SPEECH EMBEDDING WITH SEGREGATION OF PARALINGUISTIC INFORMATION FOR LOW-RESOURCE LANGUAGES

Anosha Ignatius

208054J

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science (Major Component of Research) in Computer Science and Engineering

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

November 2021

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis under my supervision.

Name of the supervisor: Dr. Uthayasanker Thayasivam

Signature of the supervisor: Date:

Abstract

Speech embeddings produced by Deep Neural Networks have yielded promising results in a variety of speech processing applications. However, the performance in speech tasks like automatic speech recognition and speech intent identification can be affected to a great extent when there is a discrepancy between training and testing conditions. This is because, in addition to linguistic information, speech signals carry para-linguistic information including speaker characteristics, emotional states, and accent. Variations in the speaker traits and states lead to compromise on performance in speech recognition applications that require only linguistic information.

Over the years, there have been various approaches that attempt to disentangle the para-linguistic information that support the linguistic information in speech. The commonly used strategy is to integrate speaker representations into speech recognition models to normalise the speaker effects. Still, it has received less attention when it comes to studies on speech-to-intent classification. Furthermore, large amounts of labeled speech data are required for these speaker normalisation techniques. Under low-resource settings, when there is only a limited number of speech samples available for training, transfer learning strategies can be used.

This study presents a speaker-invariant speech intent classification model using i-vector based feature augmentation. We investigate the use of pre-trained acoustic models for transfer-learning under low-resource settings. The proposed method is evaluated on the banking domain speech intent dataset in Sinhala and Tamil languages along with fluent speech command dataset. Experimental results show the effectiveness of the proposed method in achieving better prediction in the speech-to-intent classification model.

Keywords: speech-to-intent, speech recognition, linguistic, para-linguistic information, speaker representation

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Uthayasanker Thayasivam, for all the guidance and support extended by him. He has motivated me with his immense knowledge and abundant experience throughout my research.

I am very much grateful to Prof. Gihan Dias, who was ever willing to assist me in my research with his continuous guidance. I am also thankful to Dr. Surangika Ranathunga and Prof. Sanath Jeyasena for graciously providing me with feedback on my work.

I would like to extend my sincere thanks to my progress review panel members Dr. Charith Chitraranjan and Dr. C.R.J. Amalraj for their valuable feedback which helped me strengthen my work.

This research study was supported by Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education, Sri Lanka funded by the World Bank. I am thankful to LK Domain Registry for the financial support received to attend the SPECOM 2021 conference.

My heartfelt thanks go to my colleagues and friends who have helped me to complete my research successfully. Finally, I would like to thank my family members, who are always there to support me in all my endeavours.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
SLU	**S**poken **L**anguage **U**nderstanding
DNN	**D**eep **N**eural **N**etwork
LLD	**L**ow **L**evel **D**escriptor
MFCC	**M**el **F**requency **C**epstral **C**oefficients
LPC	**L**inear **P**redictive **C**odes
PLP	**P**erceptual **L**inear **P**rediction
GMM	**G**aussian **M**ixture **M**odel
UBM	**U**niversal **B**ackground **M**odel
RNN	**R**ecurrent **N**eural **N**etwork
APC	**A**utoregressive **P**redictive **C**oding
NPC	**N**on-autoregressive **P**redictive **C**oding