Project Title

Deciphering Silent Speech: A Multi-dimensional Neural Network Approach Using Conv3D, LSTM, and CTC for Lip-reading

Student Name

Zhenyi Li

MSc. Software Engineering 2024

TUS

##### Declaration

I hereby certify that the material, which is submitted in this thesis towards the award of MSc. Software Engineering, is entirely my own work and has not been submitted for any academic assessment other than part fulfilment of the above named award.

Future students may use the material contained in this thesis provided that the source is acknowledged in full.

Signed…………………………………………….

Date………………………………………………

##### Abstract

In this investigation, we examined the impact of image resolution on the performance of lip-reading models, with a focus on character error rate (CER) and word error rate (WER) as primary metrics. Two models were rigorously trained, one at a low resolution of 35x70 pixels and the other at a higher resolution of 70x140 pixels.The dataset we are using is one of the grid corpus datasets, which is a widely used dataset in the field of lip reading. The results were telling: the high-resolution model achieved a CER of 0.0008 and a WER of 0.0033, in contrast to the low-resolution model's CER of 0.3034 and WER of 0.3015. This indicates a significant performance leap with higher resolution inputs. However, the trade-off for precision was computational time, with the higher resolution training taking 4.319 hours, compared to 1.212 hours for the lower resolution. To translate these findings into practical insights, we have developed an online platform where users can gauge the model's accuracy in real-time scenarios. The direct comparison of CER and WER across resolutions highlights the delicate balance between model accuracy and computational efficiency in the field of automated lip-reading.

##### Acknowledgements

##### Table of Contents

[Declaration ii](#_Toc411341604)

[Abstract iii](#_Toc411341605)

[Acknowledgements iii](#_Toc411341606)

[Table of Contents iii](#_Toc411341607)

[List of Tables iii](#_Toc411341608)

[List of Figures iii](#_Toc411341609)

[Chapter 1: Introduction 3](#_Toc411341610)

[1.1 Introduction 3](#_Toc411341611)

[1.2 This uses ‘Heading 2’ style 3](#_Toc411341612)

[1.3 Research Aims and Objectives 3](#_Toc411341613)

[Chapter 2: Background Research 3](#_Toc411341614)

[2.1 Introduction 3](#_Toc411341615)

[2.2 Headings will vary based on research topic 3](#_Toc411341616)

[Chapter 3: System Design 3](#_Toc411341617)

[3.1 Introduction 3](#_Toc411341618)

[3.2 Requirements 3](#_Toc411341619)

[3.3 Architecture 3](#_Toc411341620)

[3.4 Design 3](#_Toc411341621)

[3.5 Implementation 3](#_Toc411341622)

[Chapter 4: Testing and Evaluation 3](#_Toc411341623)

[4.1 Introduction 3](#_Toc411341624)

[4.2 Testing 3](#_Toc411341625)

[4.3 Evaluation 3](#_Toc411341626)

[Chapter 5: Conclusions 3](#_Toc411341627)

[5.1 Introduction 3](#_Toc411341628)

[5.2 Reflection 3](#_Toc411341629)

[5.3 Recommendations 3](#_Toc411341630)

[References 3](#_Toc411341631)

[Glossary 3](#_Toc411341632)

[List of Abbreviations 3](#_Toc411341633)

[Appendix A: Appedix Title uses ‘Heading 6’ 3](#_Toc411341634)

[A.1 Appendix sub-title uses ‘Heading 7’ 3](#_Toc411341635)

##### List of Tables

[Table 1.3.1: When labeling a table use the ‘Table Label’ style. 1](#_Toc411344348)

##### List of Figures

[Figure 1.3.1: When labelling a figure use the ‘Figure Label’ style 1](#_Toc411344353)

[Figure 2.2.1: This is another figure heading 2](#_Toc411344354)

# Introduction

## Introduction

Lip reading is a complex visual task that involves deciphering speech from the movements of the lips, tongue, and face without auditory input. It is particularly critical for those with hearing impairments and has extensive applications in noisy environments where audio signals may be compromised. The advent of deep learning has significantly enhanced the prospects of automating this task, but achieving high accuracy depends on multiple factors including the resolution of the input data.

## Research Aims and Objectives

The core of our investigation revolves around assessing how input resolution affects the performance of advanced lip-reading models. Our chosen architectural framework combines three-dimensional convolutional neural networks (Conv3D), for their excellence in spatial feature extraction, with Long Short-Term Memory networks (LSTM), which are adept at understanding temporal sequences, and Connectionist Temporal Classification (CTC) for its effectiveness in training sequence prediction models. This holistic approach aims to harness the strengths of each component to achieve superior lip-reading accuracy.

We meticulously compare the character error rate (CER) and word error rate (WER) between models trained on datasets of varying resolutions: low (35x70 pixels) and high (70x140 pixels). These comparisons are not just statistical exercises but are aimed at understanding the intricate balance between computational demands and the fidelity of lip-reading accuracy. Such insights are pivotal for developing efficient and scalable lip-reading applications.

Furthermore, this study extends beyond theoretical analysis to practical application. The development of an online platform to test the models in real-world scenarios is a testament to our commitment to bridging the gap between academic research and practical utility. This platform will not only serve as a benchmark for our models but also as a resource for the community, facilitating further research and development in the field.

### This uses ‘Heading 3’ style

‘Heading 3’ style does not appear in Table of Contents.

#### This uses ‘Heading 4’ style

Some text here.

## Anticipated Contributions

This research is anticipated to contribute significantly to the field of automated lip reading by elucidating the nuanced impact of input resolution on model performance. By offering a detailed comparison of low and high-resolution datasets, we aim to provide guidelines for balancing computational efficiency with model precision. Additionally, the online platform we develop will serve as a valuable tool for researchers and practitioners alike, fostering further innovation and application of lip-reading technology in various domains.

When labeling a table use the ‘Table Label’ style.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

When labelling a figure use the ‘Figure Label’ style



# Background Research

## Introduction

Lip reading, traditionally termed speechreading, involves the interpretation of speech by visually recognizing the movements of the lips, face, and tongue, independent of auditory signals. This skill is essential for the deaf and hard-of-hearing community, providing a means to understand spoken language through visual cues alone. Beyond its utility for individuals with hearing impairments, lip reading is increasingly vital in noisy environments such as industrial settings or busy urban landscapes, where reliance on auditory comprehension is impractical.

## Headings will vary based on research topic

The document must use IEEE Software referencing. Here’s an example: Somerville states that “because of the lack of physical constraints, software systems can quickly become extremely complex, difficult to understand and expensive to change” [1, p. 11].

This is another figure heading

# System Design

## Introduction

## Requirements

## Architecture

## Design

## Implementation

# Testing and Evaluation

## Introduction

## Testing

## Evaluation

# Conclusions

## Introduction

## Reflection

## Recommendations

##### References

[1] I. Sommerville, *Software Engineering*. Boston: Addison-Wesley, 2011.

##### Glossary

##### List of Abbreviations

###### Appendix Title uses ‘Heading 6’

Appendix sub-title uses ‘Heading 7’