Combining Automatic Speech Recognition Models To Reduce Error Propagation in Low-Resource Transfer-Learning Speech-Command Recognition

Jazeem Mohamed Isham

(209333X)

Degree of Master of Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

March 2022

Combining Automatic Speech Recognition Models To Reduce Error Propagation in Low-Resource Transfer-Learning Speech-Command Recognition

Jazeem Mohamed Isham

(209333X)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree

Master of Science specializing in Data Science

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

March 2022

DECLARATION

I declare that this is my work and this dissertation does not incorporate without acknowledgment 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 acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic, or another 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 Master's dissertation under my supervision. I confirm that the declaration made above by the student is true and correct

Name of the supervisor: Dr. Uthayasanker Thayasivam

Signature of the supervisor: Date:

ABSTRACT

There are several applications when comes to spoken language understanding such as topic modeling and intent detection. One of the primary underlying components used in spoken language understanding studies is automatic speech-recognition models. In recent years we have seen a major improvement in the automatic speech recognition system to recognize spoken utterances. But it is still a challenging task for low-resource languages as it requires hundreds of hours of audio input to train an automatic speech recognition model.

To overcome this issue recent studies have used transfer learning techniques. However, the errors produced by the automatic speech recognition models significantly affect the downstream natural language understanding models used for intent or topic identification. In this work, we have proposed a multi-automatic speech recognition set up to overcome this issue. We have shown that combining outputs from multiple automatic speech recognition models can significantly increase the accuracy of low-resource speech-command transfer-learning tasks than using the output from a single automatic speech recognition model.

We have come up with convolution neural network-based setups that can utilize outputs from pre-trained automatic speech recognition models such as DeepSpeech2 and Wav2Vec 2.0. The experiment result shows a 7% increase in accuracy over the current state-of-the-art low resource speech-command phoneme-based speech intent classification methodology.

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Dr. Uthayasanker Thayasivam for being my supervisor and guiding me throughout the project. His expertise in this related area helped me a lot in setting the right direction for this project and obtaining the needed datasets.

I would not have achieved this without the immense support from my family. I would like to thank my parents for supporting me throughout my studies from the start the up until today. A special thanks to my wife and family for coping with me throughout my research program and encouraging me to complete it.

I also want to thank the University of Moratuwa for giving me an opportunity to participate in the MSc program and for providing the necessary resources to complete this research. This would not have been easy without the support from my workplace as well. I would like to thank Sysco LABS Srilanka for allowing me to do this part-time MSc and research while working with them. Also, my colleagues have been very understanding and allowed me to spend time on my research even during busy office schedules.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviation	Description
CNN	Convolution Neural Network
ASR	Automatic Speech Recognition
NLU	Natural Language Understanding
SLU	Spoken Language Understanding
ML	Machine Learning
DS2	DeepSpeech2 Model
W2V	Wav2Vec 2.0 Model
Exp	Experiment
WER	Word Error Rate
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
CTC	Connectionist Temporal Classification
SMBO	Sequential Model-based Optimization