

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



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

© 2024 Scrivener Publishing LLC

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

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

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	xix
Preface	xxi
Acknowledgement	xxiii
1 A Hybrid Deep Learning Model for Emotion Conversion in Tamil Language	1
<i>Satrughan Kumar Singh, Muniyan Sundararajan and Jainath Yadav</i>	
1.1 Introduction	2
1.2 Dataset Collection and Database Preparation	3
1.3 Pre-Trained CNN Architectural Models	3
1.3.1 VGG16	3
1.3.2 ResNet50	3
1.4 Proposed Method for Emotion Transformation	4
1.4.1 Architecture of Emotion Conversion Framework	5
1.4.2 Feature Alignment	6
1.4.3 Mapping of Excitation Source Parameter	6
1.5 Synthesized Speech Evaluation	7
1.5.1 Objective Evaluation	7
1.5.2 Subjective Evaluation	8
1.6 Conclusion	13
References	13
2 Attention-Based End-to-End Automatic Speech Recognition System for Vulnerable Individuals in Tamil	15
<i>S. Suhasini, B. Bharathi and Bharathi Raja Chakravarthi</i>	
2.1 Introduction	16
2.2 Related Work	17
2.3 Dataset Description	18
2.4 Implementation	19
2.4.1 ASR—Transformer Based	19

2.4.2	ASR—Attention Based	20
2.4.3	Steps Involved to Develop End-to-End Speech Recognizer Using SpeechBrain	23
2.4.3.1	Data Preparation	23
2.4.3.2	Tokenizer Training	23
2.4.3.3	Language Model Training	23
2.4.3.4	Speech Recognizer Training with Attention-Based Mechanism	23
2.4.3.5	ASR Model Built	23
2.5	Results and Discussion	24
2.6	Conclusion	24
	References	24
3	Speech-Based Dialect Identification for Tamil	27
	<i>Archana J.P. and B. Bharathi</i>	
3.1	Introduction	28
3.2	Literature Survey	29
3.3	Proposed Methodology	30
3.3.1	Architecture Diagram	30
3.3.2	Dataset Creation	30
3.3.3	Feature Extraction	31
3.3.3.1	Mel-Frequency Cepstral Coefficients (MFCC)	31
3.3.3.2	Steps for MFCC Extraction	33
3.3.4	Classification	33
3.3.4.1	Gaussian Mixture Model	33
3.4	Experimental Setup and Results	35
3.5	Conclusion	37
	References	37
4	Language Identification Using Speech Denoising Techniques: A Review	41
	<i>Amal Kumar, Piyush Kumar Singh and Jainath Yadav</i>	
4.1	Introduction	42
4.1.1	Background and Context	42
4.1.2	Related Works and Motivation	43
4.2	Speech Denoising and Language Identification	44
4.2.1	Speech Denoising	44
4.2.2	Language Identification	45

4.3	The Noisy Speech Signal is Denoised Using Temporal and Spectral Processing	45
4.4	The Denoised Signal is Classified to Identify the Language Spoken Using Recent Machine Learning Algorithm	46
4.5	Conclusion	47
	References	48
5	Domain Adaptation-Based Self-Supervised ASR Models for Low-Resource Target Domain	51
	<i>L. Ashok Kumar, D. Karthika Renuka, Naveena K. S. and Sree Resmi S.</i>	
5.1	Introduction	52
5.2	Literature Survey	52
5.3	Dataset Description	54
	5.3.1 LibriSpeech	54
	5.3.2 NPTEL Sample Dataset	55
	5.3.3 Real Time Dataset	55
5.4	Self-Supervised ASR Model	56
	5.4.1 wav2vec 2.0 Model	56
5.5	Domain Adaptation for Low-Resource Target Domain	58
	5.5.1 Data Augmentation	59
	5.5.2 Data Preprocessing	60
	5.5.3 Freeze Feature Encoder	60
	5.5.4 Fine Tuning	60
5.6	Implementation of Domain Adaptation on wav2vec2 Model for Low-Resource Target Domain	61
	5.6.1 Dataset Preparation	61
	5.6.2 Preprocessing and Tokenizing	62
	5.6.3 Freeze Feature Encoder	62
	5.6.4 Fine Tuning	63
5.7	Results Analysis	63
	5.7.1 E-Content	63
	5.7.2 NPTEL	64
	5.7.3 E-Content with NPTEL	65
5.8	Conclusion	66
	Acknowledgements	66
	References	66

6	ASR Models from Conventional Statistical Models to Transformers and Transfer Learning	69
	<i>Elizabeth Sherly, Leena G. Pillai and Kavya Manohar</i>	
6.1	Introduction	70
6.2	Preprocessing	71
6.3	Feature Extraction	74
	6.3.1 Mel-Frequency Cepstral Coefficients (MFCCs)	75
	6.3.2 Linear Predictive Coding (LPC)	79
	6.3.3 Perceptual Linear Prediction (PLP)	80
6.4	Generative Models for ASR	81
	6.4.1 Components in Statistical Generative ASR	82
	6.4.2 GMM-HMM Acoustic Models	83
	6.4.3 N-Gram Language Model	87
	6.4.4 Pronunciation Lexicon	88
	6.4.5 Decoder	88
6.5	Discriminative Models for ASR	89
	6.5.1 Support Vector Machine (SVM)	90
	6.5.2 Deep Learning in ASR	96
6.6	Deep Architectures for Low-Resource Languages	97
	6.6.1 Seq2seq Deep Neural Network	97
	6.6.1.1 Attention Mechanism	98
	6.6.2 Transformer Model with Attention	100
	6.6.3 Transfer Learning System	103
6.7	The DNN-HMM Hybrid System	104
6.8	Summary	107
	References	107
7	Syllable-Level Morphological Segmentation of Kannada and Tulu Words	113
	<i>Asha Hegde and Hosahalli Lakshmaiah Shashirekha</i>	
7.1	Introduction	114
	7.1.1 Challenges in MS of Kannada and Tulu Words	117
	7.1.2 Motivation	119
7.2	Related Work	121
7.3	Corpus Construction and Annotation	124
7.4	Methodology	127
	7.4.1 Feature Extraction	127
	7.4.2 Model Construction	128
7.5	Experiments and Results	128
7.6	Conclusion and Future Work	130
	References	130

8	A New Robust Deep Learning-Based Automatic Speech Recognition and Machine Transition Model for Tamil and Gujarati	135
	<i>Monesh Kumar M. K., Valliammai V., Geraldine Bessie Amali D. and Mathew M. Noel</i>	
8.1	Introduction	136
8.2	Literature Survey	139
8.3	Proposed Architecture	145
	8.3.1 Speech Recognition	146
	8.3.2 Text Translation Model	147
	8.3.3 Text-to-Speech Conversion	148
8.4	Experimental Setup	150
8.5	Results	151
8.6	Conclusion	152
	References	153
9	Forensic Voice Comparison Approaches for Low-Resource Languages	155
	<i>Kruthika S.G., Trisiladevi C. Nagavi and P. Mahesha</i>	
9.1	Introduction	156
	9.1.1 Aim and Objectives of Forensic Voice Comparison for Low-Resource Languages	158
	9.1.2 Forensic Voice Comparison Scenario on Low-Resource Language Datasets	159
	9.1.3 Forensic Voice Comparison Methodology	159
9.2	Challenges of Forensic Voice Comparison	162
9.3	Motivation	164
9.4	Review on Forensic Voice Comparison Approaches	166
	9.4.1 Tools and Frameworks for Forensic Voice Comparison	177
9.5	Low-Resource Language Datasets	180
	9.5.1 Kannada Corpus	181
	9.5.2 Malayalam Corpus	181
	9.5.3 Tamil Corpus	182
	9.5.4 Telugu Corpus	182
	9.5.5 Low-Resource Language Dataset Research Lab Centers	183
9.6	Applications of Forensic Voice Comparison	184
9.7	Future Research Scope	185
9.8	Conclusion	186
	References	186

10 CoRePool—Corpus for Resource-Poor Languages: Badaga Speech Corpus	193
<i>Barathi Ganesh H.B., Jyothish Lal G., Jairam R., Soman K.P., Kamal N.S. and Sharmila B.</i>	
10.1 Introduction	194
10.1.1 Motivation and Related Literature	195
10.1.1.1 Corpus	196
10.1.1.2 Foundation Models	198
10.2 CoRePool	200
10.3 Benchmarking	201
10.3.1 Speech-to-Text	201
10.3.2 Text-to-Speech	202
10.3.3 Gender Identification	203
10.3.4 Speaker Identification	204
10.3.5 Translation	204
10.4 Conclusion	206
Acknowledgment	206
References	207
11 Bridging the Linguistic Gap: A Deep Learning-Based Image- to-Text Converter for Ancient Tamil with Web Interface	213
<i>S. Umamaheswari, G. Gowtham and K. Harikumar</i>	
11.1 Introduction	214
11.2 The Historical Significance of Ancient Tamil Scripts	215
11.3 Realization Process	217
11.4 Dataset Preparation	218
11.4.1 Dataset Augmentation Process	219
11.4.2 Image Pre-Processing	220
11.4.3 Character Segmentation Process	222
11.4.3.1 Bounding Box Method	223
11.5 Convolution Neural Network	225
11.5.1 Convolution Layer	228
11.5.2 Pooling Layer	229
11.5.3 Fully Connected Layer	230
11.6 Webpage with Multilingual Translator	231
11.6.1 Why is the Web Interface Used Here?	231
11.6.2 Google Translator API	232
11.6.3 Build the Website for Deploying the CNN Model	232
11.7 Results and Discussions	235
11.8 Conclusion and Future Work	237
References	238

12 Voice Cloning for Low-Resource Languages: Investigating the Prospects for Tamil	243
<i>Vishnu Radhakrishnan, Aadharsh Aadhithya A., Jayanth Mohan, Visweswaran M., Jyothish Lal G. and Premjith B.</i>	
12.1 Introduction	244
12.2 Literature Review	246
12.3 Dataset	248
12.3.1 Tamil Mozilla Common Voice Corpus	248
12.4 Methodology	249
12.4.1 Architecture of the Method	250
12.5 Results and Discussion	251
12.6 Conclusion	254
References	255
13 Transformer-Based Multilingual Automatic Speech Recognition (ASR) Model for Dravidian Languages	259
<i>Divi Eswar Chowdary, Rahul Ganesan, Harsha Dabbara, G. Jyothish Lal and Premjith B.</i>	
13.1 Introduction	260
13.2 Literature Review	260
13.3 Dataset Description	264
13.4 Methodology	264
13.4.1 Data Preprocessing	265
13.4.2 Proposed Model Architecture	266
13.4.3 Word Error Rate	268
13.5 Experimentation Results and Analysis	268
13.6 Conclusion	272
References	272
14 Language Detection Based on Audio for Indian Languages	275
<i>Amogh A. M., A. Hari Priya, Thanvitha Sai Kanchumarti, Likhitha Ram Bommilla and Rajeshkannan Regunathan</i>	
14.1 Introduction	276
14.2 Literature Review	278
14.3 Language Detector System	282
14.3.1 Workflow	283
14.3.1.1 Data Collection	283
14.3.1.2 Data Preprocessing	283
14.3.1.3 Feature Extraction	284
14.3.1.4 Model Creation/Training	286

14.3.1.5	Evaluation and Analyzing Results	286
14.3.2	Overview of a Sample CNN Architecture	286
14.3.3	Proposed CNN Architecture	287
14.3.4	Pseudocode	288
14.3.4.1	Traversing Dataset	288
14.3.4.2	Methodology Applied on Sample Data	288
14.3.4.3	Function <code>feature_mfcc_extract</code> (audio file_name,lang)	289
14.3.4.4	Feature Extraction	289
14.3.4.5	Train Test Split	289
14.3.4.6	Model	290
14.4	Experiments and Outcomes	290
14.4.1	Dataset	290
14.4.2	Applied Methodology on Sample Data	291
14.4.3	Sample Output	292
14.4.3.1	Features Extracted	292
14.4.3.2	Model Fitting	293
14.4.3.3	Model Predicting	293
14.5	Conclusion	294
	References	294
15	Strategies for Corpus Development for Low-Resource Languages: Insights from Nepal	297
	<i>Bal Krishna Bal, Balaram Prasain, Rupak Raj Ghimire and Praveen Acharya</i>	
15.1	Low-Resource Languages and the Constraints	298
15.2	Language Resources Map for the Languages of Nepal	298
15.2.1	Language Situation of Nepal	299
15.2.2	The Constitutional Status of the Languages of Nepal	305
15.2.3	Summary of Linguistic Activities	305
15.2.4	Potential Linguistic Tasks for Low-Resourced	308
15.3	Unicode Inception and Advent in Nepal	310
15.4	Speech and Translation Initiatives	311
15.5	Corpus Development Efforts—Sharing Our Experiences	313
15.5.1	Speech	313
15.5.1.1	Speech Corpus	315
15.5.1.2	Phonetically Balanced Corpus	317
15.5.2	Machine Translation	318

15.5.2.1	Corpus	318
15.5.2.2	Corpus Used in Machine Translation Systems	320
15.5.2.3	Challenges and Opportunities	321
15.6	Constraints to Competitive Language Technology Research for Nepali and Nepal's Languages	323
15.6.1	Lack of Consolidation of Works	323
15.6.2	Lack of Funding Support From the Government for Language Technologies	323
15.6.3	Lack of Primary Datasets/Corpora on Language Technologies	324
15.6.4	Lack of Natural Language Processing Tools	324
15.7	Roadmap for the Future	324
15.7.1	Short Term	325
15.7.2	Medium Term and Long Term	326
15.8	Conclusion	327
	References	327
16	Deep Neural Machine Translation (DNMT): Hybrid Deep Learning Architecture-Based English-to-Indian Language Translation	331
	<i>Nivaashini M., Priyanka G. and Aarthi S.</i>	
16.1	Introduction	332
16.2	Literature Survey	335
16.2.1	Problem Statements	338
16.2.2	Novel Contribution of the Proposed Model	338
16.3	Background	339
16.3.1	Corpora Collection	339
16.3.2	Language Description	342
16.3.3	Neuronal Machine Translation (NMT)	344
16.3.4	Deep Learning Architecture	345
16.4	Proposed System	348
16.4.1	Corpora Used	350
16.4.2	Preprocessing	350
16.4.3	Feature Extraction Using DBN	355
16.4.4	Hybrid Deep Neural Machine Translation Model	356
16.5	Experimental Setup and Results Analysis	358
16.5.1	Building Baseline Models	359
16.5.2	Creating Back Translation Models	359
16.5.3	MT Model Parameters	359
16.5.4	Evaluation Metrics	359

16.5.5	Experimental Architecture	361
16.5.6	Automatic Evaluation	362
16.5.7	Manual Evaluation	364
16.5.8	Output Analysis	366
16.5.9	Comparison with Google Translate	368
16.6	Conclusion and Future Work	369
	References	369
17	Multiview Learning-Based Speech Recognition for Low-Resource Languages	375
	<i>Aditya Kumar and Jainath Yadav</i>	
17.1	Introduction	376
17.1.1	Automatic Speech Recognition	376
17.1.2	Issues Related to Low-Resource Languages in ASR	376
17.1.3	Multiview Learning	378
17.2	Approaches of Information Fusion in ASR	380
17.2.1	Acoustic with Lip Visual Fusion Approach for LRLs	381
17.2.2	Cross-Lingual Transfer Learning Approach for LRLs	382
17.2.3	Acoustic and Linguistic Encoder Fusion for LRLs	382
17.2.4	LRL Recognition Using a Combination of Different Features	384
17.2.5	Joint Decoding of Tandem and Combined Systems Approach for Low-Resource Setting	385
17.2.6	High-Resource Pre-Training Approach to Improve Low-Resource ASR	387
17.2.7	Multilingual Representations and Transfer Learning Approach for Low-Resource ASR	388
17.2.8	Multimodal Fusion Approach for LRL	390
17.2.9	Multimodal Fusion Approach for Machine Translation in Low-Resource Settings	391
17.2.10	Multilingual Acoustic Model Fusion Approach for Low-Resource Setting	393
17.3	Partition-Based Multiview Learning	394
17.3.1	Partitioning Methods for Audio Data	395
17.3.2	Ensemble Process of the Partitioned Audio Data	398
17.3.3	Feature Set Partition-Based Multiview Learning	399
17.4	Data Augmentation Techniques	399

17.5	Conclusion	401
	References	401
18	Automatic Speech Recognition Based on Improved Deep Learning	405
	<i>Kingston Pal Thamburaj and Karthees Ponniah</i>	
18.1	Introduction	406
18.2	Literature Review	408
18.3	Proposed Methodology	410
	18.3.1 Preprocessing Using LMS Adaptive Filter	411
	18.3.2 Feature Extraction	412
	18.3.3 Classification Using Improved Deep Learning (IDL)	416
	18.3.3.1 Recurrent Neural Network (RNN)	416
	18.3.3.2 Modified Social Spider Optimization Algorithm (MSSOA)	418
18.4	Results and Discussion	421
18.5	Conclusion	424
	References	424
19	Comprehensive Analysis of State-of-the-Art Approaches for Speaker Diarization	427
	<i>Trisiladevi C. Nagavi, Samanvitha S., Shreya Sudhanva, Sukirth Shivakumar and Vibha Hullur</i>	
19.1	Introduction	428
19.2	Generic Model of Speaker Diarization System	429
	19.2.1 Preprocessing	429
	19.2.2 Feature Extraction	429
	19.2.3 Segmentation	430
	19.2.4 Embedding	431
	19.2.5 Clustering	431
19.3	Review of Existing Speaker Diarization Techniques	432
	19.3.1 Datasets	432
	19.3.2 Review of Existing Speaker Diarization Systems	433
19.4	Challenges	438
	19.4.1 Dynamic Speaker Diarization	438
	19.4.2 Speaker Overlap	438
	19.4.3 Domain Mismatch	439
	19.4.4 When Speakers are Partially Known	439
	19.4.5 Audiovisual Modeling	439

19.5	Applications	439
19.5.1	Law Enforcement	440
19.5.1.1	Judicial System	440
19.5.1.2	Criminal Investigation	440
19.5.2	Education	440
19.5.3	Business	440
19.5.3.1	Marketing	440
19.5.3.2	Finance	441
19.6	Conclusion	441
	References	441
20	Spoken Language Translation in Low-Resource Language	445
	<i>S. Shoba, Sasithradevi A. and S. Deepa</i>	
20.1	Introduction	446
20.2	Related Work	448
20.2.1	Challenges	449
20.2.2	Overview of End-to-End Speech Translation System	450
20.3	MT Algorithms	451
20.3.1	Deep Learning MT System	451
20.3.2	GAN Network MT System	452
20.3.3	word2vec Predictive Deep Learning Model	452
20.3.4	Universal MT System	453
20.4	Dataset Collection	454
20.4.1	Parallel Data	454
20.4.2	Monolingual Data	455
20.4.3	Bilingual Data	455
20.4.4	Real Data	455
20.5	Conclusion	456
	References	456
	Index	461

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 it 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*

A Hybrid Deep Learning Model for Emotion Conversion in Tamil Language

Satrughan Kumar Singh^{1*}, Muniyan Sundararajan² and Jainath Yadav¹

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: satrughansingh@yahoo.com

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:

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

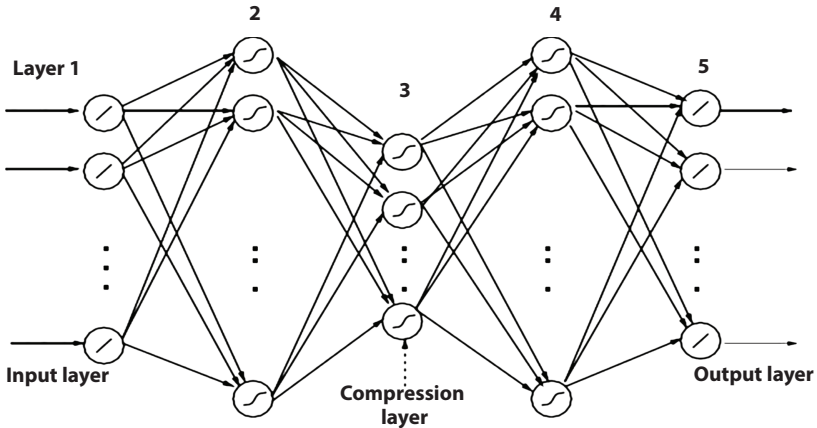


Figure 1.1 The diagram 5 layers feed forward neural network model.