(t) can be expressed as following:

$$F(t) = g\left(W^4 h\left(W^3 h\left(W^2 h\left(W^1 g(t)\right)\right)\right)\right) \tag{1.1}$$

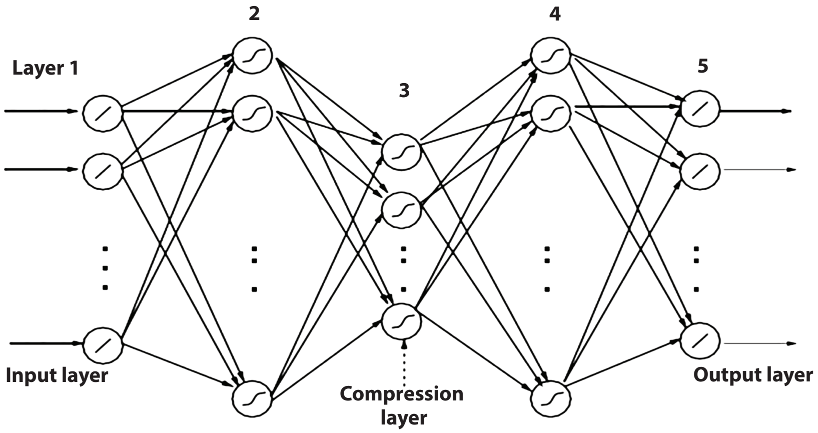

Figure 1.1 The diagram 5 layers feed forward neural network model.