

LB/TH/41/2025
TH6002

**LOW RESOURCE SPEECH INTENT
CLASSIFICATION USING MFCC FEATURES**

Anas Fathima Rifaza

219393M

Master of Science in Computer Science

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

LOW RESOURCE SPEECH INTENT CLASSIFICATION USING MFCC FEATURES

Anas Fathima Rifaza

219393M

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

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

DECLARATION

I hereby declare that this thesis is the result of my own independent work. It does not include any content that has been previously submitted for a degree or diploma at any university or institute of higher education, unless properly cited. To the best of my knowledge, all materials taken from the work of others have been appropriately acknowledged and referenced within the text. I also reserve the right to reuse parts of this work in future publications, such as journal articles or academic books.

Signature:

Date:

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: [Dr.T.Uthayasanker](#)

Signature of the Supervisor:

Date: [28 Jul 2025](#)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my MSc Research Project supervisor, Dr. Uthayasanker Thayasivam, for his invaluable guidance, continuous support, and encouragement throughout my research journey. His expertise and insights have been instrumental in shaping this work, and his unwavering support in providing the necessary resources has greatly contributed to the successful completion of my MSc thesis.

I am also deeply grateful for his constructive feedback, mentorship, and valuable suggestions, which have significantly enhanced the quality of this research. Additionally, I extend my heartfelt appreciation to my colleagues for their assistance in exploring relevant research materials and fostering a collaborative learning environment.

Furthermore, I am immensely thankful to my family—my parents, siblings, nephew, niece, and close friends—for their unwavering encouragement and support throughout this journey. Their belief in me has been a constant source of motivation.

Finally, I would like to extend my appreciation to everyone who has contributed to this endeavor, whether directly or indirectly. Their support has been invaluable in helping me navigate the challenges of my MSc studies

ABSTRACT

Speech-based user interfaces have revolutionized digital interactions, yet developing them for low-resource languages remains a challenge due to limited labeled speech data. This research proposes a Convolutional Neural Network (CNN)-based approach utilizing Mel-Frequency Cepstral Coefficients (MFCC) along with delta and delta-delta features for effective speech intent classification in Sinhala and Tamil. The methodology incorporates audio preprocessing, MFCC feature extraction, and data augmentation techniques such as noise addition, pitch shifting, and time stretching. A stratified cross-validation framework is used to ensure fair and consistent evaluation. The proposed model achieves 96.92% accuracy on the Sinhala dataset (7,624 samples) and 93.81% on the Tamil dataset (400 samples, ~0.5 hours of speech), representing a substantial improvement over prior methods. These results demonstrate the effectiveness of the CNN-based approach in capturing meaningful acoustic patterns for intent recognition in low-resource settings. The study offers a scalable, efficient solution for speech intent classification and contributes to the advancement of inclusive voice-enabled technologies.

Keywords: Speech Intent Classification, Low-Resource Languages, Pre trained Models, Convolutional Neural Network (CNN), Transfer Learning, Mel-Frequency Cepstral Coefficients (MFCC).

TABLE OF CONTENTS

Declaration	i
Acknowledgement.....	ii
Abstract	iii
Table of Contents	iv
List of Figures	vii
List of Tables.....	viii
List of Abbreviations.....	ix
Chapter 1	1
Introduction.....	1
1.1 Introduction Overview	1
1.2 Introduction	1
1.2.1 Significance of Intent Classification	1
1.2.2 Challenges in Low-Resource Settings	2
1.2.3 Proposed Solution	4
1.3 Background	5
1.4 Problem Statement	6
1.5 Objectives	6
1.6 Contributions	7
1.7 Summary	7
Chapter 2	8
LITERATURE REVIEW.....	8
2.1 Literature Overview	8
2.2 Speech Command Classification.....	8
2.2.1 Cascading ASR-NLU Models.....	8
2.2.2 Direct Speech Classification Models	9
2.3 Transfer Learning in Low-Resource Languages	10
2.3.1 Transfer Learning in ASR.....	10
2.3.2 Applications in Low-Resource Languages	11

2.4	Benchmarking in Low-Resource Speech Recognition.....	12
2.5	Gaps and Limitations.....	12
2.6	Summary	13
Chapter 3	14
Methodology	14
3.1	Methodology Overview.....	14
3.2	Proposed MFCC-CNN Architecture and Enhancements	16
3.2.1	Audio Data Collection.....	17
3.2.2	Preprocessing	17
3.2.3	Feature Extraction & Caching.....	18
3.2.4	Data Augmentation	18
3.2.5	Model Architecture and Model Training	19
3.2.6	Evaluation & Visualization.....	23
3.3	Summary	23
Chapter 4	24
EXPERIMENT	24
4.1	Data Set	24
4.2	Preprocessing.....	24
4.3	Model Implementation and Training.....	25
4.3	Hyper parameter Tuning	26
4.3	Experiment	27
4.3.1	Feature Extraction Analysis	27
4.3.2	Data Augmentation experiment	28
4.3.2	CNN architecture analysis.....	29
4.3.4	Wav2Vec2 analysis.....	29
4.4	Error Handling & Robustness	30
4.4	Scalability & Performance Optimization	30
4.5	Hardware & Computational Resources	31
Chapter 5	32
RESULT AND ANALYSIS	32
5.1	MFCC Feature analysis result for proposed methodology	32

5.2	Data Augmentation Analysis Result for Proposed Methodology.....	34
5.3	Analysis of Conv1D and Conv2D	37
5.4	Confusion matrix analysis	39
5.5	ROC Curve and AUC Evaluation for Sinhala Data – Highest Accuracy Fold	42
5.6	Loss and Accuracy Analysis for Sinhala data - Highest Accuracy Fold.	44
5.7	Classification Report analysis.....	45
5.8	Comparative Analysis of Tamil Data Performance Metrics across Folds	49
5.12	Analysis of Wav2Vec2 Performance on Tamil and Sinhala Datasets .	52
5.13	Comparison with Benchmark methodology performance	53
5.12	Summary	54
Chapter 6	55
DISCUSSION	55
6.1	Proposed MFCC Feature Result Discussion	55
6.2	Proposed Data Augmentation Result Discussion	55
6.3	Effectiveness of CNNs for Sequential Feature Classification.....	56
6.4	Performance Differences between 1D and 2D CNNs	56
Chapter 7	58
CONCLUSION	58
7.1	Conclusion.....	58
7.2	Contributions	59
7.3	Limitations.....	59
7.4	Future Work	60
7.5.1	Summary	60
References	62

LIST OF FIGURES

Figure	Description	Page
Figure 3.1:	General Workflow for Speech Intent Classification in Previous Studies	15
Figure 3.2:	Pipeline Architecture of the Proposed System, Illustrating Five Key Stages: Preprocessing, Feature Extraction, Data Augmentation, Model Training, and Evaluation.	17
Figure 3.3:	Proposed Speech Intent Classification Model Architecture	22
Figure 5.1 :	Accuracy trends across 5 folds for Tamil speech intent classification using MFCC vs MFCC + Delta + Delta-Delta features.	33
Figure 5.2 :	Key performance metrics for MFCC-only and MFCC + delta + delta-delta feature configurations (Tamil dataset).	33
Figure 5.3:	Accuracy trends across 5 folds for Sinhala speech intent classification using MFCC vs MFCC + Delta + Delta-Delta features.....	34
Figure 5.4 :	Tamil Data Accuracy Comparison: with vs without data augmentation	35
Figure 5.5 :	Sinhala Data Accuracy Comparison: with vs without data augmentation	37
Figure 5.6 :	Mean Accuracy comparison for Sinhala and Tamil Data : 1D CNN v2D CNN	38
Figure 5.7:	Confusion Matrix for Max Accuracy Fold (fold 4) - Sinhala Test Data shows majority of the samples are classified properly.....	40
Figure 5.8:	Confusion Matrix for Max Accuracy Fold (fold 5) - Tamil Test Data shows majority of the samples are classified properly.....	41
Figure 5. 9 :	AUC curve for fold 4 - Sinhala Data.....	42
Figure 5. 10 :	PR Curve for fold 4 - Sinhala Data.....	43
Figure 5.11 :	Train vs Validation Loss for fold 4 - Sinhala Data:	44
Figure 5.12 :	Train vs Validation accuracy for fold 4 - Sinhala Data.....	45
Figure 5. 13 :	Accuracy comparison across folds for Tamil Data	49
Figure 5.14 :	Correlation between Performance Metrics for Tamil Data	50
Figure 5.15:	Performance trends across folds for Tamil Data	51
Figure 5.16 :	Performance Trends Across Fold using Wav2Vec2.....	52

LIST OF TABLES

Table	Description	Page
Table 5.1:	Tamil Data Accuracy with vs without data augmentation	35
Table 5.2:	Sinhala Data Accuracy with vs without data augmentation	36
Table 5.3 :	Mean Accuracy comparison for Sinhala and Tamil Data: 1D CNN v2D CNN	38
Table 5.4:	Classification Report for max accuracy fold (Fold 5) - Tamil Data with highest accuracy 95.45%	47
Table 5.5:	Classification Report for Max Accuracy Fold (Fold 4) - Sinhala Data with highest accuracy 98%	48
Table 5.6:	Summary of results across different approaches with overall accuracy values. Gray shading indicates the accuracy of the previous benchmark methodology.	53

LIST OF ABBREVIATIONS

Abbreviation	Description
MFCC	Mel-Frequency Cepstral Coefficients
ASR	Automatic Speech Recognition
NLP	Natural Language Processing
HMMs	Hidden Markov Models
RNNs	Recurrent Neural Networks
NLU	Natural Language Understanding
OOV	Out-of-Vocabulary
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic